Robust procedures in chemometrics

Kotwa, Ewelina Katarzyna; Brockhoff, Per B.

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Kotwa, E., & Brockhoff, P. B. (2012). Robust procedures in chemometrics. Kgs. Lyngby: Technical University of Denmark (DTU). (IMM-PHD-2012; No. 284).

Robust procedures in chemometrics

Ewelina Kotwa

Kongens Lyngby 2012

IMM-PHD-2012-284

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Summary

The general aim of the thesis was to contribute to the improvement of data analytical techniques within the chemometric field. Regardless the multivariate structure of the data, it is still common in some fields to perform uni-variate data analysis using only simple statistics such as sample mean and variance. Recent instrumental developments in chemometrics often result in high-order data, for which uni-variate tools do not suffice and multivariate data analysis is required. Moreover, many multivariate models assume normality of the residuals (which in many cases is far from reality) and are not resistant towards outliers (which are known to be more the rule than the exception for empirical data). That is the reason for robust methods being a valuable tool for both semi-automated detection of outliers and model building.

The approach adapted in this thesis, can be split in two main parts: 1. applying a multivariate and multi-way data analytical frame-work in fields where less sophisticated data analysis methods are currently used, and 2. developing new, more robust alternatives to already existing multivariate tools.

The first part of the study was realized by delivering and comparing two- and three-way chemometrical methods (such as Principal Component Analysis (PCA), Parallel Factor Analysis (PARAFAC), Partial Least Squares (PLS) regression and its multi-way alternative, N-PLS) for explanatory and regression analysis of Conductivity-Temperature-Density (CTD) sea water samples. This data, so far analyzed using mostly uni-variate methods, is defined by three data modes (depth, variables and geographical location), and therefore, can benefit from introducing more complex analysis tools. The results of the study indicated superiority of the three-way frame-work, potentially constituting a novel assessment of the sea water measurements. Particularly in the case of regression models there is a clear preference towards the more complex model, delivering more reliable predictions than a classical 2-way PLS. Therefore, using multi-way data analysis tools is recommended, in order to extract the full information from multi-way data structures.

The second part of the thesis targeted qualitative properties of the analyzed data. The broad theoretical background of robust procedures was given as a very useful supplement to the classical methods, and a new tool, based on spherical PCA (S-PCA), aiming at identifying Rayleigh and Raman scatters in excitation-emission (EEM) data was developed. Moreover, the functioning of S-PCA was investigated in order to facilitate its usage among practitioners. The results show clearly that robust methods can significantly contribute to the improvement of existing analytical techniques used commonly in chemometrics, for example by providing excellent outlier detection tools. It is therefore advised to apply robust and classical procedures simultaneously, at least to determine if contamination in the data is present. For this becoming a standard procedure, further work is required, aiming at implementing reliable robust algorithms into

standard statistical programs.

Preface

This thesis was prepared at the Departament of Informatics and Mathematical Modelling in fulfillment of the requirements for acquiring the Ph.D. degree at the Technical University of Denmark.

The thesis deals with different aspects of mathematical modeling applied in the field of chemometrics. The main focus is on delivering appropriate data-driven analysis techniques. Special attention is payed to the robustness issues of the methods used within.

The thesis consists of a summary report containing a reviw of applicable robust procedures and a collection of three research papers written during the period of 2010-2013, among which one has been published and two other submitted to relevant research journals.

Lyngby, October 2012

Ewelina Kotwa

Papers included in the thesis

[A] Ewelina Kotwa, Silvia Lacorte, Carlos Duarte and Roma Tauler. Investigation of Arctic and Antarctica spatial and depth patterns of sea water in CTD profiles using chemometric data analysis. Submitted to *Polar Science*, 2012.

[B] Ewelina Kotwa, Bo Jørgensen, Per B. Brockhoff and Stina Frosch. Automatic Scatter detection in Fluorescence landscapes by means of Spherical Principal Component Analysis. *Journal of Chemometrics*, 2013.

[C] Ewelina Kotwa, Stina Frosch and Per B. Brockhoff. Spherical Principal Ccomponent Analysis. A simulation study for data containing clustered outliers. *Submitted to Journal of Chemometrics*, 2013.

Acknowledgements

Many wonderful people joined me on different stages of this journey, which started in October 2007 and led through many geographical, emotional and intellectual destinations. I would like to express here my love and infinite gratitude for being able to experience every single one of you, sharing energy, dreams, inspirations, frustrations, friendship, endless discussions, or perhaps just a few smiles... You made me feel very lucky, and because of you, these 6 years will stay in my heart as a truly *multi-dimensional* and magical experience.

I would like to thank my family (especially my wonderful parents), the closest friends and my partner for supporting me (though not always understanding), even in the most daring of my decisions.

Preparation and completion of this thesis would never be possible without my supervisors Stina Frosch and Per Brockhoff but also Bo Jørgensen. I'd like to acknowledge them for their guidance and all the help they were able to provide me.

Furthermore, my gratitude goes to prof. Roma Tauler for his unlimited patience and faith in me! Thank you!

My great appreciation is also given to Hanne, Per and Paul Eric, the best 'technical' support I know :)

Finally, a very special thanks go to all the fellow PhD students at DTU Informatics, among whom many turned to be dear friends, always ready to share their time, candies, lunches, head stands, crazy mood swings when the deadline was approaching, happy and sad times, roof over my head when needed and so much more... Julija, Piju, Roland, Peder, Philipou, Marco, Juanmi, Martin, Mahmood, Anne-Katrine, Emil, Dorini, Trygvi, Anna Helga and Javi - you are all so great! Thank YOU!

The last person I would like to acknowledge is not with us anymore but remains one of the biggest inspirations in my life, both as a passionate scientist, but foremost a warm-hearted humanist. Without him I would never be who I am now. Jean-Christophe, thank you from my heart.

Contents

Chapter 1

Introduction

1.1 Background

Understanding chemical phenomena depends on interpretation of analytical data obtained during experiments. A common practice is to treat data in the univariate manner and describe individual variables by simple location and spread statistics such as sample mean and variance. Recent instrumental developments within chemometrics field have made high-order multivariate data common, often with the number of variables exceeding the number of observations. In order to deal with this high-dimensionality of the data, uni-variate tools do not suffice and multivariate analysis is required. Methods such as fluorescence spectroscopy, chromatography, magnetic resonance, EEG etc. call for the use of *multi-way* treatment, such as Principal Component Analysis (PCA), PARAllel FACtor analysis (PARAFAC) or Tucker types of models. Moreover, these tools are more and more used outside of strictly chemometric applications, propagating to environmental science, kinetics, classification problems or sensory analysis. A common rule for all multivariate methods is that: to obtain reliable results, high data quality is needed. Gross errors, which often manifest themselves as *outliers*, are observations that appear to break the pattern or grouping shown by the majority of the data. Moreover, many multivariate models assume the residuals to be normal (or following another reasonable distribution). Today it is already well-known that the normality assumption often does not hold and the presence of outliers is more the rule than the exception when collecting and analyzing real data. The reasons for the occurrence of these erroneous data can be numerous, for example instrumental failure, transcription errors, non-representative sampling or objects belonging to other population. Usually, only complete objects (samples) are regarded as outliers, but it is equally relevant to consider variables or even individual data elements as possibly outlying. The most conventional multivariate methods such as PCA or Partial Least Squares (PLS) regression are highly sensitive to outliers due to the fact that they are based on the least squares criterion where, in the extreme case, even one outlier can have an arbitrarily large effect on the estimate, and thereby, completely offset the model. Outlier treatment is therefore mandatory prior to a proper analysis and modeling. This is usually solved in two ways: either by outlier diagnostics or by applying robust statistics. In the first case, the contaminated data are identified and expelled from the data set before building a multivariate model. However, when working with vast, multi-order data structures, visual evaluation and screening might be no longer available. In the second approach, robust estimators are used instead of the ordinary least square estimators. Robust methods remove or reduce the effect of outlying data points, allowing the remainder to predominantly determine the model. Moreover, construction of robust models allows for later recognition of outlying observations where relevant. The outliers identification is not only essential for a proper modeling but, also for understanding the reasons for unique character of these samples, which may constitute the most interesting part of the data. Therefore, robust methods are a valuable tool for both semi-automated detection of outliers by looking at the robust residuals and for model building.

1.2 Objectives

The main objectives of this Ph.D. project can essentially be categorized according to two thematic blocks, aiming at:

- proposing multivariate and multi-way analysis tools within certain chemometric areas, currently predominant in less sophisticated methodology (for example still based on uni-variate investigations);
- development and analysis of new, more robust alternatives to already existing multivariate tools such as PCA or PARAFAC models;

The first part of the study focuses on delivering an appropriate multivariate methodology for analyzing spatial and depth profiles of sea water samples and on identifying possible geographical differences within and between the Arctic and Antarctic Sea. To enable this, 2- and 3-way chemometric methods, such as PCA and PARAFAC models, were applied and their performance examined. The emphasis was also put on predicting fluorescence values, as being a natural measure of biological activity, from the other physio-chemical variables by applying and comparing the Partial Least Squares (PLS) regression technique with its multi-way alternative, N-PLS. In the second part, being the core of this work, a particular attention is paid to qualitative properties of the analyzed data. The broad theoretical background of robust procedures is presented and promoted as a very practical supplement to the classical methods. Additionally, a new tool, based on robust PCA, for identification of Rayleigh and Raman scatter in fluorescence excitation-emission (EEM) data is given as an example of successful practical application of robust work-frame in the real data situations. Finally, the functioning of Spherical PCA is examined and explained in order to encourage its usage among practitioners.

1.3 Outline

Following the Introduction, a brief preface to the chemometrics field is given in Chapter 2, including an overview of commonly encountered data types. The following chapter addresses topics of multivariate statistics in chemometrics. Tools such as PCA, PARAFAC, PLS and N-PLS will be shortly described and the results of the Arctic and Antarctic sea water investigation will be revealed. Finally, Chapter 4 refers to the statistical robustness theory, describing some of the conventional, least squares based tools confronted with their robust alternatives. Pros and cons of each approach are discussed. The chapter ends with case study results, showing an example of using a robust PCA method (S-PCA) in the practical application, and simulation results, characterizing S-PCA as an outlier detection tool. The last chapter provides a discussion and few concluding remarks, bonding the achievements obtained in this thesis.

Chapter 2

Chemometrics

Since chemometrics is a relatively new field, continuously broadening its semantic borders and not necessarily known among all statisticians and data analysts, a few words of introduction are to be given here. This chapter delivers a short resume of what is currently understood under the name 'chemometrics' together with some of the most common data characteristics and data types encountered within the field.

2.1 Definition and origins

In its most global meaning, chemometrics can be perceived as science of extracting information from chemical systems, using data-analytic tools, such as multivariate statistics, applied mathematics, and computer science [1]. Chemometrics is a highly inter-disciplinary field aiming at solving both descriptive and predictive problems in experimental life sciences, especially in chemistry, biochemistry, medicine, biology and environmental science.

The name 'chemometrics' was coined by Svante Wold in the early 1970s [1]. It was around this time, when the computer revolution reached the scientific world, allowing faster calculations and development of more computation-intensive algorithms. The computerisation progressed rapidly and already in the early 1980s a wide variety of data- and computer-driven chemical analyses were occurring. Concurrently, the more and more complex instrumental techniques emerged, such as infra-red and UV/visible spectroscopy, mass spectrometry, nuclear magnetic resonance, atomic emission/absorption or chromatography. The output of these instruments reached often thousands of measurements per sample, yielding highly complex *multivariate* data structures.

Multivariate analysis has therefore always been in the focal point of chemometric tools. Decomposition and 'data compression' methods, such as PCA or PLS regression gained an extreme popularity among practitioners in the field [2]. The reasons for this lay in the *colinearity* of these high-dimensional data sets, and implicitly, in the fact that already a low-rank linear structure can often represent the data at a satisfactory niveau. PCA, PLS and other similar projection methods proved over time to be effective tools for data exploration, visualisation and calibration for chemically interesting phenomena. Some of these tools, PLS in particular, after having been heavily used in chemometric applications, gained acknowledgement in other fields.

A more recent approach towards analysis of chemometrical data, gaining popularity in the 1980s and blooming in the late 1990s, consists of the multi-way work-frame [3]. While standard multivariate data are arranged in the two-way, matrix-like structures (a table where each row corresponds to a sample and each column to a variable, e.g. absorbance at a particular wavelength), a three-way data could be logically represented by three different indices and stored in a cube (as in the case of fluorescence spectroscopy where for each sample a set of emission wavelengths is determined for several excitation wavelength values). Moreover, it is not rare to encounter even higher-order (more than three) data sets within the chemometrics world. A proper statistical methodology was developed in order to match the conceptual requirements, yielding PARAFAC and N-PLS models (corresponding to PCA and PLS in the 2-way case) among other multi-way procedures. In some situations the 3-way structure can be analysed by ordinary 2-way tools by rearranging the data into matrices, however, it can be argued (see [2]) that in principal, 3-way techniques would be more beneficial for a 3-way data, as they preserve the inner covariance structure. Both types of methods will be described in more details in the forthcoming sections.

Chemometrics is an application driven discipline, where the continuous advances in analytical instrumental techniques constantly call for new developments of corresponding data analysis tools. While the standard chemometric methodologies are very widely used industrially, the goal and challenge for academic environments is to come up with new proposals targeting theoretical, methodological and application-wise development.

Below, two of the most common chemometrical data types will be indicated and briefly described.

2.2 Data types

2.2.1 Fluorescence spectroscopy

Fluorescence spectroscopy is a type of electromagnetic spectroscopy which detects fluorophores present in an analyte (sample). Fluorophores are the molecules able to emit light when relaxing to the ground state form an excited state. The method involves using a beam of light for exciting the electrons in molecules of certain compounds and causing them to emit light, which is measured. By analysing the different frequencies of emitted light along with their relative intensities, the particular compounds can be identified. This technique is fast, sensitive and non-invasive, and therefore, it found a broad usage in fields such as biochemistry, analytical chemistry, food and environmental science. An outcome of a fluorescence spectrometer is usually written in a so-called excitation-emission (EEM) matrix representing intensities of the compounds for certain excitation ($j = 1, ..., J$) and emission ($k = 1, ..., K$) wavelengths. If a number of samples I is considered, the whole data form a 3-way array, $\underline{\mathbf{X}}(I \times J \times K)$ (see Figure 2.1). This fundamental structure of the fluorescence data happens to be closely related to underlying assumptions of the unique PARAFAC model, which is therefore able to resolve the spectral curves and deliver the estimates of the concentrations of analytes. A chemometrician is then able to identify particular fluorophores, present in the sample, according to the peaks of the resolved

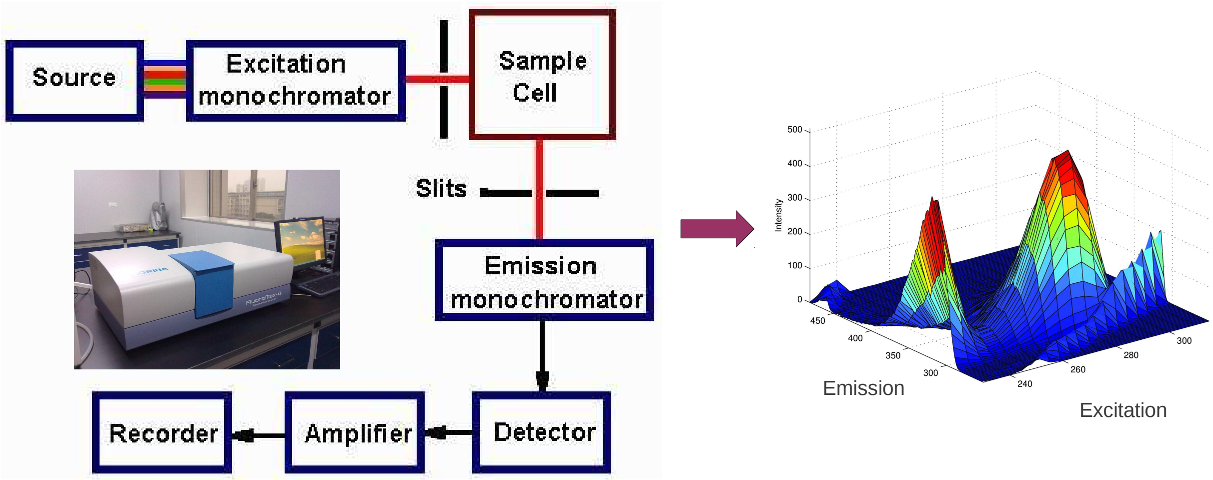

Figure 2.1: Left: a fluorescence spectrosopy 'step-by-step' scheme; right: an output of a spectrofluorometer - an EEM landscape.

spectra. This 'close relation' became the reason of great interest of applying multi-way analysis tools within this type of data (see for example [3–5]).

Even though 3-way decomposition models such as PARAFAC seem to be ideal for EEM data, a few issues usually occur, leading to complications in applying these methods practically. The most commonly encounter problem is the presence of Rayleigh (1st and 2nd order) and Raman light scatter effects. These diagonal ridges appear in the EEM landscapes (as visible in Figure 2.1) due to interactions between the molecules in the solution causing some incident light leading to deterioration of the PARAFAC results. If that is the case, Rayleigh and Raman scatters should be identified and removed from the data set, prior to PARAFAC-based analysis which was widely investigated in the literature [6–13]. Recently, automatic scatter removal techniques based on robust statistics have been proposed in [14] and the continuation of these investigations is pursued in [15] (Paper B).

2.2.2 Chromatography

Chromatographic techniques have for a goal the isolation and identification of the components in the chemical mixtures and are by far the most widely used analytical separation methods [2]. There are a number of different types of chromatography, among which gas (GC) and liquid chromatography (LC) are the most abundant, being broadly used in a variety of fields, including pure and applied sciences, forensics, and athletics, among others. The principle of chromatography is twofold: in the first step, a mixture is dissolved in a solvent (mobile phase), and secondly moved across an absorbent medium (stationary phase). The process relies on the fact that particular molecules will behave in different ways, travelling at different speeds, taking therefore different times to elute.

The output of a chromatograph takes the form of a chart, as shown in Figure 2.2 with a series of troughs and peaks. Each peak represents a substance present in the sample, and the concentrations of these substances can be determined by looking at the height and width of the peak. If chromatographic measurements are taken at different stationary phases, for different mobile phase compositions and for diverse solutes, this leads to a three-way data structure of retention factors, calling for a use of adequate analysis tools.

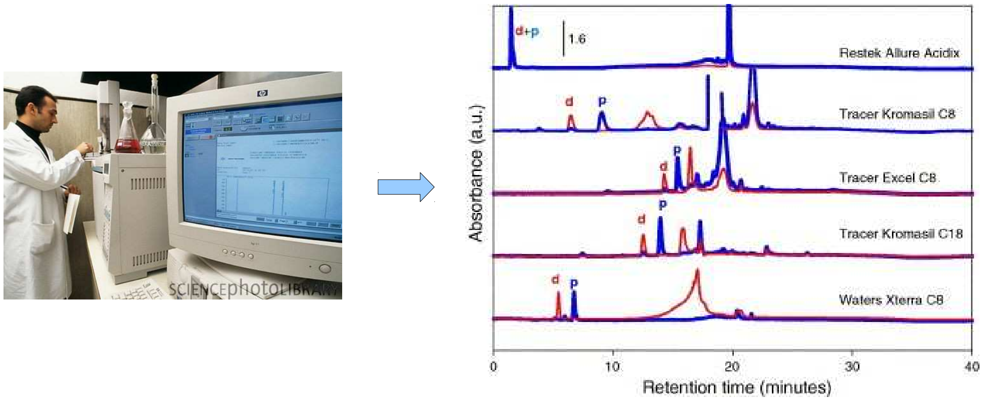

Figure 2.2: Left: a chromatographic experiment; right: resulting chromatogram.

2.2.3 Other data

Many other types of analytical instruments can produce vast data structures suitable for multivariate and multi-way analysis. Data resulting from infrared and UV/visible spectroscopy, mass spectrometry, nuclear magnetic resonance or atomic emission/absorption experiments can produce thousands of measurements per sample and are by nature highly multivariate. But these are just some of the examples contributing to the immense range of data which can benefit form the multivarite tools application.

Paper A gives an example of less common data, consisting of water samples acquired by a CTD (Conductivity-Temperature-Depth) sensor device during two oceanographic expeditions at different depths and locations. This device measures a set of water related variables (temperature, conductivity, fluorescence, etc.) at different sea levels. The resulting data can be then arranged according to three modes: sea water depth, measured variables and geographical location.

Chapter 3

Multivariate data analysis

Multivariate statistics is used when simultaneous observations and analysis of more than one variable is considered. As stated in the previous chapter, this is the case for most chemometric applications. Multivariate data analysis is therefore focused on understanding the different inter-relations between variables or samples, by using knowledge about the correlation structure present in the data. These methods offer much more powerful tools for extraction of the full information from the data, often obtained as a result of difficult and/or expensive experiments, compared to the uni-variate approach, where each variable is handled separately.

There are many different models, which can be collected under the common multivariate category, such as multivariate analysis of variance (MANOVA), multivariate regression analysis and PCA. Models, which use projection methods for representing the data by a low-rank linear structure (in chemometrics it is possible due to the high variable co-linearity), are especially popular. PCA and PLS models are among the most widely used techniques and for this reason will be briefly described below.

Another category of statistical tools, appropriate for analyzing vast amounts of data is gathered within the multi-way work-frame [2, 3] which is considered as an extension of multivariate data methods. This branch of statistics classifies the data according to number of *ways* (directions) in which each data point can be characterized. For example, a single sample of fluorescence spectroscopy can be considered as a 2-way data, as the intensities are registered according to two distinct directions: emission and excitation wavelengths. However, if the experiment is repeated for a range of mixtures, this adds a supplementary direction (samples) and the data set becomes of a 3-way nature. It is easy to imagine that four- or in general multi-way data can occur, however in this thesis only 2- and 3-way methods will be considered. A short description of PARAFAC and N-PLS models (corresponding to PCA and PLS in the 2-way case) will be given in the following sections. It is worth mentioning that 3-way data can be also analyzed by standard multivariate tools, after unfolding the cubic structure and reshaping it into a table. An example of such unfolding is illustrated in Figure 3.1. By doing so, however, one risks that the information inherent to the multi-way correlation structure will most likely be lost.

In the forthcoming part of this chapter some of the 2- and 3- way multivariate methods will be presented. PCA and PARAFAC are perhaps most widely applied exploration tools in chemometrics, whereas PLS and N-PLS find broad usage when the multivariate regression problems are considered. A practical

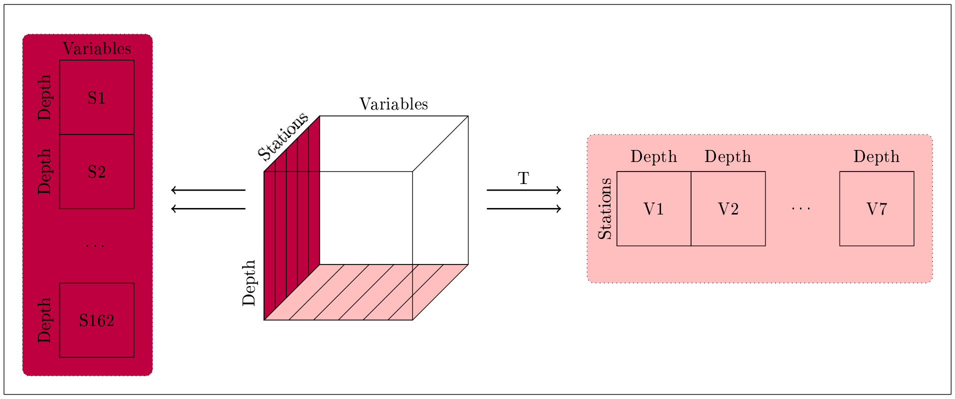

Figure 3.1: Three-way data cube and two unfolding directions: variable and station-wise.

example treating 2- and 3- way regression analysis, given in the last section, illustrates conceptual difference of both approaches and emphasizes the importance of choosing the analysis tools according to data nature.

3.1 Exploratory tools

3.1.1 PCA

Principal Component Analysis is a linear subspace-based technique, perhaps most commonly found in chemometric literature. A PCA model is presented in Equation 3.1:

$$x_{ij} = \sum_{r=1}^{R} t_{ir} p_{jr} + e_{ij} \quad i = 1, ...I;\ j = 1, ...J; \tag{3.1}$$

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} contains model's residuals. In brief, PCA maps 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 determined by maximizing the variance explained by them. Those vector products, being a linear combinations of the original variables (or objects), are called principal components and often already a small number of those components allows us to explain the data variation in a satisfactory way. More details concerning PCA method can be found in the literature (see references [16], [17] or [18]).

3.1.2 PARAFAC

A PARAFAC model, introduced by Harshman [19] and popularized by Smilde [2] and Bro [10], is a tri-linear generalization of PCA, which decomposes a data cube $\underline{\mathbf{X}}(I \times J \times K)$ into a sum of triple vector products, called loadings. The most common way of writing the model is following

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk} \tag{3.2}$$

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 number of components and e_{ijk} represents the error term. If the experimental data fulfills the tri-linear assumption (required invariability of the component profiles across the different data slices with different weighting coefficients for each slice [19], [2]), the application of a PARAFAC model is usually superior to its bilinear counterpart PCA. The reasons for this are numerous: first of all, PARAFAC takes into account that interrelations exist in all three data directions. Moreover, the problem of rotational freedom, typical to PCA is solved, as PARAFAC provides the unique solution (up to the scaling constant, sign and permutation ambiguities) [2]. In addition, the PARAFAC model resolves each mode separately, giving a straightforward physical interpretation for the depth, station and variable 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, it does not tend to over-fit, as is often the case of PCA.

In spite of benefits that may be gained from applying a PARAFAC model, some drawbacks also exist. The most important consideration is that the real data do not always conform adequately to a tri-linear assumption. In this case, the model might return degenerated solutions. Degeneracy might also occur when a large number of factors is needed and if they are interrelated (compare with the Tucker model [20]). In most of these cases, the bilinear model is still appropriate and PCA or other methods like MCR, described by [21], can be successfully applied.

3.2 Regression

3.2.1 PLS

Partial Least Squares regression is a multivariate, 2-way calibration method. Essentially, it was invented by Wold [22, 23] as a remedy for co-linearity problem in the multidimensional data and since then, it has been broadly used by practitioners in various application fields (see [24] for tutorial and [25] for more detailed description). The method approximates the $\mathbf{X}$ block by r components (called latent variables) and, at the same time, projects $\mathbf{Y}$ on those components, which are constructed to compromise between fitting $\mathbf{X}$ and predicting $\mathbf{Y}$. This can be written in a matrix notation following Ref. [2]:

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}_X; \tag{3.3}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}' + \mathbf{E}_Y. \tag{3.4}$$

$\mathbf{T}$ and $\mathbf{P}$ are again score and loaing matrices, $\mathbf{Q}$ contains regression coefficients and $\mathbf{E}_X$ an $\mathbf{E}_Y$ are the corresponding error terms. This algorithm is run sequentially, meaning that only one component is calculated at a time, and afterwards the $\mathbf{X}$ matrix is replaced by the residuals $\mathbf{E}^{-1} = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1$. A broader description of PLS and its applications can be found in [2, 23, 25].

3.2.2 N-PLS

Multi-way PLS (or N-PLS) is a generalization of Partial Least Squares regression into a higher dimension, which predicts $\mathbf{y}$ and decomposes $\underline{\mathbf{X}}$ similarly to the PARAFAC model [26–28]. This is performed by searching for a vector $\mathbf{t}$, being a linear combination of columns of $\underline{\mathbf{X}}$, which has a maximum covariance with $\underline{\mathbf{y}}$. This, [2] can be formulated in the following way

$$\underline{\mathbf{X}} = \mathbf{t}(\mathbf{w}^K \otimes \mathbf{w}^J) + \mathbf{E}_X; \tag{3.5}$$

$$\underline{\mathbf{y}} = \mathbf{t}b + \mathbf{e}_y, \tag{3.6}$$

where $\underline{\mathbf{X}}$ and $\underline{\mathbf{y}}$ are arrays of dimension $(I \times J \times K)$ and $(I \times 1 \times K)$ respectively, $\mathbf{w}^J$ and $\mathbf{w}^K$ are weighting vectors defined for modes J and K, $\otimes$ defines the Kroneker product [29] and b is the regression coefficient. As N-PLS is also a sequential method, after finding the first component, both $\underline{\mathbf{X}}$ and $\underline{\mathbf{y}}$ are being 'deflated' (replaced by residuals of the respective models), in order to recommence the algorithm.

3.3 Multivariate investigation of water samples from the Arctic and Antarctic Sea

In order to compare the performance of 2- and 3-way statistical methods, an example will be given here, presenting partial results of Arctic and Antarctic water investigations (the full survey can be found in Paper A). The analytical aim is to verify the possibility of predicting the fluorescence, being a natural measure of biological activity, from other physical variables.

3.3.1 Data

The data was collected by a *CTD sensor* (a standard measurement device used during oceanographic expeditions) and consists of variables given in Table 3.1. This common and rather simple measurement method was deployed in over 170 locations on both extremes of the globe. In each location the sensor was dropped

Table 3.1: Measured CTD variables.

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]

down to a specific depth producing the depth-profile for each variable. Eventually, a 3-way data structure was created with variable, depth and location modes. It is not uncommon to analyze this kind of output in a uni-variate way (one variable at a time and independently from the others). This treatment however, does not take into account possible underlying covariance dependencies and therefore, does not use the whole information contained in these fairly complex data structures. In order to overcome these risks, two multivariate regression models, PLS and N-PLS, will be applied and their performance compared.

3.3.2 Methodology and data pre-treatment

It is well-known that measured fluorescence will be strongly correlated to the other 'light related' variables, such as beam transmission or sea turbidity, and therefore it will be even more interesting to investigate if the biological activity might be inferred from some purely physico-chemical conditions, *e.g.* the amount of dissolved oxygen, temperature, conductivity and salinity. This will be the focus of the modeling step presented below. Before that however, the data was preprocessed (mean centering and scaling) and screened for outliers.

The problem of missing values was circumvented by interpolation.

Two models are to be applied: PLS regression and its 3-way generalization, N-PLS. For this reason data is arranged in ’Y block (predictions)’ and ’X block (explanatory variables)’. As the sea water measurements analysed in this study follow three different modes (variable, depth and location), the data cube must be unfolded (as in Figure 3.1, variable direction) and rearranged into a matrix, in order to fit the 2-way PLS model. As an alternative approach is to use a N-PLS on the whole, unchanged data cube. In order to identify the optimal number of components for both models, the Cross-Validated Root Mean Square Error (RMESCV) was calculated by means of cross-validation (200 contiguous blocks).

3.3.3 Results

The results collected in Table 3.2 show clearly that the three-component PLS model accounts for only 25% of the observed fluorescence variation, which is a very weak result. The control plot in Figure 3.2a, showing measured versus predicted Y values, indicates that the model is not able to identify the difference in data behavior within Arctic Sea and Antarctica, leading to poor predictions for each of the locations. At the same time the X-block is fully explained, as the remaining variables in the model are highly correlated. The situation is quite different in the case of multi-linear PLS. Remarkably, the model is able to explain up to 78% of the measured fluorescence and, at the same time, around 74% of the X array variability. A plot of predicted versus observed values (Figure 3.2b) confirms the obtained improvement seen ’by eye’. This result can

Table 3.2: Explained variance 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	X block	Y block
PLS	1	40.54	65.01	27.26	22.36
	2	70.30	70.47	81.42	22.64
	3	92.32	74.08	100	23.05
nPLS	1	26.66	60.40	46.00	31.56
	2	65.56	76.88	72.04	69.73
	3	72.55	85.50	83.42	79.13

be explained by the fact that the 3-way model accounts for the interrelations existing within the data locations, which could have been disregarded during unfolding of the data set.

It can be concluded that the 2-way PLS model is largely outperformed by its multi-way alternative in predicting the fluorescence values out of non-radiation related variables. It seems here, that by unfolding the data in such a manner, the 3-way correlation structure has been 'flattened' and some information lost. More importantly however, this example emphasizes how crucial the choice of correct modeling techniques is for extracting the full information from multi-order data set.

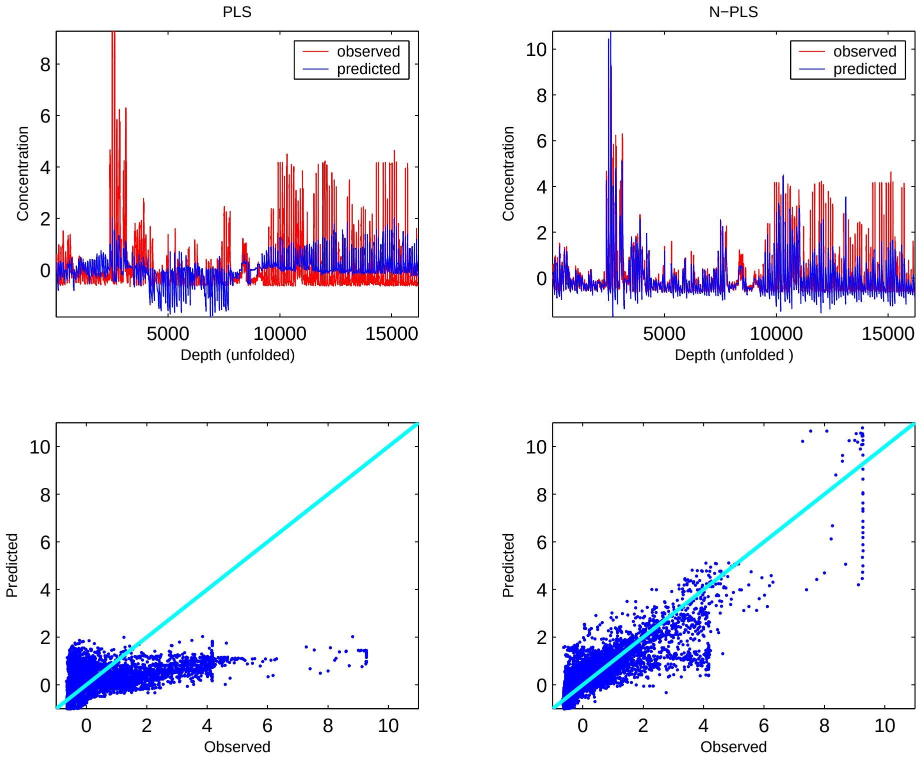

Figure 3.2: Predicted versus observed values for PLS (left) and N-PLS (right) models, with only physico-chemical variables entering X-block.

Chapter 4

Robustness procedures in chemometrics

This chapter will address concepts of statistical robustness, being the focus point of this thesis. This relatively newly defined branch of statistics (however it should be emphasized that issues discussed within that field certainly remember the days of first statistical theorems) investigates the influence of deviations from modelling assumptions on known statistical procedures. Nowadays, it is already well-known that normality assumption often does not hold nor, in addition, are the gross errors in the real data uncommon. In these situations, classical statistical tools are at risk of not being able to perform adequately, especially when multidimensional data (typical in chemometrics) is considered and *a priori* outlier detection tools are not available. The most dynamic development period of robustness theory fell on the second half of the XX century, due to contributions of many prominent statisticians, where Box, Tukey, Huber, Hampel and Rousseeuw are just few to be mentioned. Thanks to them, the theoretical background within robustness work-frame equips practitioners with various alternatives to the standard statistical procedures. On the other hand, the constant computerisation of the modern world and development of complex instrumental measurement techniques, are a bottom-less source of needs, steadily calling for new creative robust tools.

Unfortunatelly, this fairly important, especially from the practical point of view branch, seems to be still fairly neglected in academia, where general statistics courses treat the topic marginally if at all.

In the following chapter a broad theoretical background of robust statistics will be presented together with some practical examples of robustified models. Furthermore, the application of these models will be demonstrated within the chemometrics, used for the indentification of Rayleigh and Raman scattering in fluorescence spectroscopy data. Finally, results of investigations aiming at deeper understanding of spherical PCA (S-PCA) - a robut version of classical PCA - will be demonstrated. Full study concerning removing influence of scatters in fluorescence data is included in Paper B and S-PCA analysis can be found in Paper C.

