Thesis Presented to the
Technical University of Denmark

By
Stina Frosch Møller

In Partial Fulfilment of the
Requirements for the Degree

Doctor of Philosophy

Danish Institute for Fisheries Research
Department of Seafood Research

Copenhagen, Denmark – December 2005

Preface

The following pages form and constitute my thesis, submitted as a requirement for obtaining the Ph.D. degree at the Technical University of Denmark. Furthermore, it provides documentation for the work challenging the title: "*The Importance of Data Quality and Traceability in Data Mining. Applications of Robust Methods for Multivariate Data Analysis"*. The work is carried out at the Danish Institute for Fisheries Research (DIFRES), Department of Seafood Research and the Royal Veterinary and Agricultural University as a part of the projects "Quality Control and Documentations Systems in the Herring Industry. Improved Data Collection and Multivariate Data Analysis". The work was granted by the Danish Ministry of Food, Agriculture and Fisheries.

Preparation and completion of this thesis would never have been achieved, without the sincere help and moderation by my supervisor Bo Jørgensen at the Danish Institute for Fisheries Research, and co-supervisor Rasmus Bro at the Royal Veterinary and Agricultural University. The involved herring industry is thankfully appreciated for the contribution providing data and permitting knowledge of the production process.

Furthermore, my gratitude goes to librarian Søren T. Christensen and quality manager Karen Michaelsen, together with colleagues at both the Danish Institute for Fisheries Research and the Royal Veterinary and Agricultural University. Durita Nielsen and Charlotte Jacobsen from DIFRES are both gratefully acknowledged for providing the materials about fat measurement and gas chromatographic data, respectively.

A great appreciation also goes to my family, for offering help in many ways throughout the work with my thesis and for their patience and understanding, especially when bringing working laptops into any occasion, at any time.

Special thanks go to my husband Laurs Møller for editing and for always being a positive inspiration, bringing confidence that hard times doing this thesis, was worth going through.

Stina Frosch Møller
Copenhagen – December, 2005

Summary

The general aim of the thesis was to develop a documentation system and to improve the background upon which the decision-making process for quality and production control is founded within a herring processing industry. Furthermore, the possibilities of utilizing multivariate data analyses were investigated conducting data from catch to final product throughout the production chain. When generating vast amount of data, as in the case of processing herring, various samples turn out to deviate from the majority of samples, also designated outliers. Due to the nature of outliers, they posses the ability to impair analysing models based on traditional multivariate methods using least squares estimation. For that reason, possible advantages or drawbacks employing robust multivariate methods were investigated as a favoured alternative to the traditional methods.

The first part of the exploratory work was carried out as a case-study, exploiting the multiplicity of empirical and biological data, intended for quality determination in one of the leading businesses within the herring industry in Denmark. The work started out constructing a database to save all registered information, this being extended to be automatically imported, transmitted as e.g. measured weights to the database. In the case of non automatic transmission of data, the import of data to the database was manually recorded as soon as they were generated.

The preliminary screening of data demonstrated that traceability could be confirmed from vessel unto the finished marinated produce of herring with the smallest unit of traceability being a batch of topped product. This finding revealed that it was possible, at any time, to track and trace any given product back to the vessel that originally caught the fish, and do extraction of all data connected to that specific product.

Unfortunately, a great part of the multiple registrations lacked variability and suffered from uncertainties caused by the lack of traceability and/or misgivings, related to the actual registering of analysis. This, in combination with missing information of relevance, lead to that data at its present form neither had any relevance nor was representative for any further multivariate data analyses. For that reason, it was not possible to identify and link any relations between, for instance the quality characteristics of the raw material and yield, and

thereby improve the basis for the decision-making process concerned with quality and production control, within the herring processing industry.

In place of the fact that the data had to be discarded, in relation to multivariate data analyses, they proved useful in the sense that they could be informative in relation to what information needed to be improved or added to be profitable to the business. A few to mention is registration of belly bursting and waste, along with implementation of an on-line determination of fat content on single fish level and consecutive sorting of the raw material based on this fat determination. Additionally, a quality evaluating system of the marinated herring would improve the significance of the data.

Gas chromatograms of fatty acid methyl esters (GC-FAME) and of volatile lipid oxidation products (GC-ATD) from fish lipid extracts were analysed by multivariate data analysis (principal component analysis). Peak alignment was necessary in order to include all sampled points of the chromatograms in the data set. The ability of robust algorithms to deal with outlier problems, including both sample-wise and element-wise outliers, and the advantages and drawbacks of two robust PCA methods, robust PCA (ROBPCA) and robust singular value decomposition (RSVD) when analysing these GC data were investigated. The results showed that the usage of robust PCA is advantageous, compared to traditional PCA, when analysing the entire profile of chromatographic data in cases of sub-optimally aligned data. It was also demonstrated how the robust PCA method – sample (ROBPCA) or elementwise (RSVD) – depended on the type of outliers present in the data set.
The potential of removing Rayleigh and Raman scatter from fluorescence data (excitation – emission landscapes), by employing robust PARAFAC, were investigated. A PARAFAC algorithm was made robust by substitution of least squares estimation by least absolute error (*LAE*). The conclusion was that *LAE* PARAFAC cannot be considered as a confident method for handling scatter, as a result of the systematic nature of scattering. However, by taking advantage of the systematic nature of the scatter an automatic method based on robust techniques for identification of scatter in fluorescence data were developed. This method can handle both Raman and 1st and 2nd order Rayleigh scatter, and do not demand any priori visual inspection of the data before modelling.

The investigation of using robust calibration methods for prediction of fat content of fish by NIR measurements in a data set with no extreme outliers present showed that the advantages of employing robust methods for prediction was ineligible. A slightly better prediction was obtained with robust SIMPLS (RSIMPLS) compared to classical PLSR, but further investigation is needed to test the performance on an independent test set. Focusing on the drawbacks of the robust methods, especially the lower statistical efficiency and the time-consuming computations, the advantages of robust methods seems to be eliminated, when the dataset contains no obvious outliers.

Sammendrag

Formålet med dette Ph.d. projekt var at udvikle et dokumentationssystem og forbedre beslutningsgrundlaget for kvalitets- og produktionsstyring i sildeindustrien, samt at undersøge mulighederne for at benytte multivariat dataanalyse på data registreret i kæden - fra fangst til færdigt produkt. Ved generering af store datamængder, som eksempelvis i den involverede industri, vil der ofte optræde prøver, der afviger fra hovedparten af de øvrige prøver, såkaldte outliers. Hvis ikke sådanne prøver fjernes fra dataanalysen, vil de i værste fald ødelægge modellerne baseret på de traditionelle multivariate metoder, da disse er beregnet på baggrund af mindste kvadraters metode. Derfor blev eventuelle fordele og ulemper ved anvendelsen af robuste multivariate metoder som alternativ til de traditionelle multivariate metoder undersøgt.

Første del af projektet var baseret på en case, der anvendte de mangfoldige erfaringsdata samt biologiske og kvalitetsmæssige data fra en af Danmarks største virksomheder inden for sildeindustrien. Projektet blev indledt med opbygning af en computerbaseret database til opsamling af alle registrerede informationer. Undervejs i projektet blev databasen udbygget, således at mange registreringer nu automatisk overføres direkte fra f.eks. vægtene til databasen. I de tilfælde, hvor en automatisk overføring af data ikke er mulig, tastes data manuelt ind i databasen, så snart de genereres.
Ved den indledende screening af data blev det fundet, at der var sporbarhed fra kutter til færdigmarineret produkt, og at den mindste sporbare enhed var en batch af toppet produkt. Det vil sige, at det altid er muligt at spore et produkt tilbage til kutteren og udtrække alle data, der knytter sig til netop det produkt i databasen.
Endvidere viste det sig, at mangel på variabilitet i mange registreringer samt usikkerhed på grund af manglende sporbarhed og/eller usikker prøveudtagning, kombineret med direkte manglende informationer om relevante forhold, bevirkede, at data i den foreliggende form hverken var relevante eller repræsentative for en videre multivariat dataanalyse. Det var derfor heller ikke muligt at relatere nogle sammenhænge mellem f.eks. råvarens kvalitetsmæssige egenskaber og udbytte og dermed forbedre beslutningsgrundlaget for kvalitets- og produktionsstyring i sildeindustrien.
De foreliggende data kunne i stedet bruges til at påpege, hvilke informationer der eventuelt kunne forbedres, så de blev mere fyldestgørende, for eksempel kvalitetsvurderingen af

de syremarinerede sild og fedtbestemmelserne ved indføring af online fedtbestemmelse på individniveau med efterfølgende sortering, og hvilke registreringer det kunne være givtigt for virksomheden at opsamle, så som mængden af bugsprængte sild og mængden af spild.

Gaskromatografi af fedtsyre methylestere (GC-FAME) og af flygtige lipid oxidations produkter (GC-ATD), fra ekstraktioner af fiskeolie, blev analyseret ved multivariat data analyse (principal komponent analyse). En forudgående forskydning af retentionstiderne, så kromatogrammerne var sammenlignelige, var nødvendig for at inkludere alle prøvepunkter af kromatogrammet i analysen. En nærmere analyse af robuste metoders evne til at håndtere outliers, inkluderende både elementvise og prøvevise outliers, på GC data blev udført for at undersøge fordele og ulemper ved to robust PCA metoder, 'robust PCA' (ROBPCA) og 'robust singular value decomposition' (RSVD). De to metoder er robuste over for henholdsvis afvigende prøver (ROBPCA) og elementvise outliers (RSVD). Resultatet viste, at man med fordel kan bruge robust PCA sammenlignet med traditionel PCA, når man analyserer hele profiler af kromatografiske data, i tilfælde hvor der er tale om 'sub-optimal' forskydning af kromatogrammerne. Yderligere viste resultaterne, at man, afhængig af den type outliers der er tale om i datasættet, skal vælge enten prøvevise eller elementvise robuste metoder.

Muligheden for at fjerne Raman og 1. og 2. ordens Rayleigh scatter i fluorescens data (eksitations – emissions spektre) ved hjælp af robust PARAFAC blev undersøgt. Den anvendte PARAFAC algoritme blev gjort robust ved at erstatte mindste kvadraters afvigelse med mindste absolutte afvigelse (*LAE*). Konklusionen herpå var, at *LAE* PARAFAC ikke kan betragtes som værende en pålidelig metode til håndtering af scatter, hvilket skyldes den naturlige systematiske tilstedeværelse af scatter. Den systematiske tilstedeværelse af scatter kan dog udnyttes konstruktivt, og en automatisk metode baseret på robust statistik til identifikation af scatter i fluorescens data blev udviklet. Denne metode er i stand til at håndtere både Raman og 1. og 2. ordens Rayleigh scatter, og kræver ingen forudgående visuel inspektion af data.

Anvendelsen af robuste kalibreringsmetoder til prædiktion af fedtprocenten i fisk, ud fra NIR målinger i et datasæt uden ekstreme afvigende prøver, viste, at fordelene ved at anvende robuste metoder var begrænsede. En svagt bedre prædiktion blev dog opnået ved

anvendelse af robust SIMPLS (RSIMPLS) sammenlignet med klassisk PLSR, men yderligere undersøgelser er nødvendige for at teste prædiktionsevnen for uafhængige testsæt. Vender man blikket mod de robuste metoders reducerede statistiske egenskaber og den forholdsvis lange beregningstid, syntes disse at begrænse fordelene ved anvendelsen af robuste metoder i de tilfælde, hvor datasættet ikke indeholder deciderede outliers.

Abbreviations

ALS:	Alternating Least Squares
CSW:	Chilled Sea Water
ICES:	International Council for the Exploration of the Sea
LAE:	Least Absolute Error
LS:	Least Squares
LTS:	Least Trimmed Squares
MCD:	Minimum Covariance Determinant
NIPALS:	Nonlinear Iterative Partial Least Squares
NIR:	Near Infrared Reflectance
NMR	Nuclear Magnetic Resonance
PARAFAC:	Parallel Factor Analysis
PC:	Principal Component
PCA:	Principal Component Analysis
PCR:	Principal Component Regression
PLSR:	Partial Least Squares Regression
RMSEP:	Root Mean Square Error of Prediction
RSW:	Refrigerated (chilled) Sea Water
SVD:	Singular Value Decomposition
TRU:	Traceable Resource Unit

List of papers

This thesis is based on the following papers, which will be referred to in the text by their Roman numerals:

I. **How to turn data compilation and traceability in the herring processing industry into a profitable business (Viewpoint).**

Frosch Møller, S.

Danish Institute for Fisheries Research, Department of Seafood Research, The Technical University of Denmark, Build. 221, DK-2800 Kgs. Lyngby, Denmark

Submitted to Trends in Food Science and Technology

II. **Robust methods for multivariate data analysis (Review article).**

Frosch Møller, S.[1], von Frese, J.[2] and Bro, R[2].

[1]Danish Institute for Fisheries Research, Department of Seafood Research, The Technical University of Denmark, Build. 221, DK-2800 Kgs. Lyngby, Denmark

[2]Spectroscopy and Chemometrics Group, Quality and Technology, Department of Food Science, The Royal Veterinary and Agricultural University, DK-1958 Frederiksberg C, Denmark

Journal of Chemometrics, 19: 549 – 563, 2005

III. **Peak alignment and robust principal component analysis of gas chromatograms of fatty acid methyl esters and volatiles.**

Frosch Møller, S. and M. Jørgensen, B.

Danish Institute for Fisheries Research, Department of Seafood Research, The Technical University of Denmark, Build. 221, DK-2800 Kgs. Lyngby, Denmark

Journal of Chromatographic Science, 45: 169 - 176, 2007

IV. **Automatically identifying scatter in fluorescence data using robust techniques.**

Engelen, S[1]., Frosch Møller, S.[2] and Hubert, M[1].

[1]Katholieke Universiteit Leuven, Department of Mathematics, W. De Croylaan 54, 3001 Leuven, Belgium.

[2]Danish Institute for Fisheries Research, Department of Seafood Research, The Technical University of Denmark, Build. 221, DK-2800 Kgs. Lyngby, Denmark

Chemometrics and Intelligent Laboratory Systems, 86: 35 -51, 2007

Table of contents

1.0 Introduction

1.1 Background

A deeper knowledge of the relation between raw material properties, food production and the quality of food products is of great importance to the food industry as basis for production planning and product differentiation. Moreover, demands from authorities and consumers increase the product documentation and traceability. In the cases of food scandals, the industry wants to protect their brands by product and quality documentation. A system able to fulfil such needs will be of great importance to the whole food industry.
Considering fisheries and the handling of fish products, as products in any other food producing industry, there is a need to ensure optimal traceability at all stages, from processing to marketing.

The processing and handling of fish products at fisheries, generates huge amounts of data, due to great volume and high speed handling along with a range of quality measures obtained at different stages during processing. When handling such great amounts of data, multivariate data analysis is a tool that offers powerful methods, capable of analysing complex data, in a much more simple way than previously achieved (Munck et al. 1998).
Thus, it is now possible for the industry to explore and document relations that have previously only existed as “experienced personnel knowledge” and knowledge of the trade. Furthermore, multivariate data analysis can point out new and, till now, unknown relations (Bechmann et al. 1998; Nielsen et al. 1999; Nielsen et al. 2000). By integrating the multivariate techniques into the factory’s documentation system improved quality control and utilization of the herring resource can be obtained.

Today, only a limited amount of all the data collected throughout the whole production chain (raw material, intermediate products and final products) are used, even though it has been shown that it is possible to build enhanced and safer systems based on multivariate data analysis from already obtained data (Kourti et al. 1996).

When working with huge amounts of data, in both industry and research, the presence of outliers is more the rule than the exception; especially in data mining projects where data often stem from many different sources and hence are of varying quality. Outliers are observations, in this case collected data that appear to break the pattern or grouping shown by the majority of the observations. An outlier can both be a whole sample, an entire variable/measurement or just one individual measurement. The reasons for outliers are various, e.g. instrument failure, non-representative sampling, formatting errors, and/or objects stemming from other populations.

Unfortunately, most conventional multivariate data analysis methods are sensitive to outliers, due to the fact that they are based on the least squares estimate. This means that the presence of even just one single outlier in a given data set can have a large and even detrimental effect on the estimate and lead to incorrect conclusions. For that reason, it is necessary to identify outliers and decide, whether the outliers should be accommodated or rejected, in the modelling process.

The outlier problem can be solved in two ways: either by diagnostics or robust estimators (Rousseeuw & Leroy, 1987). In outlier diagnostics, the outliers are identified and expelled from the data set prior to making the multivariate model. A complication to this procedure is that it may be difficult to identify outliers, especially when multivariate data are available. Furthermore, the task gets even harder and more time-consuming, when the amount of data is huge. In the second approach, robust estimators are used instead of the ordinary non-robust least squares estimator. Robust methods reduce or remove the effect of outlying data points, allowing the remainder to predominantly determine the model. Therefore, owing to the challenges mentioned above, robust methods may be considered superior to the classical methods based on least squares and might be an excellent alternative, especially in situations where automatic and fast methods are required, as in the case of production industries. There are problems, though, with robust methods which call for some caution in their automated use as will be discussed in this thesis.

1.2 Objectives

This thesis and the objectives can roughly be split up in two main parts concerning:

1) Analysing data from the herring industry, and
2) Investigating the possibilities of using robust multivariate methods in data mining.

The link between the two parts is multivariate data analysis.

The first part of the project is built on a case study, using data from one of Denmark's largest herring industries. The traceability chain from fishing vessel to final marketed product will be scrutinized, successively analysis of the data will be performed, and the possibilities of integrating multivariate techniques into the industrial documentation system will be investigated.

The advantages and drawbacks of robust procedures for common multivariate methods, such as principal component analysis (PCA) and partial least squares regression (PLSR), will be presented by use of different kinds of data obtained from fish research in part 2.