4.1 Robustness theory

4.1.1 Why do we need robustness in statistics?

The most common assumptions of classical statistical procedures are those concerning normality (or any other reasonable parametric model) and independence of the model's residuals. Historically, considerations about the underlying error distribution go back to Gauss [30]. Instead of asking for the 'true' value (the best estimate) of n-observation sample, he formulated the problem other way round by questioning which error distribution would make the arithmetic mean optimal as a location measure. This led to the normal distribution for the error term and further sanctioning of the least squares work-frame as optimal. Additionally, the famous central limit theorem made "everybody believe in the 'dogma of normality', the mathematicians because they believed it to be en empirical fact, and the users of statistics because they believed it to be a mathematical theorem" (Frank Hampel, orally). Whereas that theorem only suggests the asymptotic normality under well-specified conditions, the empirical results show that a typical error distribution of high-quality data (data with no palpable gross errors) has longer tails than normal [31].

Moreover, already Legendre (in 1805) in his first work on the least squares noticed a need for outliers rejection rules. He was followed by many other pronounced statisticians such as Newcomb [31] (who was the first to propose a normal mixtures for modelling heavy-tailed distributions), Daniell [32] (first mathematical analysis of estimators being linear functions of order statistics and first treatment of the trimmed mean) or Jeffreys [33] (first appearance of

M-estimators in the context of robust statistics). Finally, the investigations conducted by Tukey [34], and later by Huber [35] and Hampel [36] proved that the least squares based estimators, such as standard deviation, lose their efficiency, even under tiny deviations from normality (see the example and discussion in [35]). In fact, Hampel [36] claims that these efficiency losses of least squares estimators would most likely be between 10% and 50%, which are devastating news for a classical statistician.

The above considerations concern still some 'well-behaving' data, deviating only slightly from the assumed parametric model. In the case of any empirical data, however, it is not uncommon to encounter the occurrence of gross errors. These observations often manifest themselves as outliers but not all outliers are gross errors. Some outliers can be true values and their presence might be highly interesting. For example when investigating waves hight, a 20m observation might mean a wrongly registered gross error or a tsunami phenomena which could be the most wanted information in the data set. A separate branch of statistics, called 'extreme values analysis (EVA)' is especially interested in those observations.

The concept of 'outliers', meaning observations which do not follow the pattern established by the majority of the data, is useful but not clearly defined. There are no clear boundaries defining when a point becomes an outlier and when it is still an 'ordinary' observation, and one should be aware of the continuous transition between them. It is well known, that a single, arbitrarily distant outlier may attract the least squares fit and completely offset the results. In order to mitigate this risk, different outlier rejection rules were elaborated. Even though, as a principal, the outliers should be set aside for separate analysis, the most common practice is to identify and reject these 'problematic' points and perform least squares routines on the 'cleaned' data set. This approach assumes that reasonable rejection rules are available, which is not evident in practice and according to Hampel [37] 10-20% efficiency losses compared to better robust methods should be expected. The situation becomes only more complicated when multi-variable and multidimensional data sets are taken into account, where visual outliers identification becomes infeasible.

All these 'inconveniences' urged some prominent statisticians to come up with new, more adequate solutions. Among them, inspired by Tukey [34], Peter Huber and Frank Hampel played the crucial role in building up and propagating some new concepts related to robust statistics. The most known theories include (among others): M-estimators (a generalisation of maximum likelihood estimators, given different objective functions) a small and rather full 'neighbourhood' of parametric model and the related minimax work frame; a stability theory of statistical procedures including influence curve (measuring the effects of infinitesimal perturbations) and breakdown point measure (being the largest fraction of outliers tolerated by an estimator before turning unreliable).

4.1.2 Robustness - definitions and measures.

4.1.2.1 Definition of robustness

Back in the $18th$ century, the term 'robust' referred to someone who was strong, crude and vulgar, and it was not before $19th-20th$ century when this pejorative connotation disappeared in the language evolution process, in favour of healthy, tough and strong enough to oppose difficulties [38]. The statistical meaning was coined by Box in 1953, and it was around that time when it became clear that: [39]

- it is impossible to have an accurate knowledge of underlying 'true' distribution;
- the performance of some classical procedures is very unstable, even under very small deviations from assumed model;
- even though some robust estimates are less efficient than their classical counterparts under strict normality, they show much more steady and better results once deviations from normality occur;

There is no unique definition of robustness as such. According to Box [40] robust means "insensitive to changes in extraneous factors not under test". Other criteria exist, for example related to absolute or relative efficiency of an estimate (see [39]). For Hampel, an estimate is robust, when its distribution changes little under arbitrary small variation of the underlying distribution (see [36, 41]), whereas Huber proposes somewhat broader definition, namely: "robustness signifies insensitivity against small deviations from assumptions" [35]. These 'small deviations' can be few gross errors or many smaller irregularities.

Having said all that, it seems that a consensus can be found in the following definition: Robust statistics investigates the effects of deviations from modelling assumptions on known statistical procedures (being equivalent to the stability theory of these procedures) and, if necessary, develops new, better alternatives.

4.1.2.2 Huber's M-estimators and the minimax result

This small paragraph is dedicated to Peter J. Huber and his contribution within robust estimation techniques, which, together with Hampel's later stability theory (see below), give a foundation to the modern robust statistics. In one of his first works [42] he introduced a flexible class of estimates T_n, called 'maximum likelihood type' or simply 'M-estimates'. The concept was based on the idea of replacing the squared residuals in the standard least squares work-frame, by another arbitrary function ρ yielding a minimum problem of following form

$$\sum_{1}^{n} \rho(x_i; T_n) \rightarrow min. \tag{4.1}$$

If we assign the derivative of ρ as $\psi(x, \theta) = (\partial/\partial\theta x)\rho(x; \theta)$, it implies

$$\sum_{1}^{n} \psi(x_i; T_n) = 0. \tag{4.2}$$

For T_n being a location parameter, these equations become $\sum_1^n \rho(x_i - T_n) \rightarrow min$ and $\sum_1^n \psi(x_i - T_n) = 0$, respectively. A special case, when ρ and ψ are equal to $-logf$ and $-f'/f$, with the x_i being identically distributed according to $f(x_i - T_n)$, is a standard maximum likelihood estimator of location.

Later, Huber develops his concept of rather full neighbourhood of a strict parametric model (earlier work was considering a finite number of competing parametric models). This is based on the *gross-error model* which assumes that given a parametric model $G(x)$ for known G and a known fraction of corrupted data $\varepsilon \in (0, 1)$ from the distribution $H(x)$, the modelled distribution will be of a form $F(x) = (1 - \varepsilon)G(x) + \varepsilon H(x)$. Based on this assumption, the op-

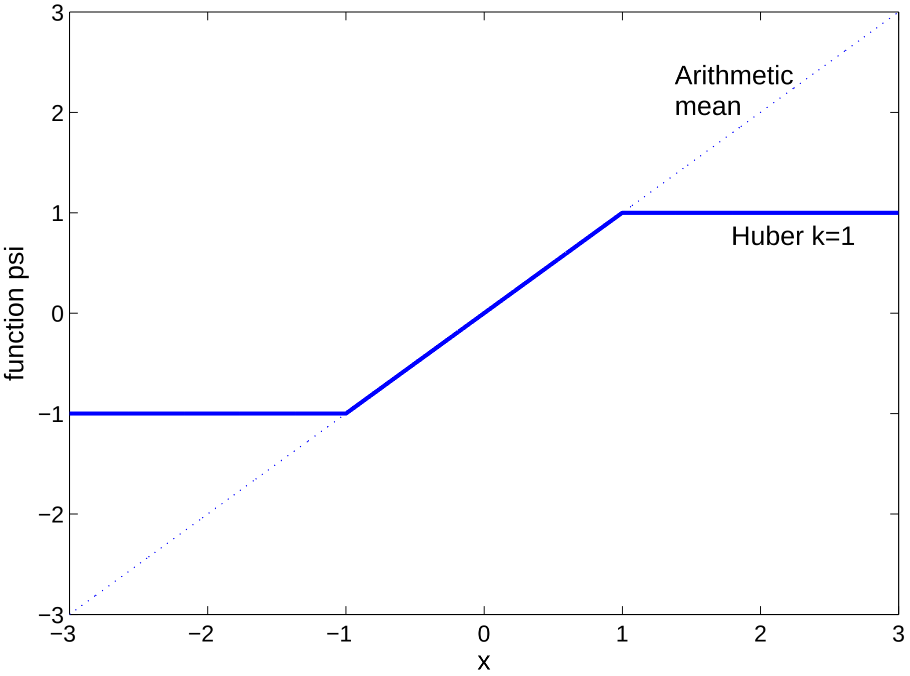

Figure 4.1: A sketch of derivatives (functions ψ) of objective functions for two estimators: a) Huber estimator (solid line) and least squares estimator (dotted). In the first case the influence of the data points far away form the center is visibly bounded.

timisation of the worst scenario over the model's neighbourhood is performed by the means of asymptotic variance of the estimator. This resulted in the *least favourable distribution* (normal in the middle and exponential in the tails) and the famous minimax result, known as a class of Huber-estimators, with $\psi(x) = max(-k, min(k, x))$, as illustrated in Figure 4.1. Huber also develops estimator for robust regression (discussed more in Section 4.1.4) and covariance matrix. His most important investigations are gathered in Reference [35] (and [43] in a more compact version).

4.1.2.3 Qualitative and quantitative robustness

Together with increasing interest and development of new robust techniques, a necessity of establishing some performance measures of these techniques became apparent. Three main reasons for deviations form parametric model will be distinguished:

1. rounding and grouping of the observation values,
2. presence of gross errors (described above),
3. an approximative nature of the model itself, for example by the means of central limit theorem.

Below, the division on qualitative and quantitative robustness will be presented, taking care of all three aspects concerning model deviations. Both approaches might be considered as components of the stability theory of statistical procedures.

Qualitative robustness. The fundamental principle of stability or continuity in the context of robustness could be expresses as follows

$$\begin{aligned} &\forall\varepsilon > 0, \exists\delta > 0, \forall G, \forall n; \\ d_*(F,G) < \delta \to\ &d_*(\mathcal{L}_F(T_n), \mathcal{L}_G(T_n)) < \varepsilon, \end{aligned} \tag{4.3}$$

where d_* is a suitable metric (for example the Prokhorov distance) in the space of the probability measures, F is a distribution of i.i.d. random variables $X_1, \cdots, X_n$, G is a probability measure (distribution) in the neighbourhood of F and $T_n = T_n(X_1, \cdots, X_n)$ is any estimate. Briefly, it means that a small deviation in the underlying distribution $F = \mathcal{L}(X)$ should cause only a small change in the performance of the estimate $\mathcal{L}(T_n)$. A *small change* is assumed to be either a small contamination in many observations (due to the grouping or rounding) or a large change in few of them (presence of gross errors).

In order to metrize the distance between the two probability measures in the metric space, a notion of distance in needed. Among many possibilities, the Prokhorov metric seems to be most attractive [43]. Prokhorov distance $d_{Pr}(F, G)$ between probability distributions (or generally, probability measures) on a measurable space $(\Omega, \mathcal{A})$, where Ω is a complete separable metric space and $\mathcal{A}$ is its Borel-α-algebra generated by the topology, is defined as

$$d_{Pr}(F, G) = inf\,\{\varepsilon > 0 \mid \forall A, F\{A\} \leqslant G\{A^{\varepsilon}\} + \varepsilon\} \tag{4.4}$$

where for every $A \subset \Omega$ its closed ε-neighborhood is defined as $A^{\varepsilon} = \{x \in \Omega \mid \underset{y \in A}{inf}\ d(x, y) \leqslant \varepsilon)\}$. It is equivalent to the set of all points (a closed ball) whose distance from A is less than ε. This expression takes care of deviations included in points 1 and 2 explicitly (stated above), and because of the fact that the Prokhorov distance leads to weak(-star) topology (convergence) [41], it also solves point 3. A much more detailed discussion concerning other types of distances and qualitative robustness definition can be found in [35].

Quantitative robustness: the breakdown point. The second aspect of the robust stability theory is a quantitative expression of how reliable a procedure under investigation is globally, or simply, how big the perturbation can be, before it turns that procedure 'useless', in the statistical meaning. In order to measure that, Hampel [41, 44] defined a breakdown point δ^* of a sequence of estimators $\{T_n\}$ as a value indicating up to which Prokhorov distance (or alternatively, which fraction of gross errors) from the parametric model, the estimator can still give some reliable information about the original distribution. It can be written

$$\begin{aligned} \delta^* = \delta^*(\{T_n\}, F) = sup\{\delta \leqslant 1 : \exists \text{ a compact set } K = K(\delta) \\ \text{which is a proper subset of the parameter space such that} \\ d_{Pr}(F, G) < \delta \Rightarrow G\{T_n \in K\} \rightarrow 1 \text{ as } n \rightarrow \infty\}. \end{aligned} \tag{4.5}$$

This expression, being asymptotic and quite 'mathematical' in nature, doesn't seem to be very practical, therefore another definition given by Donoho and Huber [45] will be presented for a finite-sample of n points stored in F, and the estimator T

$$\delta^*_n(T, F) = min\left\{\frac{m}{n}; \text{bias } (m; T, F) \text{ is infinite}\right\}. \tag{4.6}$$

Here, bias $(m; T, F) = \underset{F'}{sup} \|T(F') - T(F)\|$ means the maximal bias which can be caused by replacing any m data points by some corrupted values, stored in F'. According to this expression, if bias is infinite, this means that m points cause the estimator T breaking down.

Breakdown point can be given depending on the sample number (for example $1/n$) or as a limiting percentage value of $n \rightarrow \infty$ (e.g. 0%). If an estimator has a 0-breakdown point (which is the case for all procedures with least squares cost function), that signifies that even a single, distant enough outlier can completely spoil the model. Alternatively, a breakdown point equal to 1/2 (typical to some robust estimates such as sample median but also many others) is the highest possible value, tolerating as much as 50% of outliers. Conceptually, it is no longer possible to distinguish between the 'good' and the 'bad' points if the outlier fraction exceeds 50%. Due to its simple interpretation and very practical aspects, the breakdown point became perhaps the first index to look at, once evaluating the properties of en estimator.

Influence function: infinitesimal aspect of robustness. The last component of Hampel's stability theory is so called infinitesimal approach, measuring the local sensitivity of an estimator. The cornerstone of this approach is the famous influence function (IF), introduced in [44, 46], about which Huber said as being "the most important single heuristic tool for constructing robust estimates with specified properties" [39].

The influence function tries to describe the impact that an infinitesimal perturbation at the point x has on the estimate, standardized by the mass of the contamination, when the size $n \to \infty$. Using Hampel's notation, this can be expressed by the means of Gateaux type of derivative [44]

$$IF(x; T, F) = \underset{t \to 0}{lim} \frac{T((1-t)F + t\Delta_x) - T(F)}{t}, \tag{4.7}$$

where $t\Delta_x$ is a point mass 1 at x. In other words, IF explains what will happen if we replace the true distribution F, according to an error law, by $(1-t)F + tH$, where t is assumed to be typically between $0.01 - 0.1$. Then, the influence function will measure the asymptotic bias $t \int IF(x, F, H)H(dx)$ caused by contamination tH in the observations. From the practical point of view, one would like IF to be bounded and continuous in x. The first condition would limit the effect of potential gross errors, whereas the second would help achieving insensitivity towards small changes to the whole distribution (caused by rounding or grouping). A recently updated work concerning influence functions, together with mathematical derivations and vast theoretical background, can be found in [44].

Other measures of robustness. There exist also other properties of estimators which should be taken into account when evaluating their properties, or when comparing to an alternative estimator:

- Absolute efficiency $1/(I(F)\sigma_F^2(T))$ where $I(F)$ is the Fisher information and $\sigma_F^2(T)$ is the asymptotic variance. This value should be high for all suitable smooth shapes F, or alternatively, over a strategically selected parametric family of shapes.

- Relative efficiency $\sigma_F^2(T')/\sigma_F^2(T)$, which measures how well the robust estimate performs compared to a standard LS estimator when applied to data with no contamination.

- Equivariance, which means that a systematic transformation of the data will cause a corresponding transformation of the estimator. There exist different types of equivariance defined by the nature of transformation. Details can be found in [47].

4.1.3 Estimators. Location and spread.

If we assume $X_1, \cdots, X_n$ to be independent random variables of common distribution F_n, then any classical estimate $T_n(X_1, \cdots, X_n)$ in the least squares sense will be found as a solution to a following minimisation problem

$$\sum_{1}^{n}(x_i; T_n)^2 \rightarrow min \tag{4.8}$$

The basic idea behind this famous procedure was to optimize the fit by making the residuals very small, assuming that they would follow a certain (most often normal), well-behaved distribution. As it was stated previously, there is no reason why deviations from the assumed distribution or gross errors would not occur in the case of empirical data, leading occasionally to catastrophic results. As an alternative solution, Edgeworth [48] proposed replacing least squares by *least absolute value* criterion, in the context of regression estimator

$$\sum_{1}^{n}|x_i; T_n| \rightarrow min, \tag{4.9}$$

often being referred to as L_1 estimator, whereas L_2 notation corresponds to the least squares work-frame. This criterion was used already before by Laplace for estimating parameter of location in one-dimensional samples, yielding the sample median.

In the robust statistics theory sample *mean* and *median* (or more general L_2 and L_1 types of estimates) define the two extremities, having breakdown points equal to 0 and 1/2, respectively. However many techniques have their breakdown point within $\delta^* \in (0, 0.5)$ and different quality properties, offering therefore many possibilities for choosing a relevant analysis tool.

Various classes of estimators are available, yielding different types of estimates, such as a group of M-estimators (which were briefly described in Section 4.1.2.2), being a generalised concept of maximum likelihood estimation. Moreover, some examples of L- and R- and S-estimates, characterised by different level of robustness will be taken into account. L-estimates (well known example is the median or t-quantile range) come from linear combinations of order statistics, R-estimators are derived from the rank tests and S-estimators are based on minimisation of a scale statistic. The mathematical background and derivation techniques can be found in [35, 44, 47]. Below, some selected robust proposals of location and spread measures will be discussed for low and high-dimensional data.

4.1.3.1 Robust location and scale in low dimensions

The simplest scenario to be considered is a univariate sample $\boldsymbol{x} = (x_1, \cdots, x_n)$ where x_i are independent and identically distributed according to the unknown F. In this case, classical estimates of location and spread, μ and σ, are simply *arithmetic mean* $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and *standard deviation* $s = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2/(n-1)}$. Both of them, being derived from (4.8) have zero-breakdown point and are unbounded (see dotted function in Figure 4.1 in section 4.1.2.2). Alternatively, on the other end-point of the robustness scale stands the *sample median* and the corresponding *median absolute deviation*, defined for the ordered sample

$(x_{(1)}, x_{(2)}, \cdots, x_{(n)})$ where $(x_{(1)} <= x_{(2)} <= \cdots <= x_{(n)})$

$$\underset{i=1,\cdots,n}{median(x_i)} = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd} \\ x_{(n/2)} + x_{((n/2)+1)} & \text{if } n \text{ is even} \end{cases} \tag{4.10}$$

$$\text{MAD} = 1.483 \underset{j=1,\cdots,n}{\text{median}} |x_j - \underset{j=1,\cdots,n}{\text{median}}(x_i)|. \tag{4.11}$$

Here, factor 1.483 is a necessary correction, derived from a certain quantile of the normal distribution, in order to make MAD unbiased. The breakdown value of these estimates equals 50%, and their influence functions are bounded. The price to be paid for these properties is loss of efficiency at the normal model compared to their classical competitors and therefore many other estimates have been proposed.

One way of finding a better balance between robustness and efficiency could be obtained by applying a M-estimator, discussed in [49], which technically speaking is just a simple modification of the classical least squares estimator. It attributes a full weights to observations from the main body of the data and down-weights those coming from the tails. For the data centre it yields

$$\bar{x} = \sum_{i=1}^{n} w_i x_i \Big/ \sum_{i=1}^{n} w_i \tag{4.12}$$

where $w_i = w(d_i) = \psi(d_i)/d_i$, ψ is a bounded influence function, for example proportional to solid lined function in Figure 4.1, $d_i = (x_i - \bar{x})/s$ and s is a corresponding weighted estimate of spread. Depending on the choice of ψ, different M-estimates can be produced, having different robustness properties. Another well-known robust procedure is based on the concept of trimming, called α-

trimmed mean [35], which belongs also to so called L-estimators.

On the other hand, there exist also some good robust estimates of spread. Two commonly applied alternatives to MAD, aiming at increasing its very low efficiency at Gaussian distribution are:

$$S_n = c \text{ med}_i \ \{\text{med}_j |x_i - x_j|\} \tag{4.13}$$

and

$$Q_n = d\{|x_i - x_j|; i < j\}_{(k)} \tag{4.14}$$

with c and d being consistency factors, $k = \binom{h}{2} \approx \binom{n}{2}/4$, and $h = [n/2]+1$. Both estimates do not depend on the location, have 50%-breakdown point, are easy to compute and much more efficient (especially Q_n) than MAD. Additionally, S_n and MAD have discontinuous influence function which is overcome in the case of Q_n. A thorough analysis of the properties of S_n and Q_n can be found in [50].

4.1.3.2 Robust location and scale in higher dimensions

A straightforward generalisation of mean and standard deviation into multidimensional space results in the vector of means $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$ and the variance-covariance (scatter) matrix $\boldsymbol{S}_x = \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_i)^T/(n-1)$. In fact, each of the univariate estimates of location can be applied coordinate-wise leading to column means, column medians etc. These, however, do not take into consideration the multivariate nature of the data and possible co-dependencies within it. The situation is even more complicated in the case of scatter matrix, where such a generalisation is not even possible, therefore numerous studies aiming at robustifying these multidimensional estimates emerged, among which some will be briefly described below.

Multidimensional medians There are many concepts trying to generalize the median into higher dimensional spaces, yielding different definitions of resulting estimates. The most famous seems to be Weber's *L1*-median, also called the *spacial* median [35]. It is defined as a point, $\mu_{L1}(\boldsymbol{X})$, in the multidimensional space found by minimizing the sum of Euclidean distances from the data objects to this point. This can be expressed in the following way

$$\min_{\mu_{L1}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \mu_{L1}(\boldsymbol{X})\| \tag{4.15}$$

where $\|\cdots\|$ is the *L1*-norm. Spacial median is unique [51] and its breakdown point equals 50%. Other multidimensional medians exist, based on differently defined centres of symmetry, such as *simplex median*, *half-space median* and others. An extensive review on multidimensional medians can be found in [52, 53].

Multivariate trimming One of the first methods of calculating a robust covariance matrix is known under the name Multivariate Trimming (MVT). It is based on Mahalanobis distance, where for every iteration a fraction of samples with highest distances are removed and the new scatter matrix is estimated without them. When the estimates of the mean of the remaining 'clean' set converge, the algorithm stops and the final scatter matrix is calculated. Unfortunately the breakdown point of MVT decreases when the dimensionality grows [54].

Minimum Covariance Determinant (MCD) and Minimum Volume Ellipsoid (MVE) estimators MCD, a method of Rousseeuw [55], attempts to find a covariance matrix with minimal determinant, which would include at least h data points, where h is defined by the user and determines the breakdown point ($\delta^* = (n-h+1)/n$) of the estimator. The subsequent MCD estimates of location ($\hat{\boldsymbol{\mu}}$) and spread ($\hat{\boldsymbol{\Sigma}}$) are the arithmetic mean and covariance matrix (multiplied by consistency factor), defined on the h-subset. Moreover, in order to increase the efficiency on the finite-sample, a reweighing step was proposed by assigning to each $\boldsymbol{x}_i$ a weight w_i, for instance $w_i = 1$ if $(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}) <= \chi^2_{p,0.975}$, and $w_i = 0$ otherwise [56], obtaining

$$\hat{\boldsymbol{\mu}}_R(\boldsymbol{X}) = \left(\sum_{i=1}^{n} w_i \boldsymbol{x}_i \right) \Big/ \left(\sum_{i=1}^{n} w_i \right) \tag{4.16}$$

$$\hat{\boldsymbol{\Sigma}}_R(\boldsymbol{X}) = \left(\sum_{i=1}^{n} w_i (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_R)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_R)^T \right) \Big/ \left(\sum_{i=1}^{n} w_i - 1 \right) \tag{4.17}$$

MCD estimators have bounded influence function and can take up to $0.5n$ outliers. However, in order to assure that the robustness/efficiency ratio stays reasonable, $h \sim 0.75n$ is recommended. Due to the fact that the MCD algorithm iterates through all possible h-subsets, it might become computational exhaustive. In order to overcome this risk, faster versions of the algorithm were elaborated [57].

MVE, also proposed by Rousseeuw [58], searches for the minimal ellipsoid covering at least half of the data points, however this estimator has very low rate of convergence and hence, efficiency.

Stahel-Donoho outlyingness A conceptually similar technique, consisting of applying least squares estimates to previously 'cleaned' data as in equations (4.16) and (4.17), was presented before MCD and MVE by Stahel [59] and Donoho [60], being the first high-breakdown estimator. They introduced independently a measure called *outlyingness* of $\boldsymbol{x}_i$

$$\text{outl}(\boldsymbol{x}_i) = \underset{d}{max} \frac{\left|\boldsymbol{x}_i^T \boldsymbol{d} - \text{median}_{j=1,\cdots,n}(\boldsymbol{x}_j^T \boldsymbol{d})\right|}{\text{MAD}_{j=1,\cdots,n}(\boldsymbol{x}_j^T \boldsymbol{d})}, \tag{4.18}$$

where $\boldsymbol{x}_j^T \boldsymbol{d}$ is a projection of $\boldsymbol{x}_j$ on the direction d (it is therefore related to projection pursuit techniques) and maximum is taken over all directions. Later, relevant weights w_i are given to $\boldsymbol{x}_i$ according to outl($\boldsymbol{x}_i$) and the estimates calculated. Even-though the estimator has good robustness properties, its usage has serious limitations due to the calculation intensiveness.

4.1.4 Regression estimators

A classical multiple regression model can be expresses as

$$y_i = x_{i1}\theta_1 + \cdots + x_{ip}\theta_p + \varepsilon_i, \text{ for } i = 1, \cdots, n, \tag{4.19}$$

where y_i and $x_{i1}, \cdots, x_{ip}$ are respectively response and explanatory xariables, ε_i is the error term (in the classical case $\varepsilon_i \sim N(0, \sigma)$ is assumed) and n denotes the sample size. The aim of the regression is trying to explain a behaviour of certain response variable $\boldsymbol{y}$, by means of linear combination of some measured quantities (explanatory variables). Under this model, a vector of coefficients $\hat{\boldsymbol{\theta}}$ is to be estimated. As previously mentioned, this is usually done by optimising the least squares criterion

$$\underset{\hat{\boldsymbol{\theta}}}{Minimize} \sum_{i=1}^{n} r_i^2, \tag{4.20}$$

which, as we already know, is very sensitive to outliers, with the breakdown point $\delta^* = 1/n$ and unbounded influence function. An outlier, in the regression sense, is any point $(x_{i1}, \cdots, x_{ip}, y_i)$ deviating from the linear trend, set by the majority of the data. Rousseeuw [47] distinguishes 2 most common types: *y-direction* outliers and *x-direction* outliers (also called *leverage points*), which are denoted as group 1 and group 2 in Figure 4.2, respectively. Not all leverage points, however, have destructive influence on the fitted regression model (group 3 outliers). Alternatively, it is also possible to find the erroneous observations which fit within the data range on both x and y directions but still do not follow the data pattern and as a result, would most likely tilt the regression line (group 4). Figure 4.2 illustrates all these scenarios for the case of univariate regression, however it is clear that the difficulty in identification of outliers grow together with dimensionality of the data set.

Some of the first attempts in 'robustyfying' regression estimation were inspired by previously elaborated robust estimators of univariate location and were already briefly discussed in Section 4.1.3. L_1- or *least absolute values* regression (its objective function minimizes $\sum_{i=1}^{n} |r_i|$ with respect to $\hat{\boldsymbol{\theta}}$) was firstly proposed by Edgeworth [48]. This technique gives an improvement in comparison to L_2 fit, as its breakdown point reaches 25% for uniform or normal $x's$. Unfortunately δ^* drops to zero if 'bad' leverage points are present in the data, since this estimator bounds the influence of $y_i^{\prime}s$ and cannot handle heavily outlying

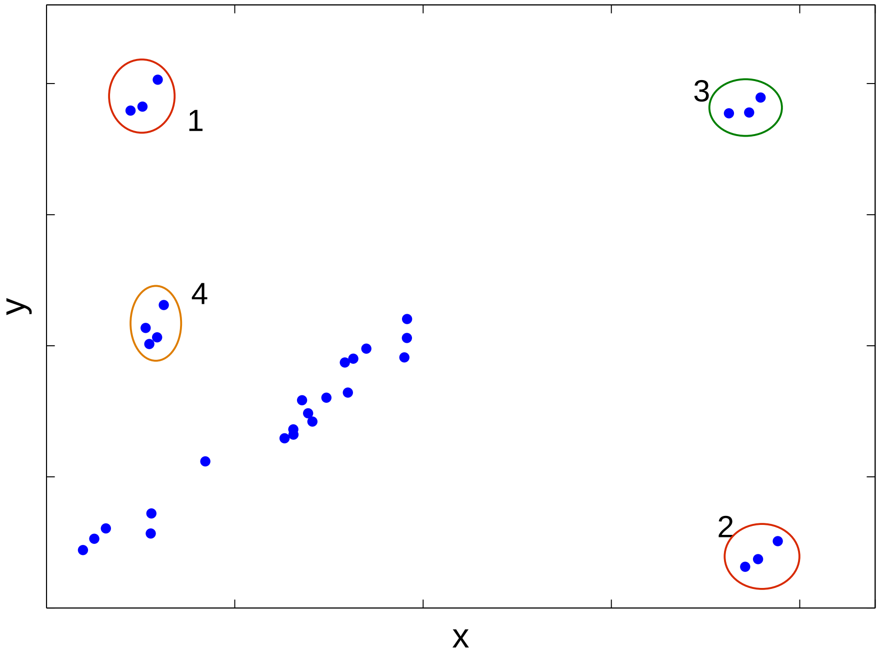

Figure 4.2: Types of regression outliers: 1. y-direction outliers; 2. x-direction outliers; 3. leverage points which do not disturb the fit; 4. points with non-outlying values on x and y but not following regression line.

$x's$. The situation is similar if M-estimator is considered. Huber arrived at his famous minimax result

$$\psi(x) = max(-k, min(k, x)), \tag{4.21}$$

called Huber's estimator, being the M-estimator for the least favourable distribution (assumed to be normal in the middle and exponential in the tails) over the neighbourhood of the model. The limiting cases, when the assumed fraction of contamination $\varepsilon \to 0$ or 1, and when $k \to \infty$ or 0 return arithmetic mean and median. Huber's method shows improved results towards both L_1 and L_2 techniques, however the influence of leverage points is still not treated and therefore $\delta^* = 0$ for long-tailed designs (in cases of 'nicely' distributed $x's$, the breakdown point is never higher than 25%).

In order to overcome the risks of destructive effects of the leverage points on regression model, the range of positive breakdown point estimators were elaborated. *Generalised* M-estimator [42], repeated median [61], least median of squares [55] and least trimmed squares [47] belong to the most known techniques, therefore they will be discussed below in more details.

4.1.4.1 Generalized M-estimators

This group of estimators, also called *bounded influence estimators*, was designed with a clear purpose of safeguarding against the outlying $\boldsymbol{x}_i$, by introducing some weighting functions $w(\boldsymbol{x}_i)$ and $v(\boldsymbol{x}_i)$ inside of the standard M-estimator expression

$$\sum_{1}^{n} w(\boldsymbol{x}_i)\psi(\boldsymbol{r}_i v(\boldsymbol{x}_i))\boldsymbol{x}_i = \mathbf{0}. \qquad (4.22)$$

Weights should be selected so that the entries coming from the main data body receive full influence, whereas outliers become down-weighted. Different sets of weights were proposed such as Huber [42], Hampel or Andrews [62] and others (see Figure 4.3), but in general, $\delta^* <= 30\%$, and it turns out that its value decreases as a function of p (number of regression coefficients), which might still be unsatisfactory in certain situations.

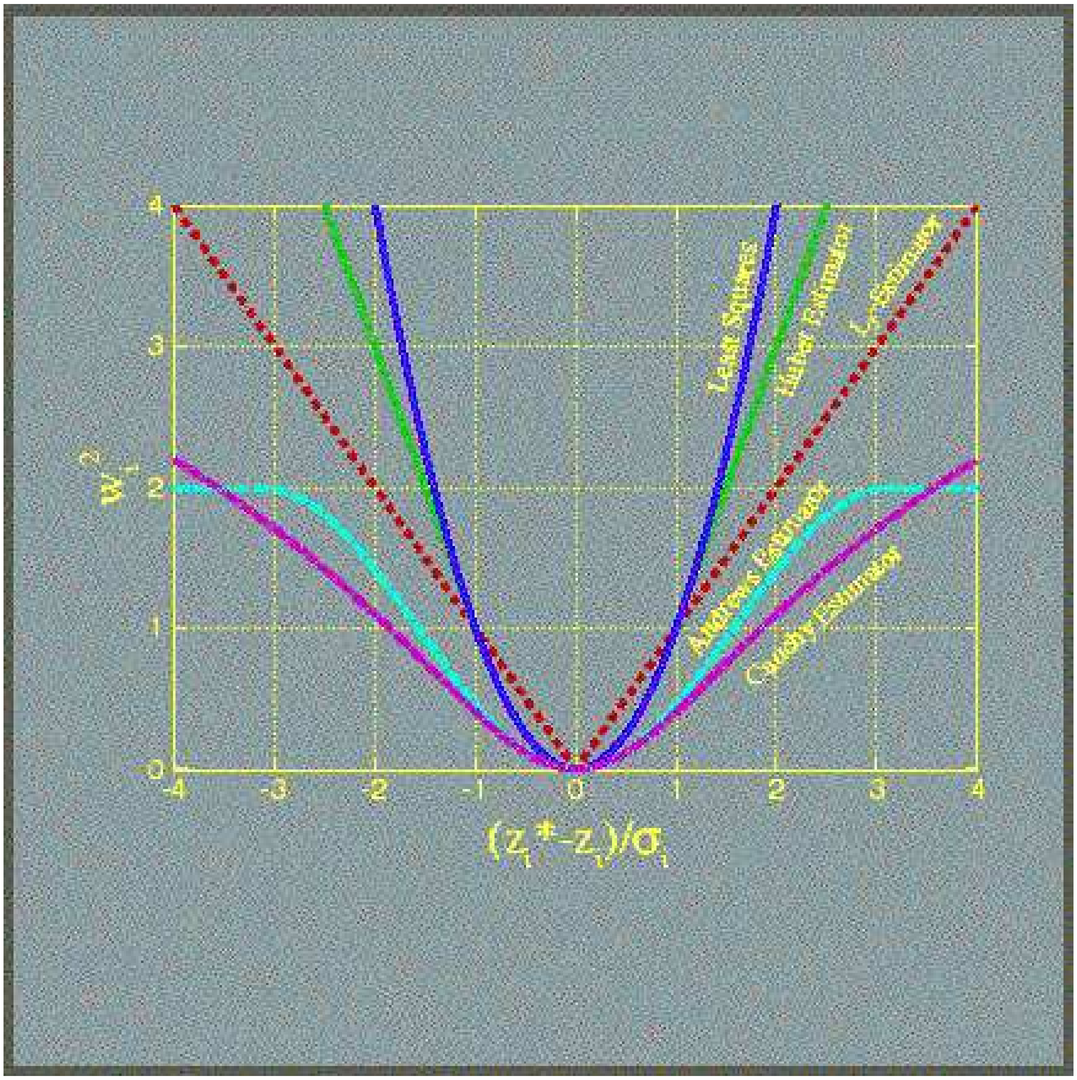

Figure 4.3: Objective functions for least squares (blue), Huber (green), L1-estimator (dashed red), Andrews (light blue) and Cauchy estimator (pink). Source: http://esd.lbl.gov/itough2

4.1.4.2 Siegel's repeated median

The first high breakdown regression estimator was proposed by Sigel in 1982 [61] under the name *repeated median* algorithm. Parameter vector $\boldsymbol{\theta}$ is defined coordinate-wise as follows

$$\hat{\theta_j} = \underset{i_1}{med}(...(\underset{i_{p-1}}{med}(\underset{i_p}{med}\ \theta_j(i_i, ..., i_p)))...), \tag{4.23}$$

for any p observations $(\boldsymbol{x}_{i_1}, y_{i_1}), ..., (\boldsymbol{x}_{i_p}, y_{i_p})$. The repeated median procedure will remain bounded whenever more than $1/2(n + p - 1)$ points will come from non-contaminated sampling, while the remaining points can be arbitrarily moved. That implies that its asymptotic breakdown value is 50%, when $n \to \infty$ and p is fixed. Since the algorithm requires iterating throughout all possible subset of p variables, the computational complexity of n data points is $O(n^p)$, which makes it fairly calculation intensive. Moreover, due to its coordinate-wise construction, the repeated median is not equvariant for linear transformations of $\boldsymbol{x}_i$ [47].

4.1.4.3 Least median of squares

Next step in the attempts for making classical regression estimates more robust was proposed by Rousseeuw [55]. He approached the task from a different angle, and instead of defining a new function of errors (replacing 'squares' by another relation), he undermined the rightfulness of the summation sign involved in the least squares expression. By saying ironically "(...) as if the only sensible thing to do with n positive numbers would be to add them", he replaced the sum with the median and arrived at the *least median of squares* (LMS) estimator

$$\underset{\hat{\boldsymbol{\theta}}}{Minimize} \; \underset{i}{med} \; r_i^2. \tag{4.24}$$

In the case of simple bi-variate regression, LMS fit corresponds geometrically to a line laying in the middle of the narrowest stripe covering $[n/2]+1$ of the data points, where $[n/2]$ denotes an integer part of the division. The breakdown point of the method $\delta^* = ([n/2]-p+2)/n$ which asymptotically equals $1/2$ when $n \to \infty$ and p is fixed. Moreover, LMS is equivariant with respect to the linear transformations on $\boldsymbol{x}_i$. The biggest two drawbacks are low efficiency when the errors actually are normal and fairly slow convergence rate $n^{-1/3}$ [47, 55].

4.1.4.4 Least trimmed squares

Another proposal of Rousseeuw, aiming at overcoming low efficiency rate of LMS is *least trimmed squares* estimator (LTS)

$$\underset{\hat{\boldsymbol{\theta}}}{Minimize} \sum_{i=1}^{h} (r_i^2)_{i:n}, \tag{4.25}$$

where $(r^2)_{1:n} <= \cdots <= (r^2)_{n:n}$ are squared and ordered residuals coming from the LMS fit and h is the number of points to be excluded, specified by the user. This technique is actually equivalent to least squares performed on non-contaminated, outlier free data, after identifying them by a robust technique. As h might vary from 0 to $n/2$, it results in different efficiency-robustness ratios, but in general the breakdown point is $(n-h+1)/n$. The algorithm converges satisfactorily as $n^{-1/2}$ and its efficiency is superior over LMS, however it is fairly computation intensive. More details about LTS can be found in [47].

4.1.4.5 Other robust estimators of regression.

Many other definitions of robust regression have been made, offering a wide range of tools for practitioners in terms of breakdown/efficiency/calculation speed ratio. For example, as a generalisation of LMS Rousseeuw and Yohai [63] introduced a class of S-estimators. Afterwards, Yohai improved the method by proposing MM-estimator [64] in order to augment the efficiency, and recently Rousseeuw came up with his MCD regression [65] for treating multivariate problems.

4.1.5 Robust multivariate models

In the previous section it was demonstrated how to safeguard a regression model from unwanted influence of outliers in the data. However, presence of outliers or deviations from assumed underlying distribution affect all kind of models with least squares cost function. In this section, some of the popular models used in multivariate data analysis, will be briefly presented together with their robust alternatives.

4.1.5.1 Robust Principal Component Analysis

If the focus of the analysis lies in summarising patterns, dependencies or differences within the data set, decomposition methods, such as Principal Component Analysis (PCA) might be of interest. PCA is a 2-way linear subspace-based technique, extensively used in many application fields. A PCA model is presented in Equation (4.26)

$$x_{ij} = \sum_{r=1}^{R} t_{ir} p_{jr} + e_{ij} \quad i = 1, ...I; \; j = 1, ...J \tag{4.26}$$

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} contains model's residuals. In brief, PCA maps objects and variables to lower dimensional spaces where exploration and visualization becomes easier. This is completed by finding a sum of vector products, called scores ($\mathbf{t}$) and loadings ($\mathbf{p}$) which are orthogonal and determined by maximizing the variance explained by them. Those vector products, being a linear combinations of the original variables (or objects), are called principal components and often already a small number is sufficient to explain the data variation in a satisfactory way. More details about the classical PCA method can be found in the literature, for example in [16–18].

An outlier in the PCA context can be defined as an observation or object having either large orthogonal (OD) or score distance (SD). Large OD means that the observation lies far away from the subspace spanned by the correct eigenvectors and high SD values indicate elements for which the projection into the model lies far from the main data bulk within that subspace. Due to the least square origins of PCA, many attempts have been made in order to improve the model quality, when data contamination is present, but essentially, two distinct approaches can be specified. The first group of methods consist of robust covariance matrix. The reasoning for that is quite straightforward, namely because PCA decomposes the covariance matrix where PC's are the eigenvectors. This implies that if the covariance matrix will be replaced by its robust version, it will also lead towards robust PCA. Some of the methods for robustifying the covariance matrix, as MVT, MVE or MCD were already discussed in Section 4.1.3.2, however their functionality is restricted to few dimensions. The second group of robust PCA techniques is based on Huber's famous Projection Pursuit (PP) [66]. PP searches for structure in the high dimensional data by projecting it into lower dimensional space, where a certain criterion would be maximized. Depending on this criterion, also called *projection index*, a particular method can be obtained. PP can be applied to high-dimensional data, due to its sequential construction, however it might be fairly computationally exhaustive. A compromise seems to be reached by combining the two methods in various ways, which often yields a faster algorithm, still functional in high dimensions. Some of the most important contributions aiming at constructing a robust PCA technique will be discussed below in more details.

Robust covariance matrix approach

As previously mentioned, replacing the classical scatter matrix by a robust version will lead to a robust PCA model. Many simultaneous studies were proposed to determine which robust estimator should be used. Some of the most commonly discussed are the M-estimators [49], S-estimators (MVE) [47] or the one-step reweighed MCD estimator [58]. Since the M-estimators are characterised by breakdown point decreasing towards zero when dimensionality grows, the high-breakdown methods are preferable. According to Croux and Haesbroek [67] the theoretical results (influence function and efficiency) favour S-estimators however low or moderate dimensionality should be kept due to extremely high computation time. Moreover if the number of variables exceeds $(n(1-\alpha))$ both

MVE and MCD break and are no longer defined, which is a serious limitation in certain application fields. Some efforts have been made to overcome this inconvenience by compressing the data by means of PCA which, as an orthogonal data transformation, will not influence the distances between the objects, and use the resulting PC's for derivation of the robust estimates [68]. Alternatively, projection-based methods may be useful in the case of high dimensional data.

Robust PCA by Projection Pursuit

Projection Pursuit represents a wide range of projection methods. In brief, PP searches for the projections in the multidimensional data space, which mostly expose outliers. This is being done by finding a direction $\boldsymbol{v}_p$ which would optimise a projection index $\rho(\boldsymbol{X}\boldsymbol{v})$ being a robust measure of spread. In fact, a classical PCA is the special case of PP with the classical variance as projection index. Therefore if a robust measure of spread will be applied, it will lead to a robust PCA. Additionally, the robustness of resulting method will also depend on the selection of the data centre estimator. One of the first algorithms was introduced by Li and Chen [69], with Hubers M-estimate as projection index and weighted mean as a data centre. This method was proven to inherit a high breakdown point of the robust scale estimator, however, appeared computational exhaustive. Later Xie et al. [70] came up with generalized simulated annealing as an optimisation procedure in calculation of PP, and used a robust measure of spread based on L_1-norm, such as MAD, still yielding a slow algorithm. Afterwards, countless studies on different variants of robust PCA aiming at speeding up the algorithms and maintaining good robustness features, were elaborated, among which, few will be described below.

Summing up briefly, PCA based on the Projection Pursuit has very attractive features. In particular, by using scale measures as MAD or Q_n, the resulting PCA inherits decent robust properties. Moreover, the calculation process can be stopped when the satisfactory amount of PC's is reached, which is not possible when PCA based on robust covariance matrix is considered. Finally, the newest algorithms allow high dimensionality situations and variables exceeding the number of objects.

CR algorithm

One of the most influential contributions in the PP-based PCA field is due to Croux and Ruiz-Gazen. They came up with a faster and relatively simple algorithm (CR-algorithm) with L_1-median as data centre and Q_n as the projection index [71]. In order to find the optimal direction, it was proposed to search within a set constituted only by the data points directions. This however means that the solution will be approximative and losses of efficiency expected, especially when the number of data objects is small ($n < p$). In this case, a new algorithm (GRID) was recently presented [72] as a complement to the previously proposed CR algorithm.

RA-PCA

An improvement of the CR-algorithm was also proposed in [73], where reflection-based algorithm for PCA (RA-PCA) was defined. The novelty of the approach consists of the initial compression of the data by the means of singular value decomposition (SVD) of $\boldsymbol{X}_{n,p}$ if $n < p$. This transformation reduces the data space to the affine subspace spanned by the n observations. Thanks to that, the procedure can deal with low- and high-dimensional data without loosing generality. Afterwards, the main idea of RA-PCA is to search for directions with maximal value of the projection index, which in this case is the Q_n estimator. In order to reduce calculation time the same procedure as in the CR method was used for initialising the choice of directions. The authors claim that the RA-PCA is more stable numerically and faster than CR method.

ROBPCA

Another, already well established proposal of robustifying the PCA model is called ROBPCA and was developed by Hubert et al. [74]. This method combines projection pursuit ideas with robust scatter matrix estimation based on the MCD estimator. As in the case of RA-PCA, the algorithm starts by reducing the data to the n-dimensional full-rank subspace by the means of PP. Here the measure called outlyingness w_i is maximised throughout each *i-th* direction

$$w_i = \underset{\|p\|=1}{argmax} \frac{\left|\boldsymbol{x}_i\boldsymbol{p}^T - \mu_{MCD}(\boldsymbol{x}_i\boldsymbol{p}^T)\right|}{\sigma_{MCD}(\boldsymbol{x}_i\boldsymbol{p}^T)}. \tag{4.27}$$

Then, a preliminary scatter matrix S_0 is constructed in order to select k components to be retained This results in a k-dimensional subspace which fits the data well. Finally, the data points are projected into that subspace and PCA is performed on that projection, yielding robust PC's. ROBPCA is location and orthogonal equivariant (the scores do not change under shift or orthogonal transformations). It can be computed relatively fast and has breakdown point of MCD estimator. Finally, as a 'by-product' ROBPCA can deliver a diagnostic plot which is a great tool to displaying and identifying outliers.

S-PCA

S-PCA is a very fast and robust version of classical PCA introduced by Locantore in [75]. This method exploits the robustness feature of the median, combined with the projection pursuit framework and is conceptually fairly simple. In the first step the robust centre of the data is defined as the L_1-median. Then, all data points $\mathbf{x}_i$ are centred and down-weighted by the inverse of these distances, which is equivalent to projection on the unit radius hyper-sphere with the centre in $L1$:

$$\mathbf{x}_i^p = \frac{\mathbf{x}_i - \boldsymbol{\mu}_{L1}(\mathbf{X})}{\|\mathbf{x}_i - \boldsymbol{\mu}_{L1}(\mathbf{X})\|} + \boldsymbol{\mu}_{L1}(\mathbf{X}) \tag{4.28}$$

where $\mathbf{x}_i^p$ is the ith data object projected onto the sphere and $\|\mathbf{x}_i - \boldsymbol{\mu}_{L1}(\mathbf{X})\|$ is the Euclidean distance to the robust centre of the data $\boldsymbol{\mu}_{L1}$. In this setting, the influence of outliers manifesting normally large distances in the denominator of the equation, will be bounded. In the next step, a classical PCA is carried out on the down-weighted data and the robust scores and loadings are obtained by projecting the original data on the resulting PCs. Further, in order to identify the outlying samples, the usual detection methods based on Robust and Orthogonal distances (RD and OD, respectively) are to be applied.

Another version of S-PCA was also defined. By taking into account different scales of the variables, data objects are projected on hyper-ellipse instead of hyper-sphere, resulting in Elliptical PCA (EPCA) [75], however this method seems to have problems with consistency.

Other approaches

Many other approaches towards robust PCA exist, for example, consisting of replacing the classical least square function by its robust alternative (idea already used in robust regression estimation). Least Trimmed Squares or M-estimator of the PCA residuals was suggested in [76]. However, since it is not the scope of this report to review all the possible robust PCA models, but rather to give a general overview of the available methods, this objective is assumed to be met.

4.1.5.2 Robust PLS

Partial Least Squares (PLS) regression is a multivariate calibration method, aiming at estimating uni- or multi-variate response (assigned $\boldsymbol{y}$ and $\boldsymbol{Y}$, respectively) by the means of high-dimensional regressors ($\boldsymbol{X}$). Essentially, it was invented by Wold [22, 23] as a remedy for co-linearity problem in the multi-dimensional data and since then, it has been broadly used by practitioners in various application fields (see [24] for tutorial and [25] for more detailed description). Two algorithms are commonly used for PLS: non-linear iterative partial least squares NIPALS [22] or SIMPLS [77], yielding the same result when univariate response is considered, but differing slightly in the multivariate case (for more details consult reference [77]). Nevertheless, both algorithms optimise a least squares criterion, and hence, are sensitive to data contamination. In this section the emphasis will be put on SIMPLS approach as it appears to be faster and more intuitive for interpretation (all components are linear combinations of the original variables), however some robust versions concerning NIPALS will aslo be mentioned.

The classical PLS problem (in SIMPLS workframe) might be expressed as a bilinear model [78]

$$\boldsymbol{x}_i = \bar{\boldsymbol{x}} + \boldsymbol{P}\tilde{\boldsymbol{t}_i} + \boldsymbol{\varepsilon}_i \tag{4.29}$$

$$\boldsymbol{y}_i = \bar{\boldsymbol{y}} + \boldsymbol{A}^T\tilde{\boldsymbol{t}_i} + \boldsymbol{\delta}_i \tag{4.30}$$

where $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ are the means of $\boldsymbol{X}$ and $\boldsymbol{Y}$, vectors $\tilde{\boldsymbol{t}_i}$ are the scores, matrix $\boldsymbol{P}$ contains the loadings and $\boldsymbol{A}$ is the regression slope matrix. Residuals of the models are stored in $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\delta}_i$. Essentially, the objective is to explain $\boldsymbol{Y}$ as good as possible and at the same time obtain a reasonable relation between $\boldsymbol{X}$ and $\boldsymbol{Y}$. For this purpose,

$$\operatorname{cov}(\tilde{\boldsymbol{Y}}\boldsymbol{q}_a, \tilde{\boldsymbol{X}}\boldsymbol{r}_a) = \boldsymbol{q}_a^T \frac{\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{X}}}{n-1}\boldsymbol{r}_a = \boldsymbol{q}_a^T \boldsymbol{S}_{xy}\boldsymbol{r}_a \tag{4.31}$$

is to be maximised with respect to normalised weight vectors $\|\boldsymbol{q}_a\| = \|\boldsymbol{r}_a\| = 1$. Here, $\tilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{X}}$ are mean-centred data matrices and $\boldsymbol{S}_{xy}$ is the symmetric cross-covariance matrix between X and Y. Then, the elements of the score matrix are found as linear combinations of the mean-centred data: $\tilde{t_{ia}} = \tilde{\boldsymbol{x}_i}^T\boldsymbol{r}_a$. SIMPLS takes the first left and right eigenvector of $\boldsymbol{S}_{xy}$ to be $\boldsymbol{r}_1$ and $\boldsymbol{q}_1$. In order to find the rest of weight vectors (for $a = 2, \cdots, k$) the orthogonality constraint, $\sum_{i=1}^{n} t_{ia}t_{ib} = 0$ is imposed and the cross-covariance matrix deflated by first calculating

$$\boldsymbol{p}_a = \boldsymbol{S}_x\boldsymbol{r}_a/(\boldsymbol{r}_a^T\boldsymbol{S}_x\boldsymbol{r}_a), \tag{4.32}$$

where $\boldsymbol{S}_x$ is the empirical scatter matrix of $\boldsymbol{X}$. Afterwards, an orthonormal base $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_a\}$ of $\{\boldsymbol{p}_1, \cdots, \boldsymbol{p}_a\}$ is constructed and the deflation performed

$$\boldsymbol{S}_{xy}^a = \boldsymbol{S}_{xy}^{a-1} - \boldsymbol{v}_a(\boldsymbol{v}_a^T \boldsymbol{S}_{xy}^{a-1}) \tag{4.33}$$

Finally the responses are regressed on k chosen components from $\tilde{\boldsymbol{T}}$ according to

$$\boldsymbol{y}_i = \boldsymbol{\alpha}_0 + \boldsymbol{A}^T \tilde{\boldsymbol{t}}_i + \boldsymbol{\delta}_i. \tag{4.34}$$

Since the PLS uses least squares criterion when calculating weights, loadings, scores and regression coefficients, it is clearly exposed to the damaging effect of outlying observations.

Two main strategies were adapted in order to circumvent this problem: 1. down-weighting of the outliers; 2. robust estimation of the covariance matrix. The first group of methods aimed at replacing all steps of classical least squares regression by a robust alternative by introducing different weighting functions to be applied to the data. This, however, led to high computational cost and loss of efficiency. An alternative solution would be replacing only selected steps resulting in 'semi-robust' approaches. Among others, the work of Wakeling and Macfie [79], Cummins and Andrews [80], Gil and Romera [81] and Pell [82] should be acknowledged. Second group of methods is related to SIMPLS, which is based on the cross-covariance $\boldsymbol{S}_{xy}$, empirical covariance $\boldsymbol{S}_x$ and the multiple regression step. Hubert and Vanden Branden [78] proposed a method called RSIMPLS, based on ROBPCA for robust estimation of $\boldsymbol{t}_i$ which are later used for the regression step. As an alternative to that, Snereels et al. [83] introduced another weighting procedure called partial robust M-regression (PRM) which uses continuous weights to diminish gradually the influence of outliers. Simulation study showed that PRM outperforms RSIMPLS in terms of statistical efficiency, however the method can so far be applied only to univariate response problems and due to reliance on GM-estimator, its breakdown point reaches at most 30%.

4.1.5.3 Robust PARAFAC model

Until now only 2-way methods were considered, where a term '2-way', not to be confused with '2-dimensional', refers to data sets which can be arranged in a matrix, having two distinct directions (modes), *e.g.* objects and variables. '3-way', consequently, denotes structures accommodated in a cubic form, where a third direction (for example location or time) is added. Practice shows that many types of empirical data can be accommodated in 3- or even multi-way structures, which creates needs for relevant data analysis techniques.

A PARAFAC model, introduced by Harshman [19] and popularized by Bro [10] and Smilde et al. [2], is a tri-linear generalization of PCA, which decomposes a data cube into the sum of triple vector products, called loadings. The most common way of writing it is following

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk} \tag{4.35}$$

where x_{ijk} is an element of the array $\boldsymbol{X}$, $\boldsymbol{A}(I \times R)$, $\boldsymbol{B}(J \times R)$ and $\boldsymbol{C}(K \times R)$ are the orthogonal matrices with elements a_{ir}, b_{jr} and c_{kr} respectively. R is the number of components and e_{ijk} is the error term. If the experimental data fulfils the tri-linear assumption (required invariability of the component profiles across the different data slices with different weighting coefficients for each slice, see [2,19]), the application of a PARAFAC model is usually superior to its bilinear counterpart PCA, which is often applied after unfolding the data cube into a matrix form. The reasons for this are numerous. First of all, PARAFAC takes into account the interrelations existing in all three data directions. Moreover, the problem of rotational freedom, typical to PCA is solved here as PARAFAC provides a unique solution (up to the scaling constant, sign and permutation ambiguities) [2]. In addition, PARAFAC model resolves each mode separately, giving a straightforward physical interpretation for the depth, station and variable 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, it does not tend to over-fit, as is often the case of PCA.

In spite of the benefits that may be gained from applying PARAFAC model, some drawbacks also exist. The most important one is that the real data do not always conform adequately with a tri-linear assumption. In this cases, the model might return degenerate solutions. Degeneracy might also occur when contaminated samples (outliers) are present in the data, due to the cost function of PARAFAC depending on the least squares criterion. This objective function can be expressed as

$$\left\| \boldsymbol{X} - \hat{\boldsymbol{X}} \right\|_F^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (x_{ijk} - \hat{x}_{ijk})^2 \rightarrow \min, \tag{4.36}$$

which is equivalent to minimisation of the squared residuals. In fact, the algorithm usually used for optimising expression (4.36) is called Alternating Least

Squares (ALS) and it consists of iterative minimisation of least squares criterion [2].

As it was described in section (4.1.5.1), many approaches toward robust PCA exist in the literature. Unfortunately, it is not the case for PARAFAC, where only few studies address the issue of robustness. In [84], Rui and Bro developed an outlier detection technique based on jack-knifing, where influence of each sample is evaluated by resolving PARAFAC model without this sample. If scores and loadings change significantly, that indicates an outlying sample. This method might be fairly computation expensive if large sample size is considered. Moreover it will suffer from so called *masking effect*, when outliers form clusters. A better method for dealing with outlying samples was proposed by Engelen and Hubert [85]. The procedure is looking for an ideal h-subset of the samples with the minimal squared residuals. The value $h \in (0.5, 1)$, defining the breakdown point of the method, is the fraction of I samples set by the user, as for MCD or LTS. The algorithm starts by applying ROBPCA on the unfolded matrix X^{IxJK} and finding the initial h-set as data points with the h smallest robust distances. Afterwards, a classical PARAFAC is performed on the h samples and scores are computed for each observation X_i together with residual distances. A new h-subset is constructed by storing samples with smallest residual distance and the whole procedure is iterated until changes in h subset become insignificant. At the end a reweighing step, for increasing efficiency is performed. The simulation study [85] shows that this method outperforms the classical PARAFAC when data contamination is present, and also, it delivers reasonable results (in terms of statistical efficiency) for the clean data scenario.

4.1.5.4 Other robust multivariate techniques

In this section some robust procedures for PCA, PLS and PARAFAC modes were discussed. However, apart of these frameworks, each model with least squares cost function is at risk when outliers or other deviations from model's assumptions occur in the analysed data. If dealing with the real (non-simulated) data, this is more of a rule than exception. Some of the ideas presented above, such as using robust cost function or robustifying a particular 'weak' point of the algorithm, find applications in other multivariate settings. Various robust proposals were given for models, such as robust Principal Component Regression (PCR) [68, 86, 87], classification tools: Quadratic Discriminant Analysis (QDA) [88] or Soft Independent Modelling of Class Analogy (SIMCA) [89–91] or a 3-way Tucker3 model [92]. For reviewing papers on robust multivariate methods used in data analysis consult references [93–96].

4.1.6 Discussion and conclusions

Empirical data often deviates from the commonly assumed normal distribution and contains corrupted observations, manifesting themselves in the form of outliers. Both types of nuisance can significantly deteriorate the performance of conventional data analysis methods based on least squares criterion. Therefore, there is no doubts that smart robust methods can largely contribute to improving data analysis tools.

A broad theoretical background of robust procedures was given here, however, since practical considerations usually diverge from purely theoretical point of view, a few important issues will be addressed and discussed here.

In theory, the ’perfect’ robust estimator is characterised by a high breakdown point, bounded influence function and high efficiency rate at normality. However, data sets with a high percentage of outliers would most likely be discarded and generation of higher quality data recommended. Therefore, robust estimators with the 50%-breakdown point might be helpful for general assessment of data quality and outlier identification, more than for actual analysis. This is also supported by the fact that high-breakdown robust methods are usually computational intensive and have lower efficiency rate then classical techniques. If the method allows, it is advisable to tune the breakdown point according to needs of the analysed data, in order to reach a compromise between good robust properties and quality of the estimate. As a contra-argument, it can be argued that if good robust techniques, delivering reliable results were at hand, the time and cost of re-running the experiments for obtaining better data quality could be saved.

A relevant question emerges here: when should robust techniques be used. To answer this, the author agrees with the opinion of Peter Rousseeuv [47], who claims that by principle both, a classical and a robust procedure should be applied. If their performance is similar (meaning no or insignificant outlier presence), the results of classical methods should be taken into account, as they yield more efficient estimates. On the contrary, if the two methods are significantly different, it is a signal that part of the data is corrupted and care should be taken in order to avoid misleading results.

To conclude, the robustness theory and methods are able to provide tremendous benefits to practitioners, dealing for example with multivariate data analysis. In order to facilitate that however, implementations of robust procedures should be available for some standard statistical data analysis programs such as R, S-plus or Matlab. Some attempts have already been undertaken resulting in few robust toolboxes (like TOMCAT or LIBRA for Matlab and *rrcov* for R), however more work is needed to propagate the usage of robust methodologies within the world of applications.

4.2 Robust PCA for automatic detection of Rayleigh and Raman scattering in fluorescence data.

As already mentioned in Section 2.2.1, the presence of Rayleigh and Raman light scatter effects leads to a deterioration of the PARAFAC performance in fluorescence excitation-emission (EEM) data. In this section, a method for identifying these scatters, based on Spherical Principal Component Analysis (S-PCA) will be given and compared to the earlier approach based on ROBPCA [14] (for the full study consult Papaer B).

The data used for this study consists of 23 laboratory prepared samples including 4 fluorophores mixed in different quantities (these are: phenylalanine, 3,4-dihydroxyphenylalanine (DOPA), 1,4-dihydroxybenzene and tryptophan). Excitation-emission landscapes were produced, measuring the emission spectra ranging from 250 to 482 nm (with 2 nm intervals), for excitation wavelengths within 230 and 315 nm every 5 nm. Visible 1st and 2nd order Rayleigh ridges

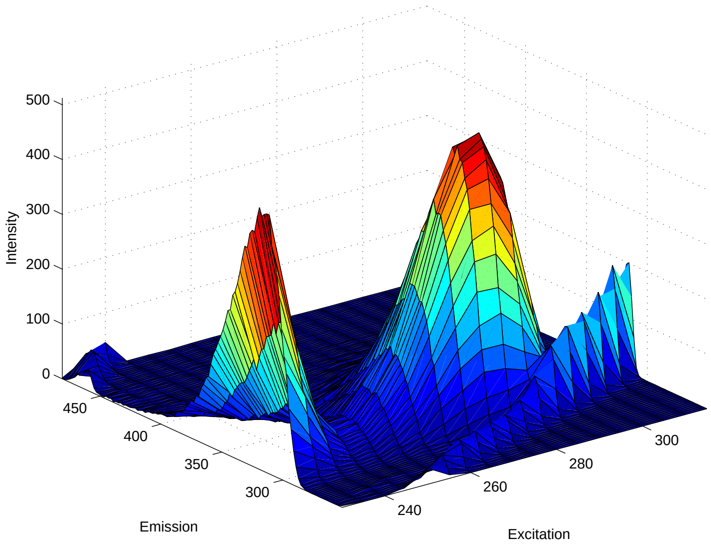

Figure 4.4: Dorit data with the visible Rayleigh scatter.

are present in every data sample, which is illustrated in Figure 4.4. The method for identification of these undesired scatter lines is based on S-PCA, assuming that the scatter is a set of outlying observations, which differ in behavior from the rest of the data. By slicing our three-way data array $\mathbf{X}$ according to the excitation and emission modes and, subsequently, transposing the resulting matrices, the scattering effect will be placed in the rows of those matrices. As a consequence, these rows can be perceived as sample-wise outliers and hence, detected by a robust PCA method applied to each matrix separately. Both, the slicing and transposing operations are illustrated in Figure 4.5.

In order to identify the outlying samples, Robust and Orthogonal distances (RD and OD, respectively) are calculated: $RD_i = \sqrt{\sum_{j=1}^{k} t_{ij}^2 / l_j}$, where $\mathbf{t}_j$ is a vector of robust scores for the jth component and $\mathbf{l}$ contains robust eigenvalues; and $OD = \|\mathbf{X} - \boldsymbol{\mu}_{L1}(\mathbf{X} - \mathbf{TP}^T)\|$, with $\|\cdot\|$ denoting the Euclidean norm

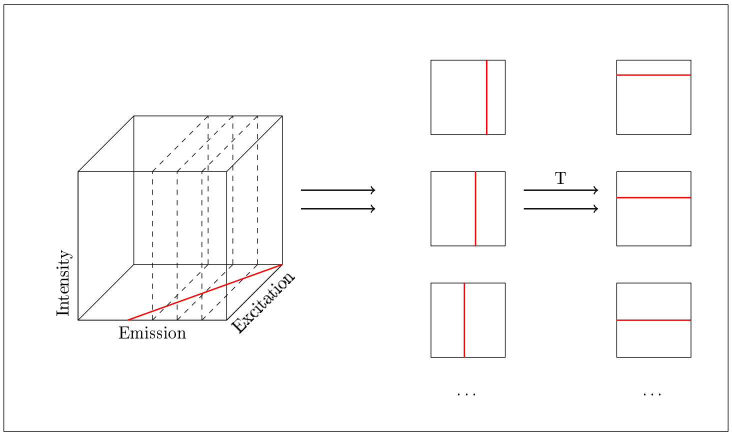

Figure 4.5: Slicing of the data cube according to emission mode (analogous operation can be performed according to excitation mode.

and k selected components being contained within $\mathbf{TP}^T$ model. Cutoff values defining outlying samples are determined by the means of robustified version z-scores, where robust location and scale estimates, such as median (med) and median absolute deviation (mad), were applied as alternatives to the standard procedures: $z = |d - \text{med}(d)|/\text{mad}(d)$.

Based on this outlier identification procedure the output of S-PCA algorithm is given in the form of a binary matrix $\mathbf{M}(J \times K)$, which will be referred to as an 'outlier mask'. This mask attributes ones to the regular observations and zeros to the elements identified as outliers and is illustrated in Figure 4.6a. It can be noticed that some regions, where the scatter ridges are present, remain undetected by S-PCA, which occasionally fails to identify the outlying samples if particular data circumstances occur. In this case, an additional interpolation step is proposed, based on the linear nature of the scatter. Firstly, observations flagged as outliers by S-PCA are projected onto a 2-dimensional coordinate system where a robust regression (here, based on the LTS) is applied. This method has a high breakdown point ensuring that the calibration line will match the majority of the data points, which in this case would be the first scatter stripe. In parallel, observations laying far from the regression line will be flagged as outliers. Afterwards, the width of the line is determined, usually by adopting the broadest band of the initially discovered pattern. The same two operations are repeated for the second scatter line, which was flagged as outliers by the previously fitted regression model. An example of the interpolated scatters can be seen in Figure 4.6b.

Figure 4.7 represents resolved spectra from PARAFAC models fitted to three different data scenarios: 1. when S-PCA was applied without the correction adjustment and the influence of undiscovered scatter is still visible in the appearance of the resolved curves; 2. with corrected scatter region, as in the Figure 4.6b, where a significant improvement of the resolved curves is observed; 3. by additionally applying a non-negativity constraint to both excitation and emission modes, resulting in 'perfectly' resolved profiles of the four investigated compounds. For comparison, some quantitative results of the PARAFAC models based on both, S-PCA and ROBPCA, for different scatter region treatment scenarios and constraints application, can be found in Table 4.1. From the indices included in the table and the resolved excitation/emission profiles (detailed ROBPCA results can be found in [14]), it is judged that the two methods perform similarly well and contribute significantly to a successful recovery of the underlying spectra within the data. The core consistency and explained variance, both having values $> 90\%$ indicate slightly in favor of the S-PCA approach, whereas the running time of the PARAFAC model and number of it-

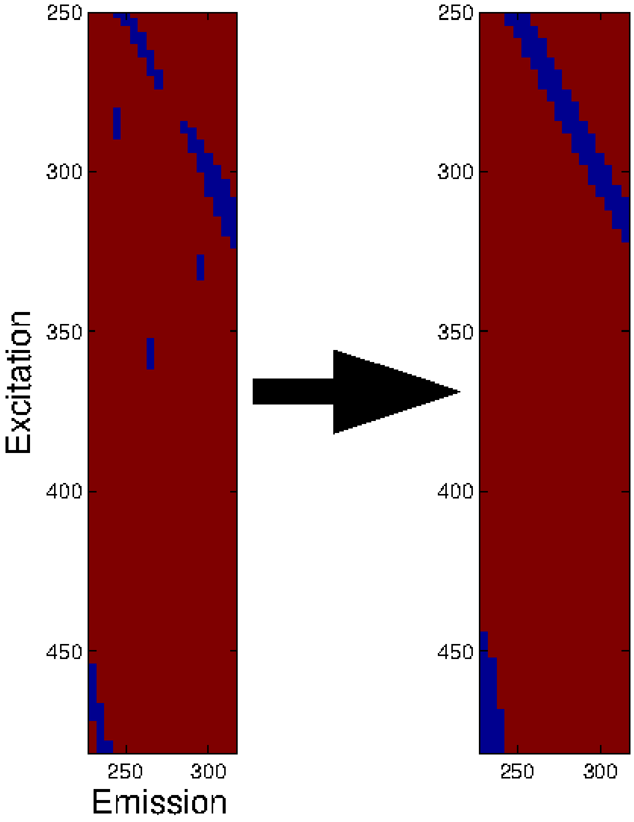

Figure 4.6: Correction undiscovered scatter: a) outlier mask after initial S-PCA routine; b) interpolated (corrected) scatter region.

Table 4.1: PARAFAC results for Dorit data expressed by the standard set of indeces. It is evident that the time required for finding the scatter by S-PCA is significantly lower than by ROBPCA.

	Name of method	Time of scatter finding [s]	Time of PARAFAC [s]	No of iter.	Core consist. [%]	Variance explained [%]
S-PCA	no correction	2.36	1.99	29.47	98.82	97.74
	correction	3.07	1.65	38.47	99.72	99.91
	correct.+ nonneg.	3.07	3.30	44.25	99.73	99.90
R-PCA	no constraints	293.28	54.94	44	99.30	99.78
	nonneg	293.30	26.89	57	99.73	99.77

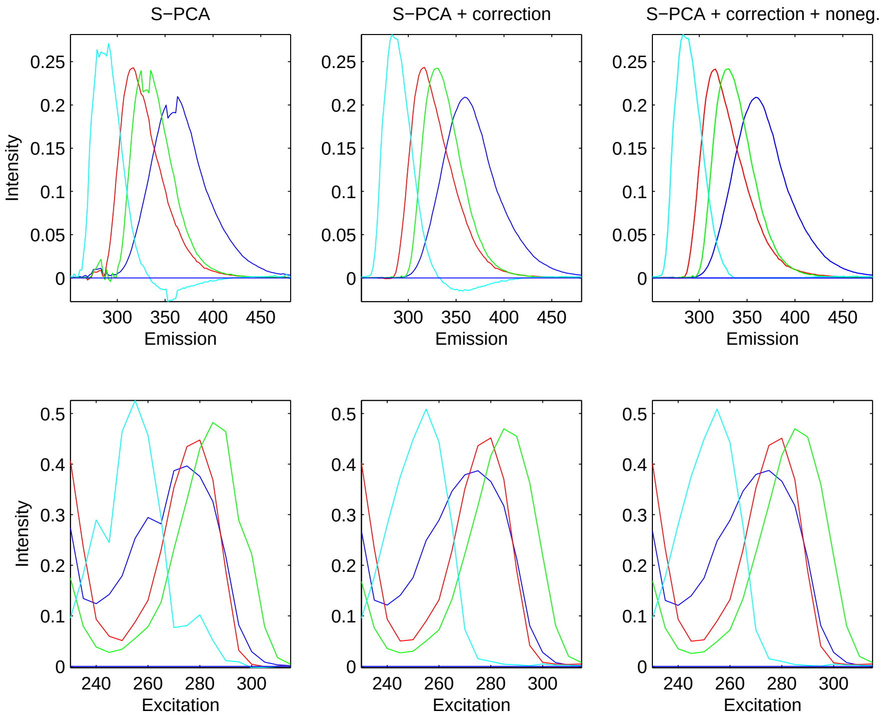

Figure 4.7: Correction of S-PCA for North Sea data: a) user chosen outlier mask after initial S-PCA routine, b,c,d) interpolated scatter regions depending on user chosen bandwidth adjustement factor set to $0, 1$ and 2, respectively.

erations required seem to be un-influenced by the scatter identification method. The obvious point, where S-PCA largely outperforms the previous method is the time required to find the scatter region, which in the case of Dorit data is around 100 times shorter than for ROBPCA.