Following section one, section two gives a short introduction to the common multivariate data analysis methods, PCA, parallel factor analysis (PARAFAC), principal component regression (PCR) and PLSR, to enlighten how these methods function and why they are interesting. Section three covers the analysis of the data from the mentioned herring industry. An introduction to outliers and robust methods can be found in section four, together with examples of how these methods employ in practice. Concluding remarks can be found in section five together with discussions of further perspectives.

1.3 Approach

Previously in the herring industry, all available data were registered on paper based forms. This makes it impossible to export information and compare data from different schemes, especially when one wants to compare much information from many schemes at one time. Therefore, before the analysis of data from the herring industry could take place, it was necessary to build a computerized database; to register all collected data, and develop a webbased user interface to log the paper based systems and various day reports. The focus in this study is limited to the production of marinated herring. Furthermore, a report tool to export data from the database to the Excel® format was developed. In this database, already registered data going back three years, were logged. These data will be referred to as historical data in the following, and make up the data used for the data analysis in section three. As can be imagined, the database was continuously extended. For the measurements/registering, and where possible, the data was logged and exported automatically. By automatic logging of data, the workload is reduced and the risk of formatting errors is limited. The development of the computer based systems was done in close collaboration between the industry and DFU-IT, to ensure a system that lives up to industrial needs, both concerning user interface and practical conditions such as a very acid and wet environment.

As the analysis progressed of the data from the herring industry, results revealed that available data lacked the ability to illustrate any advantages or drawbacks concerning robust multivariate methods. For that reason three different data sets from laboratory analysis were included in this project to investigate possible opportunities of robust methods giving different circumstances.

2.0 Multivariate data analysis methods

The following section provides an introduction to PCA, PARAFAC, PCR and PLSR, since they were intended to be applied to the data obtained from the herring industry, and furthermore, make up the background of the robust multivariate methods, managed within this thesis. First of all, a short introduction to multivariate data analysis will be given.

2.1 Multivariate data analysis

Multivariate data analysis techniques are appropriate when several response variables are measured on a sample, and repeated for many samples. The multivariate methods are often more powerful and more information about the samples can be retrieved, when analyzing complex data, compared to traditionally univariate techniques. This is due to fact that the multivariate technique utilizes the correlation among all response variables, instead of simply looking at one or a few variables at the same time. Multivariate data analytical tools handle data by extracting underlying linear independent (so-called latent) variables from the original variables.

Considering the data from the herring industry, the variables have different entities, and measurements can be as different as, e.g. catch area, fat content, size and quality measurements throughout the production chain and the samples are batches of final marinated products. In this case, we want to establish relationships, identify patterns and construct predictive models based on them, a procedure also known as data mining.

The variables do not necessarily arise from different kinds of measurement, as in the fish industry case. As is often the case, instruments produce a huge number of often highly correlated measurements per sample, as in e.g. spectroscopy and gas chromatography. In stead of simply looking at one or few wavelengths or peaks of interest, whole spectra, landscapes or chromatograms can be analyzed with multivariate data analysis. Instruments that hold the capacity of spectroscopy and chromatography

have widely been brought into play in the industry since; they are fast, non-destructive and suitable for application on-line.

The types of data, described earlier, are organized in a table – called a data matrix – in which *I* samples (observations) constitute the rows and the *J* measurements (variables), constitute the columns. This matrix can be analysed and decomposed with multivariate methods, such as PCA, PCR and PLSR. Three-way matrices also exist, when e.g. the measurement of one sample can be represented as one matrix, or when the same measurements are obtained on a time basis. Three-way matrices can be analysed by three-way methods, such as PARAFAC – an extension of the bilinear PCA into multilinear situations.

PCA and PARAFAC are qualitative methods decomposing the data into fewer components which are easier to interpret. Regression methods, as PCR and PLSR, are quantitative often used for prediction.

Common for all multivariate methods are; to obtain a good result, data should contain relevant information about the desired property, the quantitative relationship between the set of measured variables and the property of interest should exist.

2.2 Principal component analysis

PCA is the transformation of the originally *J* variable onto *A* latent variables (Hotelling, 1933; Wold et al., 1987). PCA is a commonly used method to study the multivariate data, in a model of reduced complexity, allowing for an easier interpretation and better understanding of the different sources of variations. For that reason, PCA is often the first step in the data analysis.

In PCA, a data matrix **X** is decomposed into the matrix products **TP**' and the residual matrix **E** (*Equation* 2.1).

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \qquad \textit{Equation } 2.1$$

The matrix product, **TP'**, consists of the score matrix, $\mathbf{T} = [t_1, t_2, t_3,\ldots,t_A]$, and the transposed loading matrix, $\mathbf{P} = [p_1, p_2, p_3, \ldots, p_A]$, which contains the underlying structure in the data, based on *A* latent variables or principal components. The principal component (PC) is defined as a weighted average of all the original variables. Each loading is the weight of the concerned variable, describing how this variable contributes to the PC under consideration. The loading thereby describes what type of information characterizes the samples. The associated weighted averages are the scores, describing how much of each PC the sample contains, i.e. the scores contain quantitative information about the samples. The residual matrix, **E,** contains the remaining information or noise in **X** that was not described by **TP'**.

The scores and loadings are found using a least squares approach which locate the direction, explaining the maximum quantity of variance in the original data. The second principal component is then orthogonal to the first and again maximizes the quantity of variances, not captured by the first PC. Continuing this procedure generates all the principal components, which corresponds to the eigenvectors of the empirical covariance matrix.

Different algorithms exist for finding the principal components, with nonlinear iterative partial least squares (NIPALS), and singular value decomposition (SVD) as the most common. The NIPALS algorithm is an iterative procedure that successively find the principal components, whereas as SVD computes all the eigenvectors simultaneously. SVD is numerical more stable than NIPALS. Furthermore, separations between otherwise nearly similar eigenvectors are obtained with NIPALS. On the other hand, the NIPALS algorithm can handle missing values in the data matrix, which is a common phenomenon. For a detailed description of the NIPALS and SVD algorithms, the reader is referred to Wold et al. (1987) and Jackson (1991), respectively.

2.2.1 PARAFAC

Canonical decomposition (CANDECOMP)/ Parallel factor analysis (PARAFAC) is an extension of PCA, to higher order data (Carroll & Chang, 1970; Harshman, 1970). For brevity, it will be referred to as PARAFAC in this thesis, moreover, only the three-way situations will be considered, even though the method can be extended to higher dimensions.

A decomposition of the data is made into triads or trilinear components. When the elements of a three-way array, $\underline{\mathbf{X}}$ (I x J x K), are given as x_{ijk}, $i = 1,\ldots..,I$, $j = 1,\ldots..,J$ and $k = 1,\ldots..,K$, then the structural model can be described as

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \qquad \textit{Equation } 2.2$$

where a_{if}, b_{jf} and c_{kf} denote elements of the loading matrices, **A** (I x F), **B** (J x F) , and **C** (K x F), respectively, and e_{ijk} denotes an error term for element, x_{ijk} (variation not captured by the model). F is the number of factors needed to describe the variation within the data. The model is fitted to a data set by minimizing the sum of squared residuals over **A**, **B** and **C,** by means of an alternating least squares (ALS) algorithm (Carroll & Chang, 1970; Harshman, 1970). In matrix notation, the PARAFAC model is normally written

$$\mathbf{X} = \mathbf{A}\mathbf{D}_k\mathbf{B}^{'} + \mathbf{E}_k, \quad k = 1,\ldots, K \qquad \textit{Equation } 2.3$$

where, $\mathbf{D}_{k,}$ is a diagonal matrix holding the kth row of **C,** in its diagonal, and **E** is a matrix of residuals.

The principle behind ALS is to separate the optimization problems, into conditional sub problems, and solve these in a least squares sense. Each subset of ALS fixes two of the loading matrices (**A**, **B**, and **C**), and then uses least squares regression to find the third factor matrix. The estimation of the three loading matrices is repeated iteratively, each

iteration providing a better (not worse) estimate, of one set of loadings. The overall algorithm will therefore improve the least squares fit of the model to the data. An ALS algorithm follows as:

(0) Decide the number of components, *F*
(1) Initialize **B** and **C**
(2) Estimate **A** from **X**, **B** and **C** by least squares regression
(3) Estimate **B** likewise
(4) Estimate **C** likewise
(5) Continue from 2 until convergence

If the algorithm converges to the global minimum, which is most often the case for well-behaved problems, the least-squares solution to the model is found (Bro, 1997).

The algorithms for fitting PARAFAC models are not sequential as PCA, hence refitting is necessary when, e.g. several models are being tested, as any higher number of components can not be estimated from a solution with a lower number, e.g., during outlier detection.

2.3 Multivariate regression methods

PCR (Hotelling, 1957; Kendall, 1957) and PLSR (Wold et al., 1983; Geladi & Kowalski, 1986; Martens & Næs, 1989) are multivariate regression methods, which attempt to relate multivariate data, **X**, to a reference value, **y:**

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \qquad \textit{Equation } 2.4$$

where, **b,** represents the regression coefficient and, **e,** is the variation not captured in the model. The methods can be used for analyzing data, which are strongly collinear, noisy and contain numerous **X** variables.

Typically, data in **X** are low-cost measurements that can be obtained rapidly, such as near infrared reflectance (NIR) measurements, whereas *y* data is often time-consuming and expensive reference methods. The overall purpose of the methods is to interpret the relationship between the two data sets, and to predict the *y* value in future samples. By example, fat content is of great importance for the quality of marinated herring products. Today, the fat content is measured in the laboratory by a slow and destructive method. A fast and non-destructive method for online fat determination in whole herring or herring fillets will be of great interest for the herring industry, since it will make it possible to sort the resource into much more homogenous batches and thereby optimize the production. NIR, in combination with PLSR, has shown great potential, as a fast and non destructive method for predicting the fat content in herring and herring fillets (Nielsen et al., 2005).

2.3.1 Principal component regression

PCR has become an established tool for modelling linear relations between multivariate measurements. In PCR, **X,** is first decomposed via PCA, and subsequently the scores, **T**, for a given number of components, are used as independent variables in multiple linear regression,

$$\mathbf{y} = \mathbf{Tb} + \mathbf{e} \qquad \textit{Equation } 2.5$$

relating **y** to **X**.

In situations where **X** contains a large amount of information, irrelevant for modelling **y**, PCR might fail; in view of the fact that PCR uncritically seeks the principal components, describing maximum variation in **X,** which in this case had no relevance for **y**. The worst case scenario will be when the variation, relevant for **y,** might be expressed in the higher order principal components, often regarded as noise, and normally left out of the regression.

2.3.2 Partial least squares regression

PLSR is a linear regression technique developed to deal with high-dimensional regressors by one, **y** (PLSR1), or several response variables, **Y** (PLSR2). Like PCA and PCR, PLSR is a technique for reduction of dimensionality, moreover, the PLSR technique is focused on maximizing the predictive power by guiding the decomposition of **X** during regression by the variance in **y**.

The main difference between PCR and PLSR is that in PLSR, additional loadings, called **W** (for loading weights), for **X,** are determined in a way that the covariance between **X** and **y** is put to ist maximum. After finding **W,** the belonging latent variable **T,** is found and used for regression on **y,** as described for PCR. This leads to components, which are more directly related to the variability in **y,** than by the principal components in PCR. As a result of the construction of PLSR, the PLSR technique requires fewer components than PCR (Martens & Næs, 1987; de Jong, 1993)

The most common algorithm of PLSR, considering the chemometric field, is the NIPALS PLSR algorithm. But also the SIMPLS (de Jong, 1993) algorithm is popular. In cases with only one responsible variable (**y** = 1), and no missing values, SIMPLS and PLSR1 (NIPALS) generate the same results (de Jong, 1993).

3.0 Data from a herring industry

Atlantic herring (*Clupea harengus*) is of great importance to the Danish fishing industry. As aquaculture product, herring is primarily processed into marinated and salted products. A significant share though, is also exported to semi-manufactures. The dominating herring stocks caught and processed by the industry are from the nearby seas around Denmark (the North Sea and the Baltic Sea etc). The herring, a pelagic specie, is found in large schools. As raw material the herring cannot be considered very consistent, as fish caught at the same fishing ground in the same season can have different biological origin due to mixing of different stocks, and is therefore likely to have different biochemical and functional properties as raw material. The fat content is an example of a parameter, which has revealed large variations within a catch, when considering fishing ground and season (Larsen et al., 1997; Nielsen et al., 2005).

For the last five years, since 2000, landing of herring has been decreasing or constant and marketing prices have been kept unchanged at approximately two Danish kr./kg, approximately 0.27 € for fish for human consumption (Danish Directorates of Fisheries, Ministry of Food, Agriculture, and Fisheries). Of the 200.000 to 300.000 tons of herring, landed in Denmark each year, 55 – 90 % is used for human consumption and 10 - 40 % is used as "industrially fish" and further on processed into fish meal and fish oil (Danish Directorates of Fisheries, Ministry of Food, Agriculture, and Fisheries).

The very competitive situation in the fish processing industry today means that there is an increased commercial interest in making the production more cost effective and raising the efficiency by rationalizations (Larsen et al., 1997). Furthermore, every year 10 – 20 % of the herring caught for human consumption in Denmark is discarded because of unacceptable quality and instead used for feed. This is unsatisfactory not only in terms of production cost, but also according to stock preservation. To decrease the discarded quantities of herring for human consumption and ensure a better utilization of the herring resource, better understanding of how the biological factors (fishing ground, season, fat content etc.) influence the quality and products characteristics, is necessary.

Little is known about the influence of this variation in the raw material on the quality properties of herring and especially herring products. Generally, knowledge is based on personal knowledge obtained from many years of work in the field. In accordance with marinated herring products, lipid oxidation, soft texture and belly bursting is mentioned as some of the most important quality problems by people working in the industry. Belly bursting is related to raw material quality, whereas soft texture and lipid oxidation both are related to raw material quality and the production process e.g. marinating procedure and recipes are related to soft texture and incorrect mixing of herring and marinade or too little marinade in the barrels are related to lipid oxidation.

In this study, using data from one of Denmark's largest herring industries, the chain of traceability from fishing vessel to final product will be scrutinized, as a basis for successive data analysis to gain better knowledge of the relation between the properties of the raw material and the quality of the final products. Furthermore, the possibilities of integrating multivariate techniques into the industrial documentation system, will be investigated to improve process control as well as gain better utilization of the herring resource.

3.1 The production line for marinated herring products

The production line from raw material to marinated herring products is illustrated in Figure 3.1. The marinated herring products are so-called semi-manufactured products, which are sold to other companies for final processing before the products are ready for consumption. The different production steps will be described in the following.

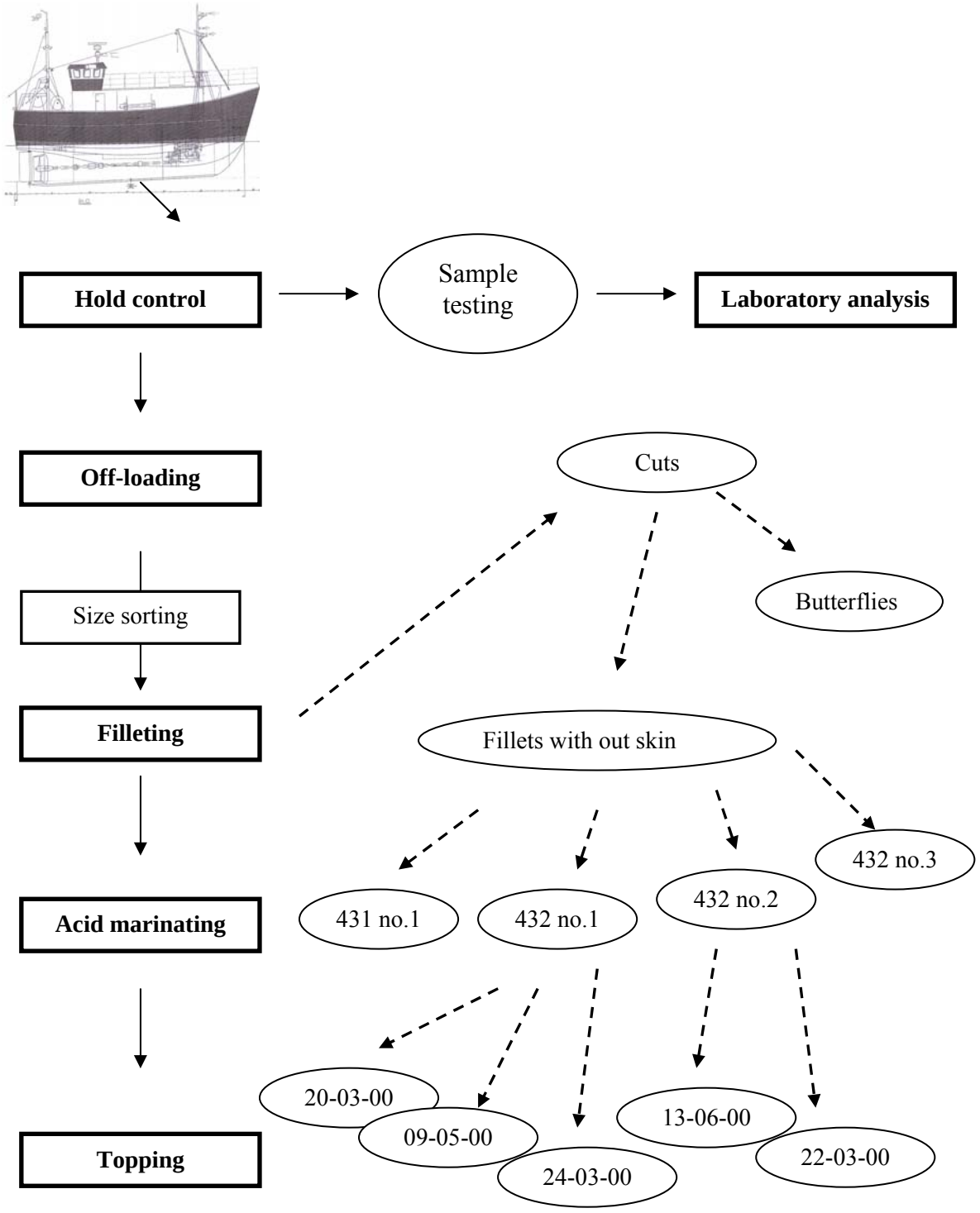

Figure 3.1. The production line: from raw material to marinated herring products from a typically Danish herring industry. The bold arrows show the production flow and the dotted arrows show examples of the different branches during the production.