4.3 Spherical PCA. A sensitivity study.

Since S-PCA method was first proposed [75], it has been used broadly within multivariate data analysis (e.g. [15, 90, 91, 97–99]) and mentioned in a few review papers concerning the chemometrics field (e.g. [94,95,100]). However, as it was shown in the previous section (see also [15]), S-PCA tends to fail in outlier identification when certain (non-specified) circumstances occure. Also Stanimirova et al. [97] mentions briefly that problems might emerge when outliers form clusters, but again the details are not known. That was the motivation for a simulation study, presented fully in Paper C, investigating the functioning scheme of S-PCA and its potential as an outlier identification tool for elliptical data and the case when outliers form clusters.

A great advantage of the S-PCA is its exceptional computation speed and relative simplicity which can be desired in many applications, especially when high dimensionality of data is considered. Figure C.3 presents a comparison of six PCA models: Classical PCA (PCA), ROBPCA [74], two versions of Projection Pursuit based PCA: by Croux and Ruiz-Gazen [101] (PP) and by Croux et al. [72] (Grid) and finally two Spherical S-PCA algorithms: traditional (SPCA) and its nested version (SPCA2). The 'nested S-PCA' will be described further. System time of algorithm execution is plotted agains the number of dimensions, for constant sample size $n = 500$. One can observe that only S-PCA methods and classical PCA have a linear time dynamics function, whereas other methods - exponential. This high computation time might cause that even a very good method could be impossible to implement in certian circumstances, for example when on-line analysis of large data set is considered.

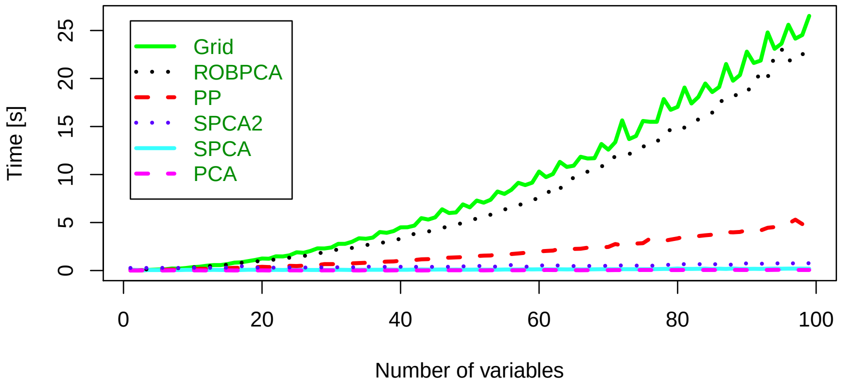

Figure 4.8: Execution time [s] of different PCA methods as a function of number of dimensions for constant number of observations $n = 500$.

The functioning scheme of S-PCA has been previously described in details in Section 4.1.5. In short, it can be split into four major parts, depicted graphically in Figure C.4, respectively:

1. find the $L1$ robust center of the data (red dot)
2. project all data points $\mathbf{x}_i$ on the unit radius sphere centered in $L1$
3. fit classical PCA to the projected data, find robust loadings and scores
4. apply the robust loadings to original data and identify outliers

Considering this functioning scheme, a sensitivity study of S-PCA was designed, aiming at investigating potential factors having influence on performance of the method. 100 samples were drawn from a bi-variate normal distribution with

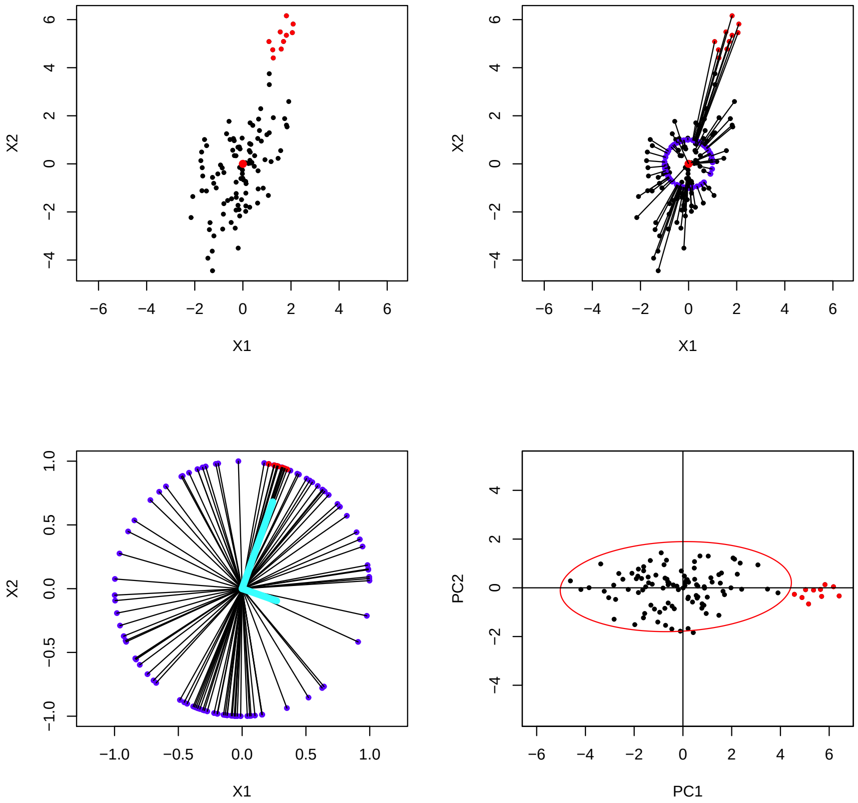

Figure 4.9: Functioning of S-PCA: a) finding a robust center of the data (here L1 median); b) projecting all data points on a sphere with unit radius; c) performing classical PCA on the projected data points - finding eigenvectors; d) applying eigenvectors to the initial data set, flagging outliers (here, by means of tolerance ellipse)

$\boldsymbol{\mu} = (0, 0)$ and

$$\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 3 \end{pmatrix}$$

and later contaminated (see Paper C for details). The factors in question were: d - direction from the center of main data cloud to the center of outlier cloud; r - distance from main data to outlying observations; ϵ - ratio of outliers. In total 16 directions $d \in \{0, 1/8\pi, \cdots, 15/8\pi\}$, three distances $r \in \{4, 6, 8\}$ and three contamination ratios $\epsilon \in \{0.1, 0.2, 0.3\}$ have been taken into account, resulting in 144 data scenarios. In an example presented in Figure 4.10 we consider four 'difficult' data scenarios, corresponding to directions $d \in \{0, 1/8\pi, 1/4\pi, 3/4\pi\}$ with $r = 6$ and $\epsilon = 0.3$. One can see that S-PCA is able to identify correctly the outlying samples only in the first case, as they are situated outside of red tollerance ellipse. At the same time, only the last panel of Figure 4.10 depicts the situation when PCA fit is well aligned (PC axes correspond to directions of highest variability in data). This is the case of good leverage outliers, which even though unrecognized, do not harm the performance of the method.

It has been concluded that location, distance and amount of contaminated samples may influence outlier identification ability of SPCA: if the outlier cloud is too close to the main data, some or all outlying samples might stay unidentified (panels b, c and d in Figure 4.10). Moreover, if the outlier ratio is high and its location not aligned with the first eigenvector, the S-PCA fit will deteriorate (panels a, b and c). Therefore, even if S-PCA is able to correctly identify the corrupted signal, its fit (loadings and scores) is often attracted towards it.

As described above, the S-PCA method functions by firstly performing classical PCA on down-weighted data, and afterwards identifying outlying samples.

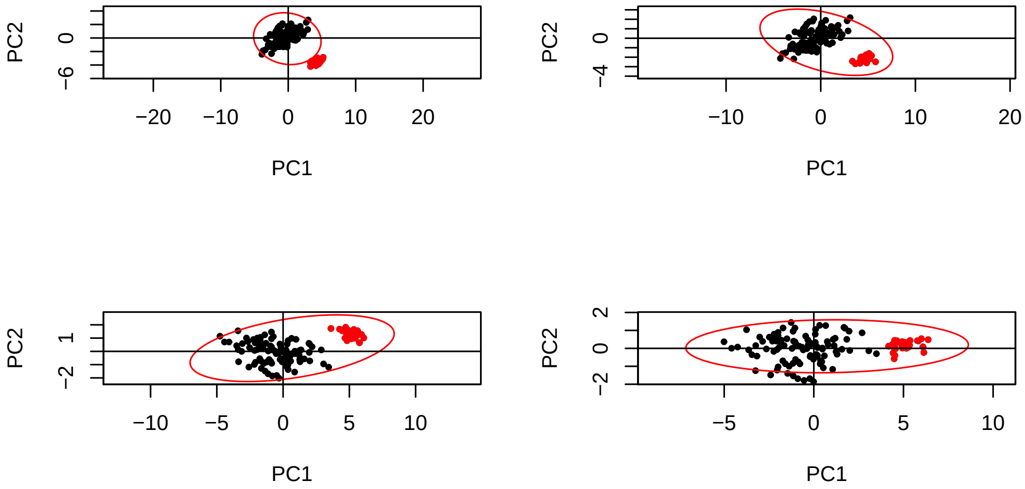

Figure 4.10: S-PCA results for first 4 directions $d \in \{0, 1/8\pi, 1/4\pi, 3/4\pi\}$, distance=8, outlier ratio=0.3

This might lead to the situation presented in Figure 4.10a, where contamination is well identified but resulting PCA loadings highly corrupted. In order to avoid that inconvenience, a simple correction called 'nested S-PCA' or 'S-PCA2' has been proposed. This correction assumes that after identification of outliers by S-PCA method, another S-PCA routine is executed, but this time on the cleaned, outlier-free data set. Then, new loadings are used for constructing robust scores by projecting them on the initial data and new distances are determined. The potential advantage of this method is twofold: in the case when ordinary S-PCA is able to correctly identify the outliers, the final fit will not be affected by them and resulting loadings will be determined correctly. Secondly, if only a part of contamination was discovered, S-PCA2 might be able to correct for that.

Figure C.13 shows 'flag plots' resulting from application of S-PCA2 and two other robust alternatives, ROBPCA and Grid, to the four 'difficult' data scenarios. The Flag Plot shows 'flags' ($\in \{0, 1\}$) which were assigned to each observation during the outlier identification procedure. Real outliers are marked with red color and identified outliers have $flag = 0$. Observations with $flag = 1$ are considered regular. Blue points denote data points wrongly qualified as outliers. All three methods return satisfactory results for $d \in \{0, 1/8\pi\}$, corresponding to directions where outlier cloud is further from the main data. One can see that S-PCA2 seems to be most sensitive, as it starts having 'problems' already for $d = 1/8\pi$, and ROBPCA most resistant, performing well even for $d = 1/4\pi$. All methods fail however in the case of good leverage outliers (panel d).

It has been concluded that S-PCA (and especially its nested version) can be a valuable asset in the toolbox of every practitioner, especially when short execution time is an important factor. It is recomended that it is applied with consciousness of possible shortcomings, such as described in this work and in Paper C

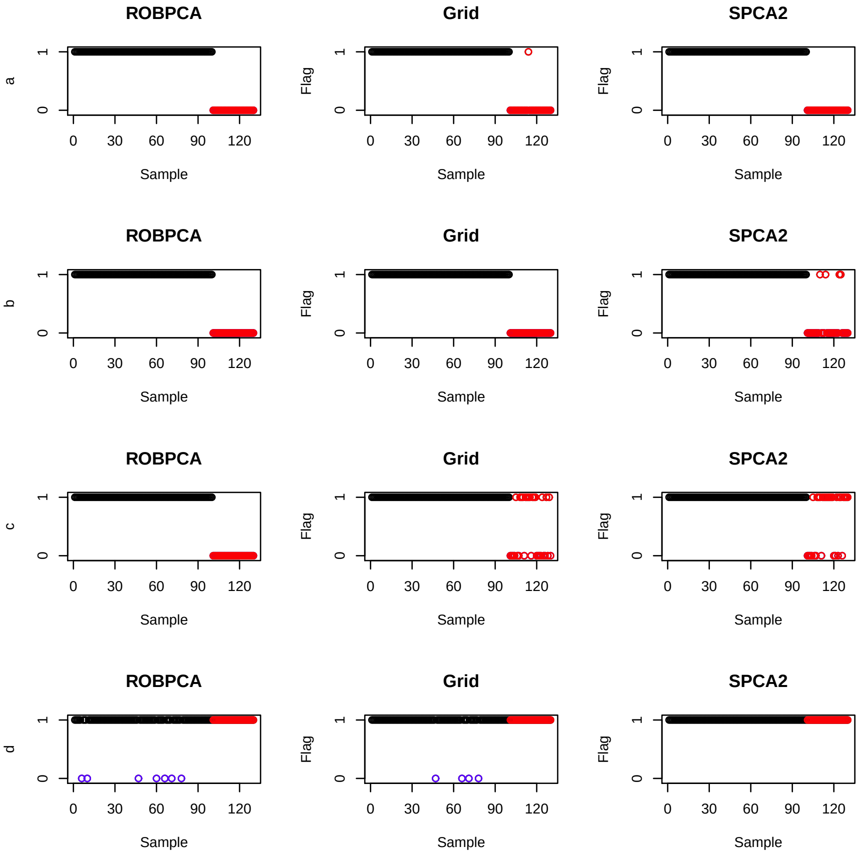

Figure 4.11: Flag plots resulting from S-PCA2, ROBPCA and Grid methods for 4 data scenarios: $d \in \{0, 1/8\pi, 1/4\pi, 3/4\pi\}$ with $r = 6$ and $\epsilon = 0.3$

Chapter 5

Discussion and conclusions

The focus of this project was twofold, targeting:

- new fields, where multivariate and multi-way data analytical tools could find their application for more efficient analysis of chemometric phenomena, by replacing less suitable data analysis techniques used currently;
- new data analytical methods, where usage of the robustness frame-work would contribute to improvement and automatisation of already existing multivariate techniques.

The first part of the study was realised by delivering appropriate multivariate tools for analysing polar CTD sea water samples, defined by three data modes:

depth, variables and geographical location. Two- and three-way chemometrical methods, such as PCA and PARAFAC models, were applied and their performance indicated superiority of the three-way frame-work, constituting a novel assessment of the sea water measurements. Moreover, fluorescence values were predicted from the other physio-chemical variables by applying PLS and N-PLS regression models. This resulted in a clear preference towards the more complex model, delivering more reliable predictions than a classical 2-way PLS.

It has been argued [2] that by using methods of lower order compared to the data order (for example using PCA on 3-way data), precious information inherent to the multi-way correlation structure are likely to be lost. On the other hand, some underlying assumptions of certain multi-way models, such as PARAFAC, might be more strict than in the case of two-way models, say PCA. The violation of these assumptions can lead to model degeneracy and make the analysis infeasible. Nevertheless, application of multi-way data analysis tools is advocated as a principle whenever the nature of the data allows.

The second part of the thesis was devoted to qualitative properties of the analysed data. The broad theoretical background of robust procedures was provided and a new tool, based on S-PCA, aiming at identifying Rayleigh and Raman scatters in EEM landscapes was elaborated, delivering a practical application of the robust frame-work in real data situations. Moreover, a simulation study of S-PCA has been performed, specifyling the outlier identifiation ability of the method and comparing it to the other robust alternatives.

It was indicated and underlined that standard least squares analysis procedures do not withstand the presence of outliers or deviations from assumed distribution, occasionally leading to dramatical loss of efficiency of the resulting estimates. The 'perfect' robust estimator is therefore characterised not only by a high breakdown point and bounded influence function, but also by high efficiency at normality condition. The arguments encouraging the removal of outliers from the data, prior to the actual analysis, have been issued. However, in the case of multidimensional data structures, visual screening for outliers is no longer possible. Furthermore, a definite removal of distant samples could result in ignoring an important phenomenon present in the data. A relevant question emerges here: when should robust techniques be used. To answer this question, the author agrees with the opinion of Peter Rousseeuv [47], who claims that by principle it is advised to apply both, a classical and a robust procedure. If their performance is similar (meaning no or insignificant outlier presence), the results of classical methods should be taken into account, as they yield more efficient estimates. On the contrary, if the two methods are significantly different, it is a signal that part of the data is corrupted and care should be taken in order to avoid misleading results.

Another point worthy attention in the discussion over robustness issues is the concept of the 50%-breakdown point. This value is assumed to be the highest possible, and if the contaminated data constitutes more than half of the observations, it is no longer possible to distinguish between good and bad data. However, one can ask: how realistic it is to encounter a data set where half of the observations are corrupted. In theory, there is no reasons to exclude such a scenario, however it is not expected to be encountered frequently in practice. Moreover, data sets with a high percentage of outlying points are often considered non-representative and therefore, are discarded. On the other hand, if good robust techniques, delivering reliable results were at hand, the time and effort of re-running the experiments for obtaining better data quality could be saved.

Many robust methods reach the 50%-breakdown point, however this often leads to less accurate estimates. If the method allows, it is advisable to tune the breakdown point according to the analysed data, in order to reach a compromise between good robust properties and quality of the estimate. Finally, if the data quality appears extremely bad, repeating the experiment generating better data quality should always be an option (as long as it is not financially- and time-exhausting), yielding a chance for more reliable results.

The studies included in Paper B and Paper C show clearly that robust methods can contribute to improving the existing analytical techniques used commonly in chemometrics. In the included papers, robust techniques (such as S-PCA) proved to be a good outlier detection tool and pattern recognition technique, and their usage can likely be spread to many other fields and applications (see for example [102]). In order to facilitate that, implementations of robust procedures should be available for some standard statistical data analysis programs such as R, S-plus or Matlab. Some attempts have already been undertaken resulting in few robust toolboxes (like TOMCAT or LIBRA for Matlab, and two R packages *chemometrics* and *rrcov*), however more work is needed to propagate the usage of robust methods among practitioners, who could benefit largely from these tools.

In order to summarise the discussion conducted above, the following final conclusions can be issued:

- in order to extract full information from multi-way data structure, multi-way data analysis tools are required;
- it is advised to apply robust and classical procedures together, in order to indicate if contamination in the data is present;
- data-driven tuning of the breakdown point in certain robust methods is recommended, to find a compromise between good robustness properties and efficiency of the estimate;
- propagating the usage of robust methods among practitioners is certainly a challenge which could be addressed in future work.

Bibliography

[1] S. Wold. Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems*, 30(1):109–115, 1995.

[2] A.K. Smilde, R. Bro, P. Geladi, and J. Wiley. *Multi-way analysis with applications in the chemical sciences*, volume 978. Wiley Online Library, 2004.

[3] R. Bro, P.R. Mobley, and B.R. Kowalski. Review of chemometrics applied to spectroscopy: 1985-95, part 3-multi-way analysis. *Applied Spectroscopy Reviews*, 32(3):237–261, 1997.

[4] C. Andersen and R. Bro. Practical aspects of parafac modeling of fluorescence excitation-emission data. *Journal of Chemometrics*, 17(4):200–215, APR 2003.

[5] R.T. Ross and S. Leurgans. Component resolution using multilinear models. *Methods in enzymology*, 246:679–700, 1995.

[6] P.D. Wentzell, S.S. Nair, and R.D. Guy. Three-way analysis of fluorescence spectra of polycyclic aromatic hydrocarbons with quenching by nitromethane. *Analytical chemistry*, 73(7):1408–1415, 2001.

[7] R.D. JiJi and K.S. Booksh. Mitigation of rayleigh and raman spectral interferences in multiway calibration of excitation-emission matrix fluorescence spectra. *Analytical Chemistry*, 72(4):718–725, 2000.

[8] Å. Rinnan and C.M. Andersen. Handling of first-order rayleigh scatter in parafac modelling of fluorescence excitation–emission data. *Chemometrics and intelligent laboratory systems*, 76(1):91–99, 2005.

[9] Å. Rinnan, K.S. Booksh, and R. Bro. First order rayleigh scatter as a separate component in the decomposition of fluorescence landscapes. *Analytica chimica acta*, 537(1):349–358, 2005.

[10] R. Bro. Parafac. tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(2):149–171, 1997.

[11] M. Bahram, R. Bro, C. Stedmon, and A. Afkhami. Handling of rayleigh and raman scatter for parafac modeling of fluorescence data using interpolation. *Journal of Chemometrics*, 20(3-4):99, 2006.

[12] Z.P. Chen and R.Q. Yu. Mitigating model deficiency in three-way data analysis by the combination of background constraining and iterative correcting techniques. *Analytica chimica acta*, 487(2):171–180, 2003.

[13] L. Thygesen, A. Rinnan, S. Barsberg, and J. Moller. Stabilizing the parafac decomposition of fluorescence spectra by insertion of zeros outside the data area. *Chemometrics and Intelligent Laboratory Systems*, 71(2):97–106, MAY 28 2004.

[14] S. Engelen, S. Frosch, and M. Hubert. Automatically identifying scatter in fluorescence data using robust techniques. *Chemometrics and Intelligent Laboratory Systems*, 86(1):35–51, 2007.

[15] E. Kotwa, B. Jørgensen, P. B. Brockhoff, and Stina Frosch. Automatic scatter detection in fluorescence landscapes by means of spherical principal component analysis. *Journal of Chemometrics*, 2013.

[16] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

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

[18] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1):37–52, 1987.

[19] R.A. Harshman. Foundations of the parafac procedure: models and conditions for an" explanatory" multimodal factor analysis. 1970.

[20] A. Smilde, R. Tauler, J. Henshaw, L. Burgess, and B. Kowalski. Multicomponent determination of chlorinated hydrocarbons using a reaction-based chemical sensor. 3. medium-rank second-order calibration with restricted tucker models. *Analytical Chemistry*, 66(20):3345–3351, 1994.

[21] R. Tauler. Multivariate curve resolution applied to second order data. *Chemometrics and Intelligent Laboratory Systems*, 30(1):133–146, 1995.

[22] H. Wold. Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach. *Perspectives in Probability and Statistics, In Honor of MS Bartlett*, pages 117–144, 1975.

[23] S. Wold, A. Ruhe, H. Wold, and WJ Dunn III. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5:735, 1984.

[24] P. Geladi and B.R. Kowalski. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17, 1986.

[25] H. Martens and T. Naes. *Multivariate calibration*. John Wiley & Sons Inc, 1992.

[26] R. Bro. Multiway calibration. multilinear pls. *Journal of Chemometrics*, 10(1):47–61, 1996.

[27] A.K. Smilde. Comments on multilinear pls. *Journal of Chemometrics*, 11(5):367–377, 1997.

[28] S. de Jong. Regression coefficients in multilinear pls. *Journal of Chemometrics*, 12(1):77–81, 1998.

[29] R.P. McDonald. A simple comprehensive model for the analysis of covariance structures: Some remarks on applications. *British Journal of Mathematical and Statistical Psychology*, 33(2):161–183, 1980.

[30] K. Gauss. Theoria combinationis observationum erroribus minimis obnoxiae (theory of the combination of observations least subject to error). english translation by gw stewart (1995), classics in applied mathematics no. 11, 1821.

[31] S. Newcomb. A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, 8(4):343–366, 1886.

[32] P.J. Daniell. Observations weighted according to order. *American Journal of Mathematics*, pages 222–236, 1920.

[33] H. Jeffreys. Theory of probability. 1961.

[34] J.W. Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 2:448–485, 1960.

[35] P.J. Huber, E. Ronchetti, and MyiLibrary. *Robust statistics*, volume 1. Wiley Online Library, 1981.

[36] F. Hampel. *Contributions to the Theory of Robust Estimation. 1968*. PhD thesis.

[37] F.R. Hampel. The breakdown points of the mean combined with some rejection rules. *Technometrics*, 27(2):95–107, 1985.

[38] S.M. Stigler. Simon newcomb, percy daniell, and the history of robust estimation 1885-1920. *Journal of the American Statistical Association*, pages 872–879, 1973.

[39] P.J. Huber. The 1972 wald lecture robust statistics: a review. *The Annals of Mathematical Statistics*, 43(4):1041–1067, 1972.

[40] G.E.P. Box. Non-normality and tests on variances. *Biometrika*, pages 318–335, 1953.

[41] F.R. Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896, 1971.

[42] P.J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[43] P.J. Huber, P. Huber, P. Huber, M. Statisticien, E.U. Suisse, P. Huber, and M. Statistician. *Robust statistical procedures*, volume 68. SIAM, 1996.

[44] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust statistics: the approach based on influence functions*, volume 114. Wiley, 2011.

[45] D.L. Donoho and P.J. Huber. The notion of breakdown point. *A Festschrift for Erich L. Lehmann*, pages 157–184, 1983.

[46] F.R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

[47] P.J. Rousseeuw, A.M. Leroy, and J. Wiley. *Robust regression and outlier detection*, volume 3. Wiley Online Library, 1987.

[48] F.Y. Edgeworth. On observations relating to several quantities. *Hermathena*, 6(13):279–285, 1887.

[49] N.A. Campbell. Robust procedures in multivariate analysis i: Robust covariance estimation. *Applied statistics*, pages 231–237, 1980.

[50] P.J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993.

[51] P. Milasevic and GR Ducharme. Uniqueness of the spatial median. *The Annals of Statistics*, 15(3):1332–1333, 1987.

[52] C.G. Small. A survey of multidimensional medians. *International Statistical Review/Revue Internationale de Statistique*, pages 263–277, 1990.

[53] J.B.S. Haldane. Note on the median of a multivariate distribution. *Biometrika*, 35(3-4):414–417, 1948.

[54] R. Gnanadesikan and J.R. Kettenring. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, pages 81–124, 1972.

[55] P.J. Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, pages 871–880, 1984.

[56] P.J. Rousseeuw, M. Debruyne, S. Engelen, and M. Hubert. Robustness and outlier detection in chemometrics. *Critical Reviews in Analytical Chemistry*, 36(3-4):221–242, 2006.

[57] P.J. Rousseeuw and K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[58] P.J. Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8:283–297, 1985.

[59] W.A. Stahel. *Robuste schätzungen: Infinitesimale optimalität und schätzungen von kovarianzmatrizen*. PhD thesis, Ph. d. thesis, 1981.

[60] D.L. Donoho. Breakdown properties of multivariate location estimators. Technical report, Technical report, Harvard University, Boston. URL http://www-stat. stanford. edu/˜ donoho/Reports/Oldies/BPMLE. pdf, 1982.

[61] A.F. Siegel. Robust regression using repeated medians. *Biometrika*, 69(1):242–244, 1982.

[62] D.F. Andrews, P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers, and J.W. Tukey. *Robust estimates of location: Survey and advances*, volume 173. Taylor & Francis, 1972.

[63] P.J. Rousseeuw and V.J. Yohai. Robust regression by means of s-estimators. *Robust and nonlinear time series*, 26:256–272, 1984.

[64] V.J. Yohai. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, pages 642–656, 1987.

[65] P.J. Rousseeuw, S. Van Aelst, K. Van Driessen, and J.A. Gulló. Robust multivariate regression. *Technometrics*, 46(3):293–305, 2004.

[66] P.J. Huber. Projection pursuit. *The annals of Statistics*, pages 435–475, 1985.

[67] C. Croux and G. Haesbroeck. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87(3):603–618, 2000.

[68] B. Walczak and DL Massart. Robust principal components regression as a detection tool for outliers. *Chemometrics and intelligent laboratory systems*, 27(1):41–54, 1995.

[69] G. Li and Z. Chen. Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and monte carlo. *Journal of the American Statistical Association*, pages 759–766, 1985.

[70] Y.L. Xie, J.H. Wang, Y.Z. Liang, L.X. Sun, X.H. Song, and R.Q. Yu. Robust principal component analysis by projection pursuit. *Journal of chemometrics*, 7(6):527–541, 1993.

[71] C. Croux and A. Ruiz-Gazen. A fast algorithm for robust principal components based on projection pursuit. In *COMPSTAT: Proceedings in computational statistics*, pages 211–217, 1996.

[72] C. Croux, P. Filzmoser, and M.R. Oliveira. Algorithms for projection–pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87(2):218–225, 2007.

[73] M. Hubert, P.J. Rousseeuw, and S. Verboven. A fast method for robust principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60(1):101–111, 2002.

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

[75] N. Locantore, J.S. Marron, D.G. Simpson, N. Tripoli, J.T. Zhang, K.L. Cohen, G. Boente, R. Fraiman, B. Brumback, C. Croux, et al. Robust principal component analysis for functional data. *Test*, 8(1):1–73, 1999.

[76] R. Maronna. Principal components and orthogonal regression based on robust scales. *Technometrics*, 47(3):264–273, 2005.

[77] S. de Jong. Simpls: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263, 1993.

[78] M. Hubert and K.V. Branden. Robust methods for partial least squares regression. *Journal of Chemometrics*, 17(10):537–549, 2003.

[79] I.N. Wakelinc and H.J.H. Macfie. A robust pls procedure. *Journal of Chemometrics*, 6(4):189–198, 1992.

[80] D.J. Cummins and C.W. Andrews. Iteratively reweighted partial least squares: a performance analysis by monte carlo simulation. *Journal of Chemometrics*, 9(6):489–507, 1995.

[81] J.A. Gil and R. Romera. On robust partial least squares (pls) methods. *Journal of chemometrics*, 12(6):365–378, 1999.

[82] R.J. Pell. Multiple outlier detection for multivariate calibration using robust statistical techniques. *Chemometrics and Intelligent Laboratory Systems*, 52(1):87–104, 2000.

[83] S. Serneels, C. Croux, P. Filzmoser, and P.J. Van Espen. Partial robust m-regression. *Chemometrics and Intelligent Laboratory Systems*, 79(1):55–64, 2005.

[84] J. Riu and R. Bro. Jack-knife technique for outlier detection and estimation of standard errors in parafac models. *Chemometrics and Intelligent Laboratory Systems*, 65(1):35–49, JAN 28 2003.

[85] S. Engelen and M. Hubert. Detecting outlying samples in a parafac model. Technical report, Katholieke Universietit Leuven, 2007.

[86] P. Filzmoser. Robust principal component regression. *Computer data analysis and modeling. Robust and computer intensive methods. Belarusian State University, Minsk*, pages 132–137, 2001.

[87] M. Hubert and S. Verboven. A robust pcr method for high-dimensional regressors. *Journal of Chemometrics*, 17(8-9):438–452, 2003.

[88] M. Hubert and K. Van Driessen. Fast and robust discriminant analysis. *Computational Statistics and Data Analysis*, 45(2):301–320, 2004.

[89] K. Vanden Branden and M. Hubert. Robust classification in high dimensions based on the simca method. *Chemometrics and Intelligent Laboratory Systems*, 79(1-2):10–21, 2005.

[90] M. Daszykowski, K. Kaczmarek, I. Stanimirova, Y. Vander Heyden, and B. Walczak. Robust simca-bounding influence of outliers. *Chemometrics and Intelligent Laboratory Systems*, 87(1):95–103, 2007.

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

[92] V. Pravdova, F. Estienne, B. Walczak, and DL Massart. A robust version of the tucker3 model. *Chemometrics and Intelligent Laboratory Systems*, 59(1):75–88, 2001.

[93] L. Yi-Zeng and O. M. Kvalheim. Robust methods for multivariate analysis – a tutorial review. *Chemometrics and Intelligent Laboratory Systems*, 32(1):1–10, 1996.

[94] S. Frosch, J. von Frese, and R. Bro. Robust methods for multivariate data analysis. *Journal of Chemometrics*, 19(10):549–563, 2005.

[95] M. Daszykowski, K. Kaczmarek, Y. Vander Heyden, and B. Walczak. Robust statistics in data analysis - a review. *Chemometrics and Intelligent Laboratory Systems*, 85(2):203–219, 2007.

[96] M. Hubert, P.J. Rousseeuw, and S. Van Aelst. High-breakdown robust multivariate methods. *Statistical Science*, 23(1):92, 2008.

[97] I. Stanimirova, M. Daszykowski, and B. Walczak. Dealing with missing values and outliers in principal component analysis. *Talanta*, 72(1):172–178, 2007.

[98] M. Daszykowski, M.S. Wróbel, A. Bierczynska-Krzysik, J. Silberring, G. Lubec, and B. Walczak. Automatic preprocessing of electrophoretic

images. *Chemometrics and Intelligent Laboratory Systems*, 97(2):132–140, 2009.

[99] V.J. Fortunato. A comparison of the construct validity of three measures of negative affectivity. *Educational and psychological measurement*, 64(2):271–289, 2004.

[100] P. Filzmoser and V. Todorov. Review of robust multivariate statistical methods in high dimension. *Analytica chimica acta*, 705(1):2–14, 2011.

[101] C. Croux and A. Ruiz-Gazen. High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1):206–226, 2005.

[102] S. Frosch and B.M. Jørgensen. Peak alignment and robust principal component analysis of gas chromatograms of fatty acid methyl esters and volatiles. *Journal of chromatographic science*, 45(4):169–176, 2007.

Appendix A

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

Authors:

E. Kotwa[1], S. Lacorte,[2] C. Duarte[3] and R. Tauler[2]

Submitted to:

Polar research, September 2012.

[1] DTU Informatics, Richard Pedersens Plads, Building 321, DK-2800 Lyngby, Denmark

[2] Institute of Environmental Assessment and Water Research, Spanish Council for Scientific Research, Barcelona, Spain.

[3] Mediterranean Institute for Advanced Studies, Spanish Council for Scientific Research, Mallorka, Spain.

Abstract

In this paper an examination of 2- and 3-way chemometric methods for analysis of Arctic and Antarctic CTD (Conductivity-Temperature-Depth) water samples was performed. A standard CTD sensor devices were used during two oceanographic expeditions (July 2007 - Arctic, February 2009 - Antarctica) within a total number of 174 locations. An output of those devices can be arranged in 3-way data structures (according to sea water depth, measured variables and geographical location). 2- and 3-way statistical tools such as PCA, PARAFAC, PLS and N-PLS were applied for exploratory analysis, spatial patterns discovery and calibration, and their performance discussed. Special importance was given to correlation and possible prediction of fluorescence, from other physical variables. MATLAB's mapping toolbox was used for geo-referencing and visualization of the results. It was concluded 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 loosing information; 3) The distinction between Arctic and Antarctic Seas was revealed mostly by PC1, related to physico-chemical properties of water samples; 4) The possibility of predicting fluorescence values from physical measurements has been confirmed when the three-way data structure was considered in N-way PLS regression.

Keywords: Arctic and Antarctica, CTD, multi-way analysis, fluorescence

A.1 Introduction

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

The main goal of this contribution is to deliver and examine appropriate multivariate tools for exploratory and regression analysis of so-called CTD (conductivity-temperature-depth) data. These data come from CTD profilers installed in surface ships navigating polar ocean waters during the open iced seasons (July-August in Arctic and January-February in the Antarctica), constituting an important source of available measurements about these areas. The resulting data, being a 3-way structure with variable, depth and location modes, is often analyzed in a uni-variate 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 the whole information contained in these fairly complex data structures. Moreover, if a large amount of variables is considered, it might become highly time consuming and simply inconvenient. It is known and broadly described in the literature, for example [Smilde et al.(2004)], that more sophisticated 2-, 3- or multi-way statistical methods, such as Principal Component Analysis (PCA) or PARAFAC model, Partial Least Squares (PLS) regression or its multi-way version N-PLS are more relevant for extracting the full information from multi-way data sets, as those obtained by the CTD sensor, which will be considered throughout this paper.

Data used in this study consists of measurements collected during 2007 ATOS I and 2009 ATOS II polar expeditions in the Arctic and Antarctic Oceans within a total number of 174 locations. Depth profiles of 7 variables, common for both polar areas were considered and included in the analysis: temperature, conductivity, salinity, oxygen, beam transmission, fluorescence and sea-point turbidity. The second objective of this study is a comparative sea water study aiming at identifying geographical differences within and between the Arctic and Antarctic Seas, considering their physical and chemical characteristics. For this sake Principal Component Analysis and PARAFAC models were applied and their results delivered and discussed.

Moreover, in the sea water investigations, measured fluorescence is an especially interesting property. It reflects the amount of chlorophyll (reflecting the maximum concentrations of biota and algae population), and hence, biological activity in the water. Therefore, a regression study, aiming at explaining and predicting fluorescence values from remaining variables is of interest here, and becomes the third objective of this paper. Methods used for attaining this objective are Partial Least Squares regression technique with its multi-way alternative - N-PLS.

The outline of the paper is following: Section A.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 A.3, followed by the results and discussion (Section A.4). Finally, the conclusions are given in Section A.5.

A.2 Experimental Data

A.2.1 Data from the Sea

Data samples used in this work were collected during two oceanographic expeditions, spanning the areas of $68°N$ - $81°N$, $20°W$-$20°E$ and $60°S$- $70°S$, $50°W$-$77°50'W$ respectively, presented in Figure A.1. During both expeditions a standard, real-time CTD sensor was used producing the depth-profile data in each of the 93 locations in Antarctica and 49 Arctic stations (for some stations, due to the position of permanent ice and current needs of scientists, the measurement was repeated, resulting in the total number of 80 samples). Since the sensors used in the campaigns were non-identical, they were measuring different ranges of variables, sometimes in different units. Eventually, 7 variables, common for both locations, were maintained and considered in the study. They are presented in Table A.1. Among them, the first four are of physical (temperature and conductivity) or chemical nture (salinity and dissolved oxygen) and the other three are related to radiation (beam transmission, fluorescence and sea turbidity).

In each location the sensor was dropped down to a certain depth (ranging form 50-1000 m, depending on specific location) resulting in replicated measurements as it was descending and during ascent. In this study, only the up-cast measurements were included as they proved to have nearly identical profiles as down-cast data, but contain a lower number of incomplete observations, and therefore will produce more reliable results. These measurements were collected continuously, approximately every few seconds, and therefore, for the sake of calculation convenience, the data set was reduced by taking averages depth-wise, resulting in

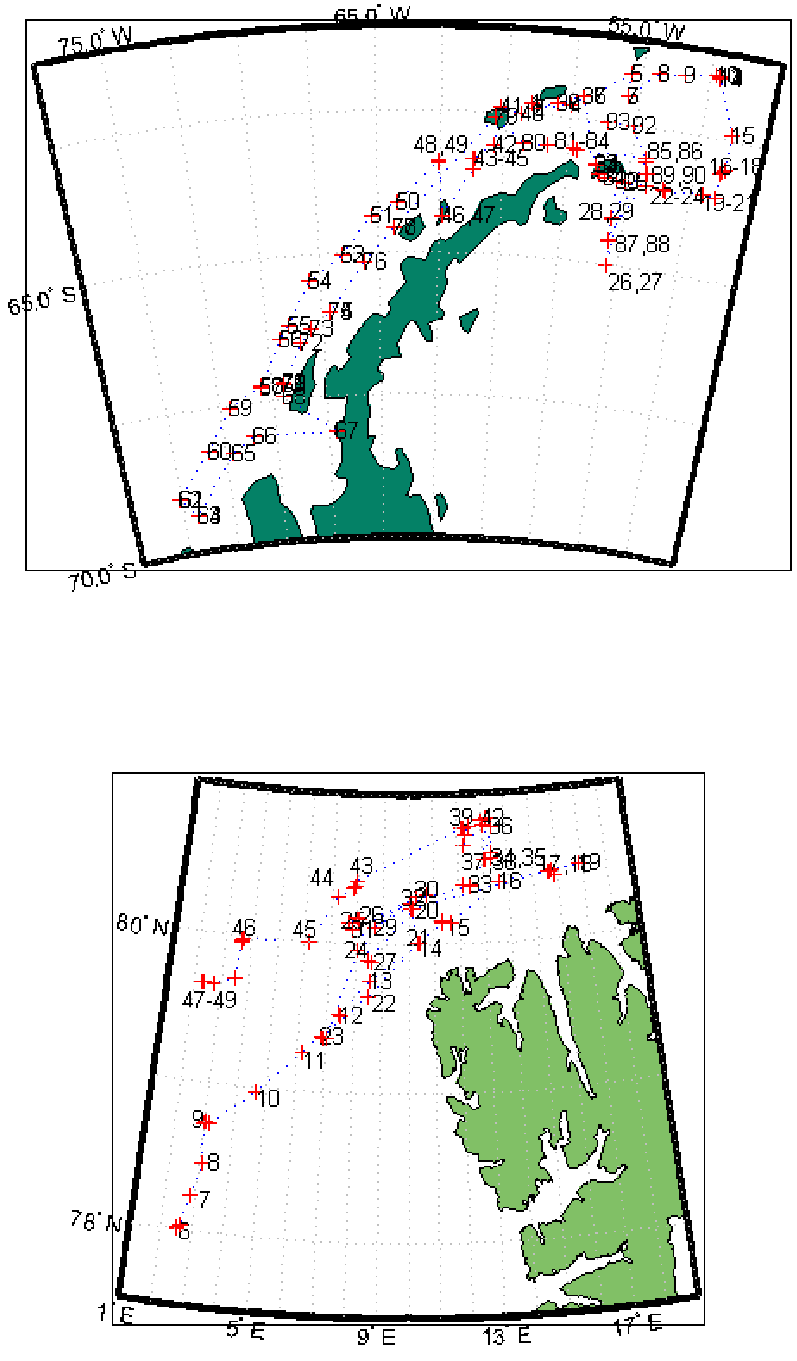

Figure A.1: Locations of CTD measurement stations in the Antarctica (left) and in the Arctic Sea (right).

Table A.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]

one observation per meter (arbitrary choice). Moreover, considering that the largest amount of changes in the measured signal was situated near the surface, we decided to disregard all data collected below 100 m depth, as most of these values were effectively constant and non-informative.

A.2.2 Arranging the data sets

The most 'natural' and straight-forward way for arranging the whole data set was a 3-way framework. In Figure A.2 one can visualize the resulting data cube, relevant for PARAFAC and NPLS models. For the PCA and PLS analysis, the data were unfolded and the two distinct unfolding directions were adapted for this study: variable (on the left) and station-wise (on the right). Moreover, due to the high amount of missing values (> 60%) we disregarded one station (number 71) from the Antarctic expedition. In addition, the first 5 stations (10 first measurements due to repetitions) were eliminated from the Arctic data. These measurements were taken 'on the way' to the final destination area, were not of significant interest to the overall study and their location did not follow a 2-D 'grid' on the map (they were alligned), which woudl cause later plotting

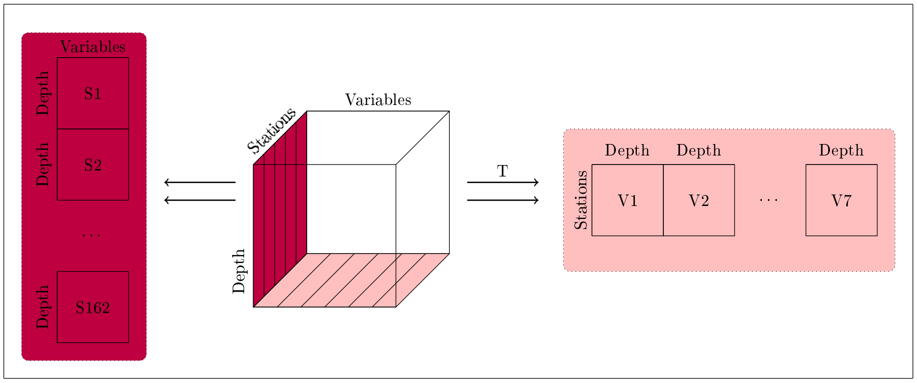

Figure A.2: Arrangement of the CTD data in a cubic structure according to three modes: depth, variables and stations. Two unfolding directions, variable- and station-wise were adapted to fit the 2-way workframe.

inconvenience. After the reduction, we arrive at the total number of location being 92 for Antarctica and 70 for the Arctic Sea (each repetition was considered as a separate measurement), resulting in the final dimension of the data, being: $[100 \times 7 \times 162]$ for 3-way methods, $[16200 \times 7]$ and $[162 \times 700]$ for both unfolded data sets, variable- and station-direction, respectively.

A.2.3 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 close to the surface is often corrupted 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 by means of a variable-wise standardization method. Standardized values exceeding 99% confidence interval were flagged as outliers and then substituted by a weighted average of neighboring values. More sophisticated methods for both, outliers removal and missing data interpolation exist in the literature ([Filzmoser et al.(2008)], [Rousseeuw et al.(2006)], [Serneels and Verdonck(2009)], [Stanimirova and Walczak(2008)]) using for example robust statistics or EM-algorithm, but for the needs of current research, the adapted methodology yields satisfactory results.

In addition, data centering and scaling was performed variable-wise. This preprocessing step was performed for the whole joined data set (Arctic plus Antarctica) and as such, it 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 them. However, since fluorescence and sea turbidity were initially given using different units for Antarctica and the Arctic Sea, in order to avoid the case where unit differences would corrupt the results, these variables were scaled prior to pre-processing step applied to all data set.

A.3 Methods

As previously stated, the objective of this work is delivering and comparing multivariate statistical tools which 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 remaining variables. For this purpose, two approaches towards data analysis were adopted. We start by considering and implementing the most commonly used 2-way chemometric techniques, such as PCA and PLS. A term '2-way', not to be confused with '2-dimensional', reffers here to data sets which can be arranged in a matrix, having two distinct directions (modes), *e.g.* objects and variables. '3-way', consequently, denotes structures accommodated in a cubic form, where a third direction (here, location) has been added. Hence, before fitting a 2-way model to the 3-way data, the analyzed cube has to be unfolded (according to one of its modes) and reshaped into a matrix (see Section A.2.2). However, it can be 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 2-way methods, such as PARAFAC and multi-way PLS (N-PLS), will be applied and discussed within this paper. Below, we explain the reasons for using particular techniques, whereas a more detailed mathematical background can be found in Appendix 1.

A.3.1 2-way methods

If the focus of the analysis lies in summarizing patterns, dependencies or differences within the data set, decomposition methods, such as Principal Component Analysis might be applied for this purpose. In brief, PCA projects the data to the lower dimensional spaces where it is easier to explore and visualize them, by means of small amount of so-called Principal Components (PCs). In order to cover the most important information contained in the polar data, two unfolding directions of the data cube (according to variable and location modes) will be taken into account. Classical and robust versions of PCA (here, ROBPCA, developed by [Hubert et al.(2005)]) will be fitted, due to the potential influence of outliers.

In parallel, in order to explain measured fluorescence by other available variables, PLS regression, being a popular 2-way calibration model, will be applied. As it was stated in Section A.1, the measured fluorescence carries information about amount of chlorophyll, and therefore, biological activity in the water. It is well known that fluorescence will be strongly correlated to other measured, ’light related’ variables, such as beam transmission or sea turbidity. One could also expect that the biological activity might be influenced by some physico-chemical conditions, *e.g.* the amount of dissolved oxygen, temperature, and salinity. Therefore, it will be interesting to compare the regression results of a PLS model when the predictors block (**X**) is constructed by all CTD variables and, as a second scenario, by the physico-chemical variables only.

In order to identify the optimal number of components for PCA and PLS models, the Cross-Validated Root Mean Square Error (RMESCV), illustrated in Figure A.3 was calculated by means of cross-validation with 81 contiguous blocks. That corresponds to one ’split’ being equal to 2 locations (200 observations for variable- and 2 for station-wise unfolded data).

A.3.2 3-way data analysis.

As the sea water measurements analyzed in this study follow three different modes (variable, depth and location), by unfolding the data cube and using 2-way techniques we risk that the 3-way correlation structure will be ’flattened’ and some information lost. PARAFAC model, which can be perceived as one of tri-linear extensions of PCA, is able to overcome this issue. Other ’extensions’ exist, such as Tucker3 model [Smilde et al.(1994)], however here we consider PARAFAC due to its simplicity and and easy interpretation of loadings. Moreover, it solves the problem of rotational freedom, typical to PCA, by providing a unique solution. Since the studies on a robust version of PARAFAC are still somewhat ambivalent, only the classical version of the algorithm will be presented. A number of components will be chosen according to four indices describing model's performance: explained variance, core consistency, number of iterations and total elapsed time, shown in Table A.2. An overview regarding component selection criteria can be found in [Bro and Kiers(2003)].

Analogously, a regression model can also be generalized to fit the 3-way data structure. In this case, we 'slice out' the **Y**-block (fluorescence) from the initial data cube. The remaining variables constitute a predictor block **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 will be considered: including all variables, or alternatively only these, related to physico-chemical properties.

A.3.3 Is 2- or 3-way better?

If the experimental data fulfills 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. Reasons for this are numerous. First of all, PARAFAC takes into account interrelations existing in all three data directions. Moreover, the problem of rotational freedom, typical to PCA is solved, as PARAFAC provides the unique solution (up to the scaling constant, sign and permutation ambiguities) [Smilde et al.(2004)]. In addition, PARAFAC model resolves each mode separately, giving a straightforward physical interpretation for the depth, station and variable 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, it does not tend to over-fit, as is often the case of PCA.

In spite of benefits that may be gained from applying a PARAFAC model, some drawbacks also exist. The most important one is that the real data do not always conform adequately to a tri-linear assumption (as in the case of oceanographic data). In this case, the model might return degenerated solutions. Degeneracy might also occur when a large number of factors is needed and if they are interrelated (compare with Tucker model). In most of these cases, the bilinear model is still appropriate and PCA or other methods like MCR, described by [Tauler(1995)], can be successfully applied.

A.3.4 Mapping method

An effective visualization tool is available when working with PARAFAC or location-unfolded PCA model. It projects the third mode location loadings directly onto a map and creates a kind of loading variability image. This map uses the method known in geo-statistics as ”kriging” and its conceptual background, which mathematically consists of random field interpolation techniques, might be technically complex, and therefore, will stay out of the scope of this paper. In brief, MATLAB’s mapping toolbox, which has been used in this work, allows to transfer the loading values according to their GPS coordinates on the specified fragment of the world’s map. Afterwards the 2D interpolation of these values is performed within the smallest convex set (results should be used with caution because they are insensitive to water/land borders) spanned by the locations coordinates. This way of representing the loadings offers an attractive tool for a better understanding and interpretation of the spacial variability in the data.

A.4 Results and Discussion

A.4.1 Data exploration

Figure A.3 shows the eigenvalues of covariance matrix and validated RMSE for variable- and station-unfolded data scenario. It can be seen that the two plots are dissimilar. In the first case two or three components seem to be enough, explaining 85% and 92% of data variability respectively, and it is clear that the model would over-fit if more components were chosen. For the latter scenario, the RMSE index decreases in value monotonously and therefore it is less evident what number of components is relevant. Eventually, in order to obtain similar variance explanation as in variable-wise case ($> 80\%$), a 3-component model is chosen.

Robust and classical PCA models show a similar performance when the station-wise unfolding of the data is considered, therefore only the results of the classical version will be discussed. The situation is different when it comes to the variable direction. Namely, 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 the Singular Value Decomposition and the resulting directions of Singular Vectors, which are likely to be attracted to the outliers still present in the data. Therefore only the output of the robust version of PCA for

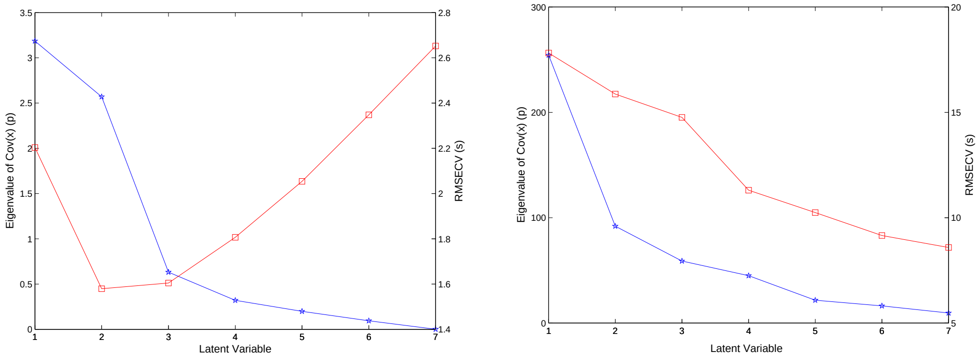

Figure A.3: Cross-Validated Root Mean Square Error (red squares) and eigenvalues of the covariance matrix X for the PCA model, for X being a) variable-wise unfolded data; b) station-wise unfolded data.

the variable direction will be finally reported.

In Figure A.4 scores (PC1, PC2, PC3) and loadings (on PC1 vs PC2 and PC1 vs PC3) for both fitted models are presented. It is quite straight forward to see the different behavior of scores corresponding to the Arctic Sea (red crosses) and Antarctica (blue circles), mostly along the PC1 and PC2 coordinates axis. In Figure A.4b, almost all Arctic locations were assigned with positive, and Antarctica with negative score values on PC1, and in Figure A.4c, red crosses have on average higher values than blue points. Another interesting property observed is the fact that the Arctic scores are more widely scattered throughout the coordinate system, which manifests higher inner variation within the that location. The third Principal Component does not seem to introduce any additional information and therefore its interpretation becomes onerous. One could argue that in order to avoid over-fitting a two component model would be more relevant here for the sufficient data explanation.

From the loading plot (Figure A.4a) it seems that there are two major types of depth profiles indicated by PC1 (blue solid line): one for physico-chemical variables and another for light related variables. By adding information from the score plot, we read that temperature, conductivity, salinity, fluorescence and sea turbidity have higher values in the Arctic Sea than in Antarctica (as their corresponding profile loadings are positive on PC1).

Moreover, two clusters are visible within the Arctic data: 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: 41-46 coming from the most North-West part covered by the expedition (apparently at the border of the ice where the ship could not move freely any longer, see Figure A.1). Therefore, we would expect them to have different characteristics (e.g. lower temperature) than the other locations investigated in the Arctic Sea. Finally, information obtained from 'variable direction' confirm that PC1, being again the component dividing the 2 polar areas, has high positive values for temperature, conductivity, salinity, fluorescence and sea turbidity. Therefore, all these variables present on average higher values in the Arctic Sea and only beam transmission would give higher scores in the Antarctic waters. On the other hand, PC2 is positively correlated with salinity and negatively with oxygen, however the geographical interpretation is more difficult as the scores from both locations are spread more evenly across this component. It seems that on average samples from Antarctica have slightly higher values on PC2 but more statistical tests should be completed in order 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 formally induces 3 different models (here only 2 were shown) and one should be careful with cross-

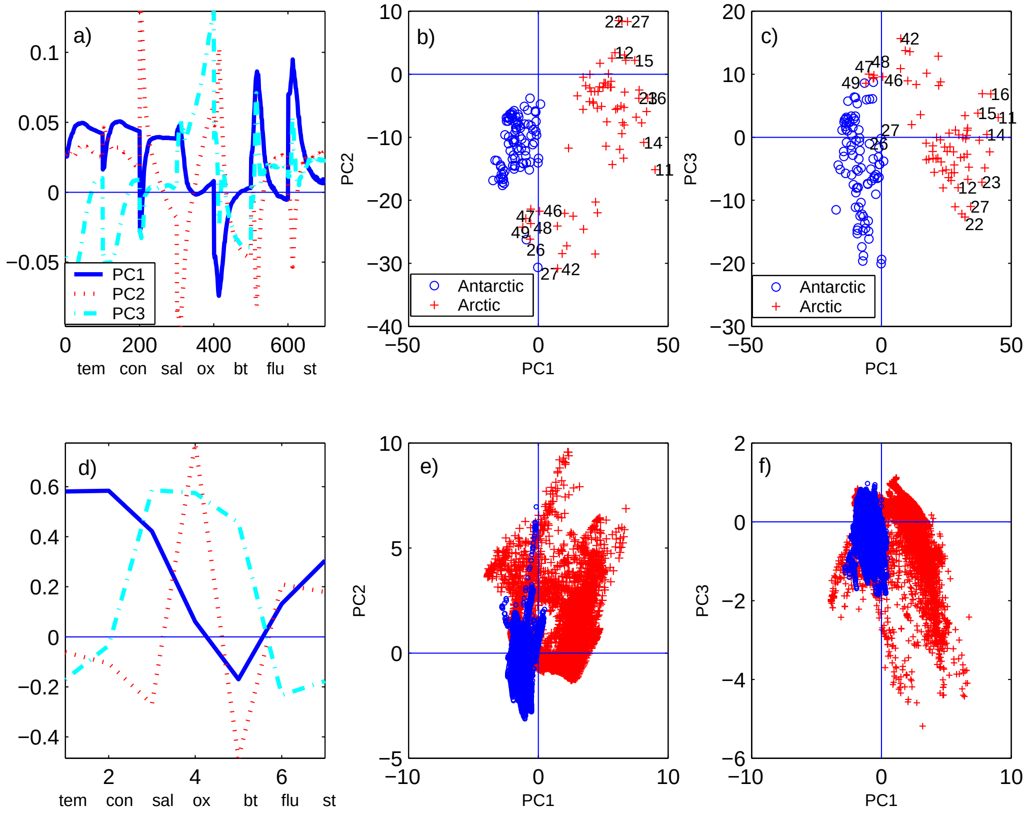

Figure A.4: Loadings and scores for PCA models on station- (up) and variable-wise (down) unfolded data.

Table A.2: Choosing amount of components in PARAFAC model according to four indices: explainced 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

interpreting their results, as there is no certitude, that for example PC1 in 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 A.2 presents four different indexes, which are normally used when determining the number of PARAFAC components. It is inferred 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% and converging fast to the minimum, as the model starts degenerating once the number of components increases. This degeneracy might be caused by the fact that the data does not follow the tri-linearity condition (see discussion in Section A.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, are given below in Figure A.5. From this picture, it becomes 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, given in Figure A.5a, describe the sea water changes observed from the surface to deeper sea water samples. The variable contributions resolved in the second mode are shown in Figure A.5b and finally, the location (geographical) profiles are presented in A.5c.

Information contained in these figures can be read as follows: first component describes the major changes occurring in 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 around 25m below the sea level, slowly declining afterwards with depth. Moreover, from image A.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 predominated by the influence of variables such as fluorescence, beam transmission, dissolved oxygen and sea turbidity with the positive drive from the first three of them. The depth profile for this component first increases to reach its peak around 15 meters below the sea level, where the average maximum of chlorophyll (DCM) is expected, and then decreases exponentially with depth. The location profile emphasizes again the differences between Arctic and Antarctica samples by attributing higher values (and also higher variance) to the Arctic area, with significant exceptions for some Antarctica samples (sample 27 being the South-East extreme station). From this, we could draw an initial conclusion that biological activity, reflected by fluorescence, is richer in the Arctic Sea. In addition, this 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 noticed, that changes on dissolved oxygen and fluorescence (biological activity) in this component are independent from changes to temperature, conductivity and salt content, in contrast to the pattern depicted by the first resolved component, where these variables were positively

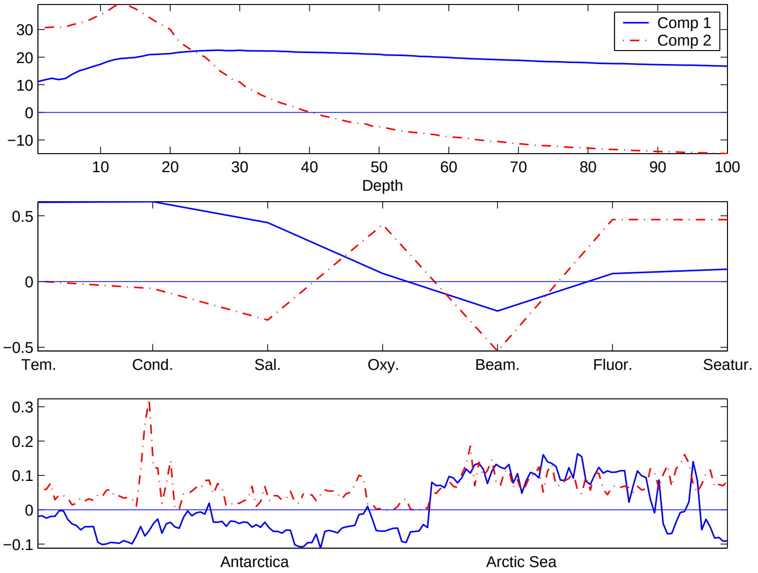

Figure A.5: Resolved profiles by 2-component PARAFAC model according to a) depth, b) variable, c) station mode.

inter-correlated. Moreover, the fact that the shapes of depth profiles for the two components are different, with their maxima around 25 and 15 meters for the first and second component 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 the time. This might be true in this case, however this kind of treatment would require fairly time consuming analysis of multiple plots, which would grow even more if higher number of variables was considered. It has been demonstrated that the more complex model structure was chosen, the easier was interpretation of its results. To sum up, the above analysis shows that: 1. similar information can be extracted by applying 2 (or 3 - not shown) PCA models or one PARAFAC model; 2. it is evident that interpretation of PARAFAC results is substantially easier, as it delivers concise information about all 3 data modes in only three figures; 3. care should be taken when choosing the number of components for both data as when this number grows it might lead to over-fitting the PCA model and degenerating PARAFAC results;

A.4.2 Map representation of the scores

In the previous section we found out 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 Antarctic and the Arctic areas, shown in Figure A.6, can then be interpreted via the variability image of those two major phenomena. By looking at the first component (two upper figures), we notice that in the Antarctica the gradient (direction in which the values grow) is pointed towards the North and in the Arctic towards the South-East. This confirms the expectation that the temperature (and other related variables) should rise when moving away from the Poles or approaching land (Svalbard). The biological activity, represented by the second component follows considerably different patterns as is illustrated in Figure A.6c. In the Antarctica region we can clearly observe an extremely high value at station 27. It is probably related to the different bio-characteristics of the region as it is located in the most South-East part covered by expedition, which might be more advantageous for biological activity. Alternatively, in the Arctic Sea one can distinguish the peaking area around stations 10-11 which might be caused by some local phenomena. In addition, the higher concentrations of bio-activity are located close to the ice border in the North. This region corresponds to relatively low temperatures, which is an observation worthy of

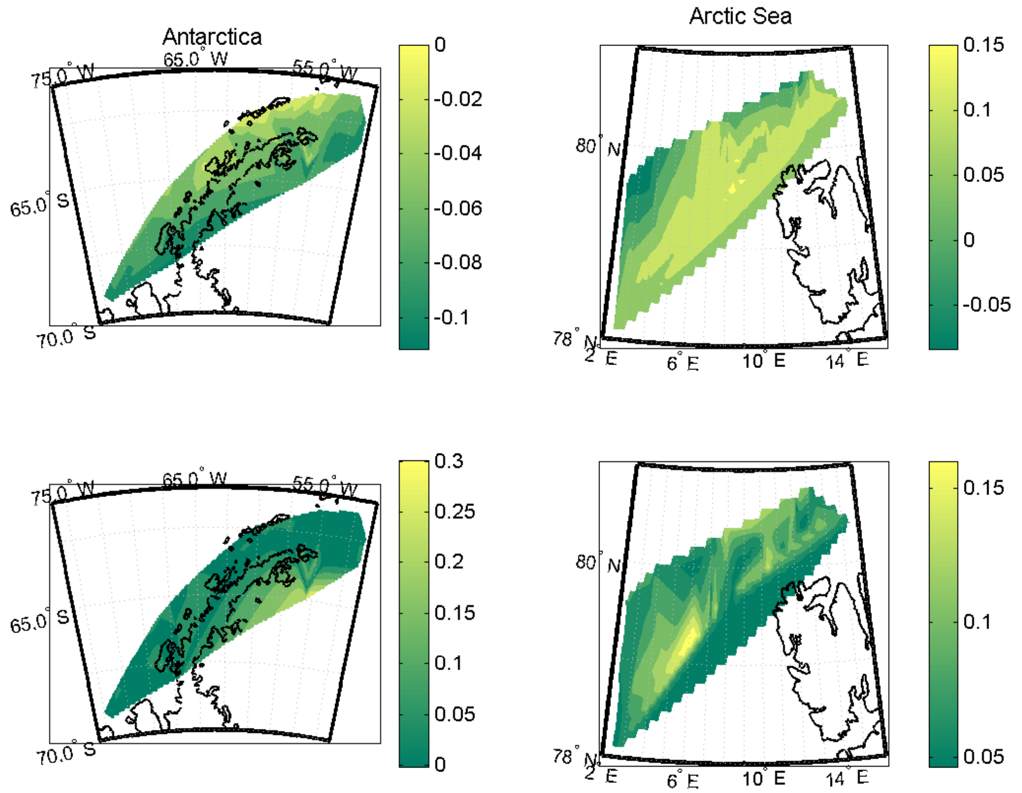

Figure A.6: Map representation of the Components 1 (up) and 2 (down) for the PARAFAC model for Antarctica (left) and the Artctic Sea (right).

note.

A.4.3 Regression

As it was previously mentioned, it is expected to discover a high correlation between radiation related variables and measured fluorescence. The most interesting results, however, will be generated when these variables are excluded from the explanatory variable X-block, which will give us information to which degree the fluorescence can be determined by physico-chemical conditions of the sea water. To start, 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) to be used, explaining respectively 92% and 82% of X- and Y-block variability (consult Table A.3). When inspecting the model's weights, which show the impact of each explanatory variable on Y, it doesn't come as a surprise that beam transmission (negative) and sea turbidity (positive) have the highest values, being both incorporated in the first latent variable. This is expected as these variables carry the light-related information, and therefore, they dominate LV1. Subsequently, the influence of temperature and conductivity was taken into account by LV2 and salinity and oxygen were manifested only in LV3. Alternatively, as a second scenario the radiation variables are removed from the predicting block. Now, the reduced, three-component model accounts for only 25% of the observed fluorescence variation, which is a very weak result. The control plot in Figure A.7a, showing measured versus predicted Y values, indicates that the model is not able to identify the difference in data behavior within Arctic Sea and Antarctica, 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 situation is quite different in the case of multi-linear PLS. Again we decided on a 3 component structure after consulting the cross-validation results. Remarkably, the N-PLS model with only physico-chemical variables as predictors, was now able to explain up to 78% of the measured fluorescence and around 74% of the X array. A plot of predicted versus observed values (Figure A.7b) confirms the obtained improvement. The complete results for both data scenarios are presented in Table A.3, from where we can conclude that 2-way PLS is largely outperformed by its 3 way alternative in predicting the fluorescence

Table A.3: Explained variance 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	X block	Y block
PLS	1	40.54	65.01	27.26	22.36
	2	70.30	70.47	81.42	22.64
	3	92.32	74.08	100	23.05
nPLS	1	26.66	60.40	46.00	31.56
	2	65.56	76.88	72.04	69.73
	3	72.55	85.50	83.42	79.13

values out of 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.

A.5 Conclusions

It has been shown that Arctic and Antarctica sea waters could be clearly differentiated, according to their CTD water samples collected during 2007 ATOS I and 2009 ATOS II polar expeditions. Two Principal Components have been identified by PCA and PARAFAC models, summarizing well the whole data set: 1st PC related to physico-chemical properties and 2nd PC accounting for 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

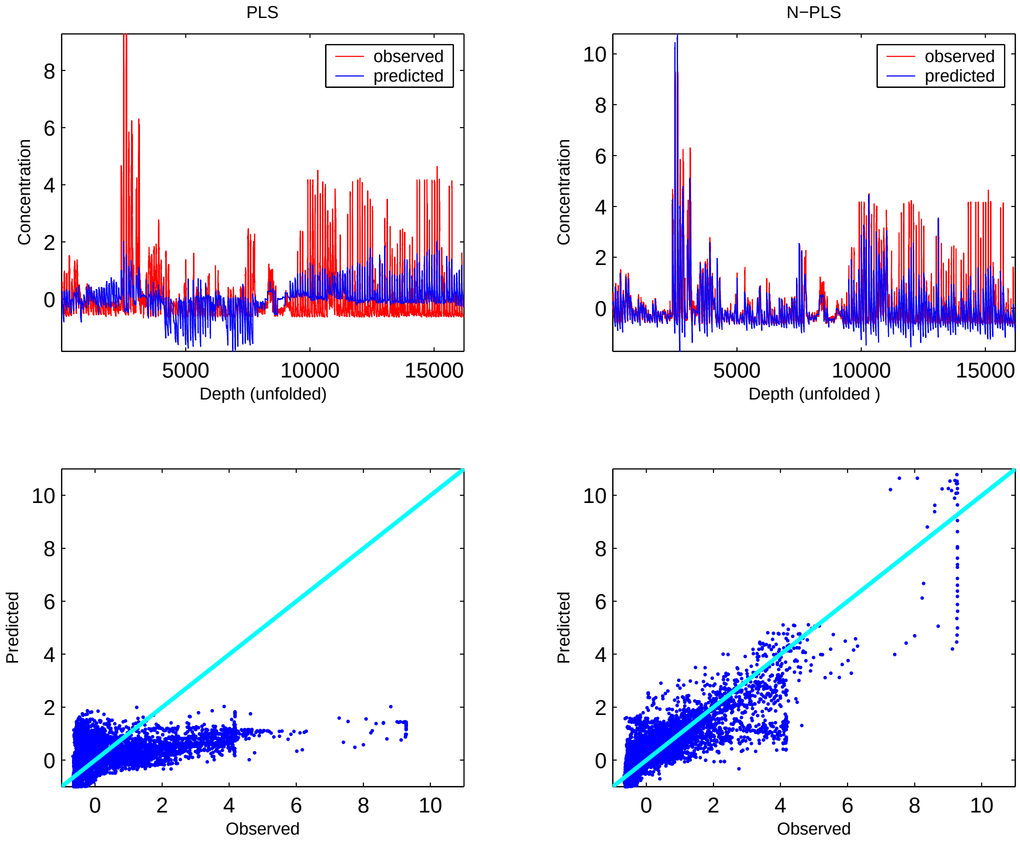

Figure A.7: Predicted versus observed values for PLS (left) and N-PLS (right) models, with only physico-chemical variables entering X-block.

measured CTD physical variables like temperature, conductivity, salinity and dissolved oxygen. This outstanding result could only be clearly revealed when the three-way data structure was considered in the regression model, and was completely hidden in the case of classical two-way unfolded PLS method.
Pros and cons related to 2- ad 3-way chemometric methods were analyzed and discussed, and the resulting conclusions can be formulated as follows:

1. similar set of information could be extracted by applying 2 (or 3 - not shown) PCA models or one PARAFAC model, however interpretation of PARAFAC results is substantially simpler, and more straight-forward

2. in general, 3-way methods will describe 3-way data better than 2-way models, if the data conforms to assumptions laying underneath 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, deteriorating significantly the result quality (see the PLS correlation model)

4. care should be taken when choosing number of components for both data, as when this number grows it might lead to over-fitting the PCA model and degenerating PARAFAC results

To sum up, recent instrumental developments within analytical chemistry, environmental sciences etc. cause that high-dimensional data sets frequently occur. This leads directly to higher requirements for data analytical tools, as often, simple statistical methods become not only highly time consuming, but in general no longer applicable to these vast data structures. Multivariate data analysis tools, such as presented in this paper, are therefore likely to become widely used in the future studies within environmental sciences inducing oceanography.

Appendix 1. Two- and three- way chemometrical methods

PCA

Principal Component Analysis is a linear subspace-based technique, perhaps most commonly found in chemometric literature. A PCA model is presented in Equation A.1:

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

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} contains model's residuals. In brief, PCA projects 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 determined by maximizing the variance explained by them. Those vector products, being a linear combinations of the original variables (or objects), are called principal components and often already a small number of those components allows us to explain the data variation in a satisfactory way. More details concerning PCA method can be found in the literature, for example in [Pearson(1901)], [Eckart and Young(1936)] or [Wold et al.(1987)].

PLS

Partial Least Squares regression is a 2-way calibration method. It approximates $\mathbf{X}$ block by r components (called latent variables) and, at the same time, projects $\mathbf{Y}$ on those components, which are constructed to compromise between fitting $\mathbf{X}$ and predicting $\mathbf{Y}$. This can be written in a matrix notation following [Smilde et al.(2004)]:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E}_X \ ; \ \mathbf{Y} = \mathbf{TQ}' + \mathbf{E}_Y \tag{A.2}$$

This algorithm is run sequentially, meaning that only one component is calculated at a time, and afterwards the $\mathbf{X}$ matrix is replaced by the residuals $\mathbf{E}^{-1} = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1$. Broader description of PLS and its applications can be found in [Wold et al.(1984)], [Martens and Naes(1992)] or [Smilde et al.(2004)].