The herring are caught with pelagic pair-trawl or purse seine, and pumped on board the fishing vessels into different holds. The fish are stored and cooled by either chilled sea water (CSW) or refrigerated chilled sea water (RSW) in the holds. To ensure correct storage of the catch on board, the temperatures in the holds are measured with regular intervals during the trip.
For all regular suppliers of herring to the industry, information about catch method, chilling method and hold capacity are known.
For every towing, information about catching ground, date of towing, duration of towing, amount of fish (herring and by-catch) and which hold(s) the towing is pumped into are registered. Furthermore, a counting sample is taken where the herring are graded into three sizes (small, medium and large), and the size is registered as piece herring per kg. For every trip the total gross amount is registered. By gross amount is meant the assumed amount of herring and by-catch including water. In Denmark the water is assumed to make up 13 % of the catch.

Arriving at the harbour, the raw herring in each hold is visually inspected for quality including freshness. The fish can be rejected as “not suited for human consumption” or “not suited for production”. The rejection “not suited for human consumption” requires the presence of the authorities and their rejection of the content. The rejection “not suited for production” means that the company can not use the herring in their production due to e.g. size, belly bursting or bad quality, caused by incorrect or too long storage on board. The outcome of the control is registered, and if rejection is necessary the reason is stated and registered. Fish passing visual inspection are transported by a conveyor belt into a chilled storage tank in the production plant.

Normally, herring (approximately 25 herring per hold) from three randomly chosen accepted holds are taken out for further quality determination in the laboratory. Before the quality determination, 20 herring from each hold are filleted in the production line. The quality determination consists of a sensory evaluation, temperature measurements of the herring and a counting test (piece herring per kg). During the sensory evaluation, colour, consistency, odour and the general quality of both the whole and the filleted herring is evaluated. A quality mark is calculated on the basis of the sensory evaluation.

Furthermore, the fat content is calculated, and the numbers of nematodes (*Anisakis larvea*) are determined (number per 5 kg). The fat content is determined for those product types (primarily fillets without skin and butterflies) which are produced from the specific trip. The fat content is not determined directly but based on a dry matter determination where 10 gram herring from a pooled sample of herring are minced and dried. The method takes advantage from experience that the sum of water and fat is constant (approx. 80 %)[1] and negative correlated, and the consequence, that the fat is a part of the dry matter. This means that the dry matter determination can be used to estimate the water content and thereby the fat content can be calculated. The dry matter determination is conducted as a single determination. The calculated fat content results are registered.

A conveyor takes the herring through size graders. In Figure 3.2 such a grading system is illustrated.

Figure 3.2. A grading system for use in the herring industry.

Afterwards, the herring are taken into fillet machines (Figure 3.3). Just before filleting the herring are "visually" inspected by an automatic vision system, removing non-herring and herring not rightly placed for filleting. The removed herring are taken back into the system for "another round", whereas the non-herring (e.g. mackerel) are discharged. Fillets pass out of the filleting machines into a conveyor system, which leads to the marinating process. During the transport the filleted herring are visually inspected for errors.

[1] However, that is not always the case (see e.g. Nielsen et al., 2005).

Figure 3.3. Fillet machine for use in the herring industry.

At each production line, counting samples are taken frequently for each product type typically fillets without skin and butterflies (both fillets, connected, with skin on), see Figure 3.4.

Figure 3.4. Fillets without skin (left) and butterflies (right).

For a fast size determination, the number of filleted pieces per 3 kg is counted. Afterwards, approximately 50 pieces of filleted products are weighed out on a laboratory weight, and the mean weight (2 decimals) +/- the standard deviation is printed. The filleted products are evaluated ("Very nice", "Nice", "Less nice" or "Deviating") in connection with the counting sample and quality errors are registered (Soft, Fungi, Red colour, Wrong cut or By-smell).

Before marinating, some fillets are pre-salted in brine (13 % NaCl) for minimum 8 -12 hours at 5 °C. To ensure a homogeneous salting, stirring takes place. If pre-salting takes place, it will be noted.

Fillets for marinating are mixed with marinade at a present ration (1.5 kg fish to 1.0 kg of marinade) in barrels or tanks. The marinade is a mixture of purified water, NaCl and acetic acid and with a pH value at approximately two. The composition of the marinade has been formulated to marinate fish and kill nematodes when stored for 35 days at 5 °C. The essential preservation factor is acid (lowering the pH), but without an adequate proportion of salt, the softening process, which is an additional effect of the acid, would proceed too far (McLay & Pirie, 1971). A sample is taken from the marinade for laboratory control to determine percentage acid, percentage salt and pH. Controlled quantities of hydrogen peroxide may be added to the marinade, if the products are for export.

The sealed barrels are rolled to ensure proper mixing of the fillets within the marinade. To avoid rancidity an insertion is used to keep the herring downwards in the marinade and thereby avoid the exposure to oxygen and subsequent rancidity. Rancidity is recognized as a yellow colouring of the fish meat.

The barrels are left to cure (i.e. to become marinated herring) for a minimum of 35 days at 5 °C. During the marinating period spot tests are taken to control the quality. Herring samples are evaluated concerning “Appearance”, “Consistency”, “Smell / Taste”, “Homogeneity” and “Specifications kept”, in addition samples are taken to measure pH and salt in both the fish and in the marinade.

After the marinating period, the barrels are topped. When topping, the content of a barrel is tipped off onto a draining board to remove excess marinade. If there is too little marinade or a bad smell is identified, it is noted. The fillets are visually inspected to remove any oxidised, yellow or badly cut fillets. Any of these quality errors are registered together with eventually foreign bodies. A sample of the marinade is taken for further analysis at the laboratory (% NaCl, % acid and pH). A counting sample is taken. Approximately 50 pieces of marinated product are weighed out on a laboratory weight, and the mean (2 decimals) + / - standard deviation is printed. The fillets are then poured into clean plastic barrels and topped up with marinade. The recipe for the

marinade can be either identical with the previous marinade or customer specific. The products are now ready for sale to other companies.

3.2 Traceability

Traceability is an important issue for a number of reasons. First of all, it is given by law within the European Commission (EC) regulation, 178/20027EC on General Food Law, issued on the 1st of January 2005. The regulation states that traceability is to be established at all stages of the food chain. This implies that it should be possible to trace and follow a food, feed, food-producing animal or substances throughout all stages of production, processing and distribution (EC Regulation 178/2002). Although given by law, there are a number of additional reasons that motivates traceability in the aspect of quality management. With effective traceability systems in place, it might bring extensive benefits to businesses, when used under proper conditions, for instance; process control, process optimization and better marketing (Paper I).
Within the fishing industry, traceability from catch to final product is furthermore necessary when links between raw material production and final product quality are investigated, as is the case of this study.

According to the ISO standard 8402 (ISO 1994), traceability can be defined as:

Traceability is the ability to trace the history, application or location of an entity, by recorded identifications.

Product traceability is first of all based on the ability to identify products uniquely. Unique identification means, according to traceability, that no other unit or component can have exactly the same, or comparable, characteristics. Unique identification and traceability in any system, hinges on the definition of what is the batch size, or using the terminology by Kim et al. (1995), the Traceable Resource Unit (TRU). The TRU size depends on what level (single fish, catch, or production day) it is possible to get specific information from. In some cases, different batches are pooled which will create new

TRU's. The first step, when implementing data analysis, is therefore to investigate the traceability chain, and decide the size of the TRU.

The traceability chain in this study includes handling from catch through processing,to final production of semi-manufactured marinated herring. Information about the subsequent history of the products is not available in this study, but plays an important role when analysing the whole traceability chain. Theoretically, it should be possible to track a single topped product back to its catching ground, but because of at least two unavoidable reasons, this is not possible here; this owing to; 1) catches from different grounds are mixed on board the fishing vessel, and 2) during off-loading, a further mixing takes place, because fish from different holds are mixed. A third problem might be that the continuous processing which means the fish from different vessels can be mixed. This is, however, more a theoretical problem than a problem in practice since only one vessel arrives at a time. The problem can be eliminated by using all herring belonging to one vessel before off-loading herring from the next vessel. During the production, the catch will be split up in different herring sizes, different cuts (e.g. butterflies and fillets without skin), different marinating procedures, and at last different batches (topping) when packing for marketing. This means that it is possible to track a specific batch back to the specific trip. The TRU, for forward traceability, will then be a batch. A very special case will be when all catches from one cruise are from the same catching ground, and thereby link and track a batch back to catching ground.
In our case, this means that the smallest TRU, for backward traceability (origin of unit/fish), is the trip. For forward traceability (depart of unit/fish), batch will be the smallest size of the TRU.

3.3 Data presentation

The different registrations included in the data analysis, obtained during production of the marinated herring, are listed in Table 3.1.

Table 3.1. List of data included in the data analysis of data from a herring industry.

Place of registration	Registration	Type of registration
Fishing vessel	Name (e-number)	Fixed value connected to the vessel
	SW-code	CSW or RSW
	Start date of trip	dd-mm-yyyy
	Place of catch	
	Date of catch	dd-mm-yyyy
Harbour	Off-loading date	dd-mm-yyyy
	Production date	dd-mm-yyyy
	Gross amount	kg
Laboratory[1]	Temperature	°C
	Counting sample	Units herring per kg
	Quality mark	Number
	Nematodes	Number per 5 kg herring
Filleting	Counting sample[2]	Gram (average over app. 50 pieces)
	Fat content[3]	%
Marinating	Acid commodity group	Digit code
	Date of salting	dd-mm-yyyy
	Percentage NaCl in the brine	%
	Marinating code	Digit code
	Date of marinating	dd-mm-yyyy
	Percentage NaCl in the marinade	%
	Percentage acid in the marinade	%
	pH in the marinade	Number
	Amount of fresh herring	kg
	Produced amount	kg
	Difference between fresh and produced amount	kg
	Appearance	Very nice / Normal / Less nice / Deviating / Bad
	Consistency	Good / Normal / Bad
	Smell/taste	Okay / Not okay
	Homogeneity	Yes / No
	Specifications kept	Yes / No
	Counting sample	Gram (average over app. 50 pieces product)
	Dispersion of counting sample	+/- a value
Topping	Date of topping	dd-mm-yyyy
	Non topped amount	kg
	Topped amount	kg
	Yellow, top	Non / Few / Some / Many
	Yellow bottom	Non / Few / Some / Many
	Loose tail	Non / Few / Some / Many
	Badly cut	Non / Few / Some / Many
	To little marinade	Non / Few / Some / Many
	Bad smell	Non / Few / Some / Many
	Foreign bodies	Non / Few / Some / Many
	Other things	Non / Few / Some / Many
	General quality	Value from 1 to 5 where 1 is highest quality

[1] Average values for the holds analysed. [2] Average values for the concerned code number. [3] Determined industrially from dry matter content of each product type produced (see page 17).

In the following, some general comments about the obtained data will be given, before a thorough examination of the data in Section 3.3.

Data was chosen on the criteria that traceability exists from start to end throughout production. This means that the starting point of the analysis is a trip, and not a towing, due to mixing of the herring from different towing onboard, and when off-loading at harbour. Not all the herring arrive at the industry by vessel, some arrive by truck. In the case of arrival by truck, not all the information before production date is obtainable and will be treated as missing values in the following data analysis.

The data base contains more information than is included in the data analysis, primary information from the vessels about e.g. hold temperature and duration of towing. But again, due to lack of traceability caused by mixing of herring from the different towing, this information is not included at this time. In addition, much of the information that should be obtained on board the vessels is very sparse, and therefore not suitable for analysis. Information about gear type is excluded since all fishing vessels with permanent relation to the industry use trawl.

The information about place of catch is imprecise, for at least two reasons. First of all, herring from different towing are mixed, as also discussed earlier, and thereby loosing traceability to exact place of catch, and secondly, the existence of different ways to specify the place of catch. It is assumed that the place of catch is an important factor in relation to the fish quality. Therefore, the place of catch is included in the data analysis for those trips where all towing are obtained within the same International Council for the Exploration of the Sea (ICES) area. By choosing ICES area, it is possible to place most of the information about catch areas registered. This was done based on maps. The ICES area division of fishing grounds covers relatively great areas, and the position of place of catch is therefore not very specific. Furthermore, it is a well know fact that due to hard competition the fishermen are not interested in coming up with precise details about the area of catch.

To make sure that traceability is present from the quality evaluation of the raw material in laboratory to end product as marinated herring, it is necessary to use a mean value of the quality character, obtained from the holds tested.

Some registrations are not directly usable in the data analysis, but can be used for calculations which then can be included in the data analysis. This holds for the many registrations of date:

- A very rough estimate for storage time onboard can be calculated as the difference between the date of the first catch and the date of off-landing.
- Duration of pre-salting can be calculated as the difference between date of pre-salting and date of marinating.
- Storage time after marinating can be calculated as the difference between the date of marinating and the date of topping.

Moreover, the production date can tell something about the effect of season (month) and year. The reason why the production date is used, and not date of catch, to evaluate the effect of season or year is that when the herring arrives to the industry by truck the date of catch cannot be obtained.

Loss during marinating of the herring can theoretically be calculated in two ways; 1) as the difference between the counting samples of the fillet products and the counting samples of the marinated products, or 2) as the difference between the un-topped amount marinated herring and the topped amount marinated herring. In practice it turned out that both methods were associated with a large degree of uncertainty.
In method 1, because the counting samples were based on spot tests. Concerning the historically data no direct link was obtainable between the product line of the fillet products and the marinated products. In method 2, because the waste amount during marinating and topping was not registered. This makes it impossible to distinguish the origin of loss, whether it is due to waste or due to marinating.

Since logging of historically data stopped mid 2003, some information about marinating and topping conducted after this date was missing.

3.4 Data analysis

The data analysis starts by univariately scrutinizing the data. This is done to get knowledge about the quality of the data and thereby clarify the relevance to the product quality, before continuing with the multivariate data analysis.

3.4.1 Overview of trips included in the data analysis

Data from 471 trips were included in the data analysis; this includes herring delivered by truck. The landing activities were highest in August, with a smaller decline in September, October and November. December showed very little delivery, followed by a smaller increase in January, February and March. A substantial decrease was seen in April, after which deliveries fully stopped in May and June before increasing again in July, see Figure 3.5.

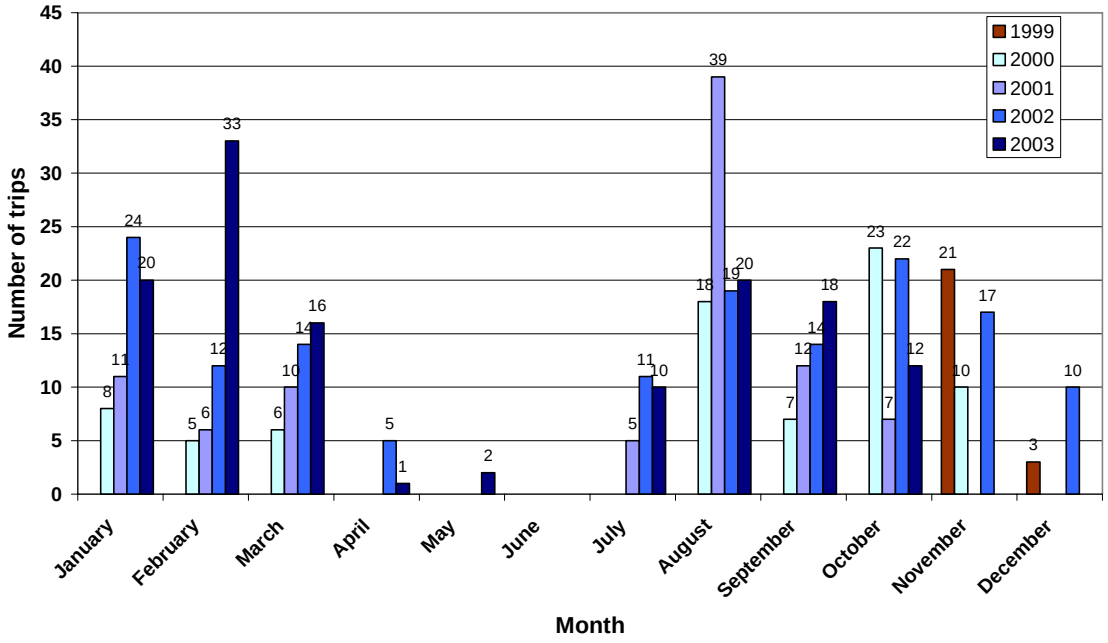

Figure 3.5. Numbers of herring landings in Denmark from mid 1999 to mid 2003 split up on month.

The number of landings each year is listed in Table 3.2. The number of landings was not directly comparable since the data base only holds information about trips conducted with vessels that were associated with the company, from mid 1999 until 2002. From 2002 to mid 2003 data also includes herring delivered by truck or other unassociated vessels. Furthermore, concerning 1999 and 2003, logging of data did not take place for the whole year.

Table 3.2. Number of landings per year of herring in Denmark.

Year	Number of trips
1999	24*
2000	77
2001	90
2002	148
2003	132*

* The period logged does not cover a whole year.