PARAFAC

A PARAFAC model, introduced by [Harshman(1970)] and popularized by [**?**] and [Smilde et al.(2004)], is a tri-linear generalization of PCA, which decomposes a data cube $\mathbf{X}(I \times J \times K)$ into a sum of triple vector products, called loadings. The most common way of writing the model is following

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk} \tag{A.3}$$

where x_{ijk} is an element of $\mathbf{X}$, $\mathbf{A}(I \times R)$, $\mathbf{B}(J \times R)$ and $\mathbf{C}(K \times R)$ are the orthogonal matrices with elements a_{ir}, b_{jr} and c_{kr} respectively. R is number of components and e_{ijk} represents the error term. If the experimental data fulfils 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. Reasons for this are numerous. First of all, PARAFAC takes into account interrelations existing in all three data directions. Moreover, the problem of rotational freedom, typical to PCA is solved, as PARAFAC provides the unique solution (up to the scaling constant, sign and permutation ambiguities) [Smilde et al.(2004)]. In addition, PARAFAC model resolves each mode separately, giving a straightforward physical interpretation for the depth, station and variable 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, it does not tend to over-fit, as is often the case of PCA.

In spite of benefits that may be gained from applying a PARAFAC model, some drawbacks also exist. The most important one is that the real data do not always conform adequately to a tri-linear assumption. In this case, the model might return degenerated solutions. Degeneracy might also occur when a large number of factors is needed and if they are interrelated (compare with Tucker model [Smilde et al.(1994)]). In most of these cases, the bilinear model is still appropriate and PCA or other methods like MCR, described by [Tauler(1995)], can be successfully applied.

N-PLS

Multi-way PLS (or N-PLS) is a generalization of Partial Least Squares regression into a higher dimension, which predicts **y** and decomposes **X** similarly to the PARAFAC model [Bro(1996)], [Smilde(1997)], [de Jong(1998)]. This is performed by searching for a vector $\mathbf{t}$, being a linear combination of columns of $\mathbf{X}$, which has a maximum covariance with $\mathbf{y}$. This, [Smilde et al.(2004)] can be formulated in the following way

$$\mathbf{X} = \mathbf{t}(\mathbf{w}^K \otimes \mathbf{w}^J) + \mathbf{E}_X \ ; \mathbf{y} = \mathbf{t}b + \mathbf{e}_y \tag{A.4}$$

where $\mathbf{X}$ is an array of dimension $(I \times J \times K)$, $\mathbf{w}^J$ and $\mathbf{w}^K$ are weighting vectors defined for modes J and K, $\otimes$ defines the Kroneker product [McDonald(1980)] and b is the regression coefficient. As N-PLS is also a sequential method, after finding the first component, both $\mathbf{X}$ and $\mathbf{y}$ are being 'deflated' (replaced by residuals of the respective models), in order to recommence the algorithm.

Bibliography

[Bro(1996)] Bro, R., 1996. Multiway calibration. multilinear pls. Journal of Chemometrics 10 (1), 47–61.

[Bro(1997)] Bro, R., 1997. Parafac. tutorial and applications. Chemometrics and intelligent laboratory systems 38 (2), 149–171.

[Bro and Kiers(2003)] Bro, R., Kiers, H., 2003. A new efficient method for determining the number of components in parafac models. Journal of Chemometrics 17 (5), 274–286.

[de Jong(1998)] de Jong, S., 1998. Regression coefficients in multilinear pls. Journal of Chemometrics 12 (1), 77–81.

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

[Filzmoser et al.(2008)] Filzmoser, P., Maronna, R., Werner, M., 2008. Outlier identification in high dimensions. Computational Statistics and Data Analysis 52 (3), 1694–1711.

[Harshman(1970)] Harshman, R., 1970. Foundations of the parafac procedure: models and conditions for an" explanatory" multimodal factor analysis.

[Hubert et al.(2005)] Hubert, M., Rousseeuw, P., Branden, K., 2005. Robpca: a new approach to robust principal component analysis. Technometrics 47 (1), 64–79.

[Martens and Naes(1992)] Martens, H., Naes, T., 1992. Multivariate calibration. John Wiley & Sons Inc.

[McDonald(1980)] McDonald, R., 1980. A simple comprehensive model for the analysis of covariance structures: Some remarks on applications. British Journal of Mathematical and Statistical Psychology 33 (2), 161–183.

[Pearson(1901)] Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (11), 559–572.

[Rousseeuw et al.(2006)] Rousseeuw, P. J., Debruyne, M., Engelen, S., Hubert, M., 2006. Robustness and outlier detection in chemometrics. Critical Reviews in Analytical Chemistry 36 (3-4), 221–242.

[Serneels and Verdonck(2009)] Serneels, S., Verdonck, T., 2009. Principal component regression for data containing outliers and missing elements. Computational Statistics and Data Analysis 53 (11), 3855–3863.

[Smilde(1997)] Smilde, A., 1997. Comments on multilinear pls. Journal of Chemometrics 11 (5), 367–377.

[Smilde et al.(2004)] Smilde, A., Bro, R., Geladi, P., Wiley, J., 2004. Multi-way analysis with applications in the chemical sciences. Vol. 978. Wiley Online Library.

[Smilde et al.(1994)] Smilde, A., Tauler, R., Henshaw, J., Burgess, L., Kowalski, B., 1994. Multicomponent determination of chlorinated hydrocarbons using a reaction-based chemical sensor. 3. medium-rank second-order calibration with restricted tucker models. Analytical Chemistry 66 (20), 3345–3351.

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

[Tauler(1995)] Tauler, R., 1995. Multivariate curve resolution applied to second order data. Chemometrics and Intelligent Laboratory Systems 30 (1), 133–146.

[Wold et al.(1987)] Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemometrics and intelligent laboratory systems 2 (1), 37–52.

[Wold et al.(1984)] Wold, S., Ruhe, A., Wold, H., Dunn III, W., 1984. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. SIAM Journal on Scientific and Statistical Computing 5, 735.

Appendix B

Automatic Scatter detection in Fluorescence landscapes by means of Spherical Principal Component Analysis

Authors:

Ewelina Kotwa[1], Bo Jørgensen[2], Per B. Brockhoff[1] and Stina Frosch[2]

Published in:

Journal of Chemometrics, vol. 27, 2013.

[1] DTU Informatics, Richard Pedersens Plads, Building 321, DK-2800 Lyngby, Denmark

[2] DTU Food, Søltofts Plads, Building 227, DK-2800 Lyngby, Denmark

Received: 17 August 2012, Revised: 26 November 2012, Accepted: 30 November 2012, Published online in Wiley Online Library: 30 January 2013

(wileyonlinelibrary.com) DOI: 10.1002/cem.2485

Automatic scatter detection in fluorescence landscapes by means of spherical principal component analysis

Ewelina Kotwa[a*], Bo Jørgensen[b], Per B. Brockhoff[a] and Stina Frosch[b]

In this paper, we introduce a new method, based on spherical principal component analysis (S-PCA), for the identification of Rayleigh and Raman scatters in fluorescence excitation–emission data. These scatters should be found and eliminated as a prestep before fitting parallel factor analysis models to the data, in order to avoid model degeneracies. The work is inspired and based on a previous research, where scatter removal was automatic (based on a robust version of PCA called ROBPCA) and required no visual data inspection but appeared to be computationally intensive. To overcome this drawback, we implement the fast S-PCA in the scatter identification routine. Moreover, an additional pattern interpolation step that complements the method, based on robust regression, will be applied. In this way, substantial time savings are gained, and the user's engagement is restricted to a minimum, which might be beneficial for certain applications. We conclude that the subsequent parallel factor analysis models fitted to excitation–emission data after scatter identification based on either ROBPCA or S-PCA are comparable; however, the modified method based on S-PCA clearly outperforms the original approach in relation to computational time. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: S-PCA; Raman and Rayleigh scatters; robustness; PARAFAC; fluorescence

1. INTRODUCTION

Fluorescence spectroscopy is a measurement technique used for providing information about fluorophores, molecules which emit light when going from excited to ground state. The method is fast, sensitive and non-invasive, and found a broad usage in fields such as biochemistry, analytical chemistry, food, and environmental science. An outcome of a fluorescence spectrometer is usually written in a so-called excitation–emission (EEM) matrix representing intensities of the compounds for certain excitation ($j = 1, ..., J$) and emission ($k = 1, ..., K$) wavelengths. If a number of samples I are considered, the whole data form a three-way array, $\underline{\mathbf{X}}(I \times J \times K)$.

It is known that parallel factor analysis (PARAFAC) is able to explain mathematically the physical behavior of low-rank fluorescence data as these data conform to the trilinearity assumption, which lies at the foundation of the model [1]. PARAFAC, being a multiway decomposition method, usually expressed as

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

returns a set of resolved spectral curves, stored in **A**, **B**, and **C** terms for each of the three data modes. An analytical chemist or a chemometrician, being mostly interested in excitation and emission modes, is then able to identify particular fluorophores, present in the analyte, according to the peaks of the resolved spectra.

Rayleigh (first and second orders) and Raman light scatter effects appear in the EEM landscapes because of the physical properties of the fluorescence technique itself, namely the interactions between the molecules in the solution causing some incident light. These diagonal ridges do not provide any additional chemical information about the investigated samples. On the contrary, as the placement of the scatter peaks varies with excitation wavelength, the low-rank trilinear model assumption is violated, which might cause a significant deterioration of the PARAFAC performance. If that is the case, Rayleigh and Raman scatters should be identified and removed from the data set prior to PARAFAC-based analysis.

Mitigation of the destabilizing influence of scattering effects was a subject of several investigations after the PARAFAC model gained popularity. Techniques such as down-weighting [2,3] or specific modeling of scatter regions [4], inserting missing values [1], interpolation [5], constraining PARAFAC decomposition [6], or inserting zeros outside of data area [7] were discussed in the literature; however, they only propose a particular treatment of the scatter signal, assuming that it is predefined according to

* Correspondence to: E. Kotwa, DTU Informatics, DTU Asmussens Alle, Building 305 2800 Kgs. Lyngby, Denmark
Email: ek@imm.dtu.dk

a E. Kotwa, P. B. Brockhoff
DTU Informatics, DTU Asmussens Alle, Building 305 2800 Kgs. Lyngby, Denmark

b B. Jørgensen, S. Frosch
DTU Food, DTU Søltofts Plads, Building 227 2800 Kgs. Lyngby, Denmark

a (subjective) visual inspection and not by statistically meaningful techniques. Only subtracting a standard [8] can be perceived as an "objective" technique; nevertheless, its usage can be hampered by several inconveniences: (i) subtraction of two large values can leave a significant residual that will still violate PARAFAC assumptions (yielding still degenerated spectra); (ii) blanc subtraction might help dealing with Rayleigh scatter but do not solve the problem of Raman scatter; and (iii) not for each EEM data it is possible to obtain a valid blanc.

A few years ago, Engelen *et al.* [9] successfully used the robust statistics framework for automatic identification of the scatter regions. That new approach proposed certain transformations of the data cube, after which the scatter signal could be considered as element-wise outliers. The method was based on ROBPCA described in [10], which provides an excellent outlier detecting tool; however, it appeared to be computationally intensive, especially if large data structures were to be considered. This became a reason to search for a faster alternative to ROBPCA, which will be presented in this paper.

In the following sections, the concept of spherical principle component analysis (S-PCA) will be described, together with its usage in identifying the scatter regions within the EEM data. First, the methodological background will be given in Section 2 and later applied to two different data sets (Section 3). Finally, the results will be discussed in Section 4.

The calculations were conducted in MATLAB using an open source package LIBRA (KU Leuven, Belgium http://wis.kuleuven.be/stat/robust/LIBRA.html/ [16 August 2012]) for robust statistics and the PLS Toolbox (Eigenvector Research Inc., Wenatchee, WA 98801, USA) for PARAFAC applications.

2. METHODOLOGY

2.1. Combination of S-PCA and PARAFAC

As previously stated, Rayleigh and Raman scattering effects might be highly disturbing when fitting a PARAFAC model to the EEM data, and therefore, their elimination techniques are of a great interest to practitioners. The method considered in this contribution combines the S-PCA for automatic identification of these undesired patterns, with PARAFAC modeling of the corrected data set for resolving EEM spectra. This approach assumes that the scatter is a set of outlying observations, which differ in behavior from the rest of the data. Following [9], if we slice our three-way data array $\underline{\mathbf{X}}$ according to the excitation and emission modes and, subsequently, transpose the resulting matrices, the scattering effect will be placed in the rows of those matrices. As a consequence, these rows can be perceived as sample-wise outliers and hence detected by some robust PCA methods applied to each matrix separately. Both the slicing and transposing operations are illustrated in Figure 1. A substantial advantage of applying the robust framework over its classical counterpart is twofold: robust models are fitted to the majority of data points, avoiding the common pitfalls of least squares-based methods, which can be largely influenced by some extreme observations (such as those contained in the scatter lines) and approximate the main data poorly. Secondly, observations that do not follow the fitted model can be flagged as outliers and can easily be localized and eliminated.

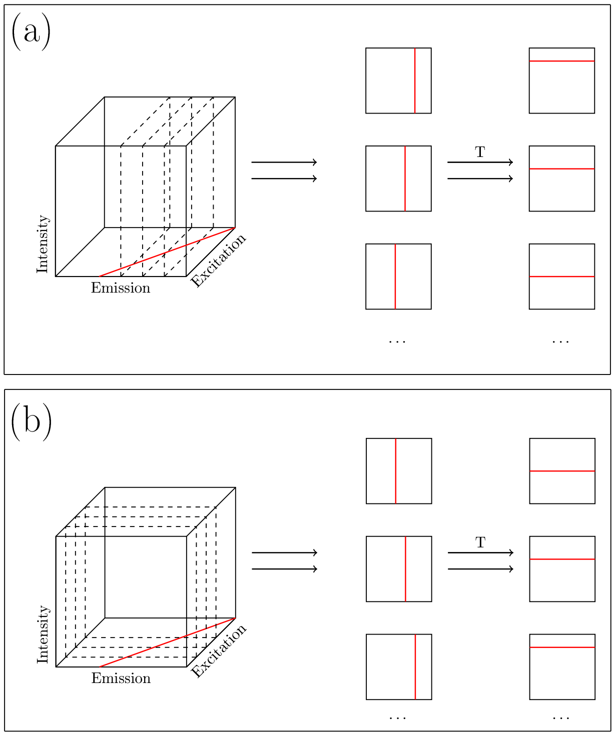

Figure 1. Slicing of the data cube according to (a) mode B (emission) and (b) mode C (excitation) in order to contain the scatter line as element-wise outliers within a set of matrices.

After having identified, the scatter lines by means of robust PCA, an adequate treatment of these areas has to be elaborated before a PARAFAC model can be employed. In this paper, three scenarios will be considered to fulfill this requirement: setting the identified scatter region to "missing values", "down-weighting", and an "interpolation" method. Additionally, nonnegativity and unimodality constraints will be applied if relevant.

2.2. PCA, ROBPCA, and S-PCA

Principal component analysis is a well-known statistical method for data compression and visualization. The main idea is to explain the data by means of preferably a small number of "artificial," orthogonal variables called principal components (PCs), being linear combinations of the genuine variables. The PCs are constructed by searching for directions of the highest variance throughout the data and involve the largest amount of information possible to be extracted by linear combinations. A more detailed description can be commonly found in the literature [11–13]. Even though PCA carries very beneficial characteristics, one must not forget that its algorithm is based on the least squares framework, and therefore, it automatically encounters some difficulties when a certain number of outlying elements appear in the data set. In fact, its breakdown point [14] equals zero, which means that in the extreme case, even one outlier can severely damage the fit and cause further errors in data interpretation.

To deal with this shortcoming, various robust alternatives were elaborated, based either on robust covariance matrix approach (e.g., M-estimators [15], S-estimator [16], or MCD [14]), the projection pursuit [17], or their combination (ROBPCA developed in [10]). For a review of the common robust methods used in data analysis, see [18,19]. Implementing ROBPCA for scatter identification was described in [9,20]; however, the fact that the method appeared computationally intensive when analyzing vast data structures encouraged a search for the faster alternative.

In this paper, we focus on S-PCA, being a very fast and robust version of classical PCA. This method exploits the robustness feature of the median, combined with the projection pursuit framework and is conceptually fairly simple. In the first step, the robust center of the data is defined as the L1-median [15], being a point in the multidimensional space found by minimizing the sum of Euclidean distances from the data objects to this point. Then, all data points $\mathbf{x}_i$ are centered and down-weighted by the inverse of these distances, which is equivalent to the projection on the unit radius hyper-sphere with the center in $L1$:

$$\mathbf{x}_i^p = \frac{\mathbf{x}_i - \boldsymbol{\mu}_{L1}(\mathbf{X})}{\|\mathbf{x}_i - \boldsymbol{\mu}_{L1}(\mathbf{X})\|} + \boldsymbol{\mu}_{L1}(\mathbf{X}) \qquad (2)$$

where $\mathbf{x}_i^p$ is the ith data object projected onto the sphere and $\|\mathbf{x}_i - \boldsymbol{\mu}_{L1}(\mathbf{X})\|$ is the Euclidean distance to the robust center of the data $\boldsymbol{\mu}_{L1}$. In this setting, the influence of outliers manifesting normally large distances in the denominator of the equation will be bounded. In the next step, a classical PCA is carried out on the down-weighted data, and the robust scores and loadings are obtained by projecting the original data on the resulting PCs. Further, to identify the outlying samples, the usual detection methods based on robust and orthogonal distances (RD and OD, respectively) are to be applied. Moreover, the Euclidean (and/or Mahalanobis) distance, itself, to the center of the data also proves to be advantageous in certain situations. The robust distance (RD_i), which is expressed as

$$RD_i = \sqrt{\sum_{j=1}^{k} \frac{t_{ij}^2}{l_j}} \qquad (3)$$

where $\mathbf{t}_j$ is a vector of robust scores for the jth component and $\mathbf{l}$ contains robust eigenvalues, is actually a Mahalanobis distance to the robust data center. On the other hand, the orthogonal distance (OD) informs how far, with respect to the space spanned by the robust model, the ith object is located.

$$OD = \|\mathbf{X} - \boldsymbol{\mu}_{L1}(\mathbf{X} - \mathbf{TP}^T)\| \qquad (4)$$

The $\|\cdot\|$ denotes the Euclidean norm, and the model $\mathbf{TP}^T$ contains k-selected components.

To establish cutoff values for all the distances, a robust version of z-scores was suggested in [21], by incorporating robust location and scale estimates such as median (μ) and median absolute deviation (σ) as an alternative to the standard procedures:

$$z = \frac{|d - \mu(d)|}{\sigma(d)} \qquad (5)$$

On the basis of this outlier identification procedure, the output of S-PCA algorithm is given in a form of a binary matrix $\mathbf{M}(J \times K)$, which will be referred to as an "outlier mask." This mask attributes ones to the regular observations and zeros to the elements identified as outliers. In the proposed implementation of the method, the user is supplied with three different masks related to the prior slicing mode of the data cube $\mathbf{X}$: mode B, mode C, and their product. The reason for this is that the algorithm will perform differently in each of these modes, and depending on the data circumstances, one of the proposed option might be more beneficial than others. A great advantage of the S-PCA approach is its exceptional computation speed and relative simplicity, which can be desired in many applications. It is known, however, that the method tends to fail when the outliers form clusters [22]. As a consequence, it might lead to the situation where not all corrupted data will be regarded as outlying samples, or opposite, some signal can be wrongly eliminated from the further analysis. In the following section, we will describe a possible solution to those shortcomings.

2.3. Correction method

As stated earlier, S-PCA will occasionally fail to identify the outlying samples if particular data circumstances occur. In the case of EEM landscapes, some regions, where the scatter ridges are present, might remain undetected. This phenomenon is especially common, when the scatter and signal intensities are on average of similar magnitude and therefore lay in a similar distance from the robust center of the data, which makes them nondistinguishable by RD and OD.

In this case, additional steps can be taken to ensure a better performance of the sub-sequentially fitted PARAFAC model. The proposed methodology uses physical properties of the scatter areas, namely their linear nature. In the first step, observations flagged as outliers by S-PCA are being projected onto a two-dimensional coordinate system. Subsequently, a robust regression (here, based on the least trimmed squares estimator [16]) is applied. This method has a high breakdown point ensuring that the calibration line will match the majority of the data points, which in this case would be the first scatter stripe. In parallel, observations laying far from the regression line will be flagged as outliers. Afterwards, a width of the line is being determined, usually by adopting the broadest band of the initially discovered pattern (default); however, user-made adjustments are also allowed if needed. The same two operations are repeated for the second scatter line, which was flagged as outliers by the previously fitted regression model. An example of the interpolated scatters can be seen in Figure 4 and 5.

There are a few points that are worthy of attention when applying this approach. First of all, unlike in the case of ROBPCA, here, the users participation is required when selecting which one of the three images (outlier masks) is to be interpolated (Figure 4) and, optionally, adjusting the width of the corrected scatter. The first choice should be dictated by considering the image that grasps relatively large amount of discovered scatter and, at the same time, low quantities of mis-identified signal observations (to ensure that the model fits the majority of the data being actually a scatter line). Secondly, the bandwidth around the regression line should be determined so that it corresponds to scatter thickness. Practice shows that it is more beneficial to remove some extra signal than leaving out erogenous observations; however, care must be taken as it is always a matter of trade-off. A more detailed algorithm for the correction method can be found in the Appendix, at the end of this work.

3. Data

3.1. Dorit

Dorit data is a set of laboratory prepared samples consisting of four fluorophores mixed in different concentrations (these are phenylalanine, 3,4-dihydroxyphenylalanine, 1,4-dihydroxybenzene, and tryptophan). This data has already been

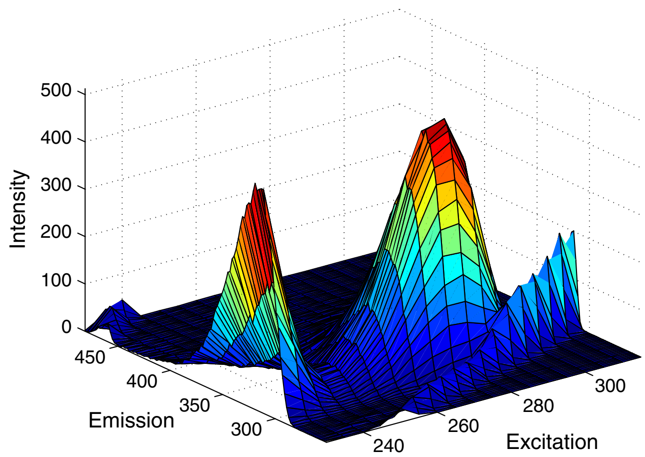

Figure 2. Dorit data with the visible Rayleigh scatter.

used in numerous investigations of the similar nature because of its “nice” properties and known components [9,23]. A total number of 27 EEM landscapes was produced, measuring the emission spectra ranging from 250 to 482 nm (with 2-nm intervals), for excitation wavelengths within 200 and 315 nm every 5 nm. The first-order and second-order Rayleigh ridges are present in every data sample, which can be clearly seen in Figure 2. No Raman scatter has been identified. The previous research suggested removing observations corresponding to excitation between 200 and 230 nm and emissions lower than 250 nm as they contain mainly missing elements. Additionally, four samples are known to be corrupted [24] and will be kept aside for further considerations as the scope of this paper is to handle scatter-like nuisance only. This leads to a (23 × 116 × 18) data array used in this work.

3.2. North Sea

The second data set used for testing the method is composed by samples of dissolved organic matter in water coming from the Dogger bank in the North Sea (see [9]). A total of 37 EEM landscapes of unknown fluorophores were produced, measuring emissions in a range of 240–600 nm every 2 nm, corresponding to the excitation wavelengths from 240–450 nm with an interval of 5 nm. The first 39 emission wavelengths were eliminated because of the presence of some artifacts, whose analysis is of no importance in this work. The resulting data cube has the size of (37 × 142 × 43). As no blank was available to diminish the scatters, the data suffers from highly disturbing Raman and Rayleigh effects that make it impossible to determine the data structure visually. Figure 3(a) illustrates this issue, and Figure 3(b) shows the data after removing the scatter ridges, where the signal becomes recognizable. Another difficulty related to this data set is a high noise-to-signal ratio, which will surely have important (and negative) impact on the results quality.

4. RESULTS AND DISCUSSION

Because both data sets used in this study were already analyzed and discussed previously, the results included in this section will only reflect the performance and comparison of the two methods (S-PCA and ROBPCA) related to the identification of scatter regions, excluding the chemical analysis of the data itself, for which appropriate references should be consulted ([9,23] for the Dorit and [9] for the North Sea data).

The Dorit data, presented in Figure 2, is one example of EEM data where the two Rayleigh scatter ridges manifest themselves: first for *Emission* = *Excitation* and second when *Emission* = 2$*$*Excitation*. A four-component PARAFAC model applied directly to the raw data returns degenerated profiles (see [9] or [20]); therefore, a scatter treatment is necessary to assure reliability of the results. Figure 4 illustrates the first output of the scatter identified by S-PCA for the Dorit data, where the user is shown three outlier masks (Figure 4(a)–(c)). As none of these masks cover the scatter region fully (leaving some kind of a “hole” in the middle of the pattern), it is advised to implement the correction step described in Section 2.3. The user is asked then to choose between one of the three images to undergo the correcting method. In the case of the Dorit data, it would be sensible to select the left-side or middle image (Figure 4(a) or (b)), as they include both scatter lines and a very low amount of wrongly identified data. In the third image (4c), the Raman scatter is not visible, which would cause that the correction algorithm, set to finding two scatter stripes, would fit the second line wrongly. The final result of the correction is visible in the most right-side image (Figure 4(d)), which is incorporated into the subsequent PARAFAC analysis. Figure 6 represents the resolved spectra from the PARAFAC models fitted to three different data scenarios: (i) when S-PCA was applied without the correction adjustment, where the influence of undiscovered scatter is still visible in the appearance of the resolved curves; (ii) with corrected scatter region, as in Figure 4(d), where a significant improvement of the resolved curves is observed; (iii) by additionally applying a nonnegativity constraint to both excitation and emission modes, resulting in “perfectly” resolved profiles of the four investigated compounds. For comparison, some quantitative results of the PARAFAC models based on both S-PCA and ROBPCA, for different scatter region treatment scenarios and constraints application, can be found in Table I. From the indices included in the table and the resolved excitation/emission profiles (detailed ROBPCA results can be found in [9]), it can be judged that the two methods perform similarly good and contribute significantly to a successful recovery of the underlying spectra within the data. The core consistency and explained variance, both having values > 90%, indicate slightly in favor of S-PCA approach, whereas the time of PARAFAC model and number of iterations seem to be uninfluenced by the scatter identification method. The obvious point, where S-PCA largely outperforms the previous method, is the time of finding the scatter region, which in the case of Dorit data is around 100 times shorter than for ROBPCA.

The new method was also challenged on a more difficult, very noisy data set of unknown fluorophores, called “North Sea” (Figure 3). The previous research (see [9]) showed that five different peaks can roughly be identified on both emission and excitation profiles, but as in the previous case, the severe first-order and second-order Rayleigh scatter ridges have to be removed before fitting a PARAFAC. Again, after applying an identification method based on S-PCA, the user chooses one of the three images to undergo the correction routine. For this data constellation, the two scatter lines were already well marked by the first part of the algorithm and, therefore, the correction focuses mostly on adjusting the thickness of those lines. Figure 5 illustrates those proceedings. Further, a five-component PARAFAC model was fitted to the resulting versions of the scatter-adjusted data set (Figure 6). Missing values and interpolation are the only presented methods for treating the identified scatter regions (“weighting” returned poor results, perhaps because of a very noisy nature of the data), and

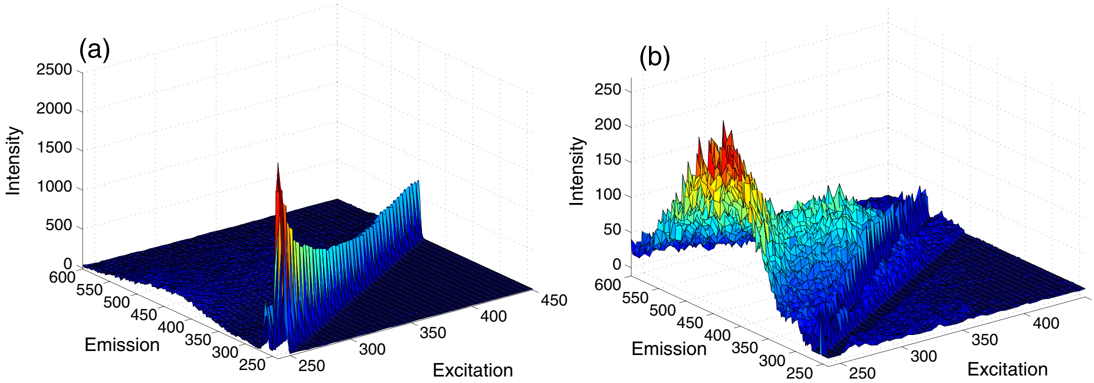

Figure 3. North Sea data (a) with the visible Rayleigh and Raman scatters and (b) after manual scatter removal.

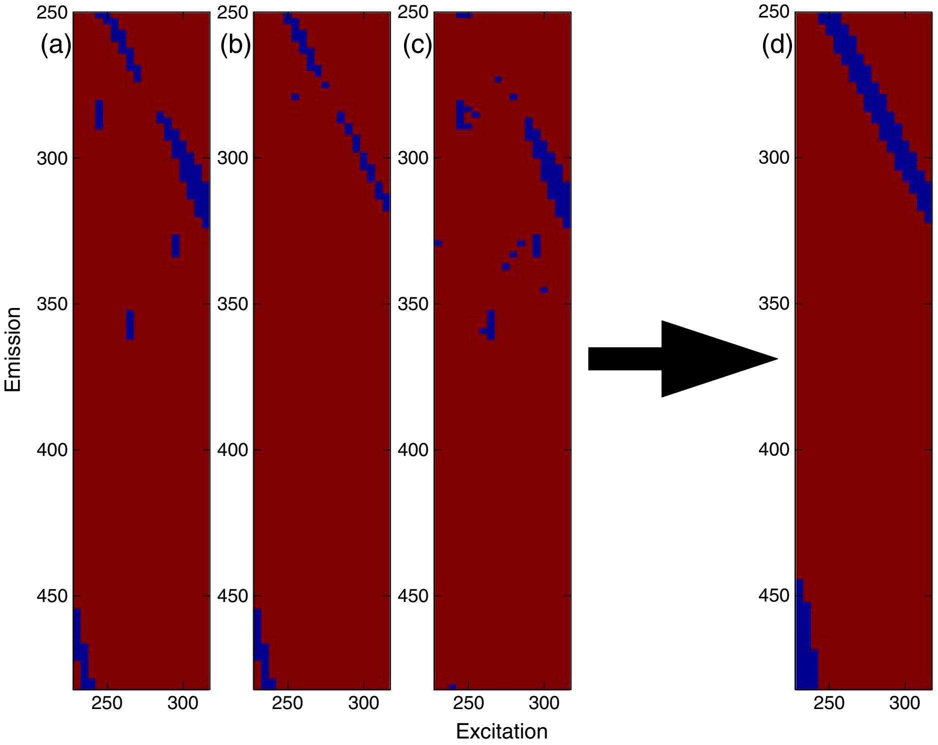

Figure 4. Correction of spherical principle component analysis (S-PCA) for the Dorit data: (a)–(c) outlier masks after initial S-PCA routine and (d) interpolated (corrected) scatter region.

Table I. PARAFAC results for the Dorit data expressed by the standard set of indices. It is evident that the time required for finding the scatter by S-PCA is significantly lower than by ROBPCA.

	Name of method	Time of scatter finding (s)	Time of PARAFAC (s)	No. of Iterations	Core consistency (%)	Variance explained (%)
S-PCA	No correction	2.36	1.99	29.47	98.82	97.74
	Correction	3.07	1.65	38.47	99.72	99.91
	Correction + nonnegativity	3.07	3.30	44.25	99.73	99.90
ROBPCA	No constraints	293.28	54.94	44	99.30	99.78
	Nonnegativity	293.30	26.89	57	99.73	99.77

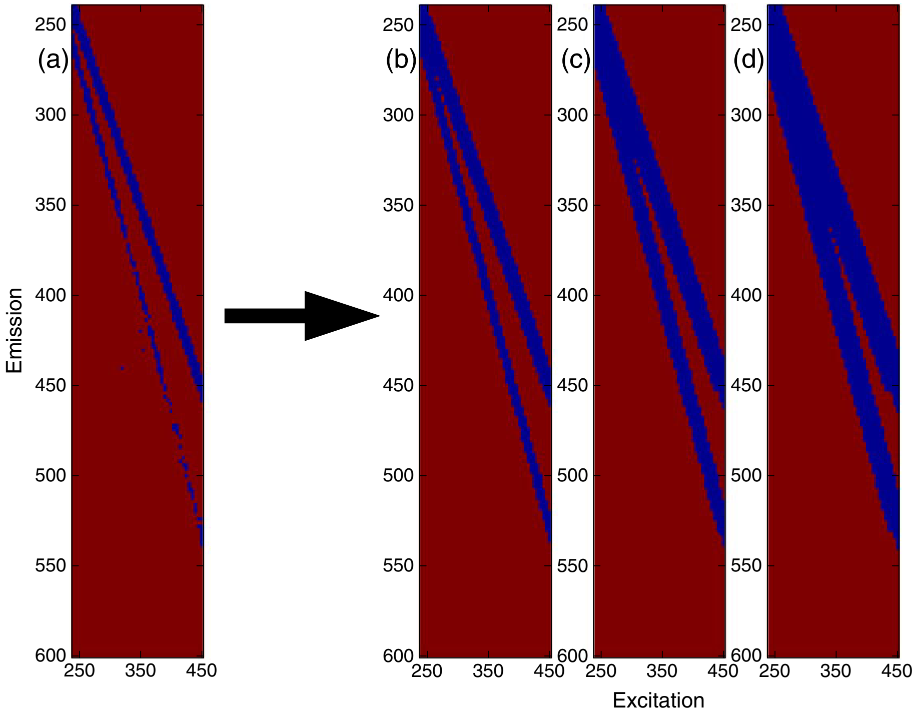

Figure 5. Correction of spherical principle component analysis (S-PCA) for the North Sea data: (a) user-chosen outlier mask after initial S-PCA routine and (b)–(d) interpolated scatter regions depending on the user-chosen bandwidth adjustment factor set to 0, 1, and 2, respectively.

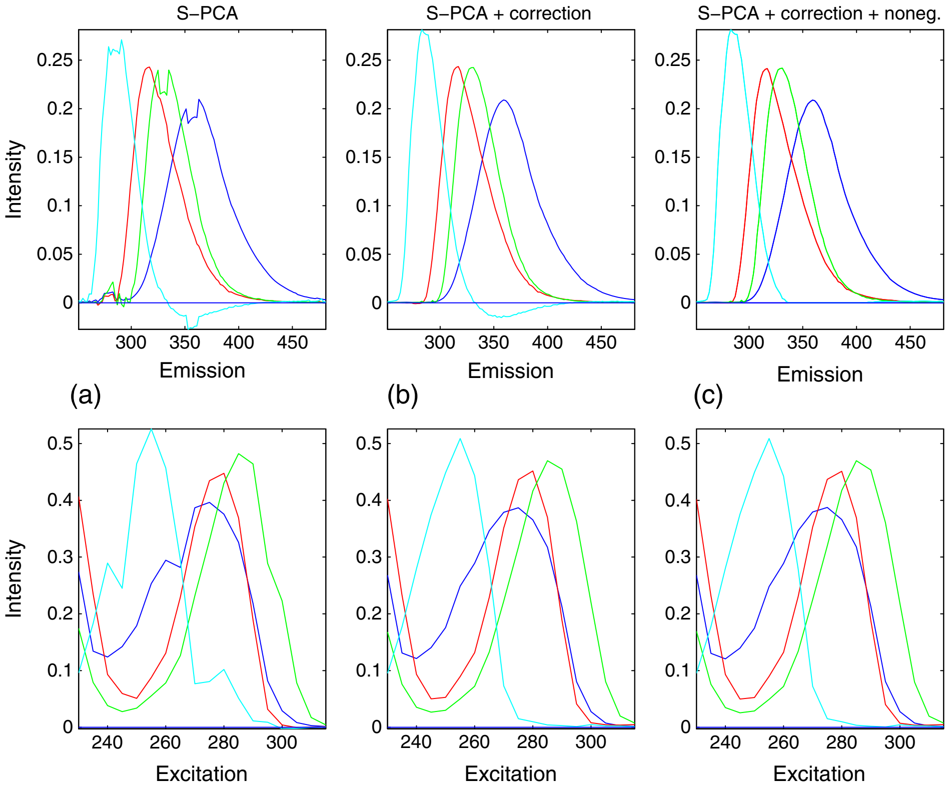

Figure 6. Parallel factor analysis resolved spectra for the Dorit data: (a) spherical principle component analysis (S-PCA) without correction method, (b) S-PCA with correction, and (c) S-PCA with correction and nonnegativity constraint.

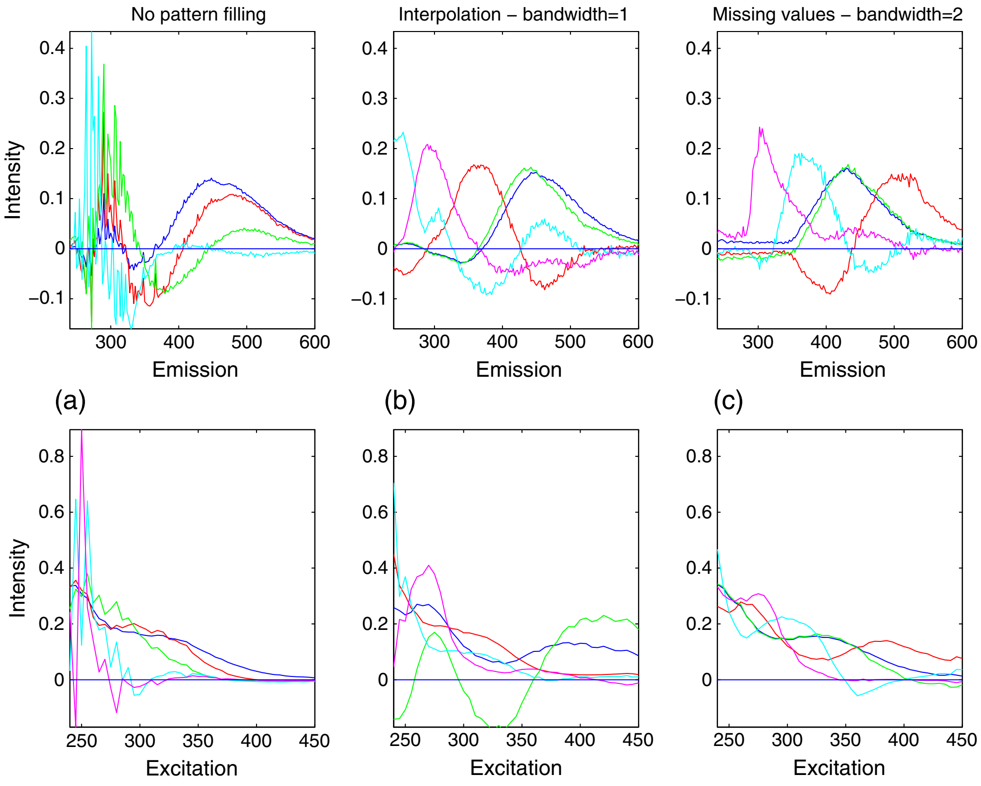

Figure 7. Parallel factor analysis resolved spectra for the North Sea data: (a) spherical principle component analysis (S-PCA) without correction method, (b) S-PCA with correction and bandwidth adjustment d = 1, and (c) S-PCA with correction and bandwidth adjustment d = 2.

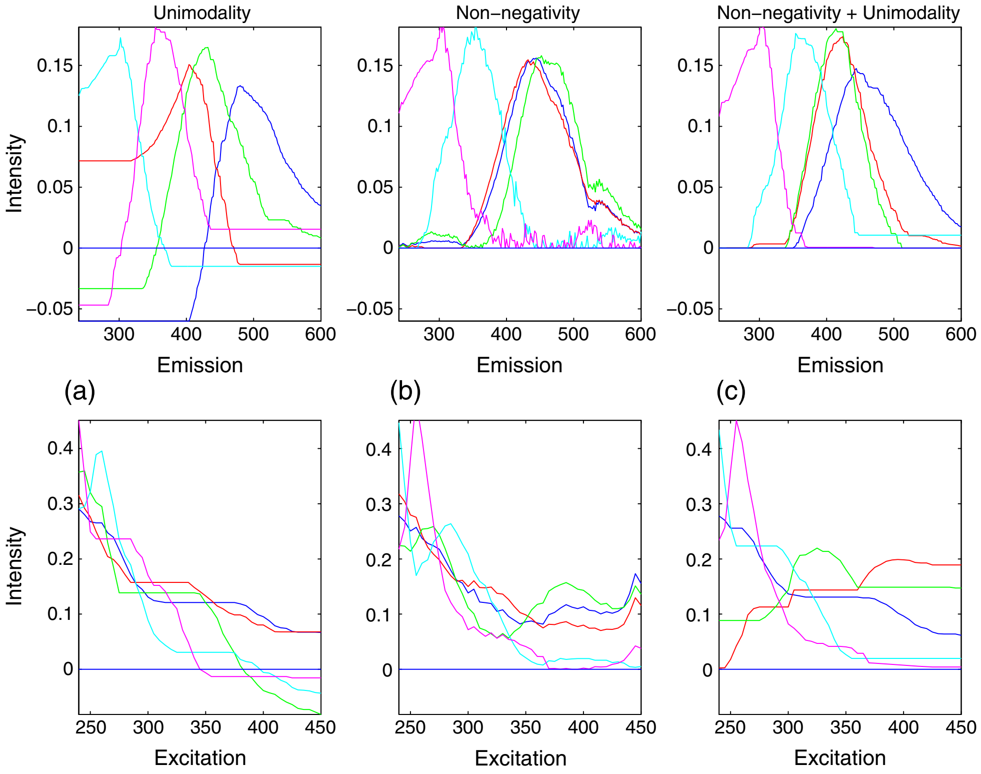

Figure 8. Parallel factor analysis resolved spectra for the North Sea data: (a) spherical principle component analysis (S-PCA) with correction and unimodality constraint, (b) S-PCA with correction and nonnegativity constraint, and (c) S-PCA with correction and unimodality and nonnegativity constraints.

Table II. PARAFAC results for the North Sea data. Negative values of core consistency indicate that PARAFAC encounters difficulties in fitting the data, which is most likely caused by the noisy character of the data. Yet, S-PCA identifies scatter region much faster than ROBPCA.

Name of method		Time of scatter finding (s)	Time of PARAFAC (s)	No of iter.	Core consistency (%)	Variance explained (%)
S-PCA	No correction	12.32	31.059	170	−41, 235.55	83.64
	Correction band = 0	13.41	43.084	170	−634.54	96.32
	Correction band = 1	13.44	48.33	107	−1306.18	97.95
	Correction band = 2	13.47	29.058	176	−804.07	98.22
	Correction + unimodality	13.47	54.257	947	−178, 821.93	97.98
	Correction + nonnegativity	13.47	39.197	275	−62, 838.01	97.98
	Correction + unimodality + nonnegativity	13.47	15.233	115	−18, 040.26	97.94
ROBPCA	No constraints	403.60	46.31	175	−1059.38	98.06
	Unimodality	403.60	41.8	636	−1, 776, 998.60	97.81
	Nonnegativity	403.60	12.90	144	−57, 907, 920.21	97.04
	Unimodality + nonnegativity	403.60	6.38	371	−50, 811, 514.17	97.77

additionally, the possible improvements by applying nonnegativity and unimodality constraints were verified. Some visualizations of the results are presented in Figure 7 and 8, and the quantitative measures, in relation to ROBPCA, are given in Table II. It is clear that because of the high noise-to-signal ratio, the PARAFAC model fits the data worse than in the case of Dorit data, which is manifested in the negative core consistency values for each model. Moreover, the reliability of the resolved spectra is somehow doubtful as it seems that four or five components could be fitted, depending on the amount of removed scatter and method used for dealing with the identified regions. Judging from the variance explanation index, it is more favorable to choose the thicker bandwidth in the correction step, as its value grows from 94% to 98% when increasing the bandwidth by a factor 2. As in the case of Dorit data, the strength of the new proposed method is recognized in the exceptionally fast algorithm for finding the scattering effect regions, which is around 13 s for S-PCA compared with > 400 s for ROBPCA. In parallel, the performance of the PARAFAC model is comparable for both underlying methods.

5. CONCLUSIONS

In this paper, we have applied S-PCA as a fast alternative to ROBPCA method for scatter detection in EEM landscapes. It is concluded that results of the PARAFAC models, fitted to the scatter-free data, based on the two techniques are comparable in performance. The proposed implementation of S-PCA is not fully unsupervised, and some user participation is required; however, it significantly outperforms the ROBPCA in terms of computation efficiency.

REFERENCES

1. Bro R. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems* 1997; **38**: 149–171.
2. JiJi RD, Booksh KS. Mitigation of rayleigh and raman spectral interferences in multiway calibration of excitation–emission matrix fluorescence spectra. *Analytical Chemistry* 2000; **72**: 718–725.
3. Rinnan Å, Andersen CM. Handling of first-order Rayleigh scatter in parafac modelling of fluorescence excitation–emission data. *Chemometrics and Intelligent Laboratory Systems* 2005; **76**: 91–99.
4. Rinnan Å, Booksh KS, Bro R. First order Rayleigh scatter as a separate component in the decomposition of fluorescence landscapes. *Analytica Chimica Acta* 2005; **537**: 349–358.
5. Bahram M, Bro R, Stedmon C, Afkhami A. Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation. *Journal of Chemometrics* 2006; **20**: 99–105.
6. Chen ZP, Yu RQ. Mitigating model deficiency in three-way data analysis by the combination of background constraining and iterative correcting techniques. *Analytica Chimica Acta* 2003; **487**: 171–180.
7. Thygesen LG, Rinnan Å, Barsberg S, Moller JKS. Stabilizing the PARAFAC decomposition of fluorescence spectra by insertion of zeros outside the data area. *Chemometrics and Intelligent Laboratory Systems* 2004; **71**: 97–106.
8. Wentzell PD, Nair SS, Guy RD. Three-way analysis of fluorescence spectra of polycyclic aromatic hydrocarbons with quenching by nitromethane. *Analytical Chemistry* 2001; **73**: 1408–1415.
9. Engelen S, Frosch S, Hubert M. Automatically identifying scatter in fluorescence data using robust techniques. *Chemometrics and Intelligent Laboratory Systems* 2007; **86**: 35–51.
10. Hubert M, Rousseeuw PJ, Branden KV. ROBPCA: a new approach to robust principal component analysis. *Technometrics* 2005; **47**: 64–79.
11. Pearson K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 1901; **2**: 559–572.
12. Eckart C, Young G. The approximation of one matrix by another of lower rank. *Psychometrika* 1936; **1**: 211–218.
13. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 1987; **2**: 37–52.
14. Rousseeuw PJ, Leroy AM. *Robust Regression and Outlier Detection*, vol. 3. Wiley Online Library: New Jersey, 1987.
15. Huber PJ, Ronchetti E. *Robust Statistics*, vol. 1. Wiley Online Library, 1981.
16. Rousseeuw PJ. Least median of squares regression. *American Statistical Association* 1984; **79**: 871–880.
17. Huber PJ. Projection pursuit. *Annals of Statistics* 1985; **13**: 435–475.
18. Møller SF, von Frese J, Bro R. Robust methods for multivariate data analysis. *Journal of Chemometrics* 2005; **19**: 549–563.
19. Daszykowski M, Kaczmarek K, Vander Heyden Y, Walczak B. Robust statistics in data analysis—a review. *Chemometrics and Intelligent Laboratory Systems* 2007; **85**: 203–219.
20. Engelen S, Frosch S, Jorgensen BM. A fully robust PARAFAC method for analyzing fluorescence data. *Journal of Chemometrics* 2009; **23**: 124–131.
21. Stanimirova I, Daszykowski M, Walczak B. Dealing with missing values and outliers in principal component analysis. *Talanta* 2007; **72**: 172–178.

22. Stanimirova I, Walczak B, Massart DL, Simeonov V. A comparison between two robust PCA algorithms. *Chemometrics and Intelligent Laboratory Systems* 2004; **71**: 83–95.
23. Riu J, Bro R. Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models. *Chemometrics and Intelligent Laboratory Systems* 2003; **65**: 35–49.
24. Engelen S, Hubert M. Detecting outlying samples in a PARAFAC model. Technical report, Katholieke Universietit Leuven, 2007.

APPENDIX: A CORRECTION METHOD ALGORITHM

Let $\mathbf{S}(J \times K)$ be a scatter pixel mask of "ones" for good observations and "zeros" for outliers, identified during S-PCA routine (Figures 4 and 5). The correction of the undiscovered scatter regions might be conducted in the following way for data including two scatter ridges:

(1) For every $s_{jk} = 0$, project a point of coordinates in $[j, k]$ onto a two-dimensional coordinate system.

(2) Fit a robust regression line (in this work based on least trimmed squares) to the projected data according to equation:

$$k_i = \theta_0 + \theta_1 j_i + \varepsilon_i \qquad (6)$$

where $\hat{\theta}_0$ and $\hat{\theta}_1$ are found by minimizing squared errors of h smallest residuals (see [14]).

(3) Store the coordinates $\left[j_i, k_i^*\right]$ of fitted regression line, where k_i^* equals to k_i rounded to the nearest integer, in vectors $\mathbf{j}_1$ and $\mathbf{k}_1$.

(4) Store the coordinates of the outliers identified by the regression model in $\mathbf{j}_0$ and $\mathbf{k}_0$.

(5) Define thickness of a bandwidth $d = \max\left\{\text{number of } k : s_{jk} = 0 \text{ for } j \in \mathbf{j}_1\right\} + band$, where *band* is a correction factor specified by the user.

(6) Fill in the scatter region in the initial matrix $\mathbf{S}$ for $j \in \mathbf{j}_1$ by setting $\mathbf{S}(j, (k - 1/2d) : (k + 1/2d))$ to zero.

(7) Repeat the same procedure for the second scatter line starting from the data stored in $\mathbf{j}_0$ and $\mathbf{k}_0$.

Appendix C

Spherical Principal Component Analysis. A simulation study for data containing clustered outliers.

Authors:

E. Kotwa[1], Stina Frosch[2] and Per B. Brockhoff[1]

Submitted to:

Journal of Chemometrics, November 2013.

[1] DTU Informatics, Richard Pedersens Plads, Building 321, DK-2800 Lyngby, Denmark

[2] DTU Food, Søltofts Plads, Building 227, DK-2800 Lyngby, Denmark

Abstract

Spherical Principal Component Analysis (S-PCA) is known as a robust and very fast version of the classical PCA. These qualities can be desired, especially when high dimensionality of data is considered. Since the method was proposed, it has been used in many applications within multivariate data analysis, and it is known that S-PCA tends to fail in outliers identification under certain circumstances. However, these circumstances were never described in a methodical way.

This paper delivers a simulation study investigating the potential of S-PCA as a fast outlier identification tool when outliers form clusters and the main data originates from an elliptical distribution. The failure reasons of the method and a possible correction named 'nested S-PCA' or 'S-PCA2' have been determined. Finally, a comparison of S-PCA and other robust PCA methods was given. *Keywords:* S-PCA, robust, simulation, outliers

C.1 Introduction

Principal Component Analysis (PCA) [1–3] is probably the mostly spread statistical method for data compression and visualization. It attempts to explain usually vast data structures by means of a preferably small number of orthogonal variables called Principal Components (PCs), being linear combinations of genuine variables. A common problem encountered by practitioners applying PCA comes from the lack of robustness related to the Least Squares foundation of the method. Its breakdown point [4] equals zero, which means that in an extreme case, even one outlier may severely damage the fit and lead to wrong data interpretation.

In order to deal with this issue, various robust alternatives were elaborated, based either on the robust covariance matrix approach (e.g. M-estimators [5], S-estimator [6], MVE or MCD [4]), the Projection Pursuit (first introduced by Li and Chen [7] and described by Huber [8], later modified by Croux and Ruiz-Gazen [9] and Croux et al. [10]), or their combination (ROBPCA developed in [11]). For a review of the common robust methods used in data analysis see [12–18].

Spherical PCA (S-PCA) is a robust and very fast version of classical PCA introduced first by Locantore [19]. The essence of the method is the classical PCA performed on data projected on a unit sphere with its midpoint in the robust center of that data. In this way the harmful influence of outlying samples, which are assumed to have large distances to the center, is bounded. Since the method was proposed, it has been used in many applications within multivariate data analysis, including chemometrics (e.g. [20–25]) and mentioned in a few review papers within the field (e.g. [13, 15, 18]).

Some simulation studies [26, 27] aiming at deepening the understanding of the method have also been performed. First simulations were run by Boente and Fraiman [26], indicating that if data has elliptical distribution, the estimates of eigenvectors are consistent. Moreover, authors claim that S-PCA will be resistant for any contamination model with underlying elliptical distribution. Additionally, some statistical properties (equivariance, influence function and efficiency) were studied by Croux [28]. The covariance matrix estimator is orthogonal equivariant and has bounded Influence Function. The efficiency of the method depends on the parameter gamma γ and it becomes higher for more spherical distributions.

Maronna in [29] used S-PCA as one of the alternatives to his newly proposed method, showing relatively good behavior of S-PCA in terms of efficiency (using Mean Prediction Error as a criterion). Wilcox [27] compares several robust PCA methods among which S-PCA) using a generalized variance criterion. Finally S-PCA, together with other robust methods were incorporated into two R packages: *chemometrics* and *rrcov* described in [18].

A great advantage of the S-PCA is its exceptional computation speed and relative simplicity. This can be desired in many applications, especially when high dimensionality of data (including problems with $p > n$) is considered. It is known, however, that the method tends to fail under certain circumstances. For example Stanimirova [21] mentions that problems might occur when outliers form clusters. Recently, Kotwa et al. [25] used S-PCA to identify disturbing scatter effects in EEM fluorescence landscapes. Their findings show that an additional correction step is needed, as S-PCA itself fails in discovering outlying elements (scatter) in certain situations. In both cases, it remains unclear which situations they are.

In this paper we investigated the potential of S-PCA as a fast outlier identification tool when outliers form clusters and the main data originates from an elliptical distribution (a 'standard' case). We determined failure reasons of the method and proposed a possible correction called 'nested S-PCA' or 'S-PCA2'. Finally, a clarification of advantages and disadvantages of S-PCA compared to some other robust PCA methods were given.

C.2 Case study

In order to show the performance of different robust PCA methods and introduce issues which will be discussed further in this paper, a simulation case-study example will be presented. Six PCA models were applied to two contaminated data sets: Classical PCA (PCA), ROBPCA [11], two versions of Projection Pursuit based PCA: by Croux and Ruiz-Gazen [9] (PP) and by Croux et al. [10] (Grid) and finally two S-PCA algorithms: traditional (SPCA) and its nested version (SPCA2). Both S-PCA methods will be described in details in Section C.3.

Two data sets were constructed from two multivariate normal distributions $\boldsymbol{X}_1 \sim N(\mathbf{0}, \boldsymbol{\Sigma_1})$ and $\boldsymbol{X}_2 \sim N(\mathbf{0}, \boldsymbol{\Sigma_2})$, with number of samples $n_1 = n_2 = 500$ and number of variables $k_1 = 5$ and $k_2 = 100$. Covariance matrices $\boldsymbol{\Sigma_1}$ and $\boldsymbol{\Sigma_2}$ were calculated in the following way: $\boldsymbol{\Sigma_1} = \boldsymbol{Y_1}^T \boldsymbol{Y_1}$ and $\boldsymbol{\Sigma_2} = \boldsymbol{Y_2}^T \boldsymbol{Y_2}$, where $\boldsymbol{Y_1}$ and $\boldsymbol{Y_2}$ were drawn from multivariate standard normal distributions of corresponding dimensionality. Subsequently, both data sets were contaminated by adding 20% degenerated samples from a generic contamination distribution $\boldsymbol{X_i}' \sim N(\boldsymbol{\mu_i}', \boldsymbol{\Sigma_i}')$, where $\boldsymbol{\mu_i}' = s_i(1, ..., k_i)$ and $\boldsymbol{\Sigma_i}' = 0.5\boldsymbol{\Sigma_i}$, for $i \in \{1, 2\}$. Here, the outlier covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{i}}'$ correspond to scaled version of data covariances. This is to ensure that outliers form a cluster - a cloud of points smaller than the main data bulk. A scalar s_i was set to $s_1 = 1$ and $s_2 = 4$, which corresponds to 'small' and 'large' distances of the outlier cluster towards the main data distribution.

Figures C.1 and C.2 depict the outlier identification process for each of applied methods by a *Diagnostic Plot* of score (SD) and orthogonal (OD) distances and a *Flag Plot*. In the Diagnostic Plot the introduced outliers are marked by red and a horizontal and vertical line determines the cut-off values for each of the two distances. Data points having distance values exceeding particular cut-off value are identified as outliers (the cut-off values are described in details further in Section C.3.2). The Flag Plot shows 'flags' ($\in \{0, 1\}$) which were assigned to each observation during the outlier identification procedure. Again, 'real' outliers are marked with red color and identified outliers have $flag = 0$. Observations with $flag = 1$ are considered regular. Blue points denote data points wrongly qualified as outliers.

In brief, it is shown that Classical PCA, as expected, fails completely in identifying outliers for both data scenarios, as all 'red points' lay within the area demarked by the cut-off value lines and the correposonding flags equal 1. The best method (able to determine all outlyig samples) when $k = 5$ seems to be ROBPCA and for $k = 100$ the Grid-PCA, however, in both cases S-PCA2 performs nearly as good (though, it has somewhat higher amount of non-outliers concidered as contamination). Interestingly enough, ROBPCA breaks when dimensionality grows (in this set-up when $k > 36$, results not shown), which is not surprising as this method is designed for low dimensional data. The situation is quite opposite in the case of Grid-PCA, which wrongly qualifies outliers when

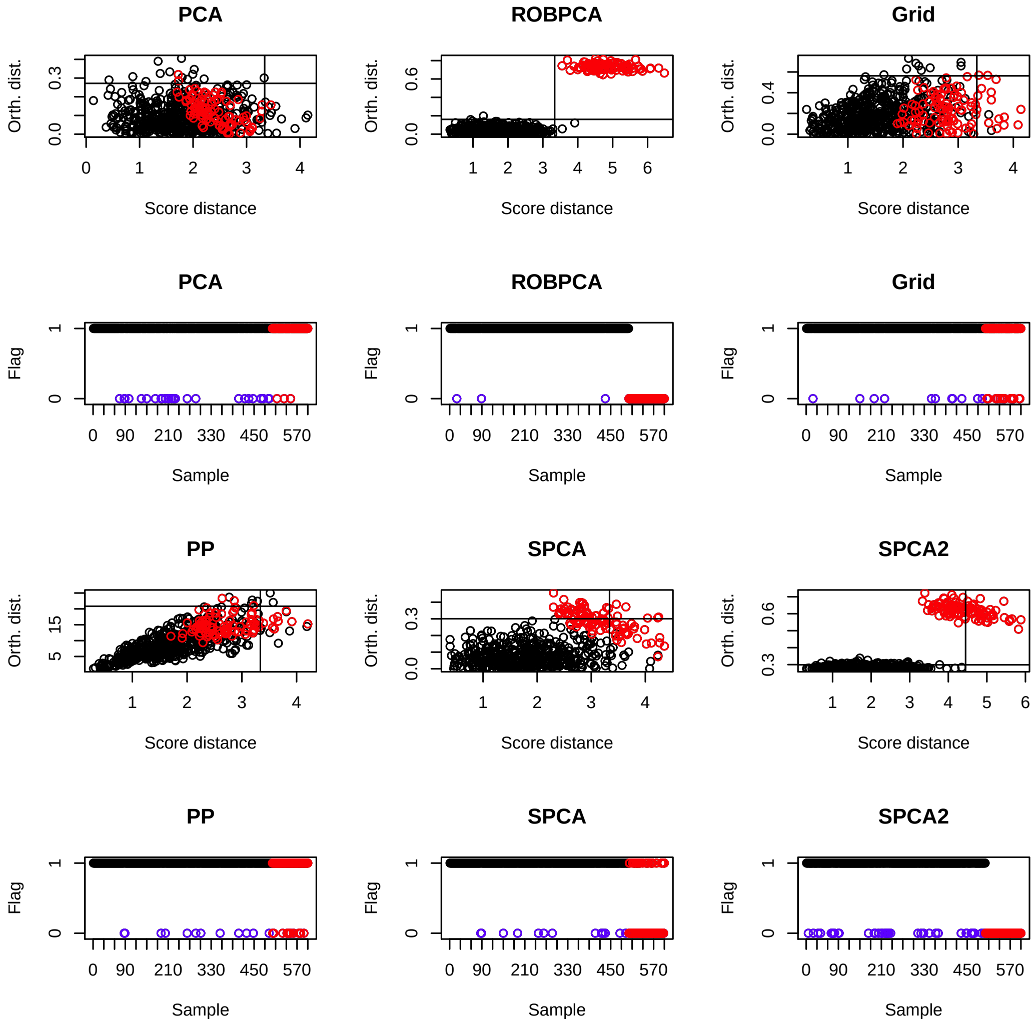

Figure C.1: Diagnostic and Flag plots for $\boldsymbol{X}_1$ and $\boldsymbol{X}_1^{'}$ $(k = 5)$

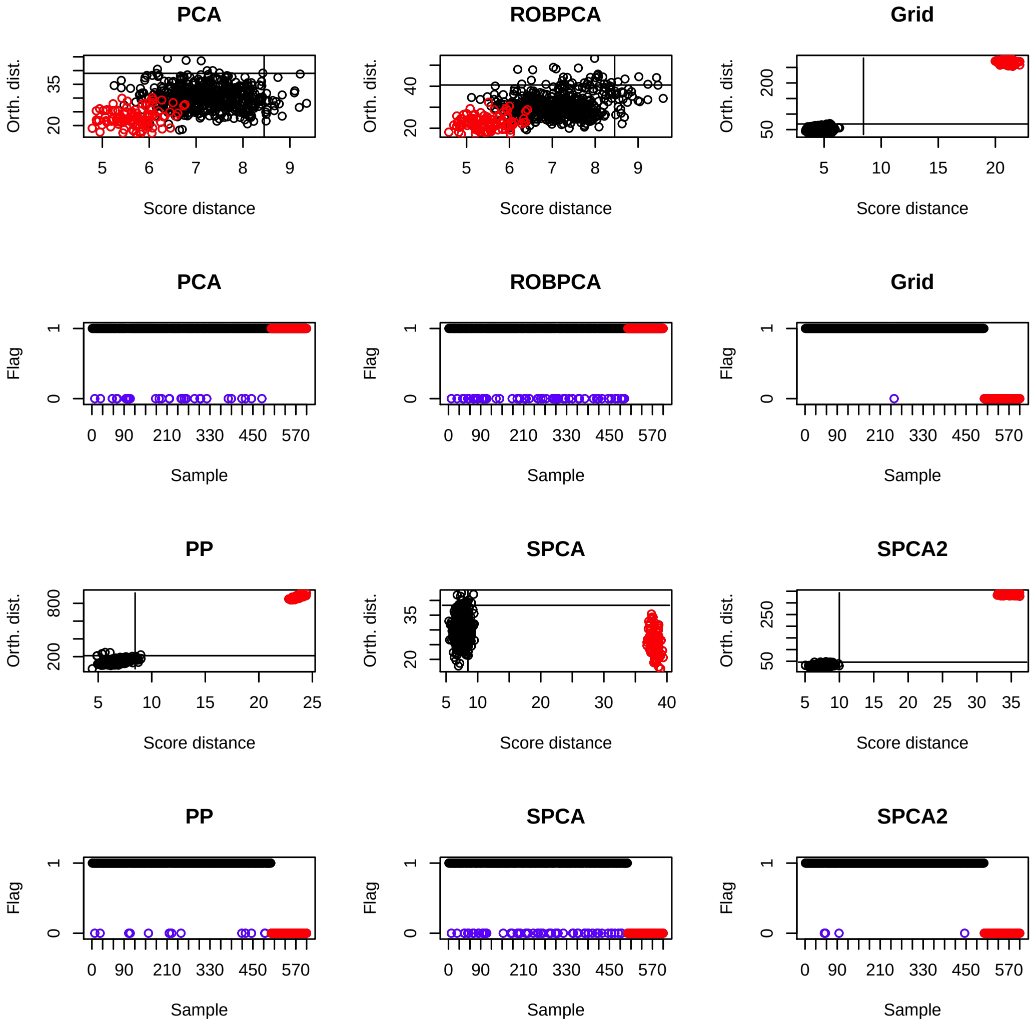

Figure C.2: Diagnostic and Flag plots for $\boldsymbol{X}_2$ and $\boldsymbol{X}_2^{'}$ $(k = 100)$

Table C.1: Running system Time [s] for 6 PCA methods performed on data with number of dimensions k=5 and k=100 and the constant sample size $n = 500$.

	PCA	SPCA	SPCA2	ROBPCA	Grid	PP
k=5	0.01	0.08	0.34	0.09	0.09	0.06
k=100	0.07	0.46	0.58	16.02	20.10	3.29

number of dimensions is small (this might also be related to the small distance of outlier bulk to the rest of the data). The ordinary S-PCA seems to suffer from similar factors and fails to identify all the corrupted observations when data and contamination lie close to each other. Table C.1 reveals the system computation time for each applied method, which is comparable when $k = 5$ but grows significantly together with dimensionality for ROBPCA and Grid-PCA. Moreover, Figure C.3 shows the dynamics of computational time as a function of number of dimensions given constant sample size: linear for S-PCA and exponential for ROB- and Grid-PCA algorithms.

This simple example shows that PCA based on spherical projection has a great potential when a robust and quick data compression method, especially for vast data structures, is needed. For practitioners being able and willing to use it, more information about performance of S-PCA is needed, which will be the topic and objective of the rest of this paper.

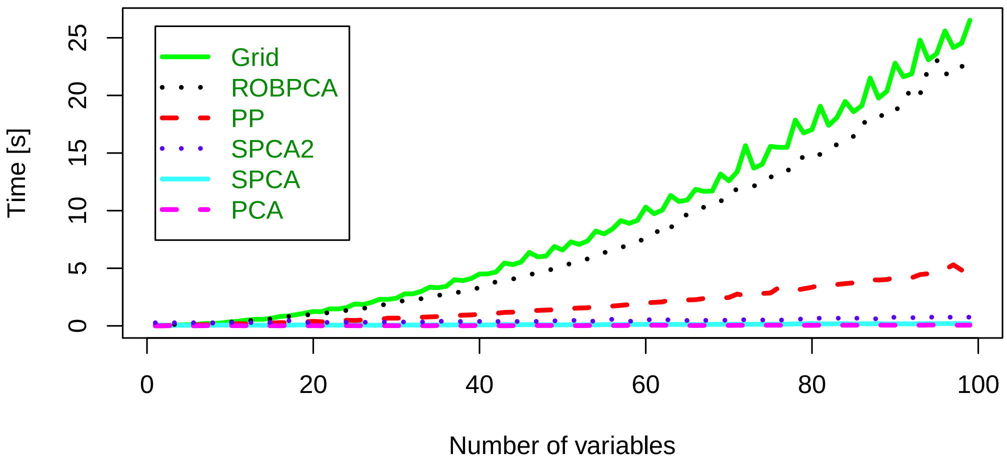

Figure C.3: Execution time [s] of different PCA methods as a function of number of dimensions for constant number of observations $n = 500$.

C.3 S-PCA. How does it work?

C.3.1 $L1$-median and spherical projection

S-PCA, being a very fast and robust version of classical PCA, belongs to the so-called ’projection methods’. It combines the robustness feature of the median with the projection pursuit framework and is conceptually fairly simple. In the first step, the robust center of the data is determined as the L1-median [5]. L1-median is defined as a point in the multidimensional space found by minimizing the sum of Euclidean distances from the data objects to that point. Subsequently, all data points $\mathbf{x}_i$ are centered and down-weighted by the inverse of these distances, which is equivalent to projecting each data point onto a unit radius hyper-sphere with the center in $L1$:

$$\mathbf{x}_i^p = \frac{\boldsymbol{x}_i - \boldsymbol{\mu}_{L1}(\mathbf{X})}{\|\mathbf{x}_i - \boldsymbol{\mu}_{L1}(\mathbf{X})\|} + \boldsymbol{\mu}_{L1}(\mathbf{X}) \tag{C.1}$$

where $\mathbf{x}_i^p$ is the ith projected data object and $\|\mathbf{x}_i - \boldsymbol{\mu}_{L1}(\mathbf{X})\|$ is the Euclidean distance to the robust center of the data $\boldsymbol{\mu}_{L1}$. In this setting, the influence of outlying samples manifesting normally large distances in the denominator of the equation (C.1), will be bounded. In the next step, a classical PCA is carried out on the down-weighted data and the robust scores are obtained by projecting the original data on the resulting robust loadings. Finally, the outlier detection is conducted, based on robust center of the data, eigenvalues of the robust covariance matrix, robust scores and loadings.

In short, we propose describing the functioning scheme of S-PCA as a four step algorithm, depicted graphically in Figure C.4, respectively:

1. find the $L1$ robust center of the data (red dot)
2. project all data points $\mathbf{x}_i$ on the unit radius sphere centered in $L1$
3. fit classical PCA to the projected data, find robust loadings and scores
4. apply the robust loadings to original data and identify outliers

C.3.2 Outlier identification

A popular method for determining outlying elements consists of Score and Orthogonal distances (SD and OD, respectively). The score distance (SD_i), also

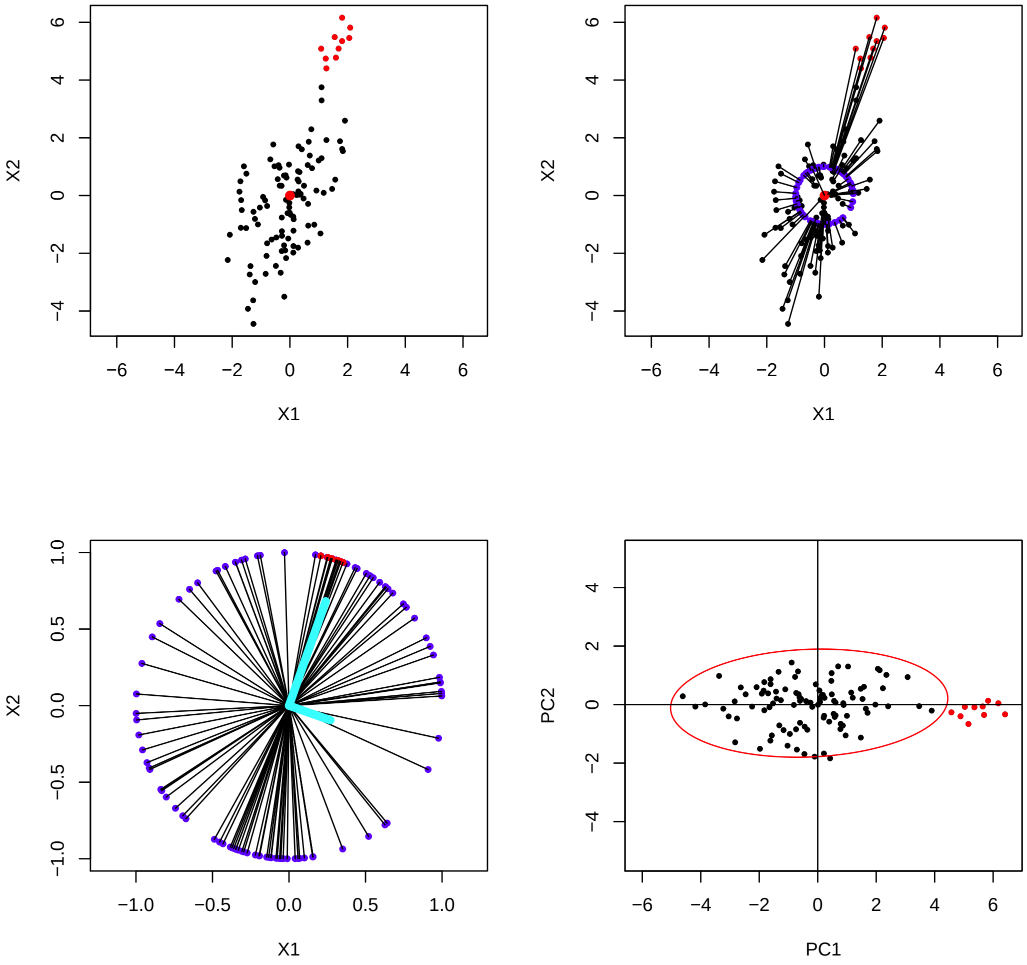

Figure C.4: Functioning of S-PCA: a) finding a robust center of the data (here L1 median); b) projecting all data points on a sphere with unit radius; c) performing classical PCA on the projected data points - finding eigenvectors; d) applying eigenvectors to the initial data set, flagging outliers (here, by means of tolerance ellipse)

known as robust distance, expressed as:

$$SD_i = \sqrt{\sum_{j=1}^{k} \frac{t_{ij}^2}{l_j}} \tag{C.2}$$

where $\mathbf{t}_j$ is a vector of robust scores for the jth component and $\mathbf{l}$ contains robust eigenvalues, is actually a Mahalanobis distance to the robust data center. On the other hand, the Orthogonal Distance informs how far, with respect to the space spanned by the robust model, the ith object is located.

$$OD_i = \|\boldsymbol{x}_i - \boldsymbol{\mu}_{L1} - \boldsymbol{P}\boldsymbol{t}_i\| \tag{C.3}$$

Here, $\|\cdot\|$ denotes the Euclidean norm and the model $\mathbf{TP}^T$ contains k selected components. A data point exceeding a certain cut-off value of at least one of distances is automatically flagged as an outlier.