From the data included in this study it should be possible to both analyse the effect of season and year to year variation on raw material and product quality, since data are well represented throughout the year, and data from three whole years was included.

3.4.2 Data obtained in connection with the trip

For 183 trips, it was possible to identify the place of catch by ICES areas. The primary places of catching ground were 4A: north North Sea (72 trips) and 4B: central North Sea (48 trips) followed by 3A: Kattegat/Skagerrak (27 trips) and 2A: Norskehavet (21 trips). Herring caught in the North Sea are primarily winter and autumn spawning, whereas herring from Kattegat and Skagerrak is spring spawning (Jensen 1949; Rosenberg & Palmén 1982; Slotte 1998; Johannessen & Jørgensen 1990). Catching ground might indirectly be important for the quality, due to mixing of herring stocks, resulting in different biochemical and functional properties of the raw material. For herring spawning in autumn, their fat content will increase rapidly during the early part of the summer, reach a maximum fat content in late summer and deplete strongly during spawning time. The fat content can vary between 1 and 30 % during the year. Furthermore, a phenomenon such as off-flavours can often be related to fishing ground, since certain localities evidently relates to variations in flavour (Karl & Münkner, 2002). Several of these off-flavours can be attributed to the feeding on different compounds or organisms e.g. the larvae of *Mytilus* spp. which causes a bitter taste in herring. Marine algae, sponges and Bryozoa forms volatile bromophenolic compounds which causes an iodine-like flavour. An oil taint might be found in the fish flesh in areas with off-shore activities, or in areas with large oil spills (Huss, 1995). When investigating the effect of fishing ground on the sensory quality, i.e. appearance, odour, flavour and texture of marinated herring products, Nielsen et al. (2003) found no differences in sensory quality, which could be ascribed to fishing ground. However, a more detailed and uniform specification of fishing ground might still be useful considering traceability and valuable in cases with off-flavours or pollution.

The storage time onboard the vessel can roughly be estimated for 231 trips. The involved trips are evenly distributed both during the year and between years, and no

systematic effect between years is observed (results not shown). From Figure 3.6, it can be seen that the storage time on board approximately follows a normal distribution with 3 days as mean value. Storage values calculated as 11 and 32 days are regarded as outliers and should be kept out of further analysis. The maximal shelf life of herring on ice is 2 to 12 days (Hansen et al., 1970; Kolakowska et al., 1992), depending on the fat content and enzymatic activity. Herring with low fat content and low enzymatic activity (winter herring) have longer shelf-life than fat and feeding herring (summer herring). The effect of time / temperature storage conditions on product shelf-life has shown to be cumulative (Charm et al., 1972). Findings show that maintaining a continuous monitoring, and control of the storage temperature and keeping the fishing trips as short as possible is crucial. Deterioration due to enzymatic activity is a risk, since the herring are stored un-gutted, but the primary reason to spoilage of fatty fish, as in the case of herring, is due to oxidation. The duration of the trip should therefore be as short as possible. Furthermore, fast cooling of the catch and a constant low temperature should be kept to maintain appropriate quality. Unfortunately, the temperature measurements from the vessels in this study were very sparse and not suited for inclusion in the analysis. A suggestion for continuous temperature control during storage on board would therefore be the use of automatic temperature loggers. Needless to say, some practical conditions about placing should be considered before implementing by reason of heterogeneous temperatures in the hold.

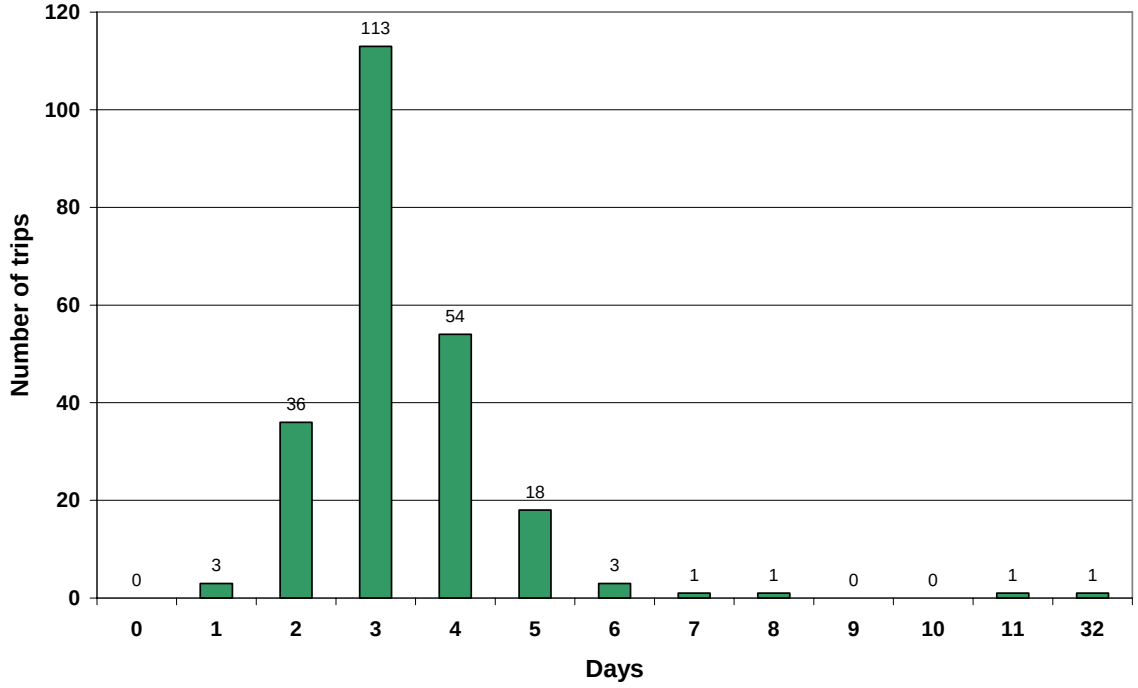

Figure 3.6. The storage time (days) of herring on board the vessel calculated for 231 trips.

The gross amounts (kg) for 343 trips are available. The mean value is 187 233 kg + / - 72 767 kg, the minimum value is 18 965 kg and the maximum value is 361 050 kg. The wide range between minimum and maximum amount is owing to the different holding capacities between RSW and CSW vessels. RSW vessels have larger holding capacities (615 + / - 150 m^3) than CSW vessels (324 + /- 60 m^3). The mean values, when dividing up in samples from RSW and CSW vessels, was 226 127 kg + / - 59 002 kg and 134 329 kg + / - 53159 kg, respectively.

3.4.3 Quality evaluation of the raw material

Results, given as mean values from the quality evaluation of the raw material, are listed in Table 3.3. All values are average values covering the number of holds tested from one vessel.

Table 3.3. Minimum, maximum, mean and standard deviation for determination of: nematodes (pieces per 5 kg.), temperature (°C), counting sample (pieces of whole herring per kg) and quality mark when evaluating the herring in the laboratory.

Measurement	Minimum value	Maximum value	Mean value	Standard deviation	Number of samples included
Nematodes (pieces per 5 kg.)	0	34	9.6	6.0	443
Temperature (°C)	-2.10	13	-0.43	1.18	462
- CSW*	-2.1	4.5	-0.08	1.10	163
- RSW*	-1.8	1.7	-0.95	0.49	226
Counting sample (pieces of whole herring per kg)	0	27	6.6	1.8	431
Quality mark	0	9.7	8.0	1.0	459

*The samples landed with vessels with association to the company.

The nematodes are in the range 0 to 34 pieces per 5 kg. Only 12 determinations out of 443 have values over 22 nematodes per 5 kg, 41 determinations have values between 17 and 22 nematodes per 5 kg, whereas the rest of the determinations (390) are evenly distributed, with values between 0 and 17 nematodes per 5 kg. *Anisakis larvae* are found almost ubiquitously in the intestines of herring from Nordatlanten, Skagerrak and Kattegat (Jessen, 1987). The herring most commonly get infected with *Anisakis larvae* during feeding with krill (Podolska and Horbowy, 2003). The larvae are typically found in the intestine, but can migrate to the flesh. Therefore they make up a possible infection risk in human consumption if not killed during the marinating process (Jessen, 1987). The occurrence of nematodes is highest during the spawning period and increases by age (Karl & Münkner, 2002; Podolska & Horbowy, 2003). The time estimated to kill *Anisakis larvae* in marinated herring products topped in a marinade of 5 % acetic acid and 10 % NaCl is 35 days (Karl et al., 1995). This estimate is coherent with the customary marinating time seen in Danish herring industries. The products included in this study were all stored for at least 35 days (results not shown). The recommendation for ensuring the inactivation of nematodes in fat herring, includes rolling of the barrels

at regular intervals to avoid a concentration gradient within the barrels (Karl et al., 1995).

Considering temperature determination in the herring, a value of 13 °C or above is a mistake since such high temperatures are unrealistic when working with fresh herring stored either in RSW or CSW. Observations with such values should be excluded from further multivariate data analysis. Temperatures measured that high were either due to wrong typing or to long storage time without chilling in the laboratory, before measuring. For all samples the mean value is -0.43 °C +/- 1.18 °C. The effect of cooling system used on board the vessels is reflected in the temperature values measured in herring from 226 RSW and 163 CSW fishing vessels. Herring cooled with RSW had a lower temperature than herring cooled with CSW, the mean values are -0.95 °C and -0.08 °C, respectively. The temperature interval is wider for vessels using CSW (-2.1 °C to 4.5 °C) than vessels using RSW (-1.8 °C to 1.7 °C) and more samples from CSW vessels have measured higher temperatures, see Figure 3.7. Studies from both Smith et al. (1980) and Hattula et al. (2002) shows, that the effects on quality from storage in CSW and RSW are similar when the temperature is kept low (app. 0 ºC). In both situations, off-flavours will develop in the herring, if the seawater is not renewed (Smith et al., 1980).

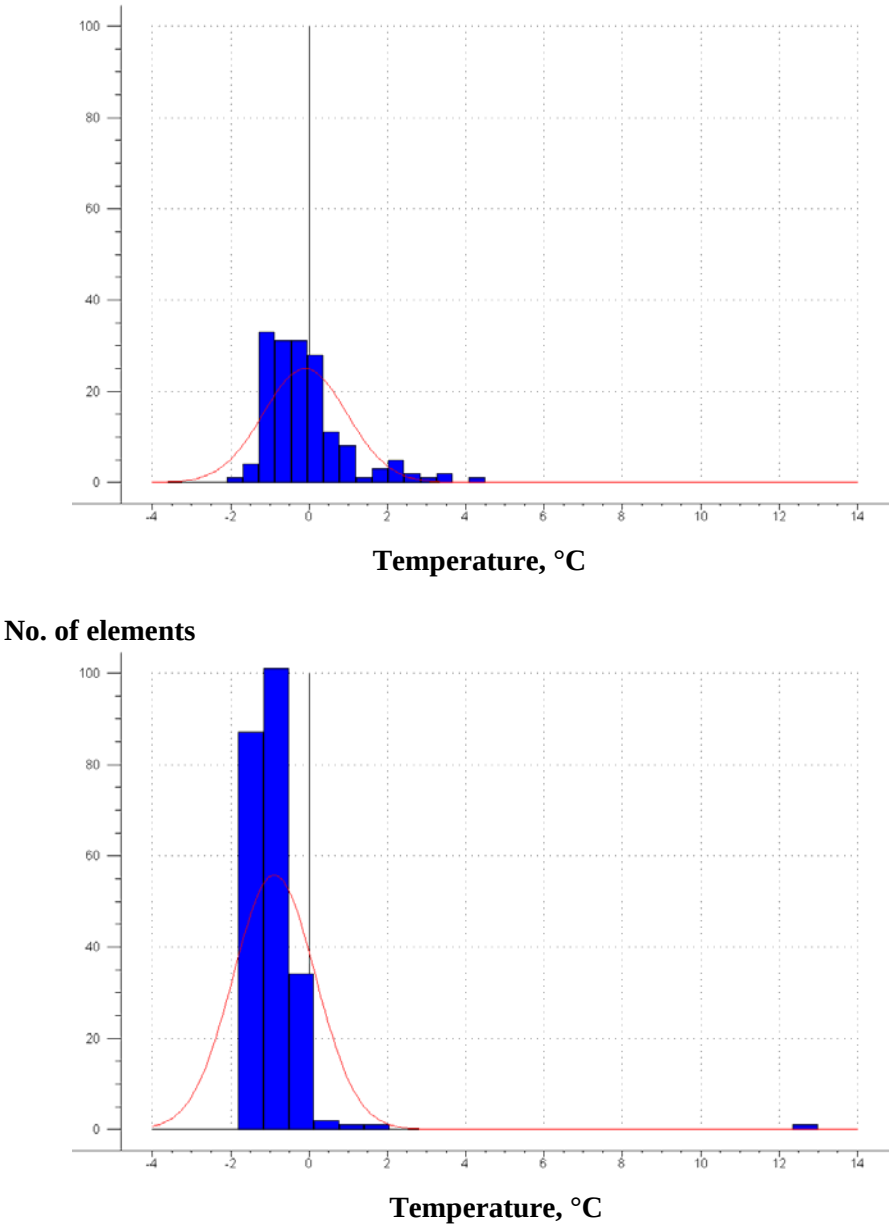

Figure 3.7. Histogram plots of the temperature measured in the whole herring cooled with CSW (upper) and RSW (lower) onboard the vessel.

Counting samples (pieces herring per kg) for 431 samples were included in the analysis. The counting samples followed a normal distribution with mean value around 6.6 herring per kg, see Figure 3.8. The samples marked by a circle in the figure are outliers as counting of samples that holds 1 and 2 herring per kg as well as 27 herring per kg are unrealistic, and should be excluded from the dataset before further analysis. The size has to match with the corresponding product. For a predefined body weight (giving fillets weighing above 25 g) Nielsen et al. (2003) found an effect of body weight on the

sensory quality of marinated herring products. An increase in body weight was accompanied by an increase in the quality parameters: firmness, juiciness and elasticity and a decrease in gritty texture in products produced immediately post-mortem.

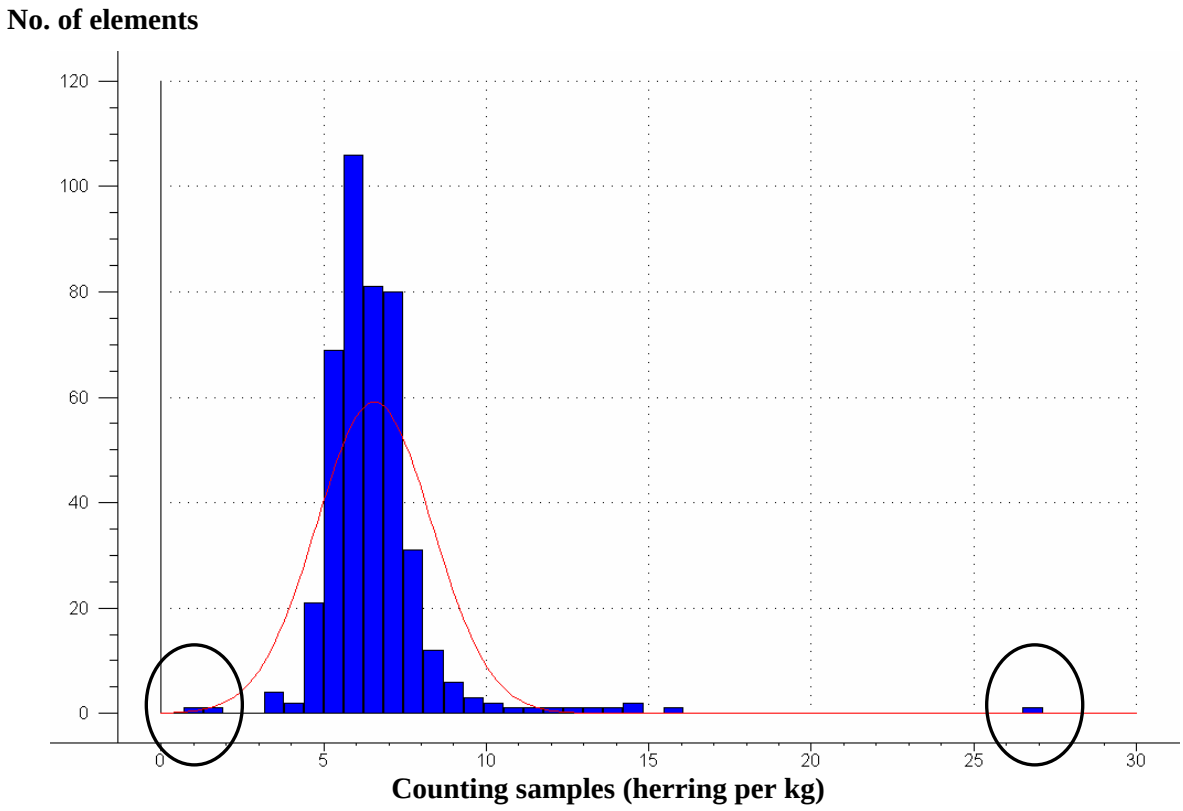

Figure 3.8. Histogram plots of the counting samples of whole herring obtained from the quality determinations of the raw material. Outlying values are marked with a circle.

Figure 3.9 show a histogram plot of the quality mark of the raw material. The quality marks for the 459 determinations follow a normal distribution, with a mean value around 8.0. The highest obtainable value is 10. The sample with a value of 0 was an outlier and consequently excluded from the data set. Quality marks below 4 should not appear in practice, since such low values reflect a very poor quality, not acceptable for further production (Michaelsen K, personal communication). The quality marks reflect variation in the data set, even though it was not possible to relate quality to season.

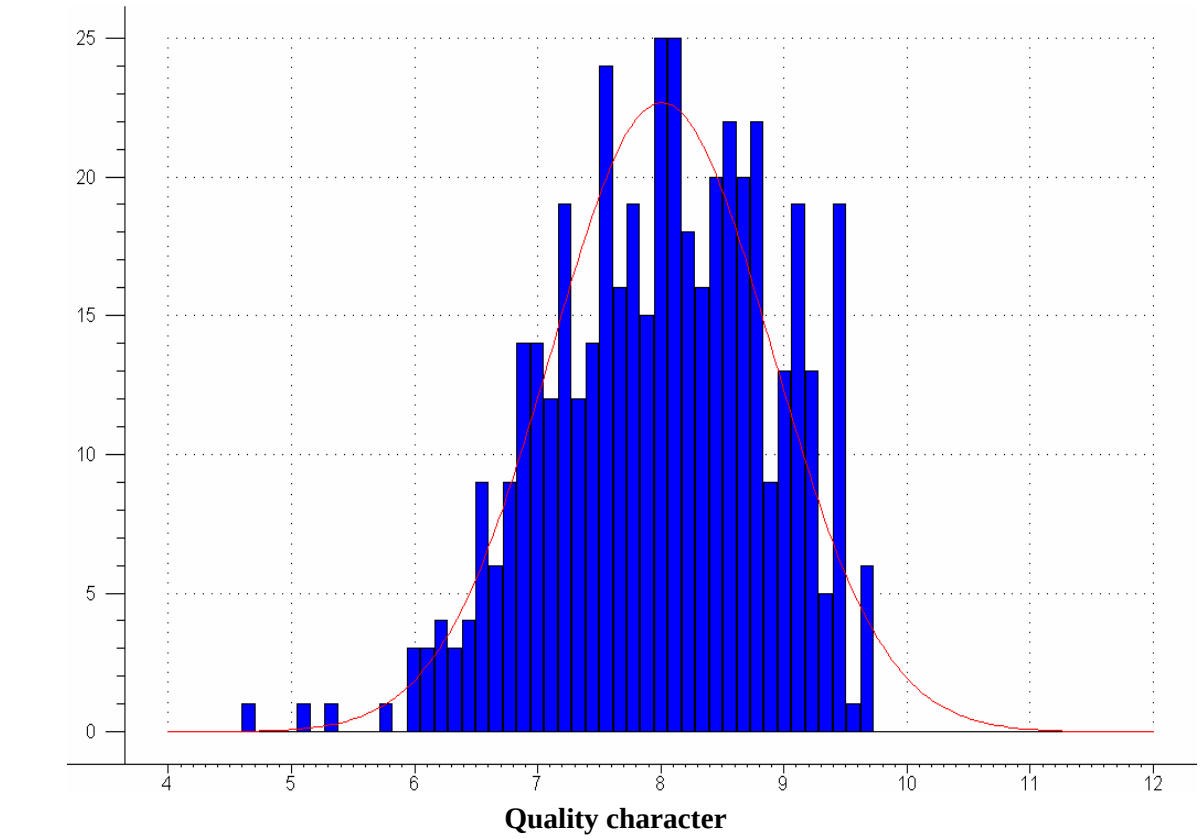

Figure 3.9. Histogram plots of the quality marks obtained from the quality determinations of the raw material.

The calculated fat content is primarily determined for fillets without skin and butterflies. Owing to the procedure for fat determination, where 10 gram of the actual product type was minced and dried, only one calculated fat determination exists for each product type (fillets without skin and butterflies). This value for e.g. butterflies then represents the fat content in all butterfly products, produced from that specific cruise. The calculated fat determinations as function of production date are plotted in Figure 3.10 for 288 samples of fillets without skin and 382 samples of butterflies. The calculated fat content varies according to season, with the highest values around August and lowest values around March. The variation in calculated fat content is in accordance with feed availability and follows the cycle of maturation. The fat content increases from juvenility to mature herring. Furthermore, fat content decreases rapidly during spawning, followed by a subsequent increase after spawning (Iles, 1964). A broad variation of calculated fat

content is also observed within the same month, indicating that the raw material is very heterogeneous, among others, due to the different catching grounds. A substantially part of the fat depots are located in the subcutaneous tissue, explaining the generally higher fat content in butterflies, compared to fillets without skin, given that a part of the fat is removed with skin.

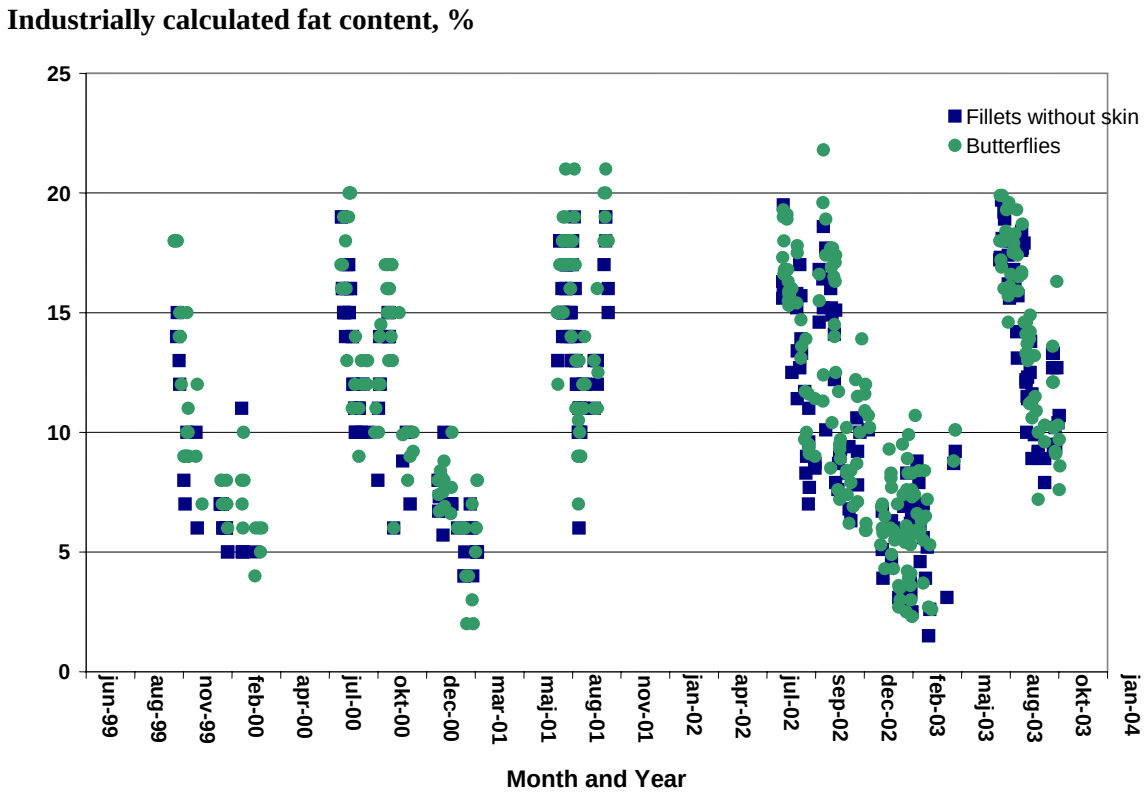

Figure 3.10. Calculated fat content coloured by product type (fillets without skin and butterflies) versus production date.

Fat determinations on single fish level by Bligh and Dyer extraction from approximately 50 herring from 4 trips included in the analysis, shows great variation within a trip, see Table 3.4. These findings are in accordance with results obtained by Larsen et al. (1997), showing large variations in fat content within same catches conducted by commercial vessels in the North Sea. The fat content is a very important quality parameter in view of several herring products, including marinated herring manufactures, where a fat content of minimum 8 % is desirable (Karl & Münkner, 2002). Nielsen et al. (2003) found that the fat content had a very clear influence of the

sensory properties of the marinated herring. High lipid content results in fillets with higher intensities of the characteristic herring odour and flavour. Furthermore, they were juicier and gave a more fatty mouth feel than fillets from leaner herring. Lean herring had higher intensities of sweet odour and flavour, were firmer and had a higher intensity of gritty texture. The results illustrate that the way the fat content is calculated today in the industry, do not reflect the great variation in fat content within a catch. On-line measuring of fat content and subsequently sorting will provide more homogeneous products according to fat content, thus improve process optimization.
Due to the relatively imprecise fat values in the data, analysis that includes fat content as calculated today will be associated with uncertainty.

Table 3.4. Fat content (%) determined on single fish level by Bligh and Dyer extraction and calculated on batch level based on dry matter determinations.

Time	Fat content, %					
	Single level (Research)*					Batch level (Industry)**
	Mean	SD	Median	No. of samples	Range	
May	6.93	1.82	6.65	48	3.99 – 13.08	8.8
September	10.09	2.99	9.97	57	3.85 – 17.39	11.7
November	6.49	2.72	6.19	50	1.40 – 16.51	7.9
February	4.48	2.04	3.46	50	2.01 – 12.45	3.0

* Bligh & Dyer extraction, ** Dry matter determination (see page 17).

3.4.4 Evaluation of the marinated products

Data obtained during the marinating process included data from 1351 products; 1162 products had been pre-salted, 119 products were marinated directly, while the information about pre-salting was missing for the last 70 products. Pre-salting improves the strength of the fillets and leach blood and other impurities (Jessen, 1987). The duration of the pre-salting depends on the pre-salting process, which again is dependent on the fat-content in the herring. In a study by Birkeland et al. (2005) the effect of different brine conditions (NaCl concentration: 10.0 %, 16.5 % and 25.5 %; storage temperature: 3.5 °C and 17.5 °C; skin-on versus skin-off) on weight gain during storage

were investigated. It was shown that the weight gain in herring fillets increases during brining. At storage temperature at 17.5 °C equilibrium between the brine and the interior muscle tissue of the herring fillets was reached after 1 to 2 days. For storage temperature at 3.5 °C this equilibrium was not reached after 7 days storage causing an influx of salt and water to the fillets. In general, the highest weight gains were obtained for brines with 10.0 % NaCl and fillets without skin. The average pre-salting time in this study were 1.2 days at 5 °C.

During the marinating process, spot tests were taken to evaluate the product quality. The parameter "Smell / taste" is evaluated by "Ok" or "Not ok". All 1228 samples evaluated were evaluated as "Ok". This parameter can then be excluded from the following data analysis, since it does not tell anything about the product. The results from the other evaluations "Appearance", "Consistency", "Homogenity" and "Specifications kept" are presented in Figure 3.11.

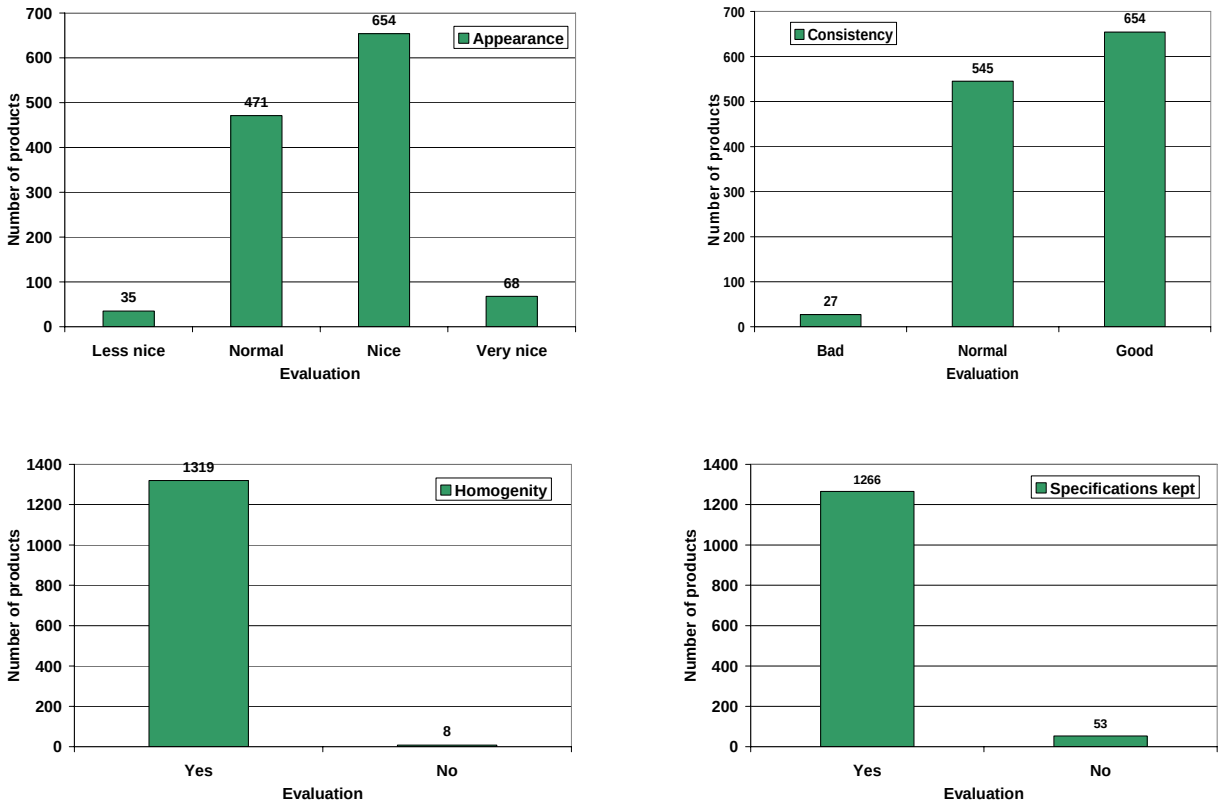

Figure 3.11. Evaluation (Appearance, Consistency, Homogeneity and Specifications kept) of the results from the spot test obtained during storage of the marinated herring products.

Concerning “Appearance”, the primary part of the products was judged as “Normal” (38.4 %) or “Nice” (53.3 %), whereas “Very nice” and “Less nice” were only used for a minor part of the products, 5.5 % and 2.9 % respectively.

The results obtained from the evaluation of “Consistency” reflects that something is wrong with the scale, since most of the products (53.3 %), obtain the best evaluation “Good”, while only 2.2 % of the products was judged as “Bad”. The remaining part of the products was evaluated as “Normal”. With a more accurate scale it would be expected that most of the products would be evaluated by the mean value, as “Good”. Furthermore, it seems like the difference between “Bad” and “Normal” was bigger than the difference between “Normal” and “Good” – the scale was not used equally for the different characters.

Only a minor variation was observed in the products, with respect to “Homogeneity” and “Specifications kept”. Out of 1327 products evaluated, only 8 were evaluated as “No” with respect to “Homogeneity”, and out of 1319 products evaluated for “Specifications kept” only 53 were evaluated as “No”. Moreover, both of these parameters were more process dependent than depending on the actual quality of the raw material. These evaluations were therefore not relevant for further data analysis, when analysing the effect of raw material quality on the final product quality.

The results from the evaluation of the marinated products showed very sparse variability in the parameters “Smell / Taste”, “Homogeneity” and “Specifications kept”. For that reason these parameters were not suitable for further multivariate data analysis. The evaluation of the parameter “Consistency”, indicated that the scale should be redefined. Only the parameter “Appearance” seemed to be suitable for further analysis, with that in mind that the results would be based on spot tests, and therefore conducted with some degree of uncertainty.

3.4.5 Quality determination at topping

During the marinating process the fillets lose weight due to removal of water from the flesh caused by coagulation of proteins induced by the salt in the marinade (Somers, 1975). In general, the weight loss is around 20 % of the weight depending on the fish quality (Herborg, 1978). The loss during the marinating process increases with decreasing fat content (Jessen, 1987). In practice, also a major weight loss is observed for fat summer herring. The fat "melts off" the herring, and drift to the top of the barrels. An explanation for this "melting off" is that the fat in the fat summer herring are not incorporated into fish muscle as it is primarily stored subcutaneous. Additionally, the storage time also influences the weight loss to a certain limit: the longer storage, the higher weight loss. Theoretically, the loss during marinating can in our case be calculated in two ways as described in section 3.3 (page 24); 1) as the difference between the counting samples of the fillet products and the counting samples of the marinated products, or 2) as the difference between the un-topped amount and the topped amount. Also described in section 3.3 (page 24), it turned out that both methods were connected with large uncertainty. Because, in method 1 no direct link existed between the product line of the specific fillet products, and the marinated products. This means that the counting sample used for the fillet products, is a mean value of all the counted samples conducted for that specific product type (e.g. fillets without skin and butterflies), and do not account for the different sizes of the herring. Method 2, because the waste amount during marinating and topping is not registered. That made it impossible to distinguish between losses, due to waste or marinating. An improved system to trace the source of the fillet product is necessary to connect counting samples of fillet products with the counting samples of marinated products. In addition, registration in relation to the amount wasted, needed to be introduced.

The topped products are evaluated according to "Yellow, top", "Yellow, bottom", "Loose tail", "Badly cut", "Too little marinade", "Bad smell", "Foreign bodies" and "Other things". The evaluation was differentiated into "No", "Some" and "Many". The obtained results for the evaluated products are illustrated in Figure 3.12. The variation

in data was very sparse; some parameters such as “Too little marinade” and “Bad smell” were almost not used, and therefore not suitable for further multivariate data analysis.