Two rules were taken into consideration in order to define cut-off values: $\sqrt{\chi^2_{n-1,0.975}}$, where n is a number of observations (for SD) or selected PCs (for OD) [11], or a robust version of z-scores (suggested in [21]). These scores are constructed by incorporating robust location and scale estimates such as median (μ) and median absolute deviation (σ) as an alternative to the standard procedures:

$$z = \frac{|d - \mu(d)|}{\sigma(d)} \tag{C.4}$$

Then, a default cut-off value equal to three can be used for identification of observations with large distances. Both approaches give similar results, however it seems that the z-sores cut-off value tends to be slightly higher.

C.3.3 Weak points

Previously, in Section C.2, we saw that even though S-PCA (and especially nested S-PCA) provided decent robust results, it failed in identifying outliers in certain situations, and only a part of contamination was discovered (see Figure C.1). Judging from the functioning scheme presented in Figure C.4, the performance of S-PCA might be hampered by clustered outliers if at least one of the following situation occurs:

1. a bulk of outliers affects significantly the location of $L1$ median
2. a number of outliers is high enough to attract the classical PCA fit on projected data and, as a consequence, corrupt the resulting loadings
3. an outlier cloud is too close to the main data so that the two distances (SD and OD) do not succeed in identifying it

These three factors will be in the focus point of the S-PCA sensitivity study conducted in the following sections.

C.3.4 Nested S-PCA

In the *rrcov* R package introduced in [18], the S-PCA method is imprenented as described above: by first performing Classical PCA on down-weighted data, and afterwards identifying outlying samples. This means that contaminated observations are neutralized distance-wise, however not direction-wise. They still are able to attract the Classical PCA fit and corrupt the out-coming loadings, even if the outliers are identified correctly. In order to avoid that inconvenience, we propose a simple correction, which will further be referred to as 'S-PCA2' or 'nested S-PCA'. The correction assumes that after identification of outliers by the S-PCA method, another S-PCA routine is executed, this time on the cleaned, outlier-free data set. Then, loadings are used for constructing robust scores by projecting them on the initial data and new distances are determined. The potential advantage of this method is twofold: the actual PCA results will be more reliable as they will be less (or at all) affected by outlying samples direction-wise, and secondly, if simple S-PCA was able to identify only some part of outliers, S-PCA2 might be able to correct for that (as in Figure C.1).

C.4 Simulations

C.4.1 A bi-variate design.

In order to clearly present functioning of S-PCA, a simple bi-variate design is considered: a number of observations $n = 100$ is drawn from a bi-variate normal distribution with $\boldsymbol{\mu} = (0, 0)$ and

$$\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 3 \end{pmatrix}$$

following a contamination model (as in [30]):

$$(1-\epsilon)N_p(\mathbf{0}, \Sigma) + \epsilon N_p(\widetilde{\boldsymbol{\mu}}, \widetilde{\Sigma}) \tag{C.5}$$

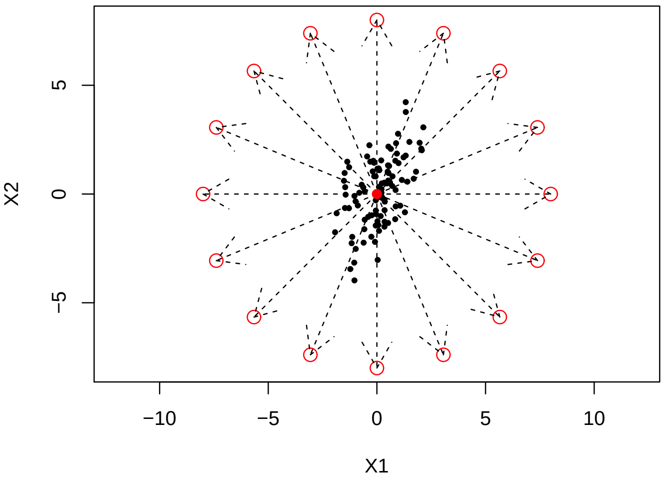

Figure C.5: Directions for the placement of outliers bulk center.

where ϵ stands for contamination ratio. In this experimental design, it has been decided to introduce a cluster of outliers by varying the following parameters: d - direction from the center of main data cloud to the center of outlier cloud; r - distance from main data to outlying observations; ϵ - ratio of outliers. In total 16 directions $d \in \{0, 1/8\pi, \cdots, 15/8\pi\}$ (presented in Figure C.5), three distances $r \in \{4, 6, 8\}$ and three contamination ratios $\epsilon \in \{0.1, 0.2, 0.3\}$ have been taken into account. The distances were chosen so that they correspond to a small, medium and large distance of outlier bulk to the main data. The selection of outlier ratios can be motivated by the fact that these scenarios will be of highest interest for practitioners. From the experience a common practice would most likely be rejecting a data set where contamination constitutes more than 1/3 of the data. As performance measures in the simulation study, following criteria are considered (similar to [30]):

1. γ - an angle between the first eigenvector of the investigated method and first eigenvector of classical PCA in the case of uncontaminated data

2. NIO (not identified outliers) - outlying points which have not been recognized by the classical or robust PCA method

3. WIP (wrongly identified point) - data points which have been wrongly flagged as outliers

Optimal values of NIO, WIP and γ are 0. Time of execution is disregarded in this study due to a small data size. Other criteria exist and have been used in the literature, i.e. relative generalized variance estimate [27] or relative prediction error and relative estimation error [29]. However, for this study the three selected criteria will sufficiently describe differences between the applied models.

C.4.2 Sensitivity study.

The S-PCA algorithm was applied to 144 (3x3x16) contaminated data sets. Some selected results obtained from the analysis are presented in Figures C.6, C.7, C.8 and C.9 and Table C.2. The results lead to the following conclusions: in the case of data following elliptical distribution, assuming constant distance to the outlier cloud, the direction d indicating location of the outlier cloud has an influence on S-PCA results. The most sensitive scenario, likely to result in un-identified outliers, is the case of good leverage points, situated closest to the main data bulk. Secondly, the distance r also impacts method's performance as the S-PCA fails to identify outlying samples, when their cloud is situated

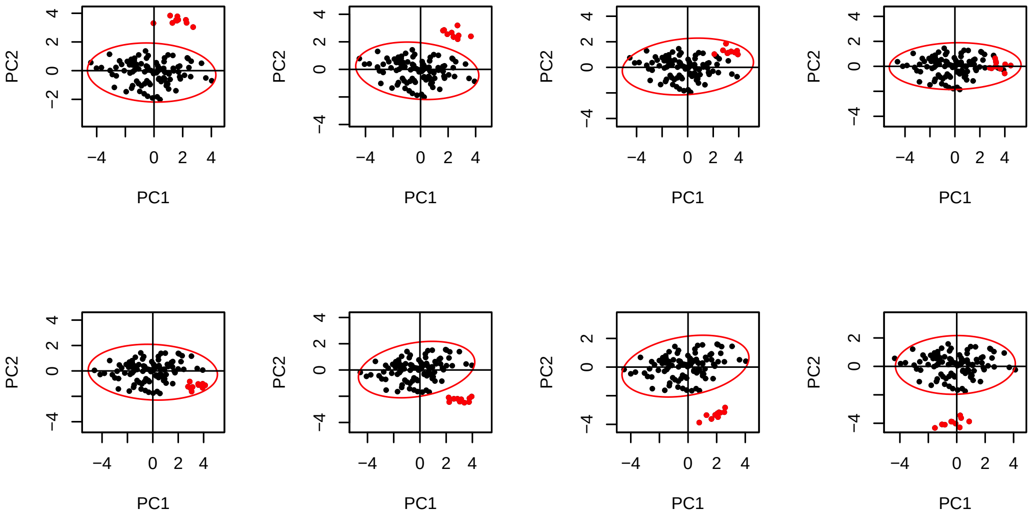

Figure C.6: S-PCA results for first 8 directions, distance=4, outlier ratio=0.1

'too close' to the main data distribution (compare Figure C.6c, d and e with the corresponding panels of Figure C.7, where the distance has been increased). Finally, the amount of outliers can have a major influence on the S-PCA fit. Figure C.8 and Figure C.7 present the same data scenario, where only the ratio of outliers has been increased to $\epsilon = 0.3$. Also, the resultis in Table C.2 indicate clearly that the angle γ is much larger when amount of outliers grows, reflecting corrupted results.

C.4.3 Failure reasons

Above, three parameters which influence the performance of S-PCA have been identified: amount of outliers contaminating the data, their position and distance towards the main data bulk. These factors will affect different parts of the

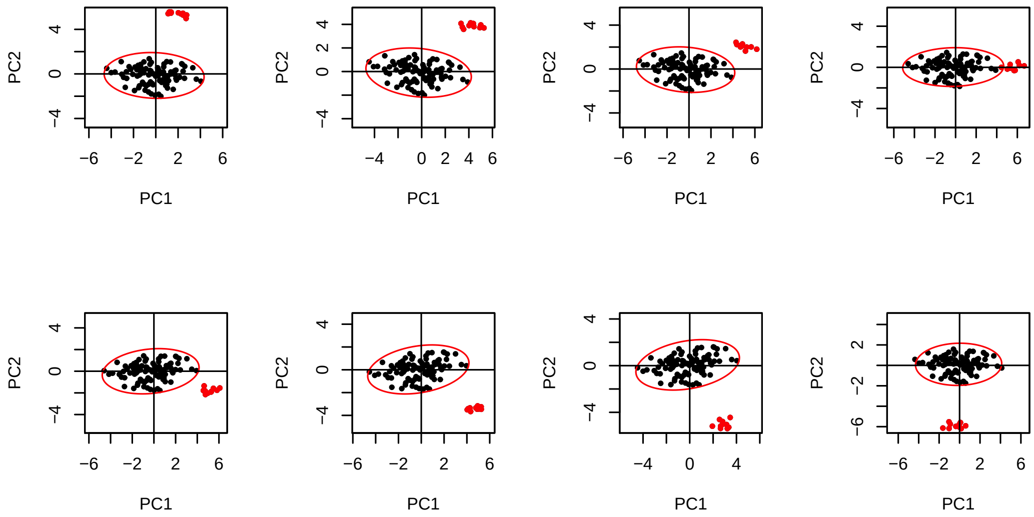

Figure C.7: S-PCA results for first 8 directions, distance=6, outlier ratio=0.1

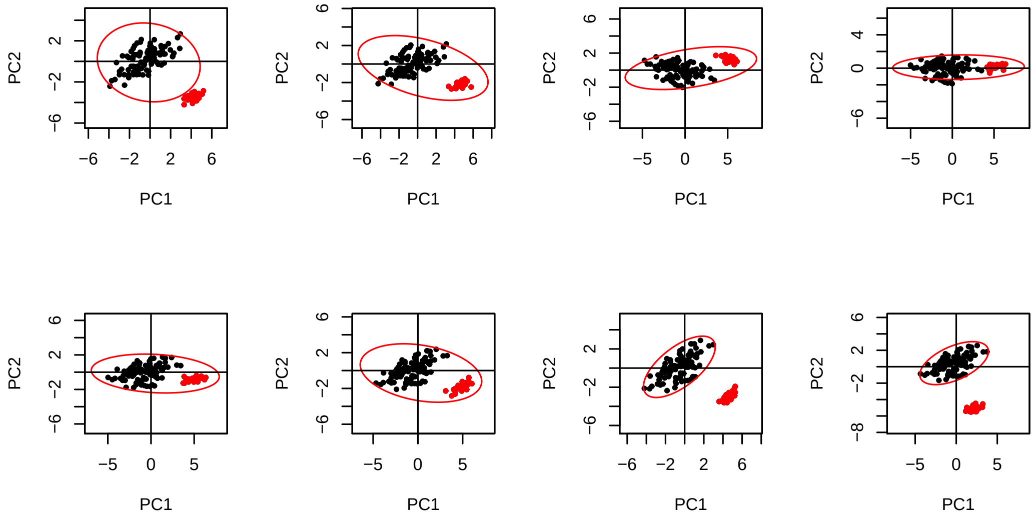

Figure C.8: S-PCA results for first 8 directions, distance=6, outlier ratio=0.3

Table C.2: Simulation results for S-PCA.

	1/8Pi	1/4Pi	3/8Pi	1/2Pi	5/8Pi	3/4Pi	7/8Pi	Pi
d=4, e=0.1								
Angle	0.0489	0.0946	0.0736	0.0223	0.1025	0.1637	0.1684	0.0061
NIO	0.0000	0.0000	10.0000	10.0000	9.0000	0.0000	0.0000	0.0000
WIP	0.0000	0.0000	0.0000	2.0000	0.0000	0.0000	0.0000	2.0000
d=6, e=0.1								
Angle	0.0369	0.1019	0.0795	0.0268	0.1006	0.1639	0.1715	0.0019
NIO	0.0000	0.0000	0.0000	2.0000	0.0000	0.0000	0.0000	0.0000
WIP	0.0000	0.0000	0.0000	3.0000	0.0000	0.0000	0.0000	2.0000
d=6, e=0.3								
Angle	0.5390	0.4286	0.1771	0.0220	0.2616	0.2916	0.0952	0.3414
NIO	9.0000	29.0000	30.0000	30.0000	28.0000	23.0000	0.0000	0.0000
WIP	0.0000	0.0000	0.0000	9.0000	0.0000	0.0000	0.0000	0.0000
d=8, e=0.3								
Angle	0.5442	0.4187	0.1749	0.0176	0.2524	0.3053	0.0916	0.3421
NIO	0.0000	0.0000	0.0000	4.0000	0.0000	0.0000	0.0000	0.0000
WIP	0.0000	0.0000	0.0000	9.0000	0.0000	0.0000	0.0000	0.0000

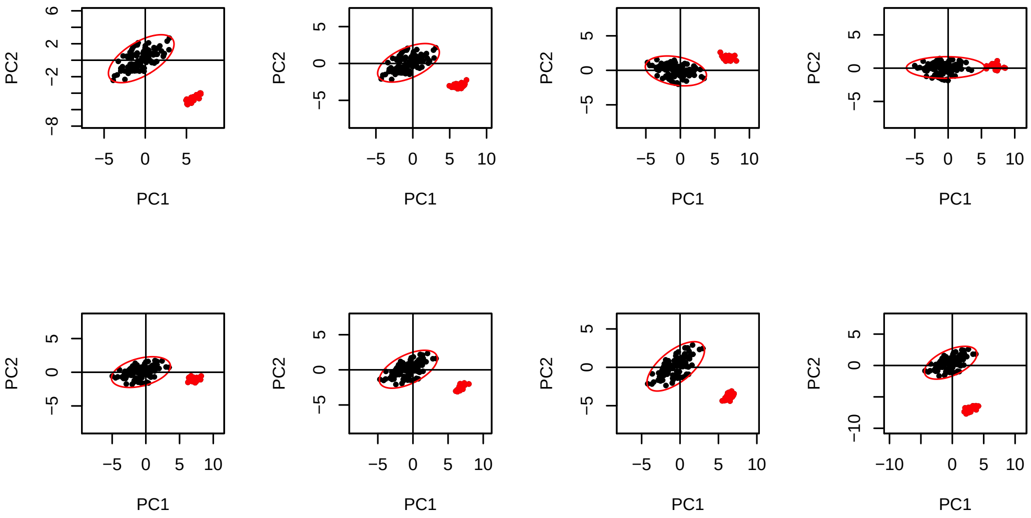

Figure C.9: S-PCA results for first 8 directions, distance=8, outlier ratio=0.3

algorithm, when 'difficult' circumstances occur. We are in particular interested in the behavior of the $L1$ center, robust loadings, as well as score distance.

As a case study we use the 'difficult data' from section C.4.2, corresponding to $r = 6$, $\epsilon = 0.3$ and $d \in \{0, 1/8\pi, 1/4\pi, 3/8\pi\}$, where for three out of four cases outliers remained unidentified and the fit was corrupted. Figure C.10 shows the four outlier scenarios considered, and Figure C.11 depicts a shift of the $L1$ center (blue dot) in relation to the 'real' center of not contaminated data (red dot). One can observe that the $L1$ median is attracted by clustered outliers, depending on their location. Moreover, the blue line stands for the first PC, resulting from the S-PCA fit. Here, it becomes apparent that location of high ratio of clustered outliers has a major influence on the resulting PCA fit. This could surely be adjusted, if the method would succeed in identifying these erroneous points correctly, by disregarding the outliers and repeating the

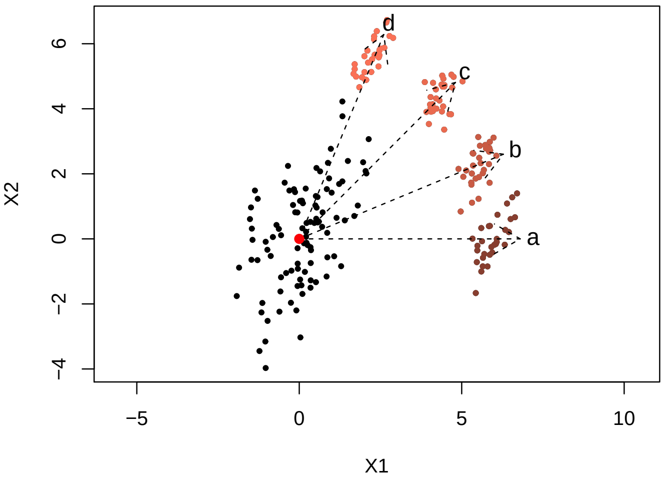

Figure C.10: Directions for the placement of outliers center for the case study: a) 0, b) $1/8\pi$, c) $1/4\pi$, d) $3/8\pi$

PCA routine on cleaned data (nested S-PCA). In order to determine outlier identification ability of the method, we will look at the Score (Robust) distance. Since Orthogonal Distance informs about how far a particular data point lies from a subspace spanned by the fitted model, it will not be reliable here as the considered fit is burdened with an error and therefore corrupted. In this case, it is the SD which can be used for detecting the outlying observations. Figure C.12 presents outlier identification by the means of SD, for data containing different ratios of outliers. Here, it is obvious that the amount of outliers has a direct influence on the determined robust distances to the $L1$ center. One should keep in mind that the center itself will change its location depending on the amount of clustered outliers causing the distance change. Additionally, for construction of SD distances, we scale the PCA scores using (corrupted) eigenvalues obtained from the previous PCA fit. Both reason lead to situations where the clustered outliers remind unidentified.

To sum up the above breakdown study of S-PCA, we can conclude the following:

1. Location, distance and amount of contaminated samples may influence outlier identification ability of S-PCA: if the outlier cloud is too close to the main data, some or all outlying samples might stay unidentified (especially good leverage points), if the outlier ratio is high and its location not aligned with the first eigenvector, the S-PCA fit (reflected in γ) will deteriorate.

2. Even if S-PCA is able to correctly identify the outliers, the resulting loadings and scores are often attracted corrupted by the contamination, therefore a relevant correction, based on outliers removal and re-fitting the model on clean data is necessary (for example nested S-PCA).

C.4.4 Final case study - nestes S-PCA

In previous sections we discussed possible breakdown reasons of the S-PCA and introduced a correction which could partialy improve the inefficiency of the method. Here, the final comparison of nested S-PCA together with two alternative approaches, ROBPCA and Grid PCA (which showed most reliable results in Section C.2), are examined, taking into consideration outlier identification ability. Again, the 4 data scenarios ($r = 6$, $\epsilon = 0.3$, $d \in \{0, 1/8\pi, 1/4\pi, 3/8\pi\}$), presented in Figure C.10 will be considered. The results in Figure C.13 show how well each of the three robust techniques is able to identify the simulated contamination (shown in red), by plotting resulting flags associated to each ovservation ($flag = 0$ means that the data point was qualified as outlying).

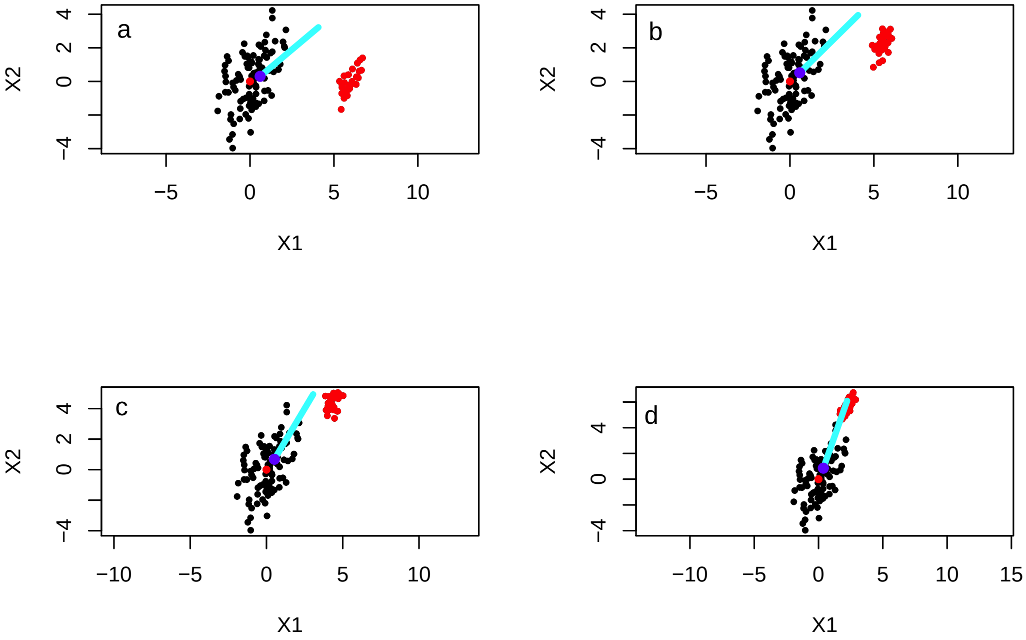

Figure C.11: L1 center and loadings study of S-PCA for the data containing outliers situated in defferent locations towards the main data bulk: a) $d = 0$; b) $d = 1/8\pi$; c) $d = 2/8\pi$; d) $d = 3/8\pi$

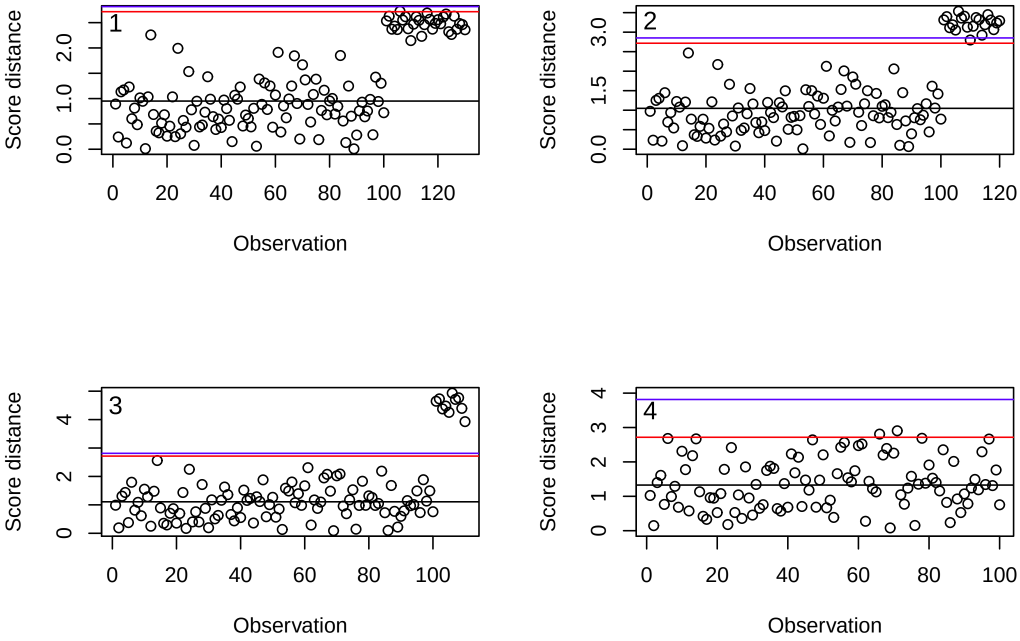

Figure C.12: Robust (score) distance of data points to the L1-median for data containing outlier ratio equal to a) $\epsilon = 0.3$; b) $\epsilon = 0.2$; c) $\epsilon = 0.1$; d) $\epsilon = 0$; black line is the median of SDs, blue and red lines are the cut-off values according to the robust z-scores and Chi-squared criteria, respectively.

All three methods return satisfactory results for $d \in \{0, 1/8\pi\}$, corresponding to directions where the outlier cloud is further from the main data. One can see that S-PCA2 seems to be most sensitive, as it starts having 'problems' already for $d = 1/8\pi$, and ROBPCA most resistant, performing well even for $d = 1/4\pi$. All methods fail however in the case of good leverage outliers ($d = 3/8\pi$, see Fig. C.10).

C.5 Discussion and conclusion

In this work we investigated potential applications of Spherical PCA as an outlier identification tool, focusing on multivariate elliptical data and outliers forming clusters. The functioning scheme of the method and its weak points have been identified, leading to better understanding of circumstances where S-PCA could be applied, and when it would fail. Simulations performed in this study show that S-PCA is able to identify contaminated observations in data, however its functionality is subject to specific data circumstances. Location, distance and amount of contaminated samples were found to have a large impact. If outliers are too close to the main data cloud, some or all contaminated samples might stay unidentified (especially good leverage points). Moreover, if the outlier ratio is high and its location not aligned with the first eigenvector, the resulting loadings can be corrupted, and hence, the S-PCA fit will deteriorate. A relevant correction, called S-PCA2 (or nested S-PCA), has been proposed in order to account for deteriorated fit of S-PCA and to increase overall applicability of the method.

S-PCA and S-PCA2 were also compared to other robust versions of PCA in

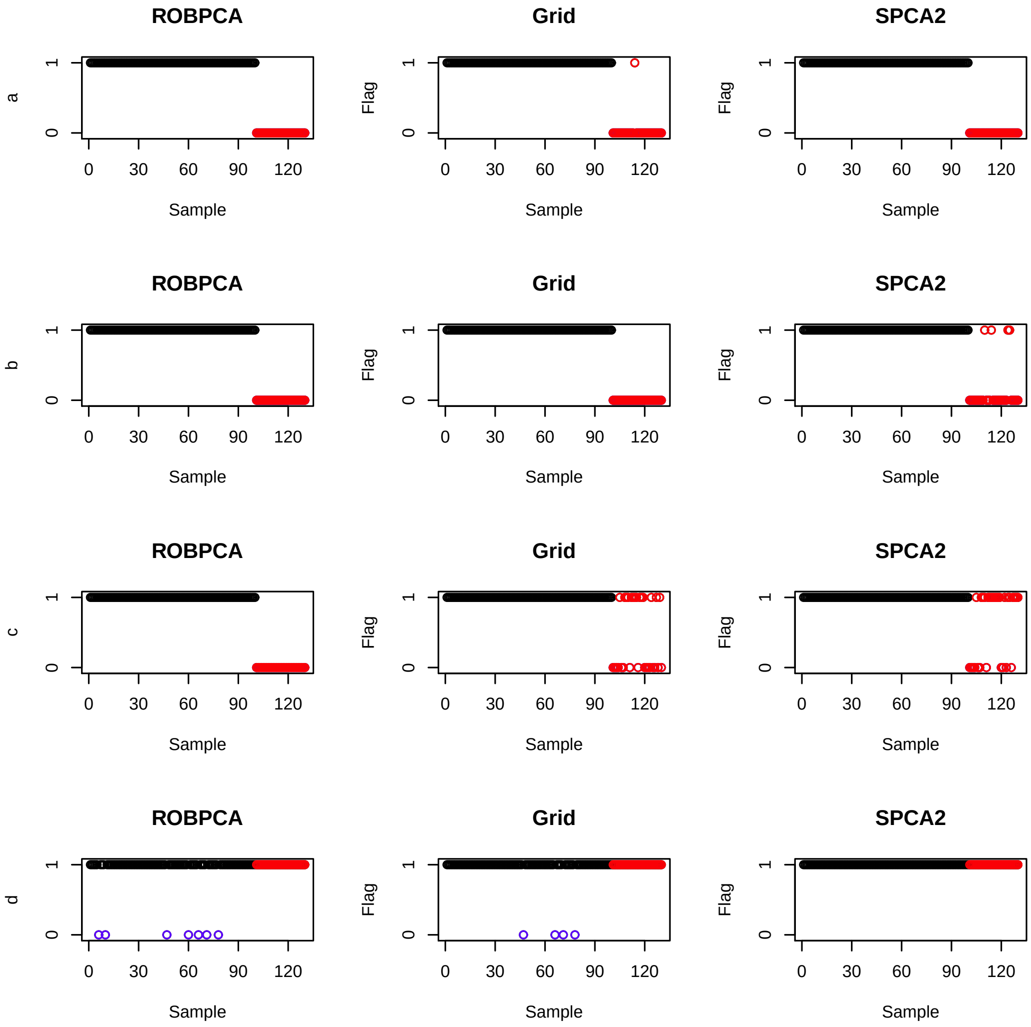

Figure C.13: Flag plots resulting from S-PCA2, ROBPCA and Grid methods for 4 data scenarios: $d \in \{0, 1/8\pi, 1/4\pi, 3/4\pi\}$ with $r = 6$ and $\epsilon = 0.3$

terms of outlier detection ability, reliability of the fit and time. The study showed that none of the methods was relevant for all data circumstances, having each its limitations and functionality field. For example, ROBPCA being a very reliable method in low dimensions can not be used for vast data structures. Grid PCA seems to be applicable in opposite situations as shown in Figures C.1 and C.2, however, when 'big' data are considered, it becomes highly computation-expensive, with an exponential execution time growth. Since nowadays data structures with number of variables > 1000 are no longer uncommon and robust PCA is often used as part of larger algorithms as an outlier detecting tool, the high performance time can constitute a serious limitation.

We conclude that S-PCA (and especially its nested version as proposed here) can be a valuable asset in the toolbox of every practitioner, especially when short execution time is an important factor. We highly recommend that S-PCA is applied with consciousness of possible shortcomings, such as described in this work.

Bibliography

[1] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

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

[3] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1):37–52, 1987.

[4] P.J. Rousseeuw, A.M. Leroy, and J. Wiley. *Robust regression and outlier detection*, volume 3. Wiley Online Library, 1987.

[5] P.J. Huber, E. Ronchetti, and MyiLibrary. *Robust statistics*, volume 1. Wiley Online Library, 1981.

[6] P.J. Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, pages 871–880, 1984.

[7] G. Li and Z. Chen. Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and monte carlo. *Journal of the American Statistical Association*, pages 759–766, 1985.

[8] P.J. Huber. Projection pursuit. *The annals of Statistics*, pages 435–475, 1985.

[9] C. Croux and A. Ruiz-Gazen. High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1):206–226, 2005.

[10] C. Croux, P. Filzmoser, and M.R. Oliveira. Algorithms for projection–pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87(2):218–225, 2007.

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

[12] L. Yi-Zeng and O. M. Kvalheim. Robust methods for multivariate analysis – a tutorial review. *Chemometrics and Intelligent Laboratory Systems*, 32(1):1–10, 1996.

[13] S.F. Møller, J. von Frese, and R. Bro. Robust methods for multivariate data analysis. *Journal of Chemometrics*, 19(10):549–563, 2005.

[14] P. J. Rousseeuw, M. Debruyne, S. Engelen, and M. Hubert. Robustness and outlier detection in chemometrics. *Critical Reviews in Analytical Chemistry*, 36(3-4):221–242, 2006.

[15] M. Daszykowski, K. Kaczmarek, Y. Vander Heyden, and B. Walczak. Robust statistics in data analysis - a review. *Chemometrics and Intelligent Laboratory Systems*, 85(2):203–219, 2007.

[16] M. Hubert, P. J. Rousseeuw, and S. Van Aelst. High-breakdown robust multivariate methods. *Statistical Science*, 23(1):92, 2008.

[17] S. Serneels and T. Verdonck. Principal component regression for data containing outliers and missing elements. *Computational Statistics and Data Analysis*, 53(11):3855–3863, 2009.

[18] P. Filzmoser and V. Todorov. Review of robust multivariate statistical methods in high dimension. *Analytica chimica acta*, 705(1):2–14, 2011.

[19] N. Locantore, JS Marron, DG Simpson, N. Tripoli, JT Zhang, KL Cohen, G. Boente, R. Fraiman, B. Brumback, C. Croux, et al. Robust principal component analysis for functional data. *Test*, 8(1):1–73, 1999.

[20] M. Daszykowski, K. Kaczmarek, I. Stanimirova, Y. Vander Heyden, and B. Walczak. Robust simca-bounding influence of outliers. *Chemometrics and Intelligent Laboratory Systems*, 87(1):95–103, 2007.

[21] I. Stanimirova, M. Daszykowski, and B. Walczak. Dealing with missing values and outliers in principal component analysis. *Talanta*, 72(1):172–178, 2007.

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

[23] M. Daszykowski, M. Wróbel, A. Bierczynska-Krzysik, J. Silberring, G. Lubec, and B. Walczak. Automatic preprocessing of electrophoretic images. *Chemometrics and Intelligent Laboratory Systems*, 97(2):132–140, 2009.

[24] V. J. Fortunato. A comparison of the construct validity of three measures of negative affectivity. *Educational and psychological measurement*, 64(2):271–289, 2004.

[25] E. Kotwa, B. Jørgensen, P. B. Brockhoff, and S. Frosch. Automatic scatter detection in fluorescence landscapes by means of spherical principal component analysis. *Journal of Chemometrics*, 2013.

[26] G. Boente and R. Fraiman. Discussion of robust principal components for functional data by locantore et al. *Test*, 8:1–28, 1999.

[27] R. Wilcox. Robust principal components: A generalized variance perspective. *Behavior research methods*, 40(1):102–108, 2008.

[28] C. Croux. Discussion of robust principal component analysis for functional data. *Test*, 8(1):1–73, 1999.

[29] R. Maronna. Principal components and orthogonal regression based on robust scales. *Technometrics*, 47(3):264–273, 2005.

[30] M. Hubert, P. Rousseeuw, and T. Verdonck. Robust pca for skewed data and its outlier map. *Computational statistics & data analysis*, 53(6):2264–2274, 2009.