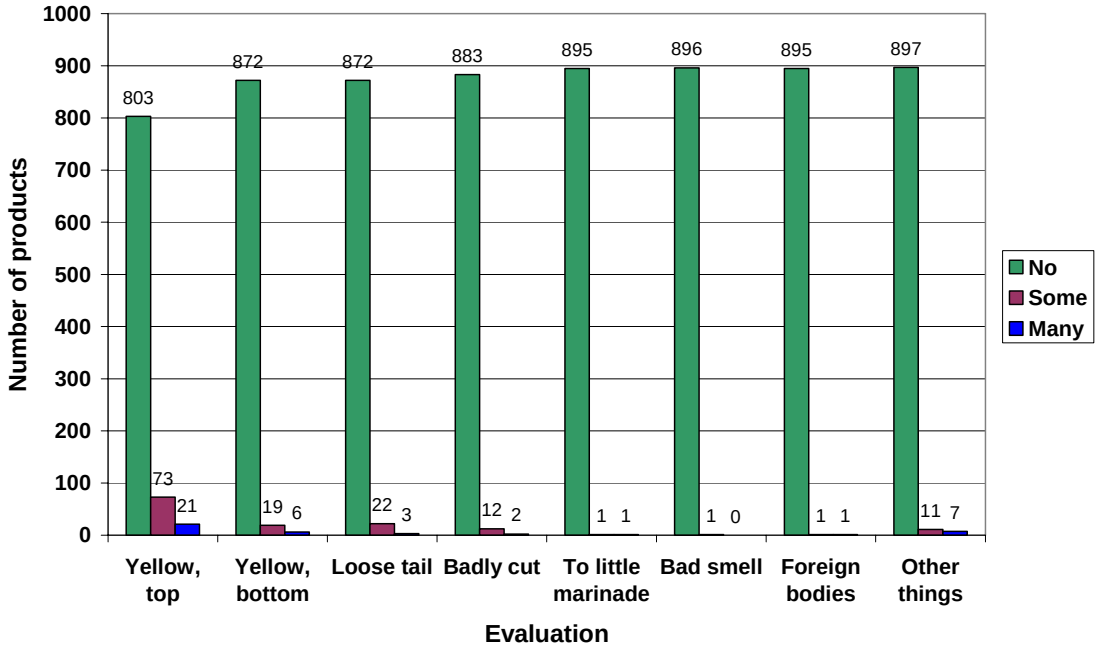

Figure 3.12. The results obtained from the evaluation of the topped herring products after marinating.

The results from the general quality assessment of the final products are presented in Figure 3.13. It clearly appears that there was almost no variation within this parameter. Out of the 890 evaluated products, only 21 products were evaluated as lower quality, the rest of the products were evaluated as being of best quality. This indicates that the evaluating procedure was not optimal and / or that the final product quality was independent of the quality of the raw material. Both scenarios seem to be right: The quality range in products evaluated as being of the best quality is much broader than in the other groups (Michaelsen, K. personal communication). A study from Nielsen et al (2003) has shown that when herring are processed immediately *post mortem,* then the variation in the products is so little that the consumers mostly will not notice it. They concluded that this might either be caused by the fact that no differences in the products or also the acetic acid or salt containing brine used for the marinating, mask any

differences. A new method to evaluate the marinated herring products reflecting the relevant quality parameters would be appreciated. A constraint for the method to be successful is that the method should be easy and fast to carry out for one person, and that the method is independent of the person doing it.

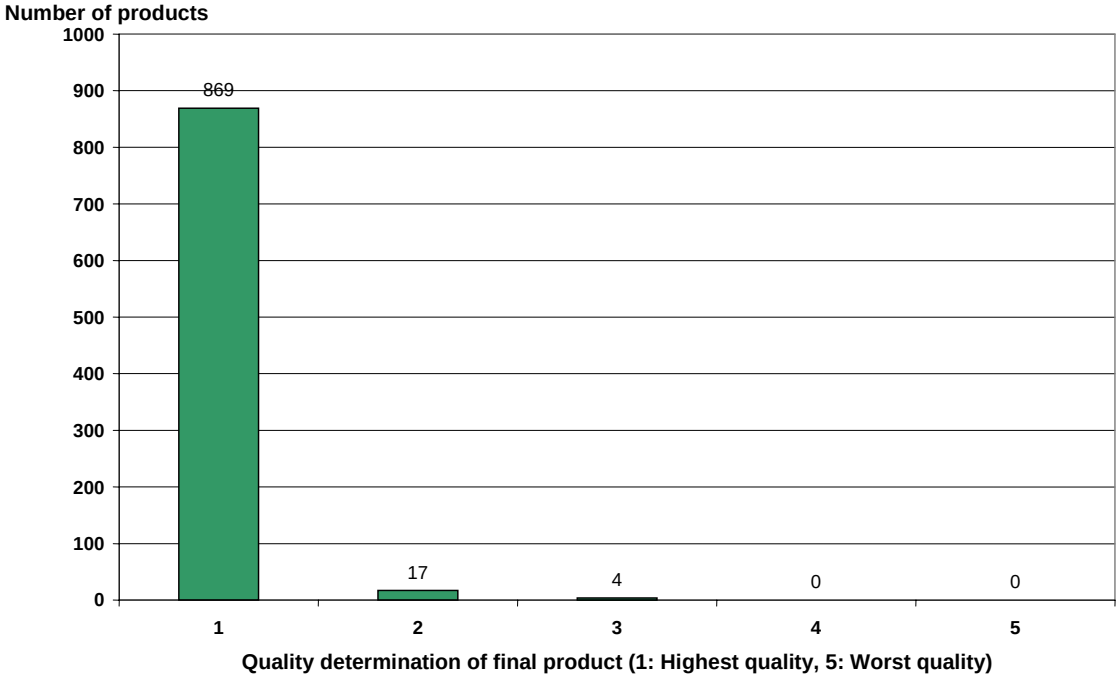

Figure 3.13. The results obtained from the general assessment of the final marinated products.

The distribution of the deviating products (products with quality 2 and 3 in the final quality determination) is illustrated in Table 3.5. The deviating products originate from months with a high production rate (August and September) and when the production was started again after the summer leave. The deviating products did not originate from the same fishing vessel (trips) or marinating batches – other products from the same vessel (trip) or marinating batch obtained the best quality.

Table 3.5. Distribution of the deviating products (products with a lower quality than 1) from the general quality assessment of the final marinated herring products.

	Year		
	2000	2000	2001
Month	Character 2	Character 3	Character 2
August	11	4	4
September			1
October	1		

As a result of very little variation in the quality assessment of the final products, it was not possible to use this parameter in multivariate data analysis. Regardless of the quality of the raw material, the quality of the final product would be acceptable.

3.5 Multivariate data analysis of data from the herring industry

Albeit, the initial screening did not reveal any promising findings for further multivariate data analysis, several attempts to find information in the data were made. In the following some of these results will be presented.

At first, a PCA model on the data related to the raw material was carried out. The variables included were: number of nematodes, counting sample, temperature, quality mark and calculated fat content, to investigate a pattern due to date of catch (month or year), place of catch and/or specific cooling method. As pre-processing all variables were mean centered and scaled to unit standard deviation (autoscaled). The most extreme outliers were initially removed, and the model validated with randomly chosen segments, consisting of 10 samples each. There was no clear break in the variance curve, and the explained variance for a four component model was 88.0 %, compared to 33.8 % for the explained validated model, using four PCs. This low validated variance indicates that the pattern in the data is not very strong.

In Figure 3.14, the score plot of PC2 versus PC1 is shown with samples coloured according to the cooling method. There is a tendency that the samples cooled with RSW

lies to the left in the score plot, while the samples cooled with CSW lies to the right. This is in accordance with the corresponding loading plot (Figure 3.15), which illustrates that samples to the left have a lower temperature than samples to the right – the more the samples appear to the right, the higher the temperature. However, this was also expected since RSW is expected to cool better than CSW. The temperature seems to covary with the quality mark – a low temperature gives a high quality mark, and vice versa, which seems reasonable. Also a covariation is observed between counting sample and quality mark, a high counting sample (small herring) gives a low quality mark. A not so straightforward reasoning is the connection between a high temperature and a high counting sample, and the connection between a high quality mark and high number of nematodes. A high number of nematodes would normally be expected to influence the quality negatively, as nematodes are undesirable. PC1 seems to describe a combination of all the variables except the calculated fat content. PC2 seems to describe the calculated fat content and nematodes. The conclusions drawn from this PC are however in doubt according to the weak model. Samples with a negative score value have a high calculated fat content where as a positive score value indicated a high number of nematodes. This could be explained by the fact that lean fish have more meat where the nematodes are to be found. Neither for PC2 nor PC1 and any other combination of higher order PCs, connections that could link quality mark and place of catch and/or date of catch (neither year nor months) were observed.

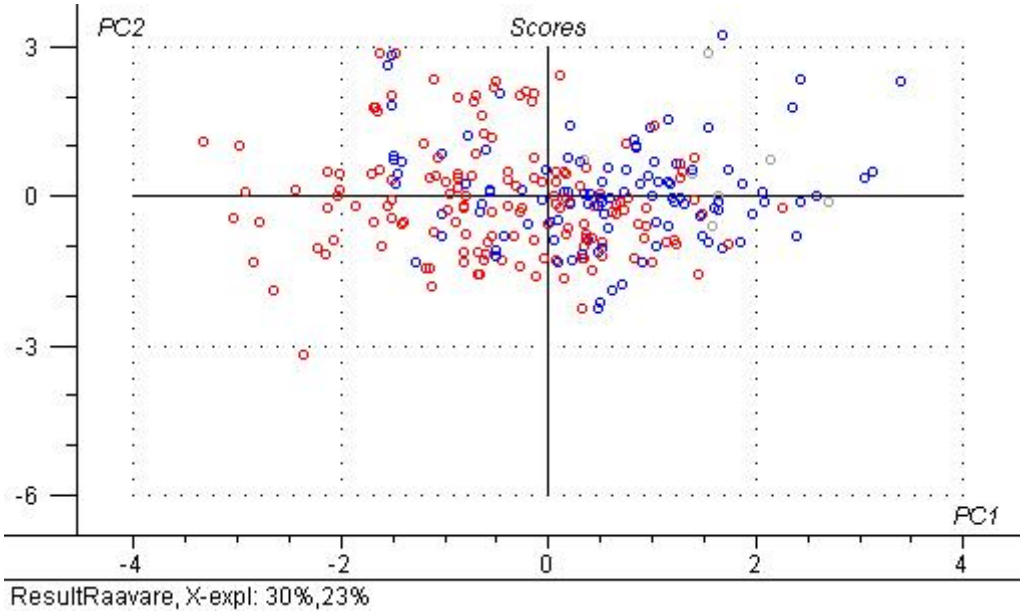

Figure 3.14. PCA scores; PC2 *versus* PC1 from a PCA model of a data matrix related to the raw material. The samples are marked according to chilling method: RSW (Red), CSW (Blue) and unknown (Grey).

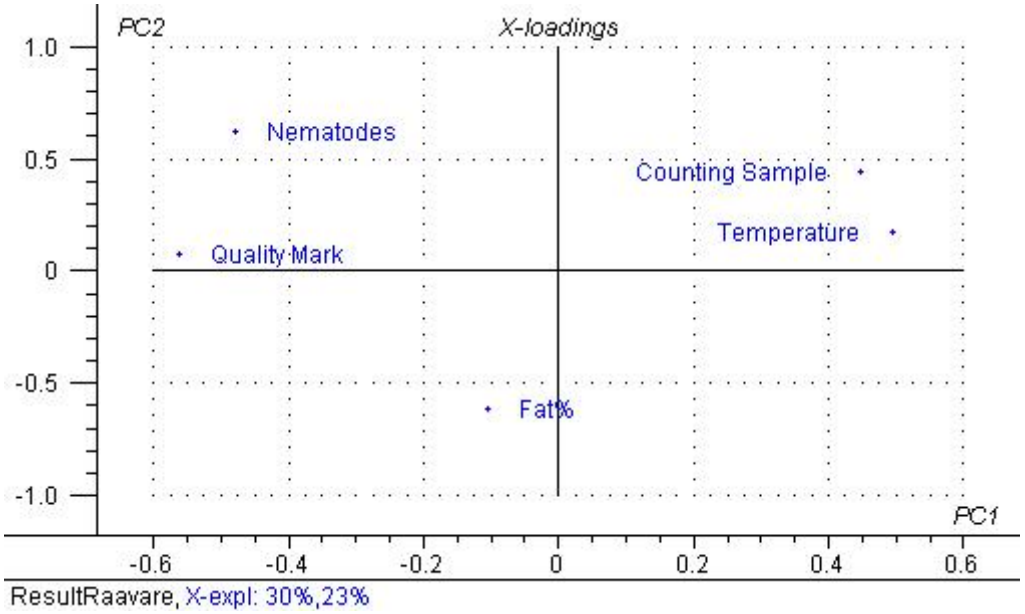

Figure 3.15. PCA loadings; PC2 *versus* PC1 from a PCA model of a data matrix related to the raw material.

To investigate the correlation between raw material properties, in combination with the handling during production (e.g. product type, pre-salting and duration of marinating), and the 'value' of the final quality, a PCA model was conducted. When the variables were expressed by statements such as "Yes" or "No", they were included as binary

numbers (-1/1). As a start, 856 samples were included in the model, but 49 samples were removed caused by outlying properties. This however, only improved the explained variance slightly. Together the first two PCs described 18 % of the explained variance. The plot of PC2 versus PC1, for a PCA model with auto scaled variables, appears as illustrated in Figure 3.16. The samples are marked according to product type. The corresponding loading plot is illustrated in Figure 3.17. A combination of the first and second PC discriminates between the two product types. Butterflies were characterised by higher scores for counting samples, both for the cut and marinated products, and higher acid percentage in the marinade, a finding that can be related to the recipe of the marinade. The opposite was observed for the fillets without skin. All of these parameters were related to the production and process, and did not reflect relations to quality. The two first PCs were also used to indirectly describe the cooling method as RSW vessels have higher capacity than CSW vessels, or when herring arrived by truck (Figure 3.18). The second PC was also used to discriminate between the final product qualities, characterising samples having a lower quality than 1 with negative score values (Figure 3.19). From the loading plot it was not possible to determine which quality parameters that described these samples. What was common for those samples was that they were primarily caught and marinated in august 2000. However, as also described in section 3.4.5, other products from the same marinated batches, obtained the best quality. No other combination of any higher order PCs reflected a correlation between raw material quality and the final product quality. Hence, the PCA supported the initial findings when screening the data that the data at hand did not perform successfully in respect of analysing the influence of the raw material quality on the final product quality.

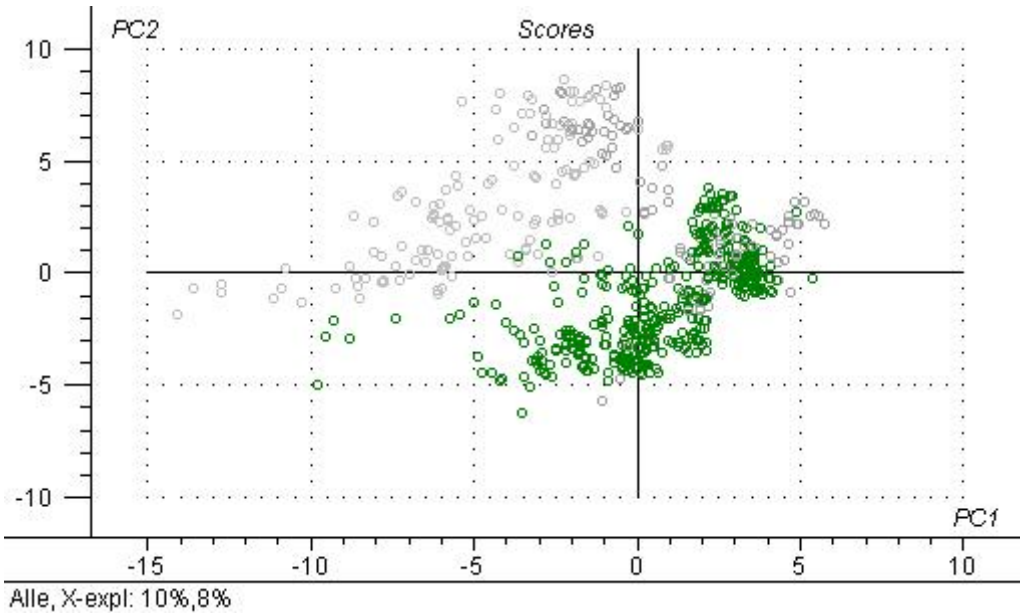

Figure 3.16. PCA scores; PC2 *versus* PC1 from a PCA model of a data matrix related to raw material and the production of marinated herring. The samples are marked according to product type: Butterflies (Grey) and Fillets without skin (Green).

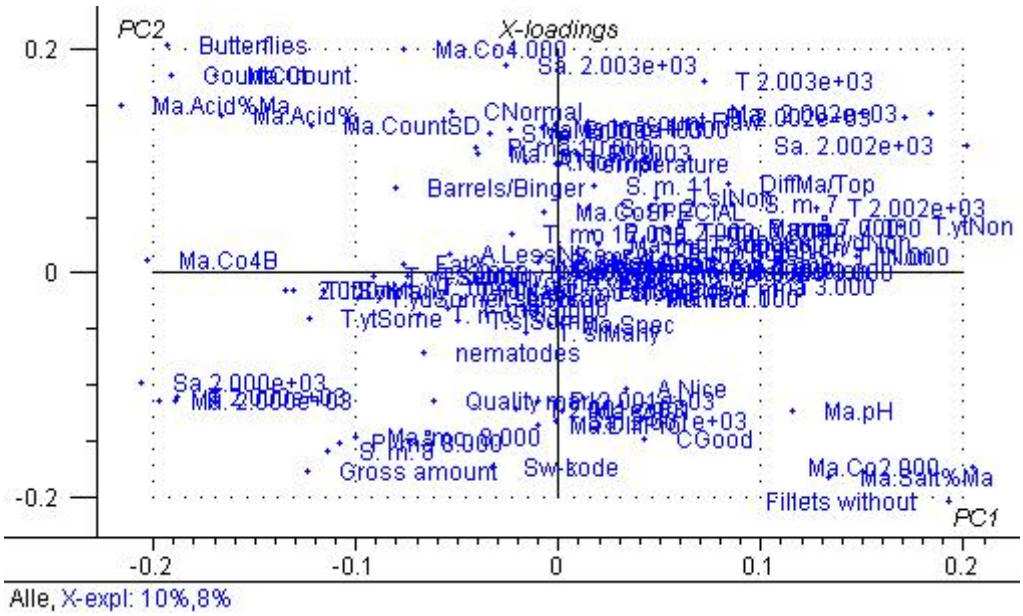

Figure 3.17. PCA loadings; PC2 *versus* PC1 from a PCA model of a data matrix related to raw material and the production of marinated herring.

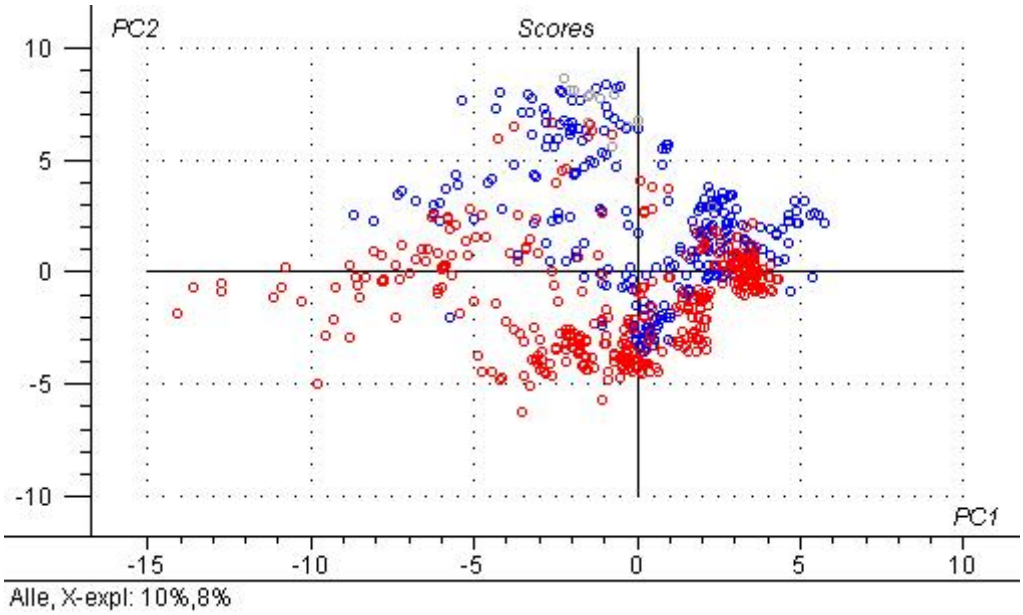

Figure 3.18. PCA scores; PC2 *versus* PC1 from a PCA model of a data matrix related to raw material and the production of marinated herring. The samples are marked according to cooling method: CSW (Red), RSW (Blue) and Missing information (Grey).

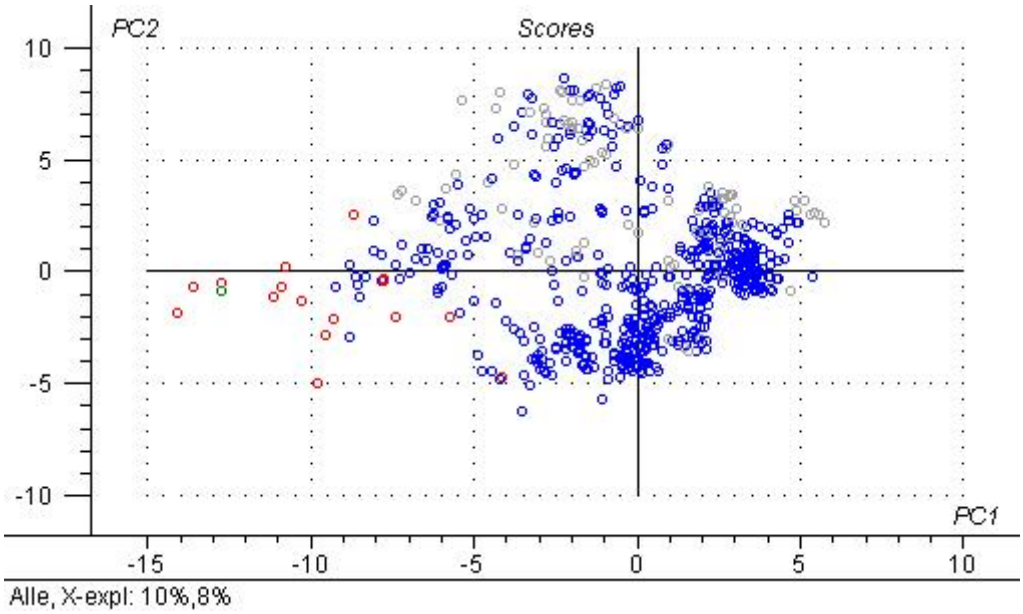

Figure 3.19. PCA scores; PC2 *versus* PC1 from a PCA model of a data matrix related to raw material and the production of marinated herring. The samples are marked according to final product quality: Quality 1 (Blue), Quality 2 (Red), Quality 3 (Green) and Missing information (Grey).

3.6 Additional measurements

The initial screening of the data, resulted in suggestions of some additional measurements/registrations and improvements of already existing measurement/registrations. The suggestions will be listed here and a deeper explanation of some of them follows below:

- Temperature loggers on board the fishing vessels
- Uniform and precise way of specifying the place of catch
- Registration of belly bursting
- Improved traceability between counting samples before and after marinating
- Registration of waste amount during marinating
- On-line fat measurement on single herring level
- Improved quality evaluation of the final product

A uniform and precise way of specifying the place of catch will make it possible to trace the herring to catching ground. This will obviously not solve the problem with mixing on board the vessel and during landing, but in most situations all catches within a trip were from the same area. What turned out to prevent the traceability back to catching ground in this study was that restructuring at the industry cut the belonging between vessels associated with the industry and the industry.
Unfortunately, this also ruined the possibility to improve the registrations obtained on board the vessels and complicated the information transferred between vessel and industry, as they are now two individual companies.

Even though belly bursting is mentioned as a quality problem, related to raw material, the amount of belly bursted herring was not registered in the industry. Belly bursting is related to season and occurs mainly in feeding herring because of high enzymatic activity (Kolakowska et al., 1992). A cell for registration of the amount of belly bursted herring was included in the data base.

The way the fat content was calculated in the industry, as one value for each product type determined on a pooled sample, does not reflect the great variation within fat content in a catch of herring. The fat content is a very important quality parameter in a range of herring products, including marinated herring products for which the desirable fat content is at minimum 8 % (Karl & Münkner, 2002). Introduction of on-line measuring of fat content and subsequent sorting according to fat content will provide a more homogeneous product according to fat content and improved the possibilities for process optimization. In a study by Nielsen et al. (2005), comparing solvent extraction, Torry Fish Fat Meter, NIR and nuclear magnetic resonance (NMR) for fat analysis, the NIR technique showed the highest potential as a production line measurement for fat determination. Such an instrument should meet certain criteria e.g. be fast (at least 5 determinations per second), non-destructive, able to measure on whole herring or fillets and perform stable in a wet and acid environment. To the author's knowledge, an improved instrument as such is not, for the time being, available to the herring industry. The loss during marinating was a very important parameter, especially in consideration of product optimization and economics. A registration system was implemented to improve the traceability between counting samples of the fillet products and counting samples of the marinated products. Furthermore, cells for registration of the amount of waste (kg) and the reason for waste were included in the data base. Future on it should then be possible to calculate the loss during marinating, caused by the marinating process, and relate this to the information obtained on the raw material.

An improved method for quality determination of the final product reflecting the actual differences is hardly needed. The method needed to be fast and easy to carry out to ensure optimal success. According to the industry they have not found a better method yet to replace the method included in this study.

3.7 Concluding remarks

The data analysis indicated that the historical data were not suitable for further multivariate data analysis, by reason of lack of variability and / or lack of traceability on

the needed level in a range of essential measurements / registrations, such as calculated fat content and final product quality. This is not unique for historical data, since this sort of data are often obtained for other reasons than the objectives of the present study. In this study, many of the historical data reflected quality related to the process e.g. cutting procedure and marinating procedure, rather than quality related to the raw material. In addition, the method for final product quality determination does not reflect the variation in the products. Therefore it may not be relevant and / or representative for the ongoing purpose, which is to relate raw material quality to the final product quality, to continue with these data.

On the other hand, the historically data can be used to point out which types of measurements are missing and which need to be improved, to be informative in the sense of process control and process optimization within the herring industry.

Now, the main part of data logged will automatically be saved into the data base, and thereby reducing the uncertainty related to converting written registrations on paper typed into the database, as was the case for the historically data.

4.0 Applications of robust multivariate methods

Outliers are observations that appear to break the pattern or grouping shown by a majority of observations. Presence of outliers is more the rule than the exception when working with experimental data with many observations and / or variables, as is often the case in many branches of chemometrics, both in industry and research. Large amounts of data makes visually based evaluation and screening for outliers difficult. There are various reasons for outliers, e.g. instrument failure, non-representative sampling, formatting errors, and objects stemming from other populations. Usually, only complete objects ($\mathbf{x}_{i.}$) are considered as outliers, but it is equally relevant to look for outliers in variables ($\mathbf{x}_{.j}$) and even individual data elements (x_{ij}). Most conventional multivariate methods are sensitive to outliers due to the fact that they are based on arithmetic means, covariance matrices and least squares (*LS*) fittings or similar criteria. Even a single outlier can have a large effect on the estimate and deteriorate the model. Therefore, it is necessary to 1) identify outliers and 2) decide whether outliers should be accommodated or rejected in the modelling process.
The aim of any robust method is to reduce, or remove the effect of outlying data points and allow the remainder to predominantly determine the results. Robust methods are helpful for both semi-automated detection of outliers, by looking at the robust residuals and for model building. When no outliers are present in the data set, the result from a robust method should be consistent with the result from the corresponding non robust method – the method based on the *LS* estimation. Robust methods provide a powerful methodology, extending a conventional 'manual' analysis and eliminate outliers by using exploratory methods and 'conventional' outlier diagnostics.

As noted by Gnanadesikan (1977), the consequence of outliers in multivariate data is intrinsically more complex than in the univariate case. A multivariate outlier can distort measures of location and scale, and thereby also those of covariance structure. As a result the modelling methods may describe the shape of the majority of the data incorrectly, and conclusions drawn can be misleading. An additional complication is that it is much more difficult to identify multivariate outliers. A single univariate outlier may be detected graphically, a task not that straightforward in higher dimensions. Many

multivariate methods work well for identifying single outliers, but when there are many outliers masking and swamping effects may occur. The masking effect means that some outliers are unnoticed because, the presence of other outliers masks their misleading influence (Ryan, 1997; Galpin & Hawkins, 1987). The swamping effect consists of wrongly identifying/diagnosing an observation as an outlier, because of the presence of other outliers (Hampel et al., 1986).

Much focus has been put on making the common chemometric techniques, such as Principal Component Analysis (PCA), Principal Component Regression (PCR) and Partial Least Squares (PLS) regression, more robust against outliers using robust estimates to replace the non robust *LS* estimate. Rousseeuw & Leroy (1987) presented an overview of robust estimates in regression and outlier detection, and Maronna & Yohai (1998) described recent advances in robust estimation in multivariate location and scatter estimation. Liang & Kvalheim (1996) wrote a review of the robust methods for multivariate analysis until 1996. Hubert et al. (2005b) described the *minimum covariance determinant* (*MCD*) and *least trimmed squares* (*LTS*) estimators for location, scatter and regression, and the recently developed robust methods for multivariate data analysis based on these estimators. Paper II is a review of robust methods for PCA, PCR, and PLSR, together with an introduction to the robust estimates for regression, location and covariance used in the robust multivariate methods, discussed in the paper.

In section 4.1, a short introduction to outliers and their effect on least squares estimation of location, scatter and regression will be given, followed by examples of applications of the robust methods for PCA, PLSR and PARAFAC is given.

4.1 Outliers

As stated in the beginning of this chapter; outliers can be defined as observations that appear to break the pattern or grouping shown by a majority of observations.

The data are assumed to be stored in an n x p data matrix $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$, with $\boldsymbol{x}_i = (x_{i1}, \ldots x_{ip})'$ the i^{th} observation, as described in section 2.1. The common estimates for the multivariate location $\hat{\boldsymbol{\mu}}_0$ and scatter matrix $\hat{\mathbf{C}}_0$ are the arithmetic mean and classical covariance matrix, respectively. However, it is well-known that these estimates will be influenced by the occurrence of outliers. Classical illustrative examples, showing their sensitivity to outlying samples, are given in e.g. Rousseeuw & Leroy (1987) and Maronna & Yohai (1998). To get reliable results that can persist possible outliers, robust alternatives such as *Stahel-Donoho* (Stahel, 1981; Donoho, 1982) and *MCD* (Rousseeuw, 1984) estimates of location and scatter can be used. For more information about robust estimators for estimating multivariate location and scatter, see Paper II.

In multiple linear regression models, it is assumed that also a response variable y is measured.
For all observations $(\mathbf{x}_{i.}, y_i)$ with $i = 1, \ldots, n$, it holds that

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + r_i \qquad \textit{Equation } 4.1$$

with errors r_i. The classical least squares method to estimate $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ is extremely sensitive to outliers. The reason for *LS* not being resistant to outliers follows from the properties of the objective function for *LS* procedures. The objective function to be minimized is the sum of the squared residuals:

$$\underset{\hat{\beta}}{\text{minimize}} \sum_{i=1}^{n} r_i^2 \qquad \textit{Equation } 4.2$$

in which the residuals r_i are given by

$$r_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip} \qquad \textit{Equation } 4.3$$

where $y_i \left(i = 1,..., n\right)$ are the corresponding values of the dependent variables, $x_{ij} \left(i = 1,..., n; j = 1,..., p\right)$ the values of the explanatory variables, and $\hat{\beta}_j = (j = 1,..., p)$ is the *LS* estimate of the parameters. This means that a relatively large outlier will exert an inappropriately large influence on the *LS*-estimate as will be illustrated in the following.

Three categories of outliers can be considered in cases of regression: 1) "Good" leverage points, which are observations isolated from the major part of the observations in the data matrix **X** that still follows the same regression model, 2) "Bad" leverage points, which in addition to being isolated from the major part of **X,** deviate strongly from the regression model defined by the other observations and 3) Outliers that are not leverage points, but have large *y* prediction residuals in calibration, and are therefore referred to as high *y* residual outliers or vertical outliers. Figure 4.1 illustrate the three outlier types, where high *y* residual observations are marked with a "1", "2" represent good leverage points, and bad leverage points are marked with "3". In robust analysis, the good leverage points are usually not denoted as outliers, as they are not harmful to the regression model, but merely reflect an "unfortunate design". These three types of outliers can occur both during model fitting and during predictions with a previously established model.

Both the high *y* residual outliers and the bad leverage points affect the calibration model by distorting the least squares model to a certain degree, and should be eliminated.

Generally, outliers are not necessarily wrong measurements, but could also indicate samples belonging to another group than the majority of the data. To get reliable results robust estimates for regression, such as *least median of squares* (Rousseeuw, 1984) and *LTS* (Rousseeuw, 1984), are needed (see Paper II for examples and descriptions of robust estimates for multivariate regression).

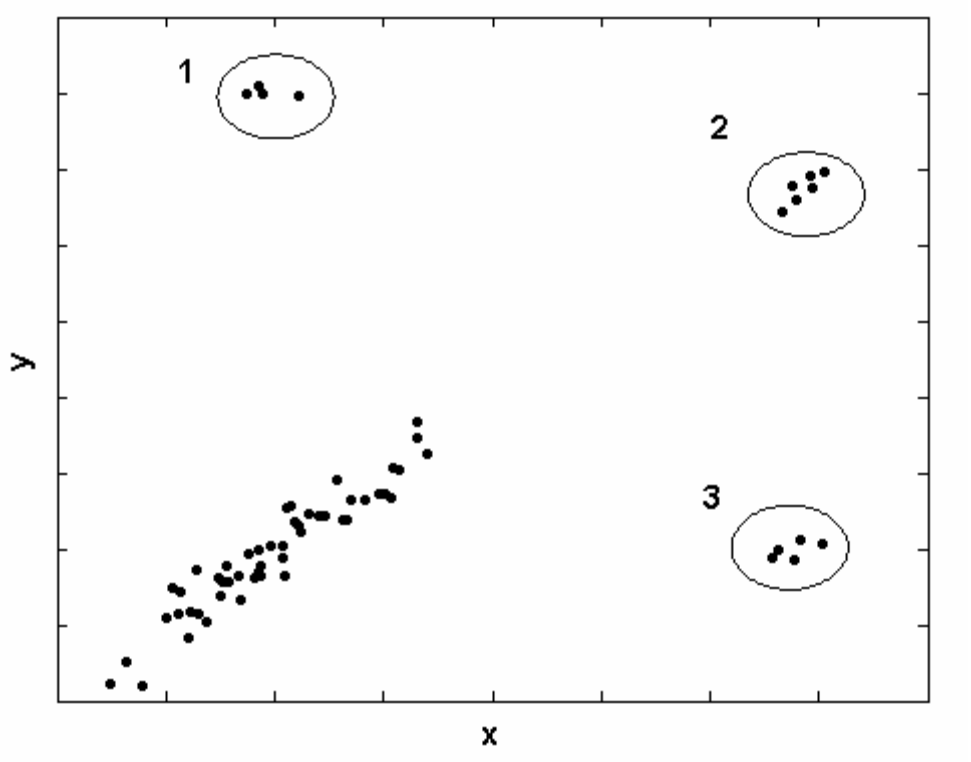

Figure 4.1. High y residual outliers (1) and leverage points (Good leverage points are denoted "2" and bad leverage points are denoted "3").

The robustness of the estimators can be quantified in different ways most commonly using two diagnostics: breakdown point and influence function. The breakdown point ε^* (Hampel, 1971) is a very useful measure of robustness, when comparing different robust methods in various situations. The finite-breakdown point can loosely be defined (Donoho & Huber, 1983) as the smallest fraction of samples (with respect to n), that can render the estimator useless. The breakdown point of the classical sample mean and the covariance matrix is $1 / n$, the lowest possible, meaning that one outlier is sufficient to ruin the sample mean or covariance matrix. Estimators with ε^* = 50 %, the highest possible breakdown point, are called high breakdown point estimators. The influence function (Hampel et al., 1986) tries to quantify the influence from an infinitesimal outlier on the estimate. Thus, in principle this allows for a more detailed quantitative comparison of different robust methods under a single outlier. A fundamental question here is, if the influence function is bounded, i.e. if a single outlier can lead to a breakdown of the estimator.

Another concept often used in connection with robust estimators is the asymptotic efficiency. Efficiency is the ratio of the mean square error from a robust estimator to the mean square error from an ordinary least squares estimator, when applied to a data set that is sufficiently normal and embrace no outlying samples (Ryan, 1997).

Multiple linear regressions, as well as estimation of sample mean and covariance, are the cornerstones of multivariate data analysis methods such as: PCA, PCR and PLSR (Rousseeuw & Leroy, 1987; Maronna & Yohai, 1998). The former underlying techniques are not resistant to outliers, as they are based on *LS* techniques. Such analysis is therefore extremely sensitive to outlying samples, and the conclusions drawn may be adversely affected by the outliers and are often misleading. Consequently, substituting the classical estimates with robust alternatives is often the basis for obtaining robust versions of the latter multivariate data analysis methods.
Many of the approaches proposed in the literature for multivariate data analysis, especially the older methods, rely on complex and often on very computer intensive calculations to carry out the analysis. Furthermore, some approaches such as the methods based on replacing the classical covariance by a robust estimator can not handle situations with more variables than samples, which are often the case in multivariate data analysis. One of the motivations behind the investigations of robust multivariate methods is the challenge to implement techniques fairly easy to handle to unskilled personnel within the industry. The methods applied in section 4.2.1 and section 4.3.1 are therefore chosen on the conditions that they should be computational feasible, capable of handling high dimensional data and the algorithms available.

4.2 Robust PCA

Classical PCA is often estimated using the eigenvectors (eigenvalues) of the sample covariance matrix. An outlier in PCA context can then be defined as observation/sample that lies far away from the subspace spanned by the correct k eigenvectors, and/or for which the projection into the model lies far from the remainder of the data within the subspace (Martens & Næs, 1989). The most intuitive and appealing way of robustifing

PCA is to replace the classical covariance matrix by a robust scatter matrix, via robust estimators of location and scale (Maronna, 1976, Campbell, 1980, Devlin et al., 1981, Rivest & Plante, 1988, Daigle & Rivest, 1992, Croux & Haesbroeck, 2000). A different approach to robust PCA uses projection pursuit techniques; searching for structure in high dimensional data by projecting these data into a lower-dimensional space, which maximizes a robust measure of spread, instead of the variance as in the classical approach (Ruymgaart 1981, Li & Chen, 1985, Ammann 1989, Galpin & Hawkins 1987, Xie et al 1993, Croux & Ruiz-Gazen 1996, Hubert et al. 2002). Recently, a combination of the above two approaches were proposed, using the projection pursuit part for initial dimension reduction, followed by the robust scatter estimators applied to this lower dimensional data space (Hubert et al., 2005a). All approaches so far consider the entire samples, $\mathbf{x}_i$, as outliers, but methods capable of handling elemental outliers, x_{ij} also exist. These methods are based on adjustments to the internal computations of the SVD algorithm, replacing the least squares criterion with a robust estimate (Hawkins et al. 2001, Liu et al. 2003, Croux et al. 2003). For a review of robust PCA, the reader is referred to Paper II.

4.2.1 Application of robust PCA

Methods for analysing chromatographic data often relies on subjective peak detection and peak areas, and on integration parameters which, if not properly set, may cause great errors in the calculated peak areas. Implications of the data extraction method are thus incorporated into the further analysis, often based on PCA. Other drawbacks concerning the manual peak area analysis caused by the selection of a subset of peaks are loss of information, regarding peak shapes and the absence/presence of peaks. Alignment of the chromatograms to correct for retention time shifts is necessary before turning into any multivariate data analysis. Variations are thus not dominated by shifts between variables, but by different levels of the variables (chemicals) as they ought to.

In Paper III, the possibility of using all collected data points from the chromatograms in PCA, combined with correlation optimization warping (Nielsen et al., 1998; Tomasi et al., 2004) as pre-processing are illustrated. Because of an outlier problem, concerning

both sample-wise and element-wise outliers, the advantages and drawbacks of two robust PCA methods, ROBPCA (Hubert et al., 2005a) and robust SVD (Hawkins et al., 2001), for analysing gas chromatographic data are investigated. The methods are robust against outlying samples and outlying elements, respectively (Paper II). The background for choosing RSVD was that misalignment may be dealt with by using this method only excluding outlying elements. This means that it is not necessary to exclude whole samples due to misalignment in some part of the chromatograms, as is the case in ROBPCA, because the properly aligned parts of the chromatograms are still available for analysis. By using RSVD it should be possible to obtain reliable results from the PCA analysis using the entire chromatogram without optimal alignments of the chromatograms.

The analyses were performed on two data sets differing in quality. The first set of data was obtained from gas chromatograms of fatty methyl esters (GC-FAME), data which were well behaved, in the sense that outliers are expected to be caused by insufficient peak alignment only since the method by itself is highly robust. The second data set consisted of volatile lipid oxidation products, collected by a dynamic head-space (GC-ATD). These data had a relatively higher risk of artefacts due to a more complex procedure and unstable products which results in larger sample differences and peak shifts. Data were kindly provided by the lipid group (att. C. Jacobsen) of the institute for Fisheries Research.
In the present case, samples of fish oil from farmed rainbow trout, fed two different diets were included. The samples included were frozen at -20 °C, -30 °C or -80 °C for 0-24 months.

In addition, to the alignment pre-processing of the chromatograms prior to PCA, baseline correction and normalisation were necessary to remove variations unrelated to chemical compositions (Paper III).

The PCA can explain the relationship between the different feeding types, measured as the fatty acid composition (GC-FAME) of the fish meat in the case with data of high quality (good alignment of the chromatograms). Fish feed vegetable oil contained

higher amounts of 18:1(n-9), 18:2(n-6) and 18:3(n-3), and lower amounts of 14:0, 16:1(n-7), 20:4(n-3), 20:5(n-3), 22:1(n-11), 22:5(n-3) and 22:6(n-3) than fish feed fish oil. The core plot of PC1 versus PC2, both from traditional PCA and ROBPCA and PC2 versus PC3 for RSVD, are shown in Figure 4.2 (first row). Centering of the data was not built in this RSVD algorithm, as is the case for ROBPCA, meaning that the first PC explained the centring of the data, and was for that reason not interesting.

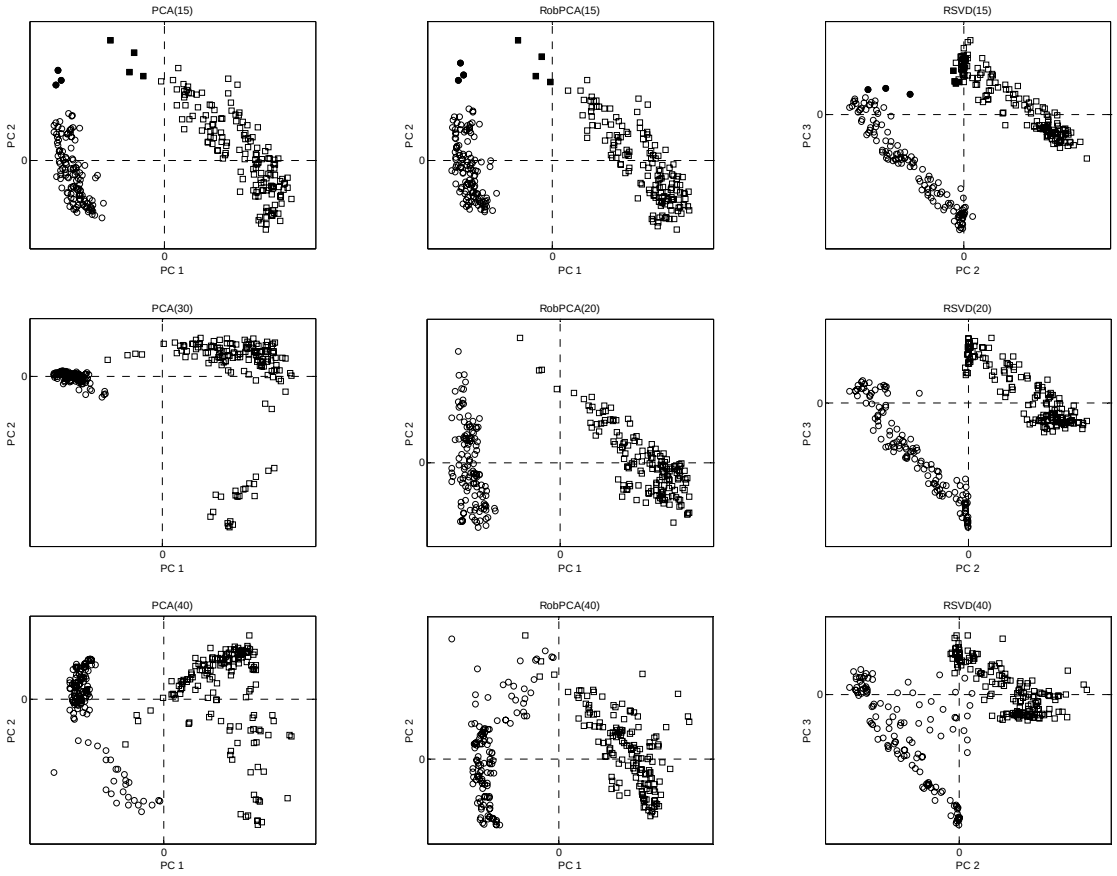

Figure 4.2. PCA scores; PC2 *versus* PC1 for classical PCA (leftmost column) for ROBPCA (middle column), and PC3 *versus* PC2 for RSVD (rightmost column). The quality of the alignment is decreasing from the first row and down. The samples are marked according to oil type in the feed: vegetable oil (○) and fish oil (□). A few 'extreme' samples are marked with filled symbols in the first row.

When data were of high quality (good alignment of the chromatograms), there were no difference in the score plot between the results obtained with traditional PCA or ROBPCA. In none of the two models (traditionally PCA and ROBPCA), PC2 was correlated to the variation that was investigated, but was primarily caused by biological variation within the groups. No other meaningful groupings were found in higher order

PCs. A difference in a part of the chromatographic profile was especially pronounced for the extreme samples with high score values in PC2, in both traditionally PCA and ROBPCA (filled symbols). These extreme samples were only outlying in a part of the chromatogram (less than 50 % of the variables), and could therefore be excluded by the RSVD. This ability of the RSVD method to exclude outlying elements reveals an even better grouping obtained with RSVD than with classical PCA and ROBPCA.

The score plots in Figure 4.2, illustrate the effect of reduced data quality (87.2 %: first row, 86.0 %: second row and 79.6 %: third row) of the three different procedures of principal component analysis. The evaluation of the data quality was based on the explained variance for a one component PCA mode, fitted to normalized un-centred data, aligned with different warping parameters and tested as proposed by Christensen et al. (2005). With decreasing data quality (from 87.2 % to 67.0 % explained variance) the clustering, according to different types of oil in the feed, was observed for all three methods of data of high quality, although the clearest clustering obtained was attributable to the two robust methods. With decreasing data quality, i.e. 79.6 % explained variance (Figure 4.2, second row) and below in this case, the plot got more unclear, regardless of what PCA method was used to analyse the warped data. This clearly illustrated that data, and thereby the warping, needed to be of a certain quality to obtain reliable results. The robust methods can not remedy problems with large shifts in retention time.

In the more difficult GC-ATD data set, a grouping according to storage temperature (-20 °C versus -80 °C) was obtained with both traditionally and robust PCA for samples stored for 24 months, see Figure 4.3. The clearest grouping was observed with RSVD (Figure 4.3, bottom), attributable to a non optimal alignment, resulting in a relatively large number of outlying variables in a majority of the samples. If more than 50 % of the variables are outlying compared to a majority of the chromatograms, a robust procedure to handle the samples, such as ROBPCA, was needed. Three such clearly outlying samples were separated from the other samples along PC2 (PC3 for RSVD). With ROBPCA the three outliers were excluded from the modelling step, and placed closer to the other samples. Additionally, the variation accounted for by the PC2 scores

(PC3 for RSVD) was due to variation within each grouping of storage time, reflecting the biological variation of the groups of fish. It was not possible to identify other patterns in the data by plotting other combinations of principal components.

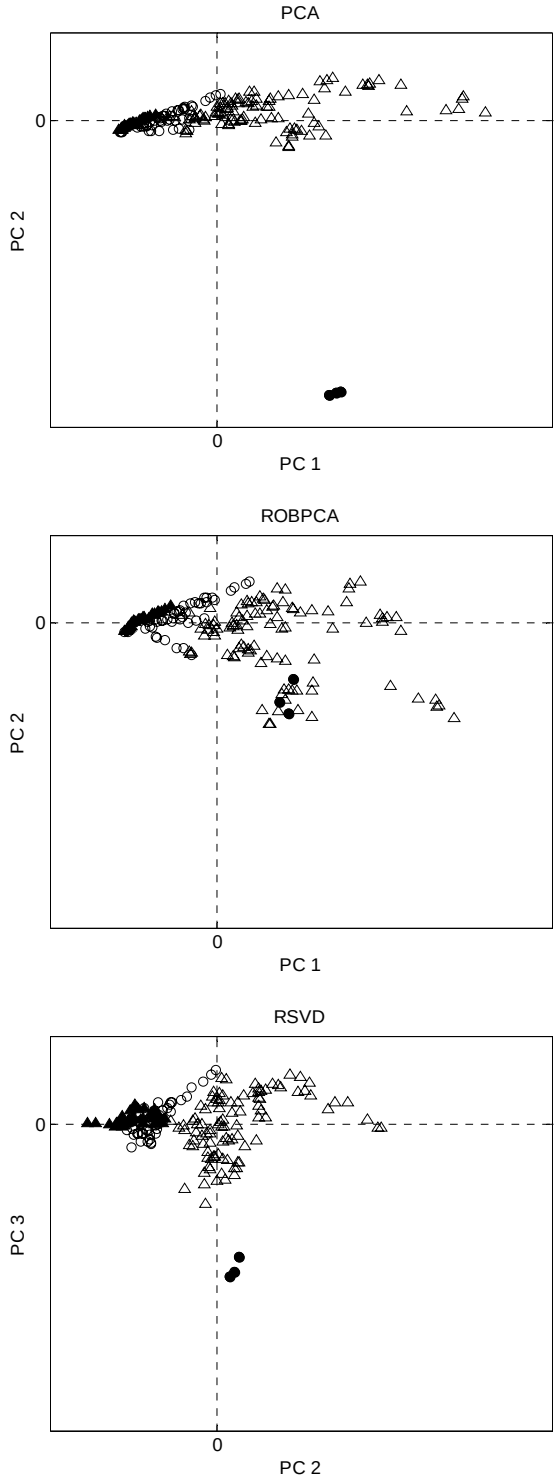

Figure 4.3. PCA scores; PC2 *versus* PC1 for classical PCA (left) and ROBPCA (middle), PC3 *versus* PC2 for RSVD (right) when the models were fitted to aligned data. The samples are marked according to storage temperature: -20 ºC (Δ), -30 ºC (○), and -80 ºC (▲). Three outliers (all -30 ºC samples) are marked with filled circles.

This study demonstrates that the usage of robust PCA is advantageous compared to traditional PCA, when analysing the entire profile of chromatographic data in cases of not perfectly aligned data. Which method of robust PCA to chose – sample or elementwise – depends on the type of outliers that would be expected. When outliers, deviating in the entire profile are present in the data set, ROBPCA are preferably compared to RSVD, which only can handle up to 50 % of the outlying elements in each data vector. When the data set is not perfectly warped - meaning that all peaks are not perfectly warped and outlying elements exist, the RSVD method is to be preferred.

4.3 Robust PLSR

Classical PLS regression makes use of ordinary least squares regression steps in the calculation of weights, loadings, scores and regression coefficients. Since outliers in **X** (leverage points) and / or **y** or **Y** (vertical outliers or high *y* residual outliers) variables highly influence the *LS* estimates in multivariate regression, the PLSR model may be hampered and unreliable. Therefore, several robust alternatives to classical PLSR have been developed.

The first authors to propose a robust version of PLSR were Wakeling & MacFie (1992) who replaced all the univariate regression steps in the PLS2 algorithm by robust alternatives. The drawbacks are high computational cost and lower efficiency of the regression steps. Following the idea of Wakeling & Macfie (1992), Griep et al. (1995) carried out a comparison among three different methods of robust regression and studied their incorporation into the PLSR1 algorithm when replacing the regression step for the weight vector **w** with three different methods of robust regression. Their empirical results indicate that the best option is to use *IRLS* compared to *LMS* and Siegels *RM* (Siegel, 1982). Methods based on iteratively reweighted algorithms have been proposed by Cummins & Andrews (1995) and Pell (2000). These algorithms are no longer prone to high computational cost, but can not withstand leverage points and are only valid for PLSR1 regression. In Gil & Romera (1998) a robust PLSR1 method is obtained by robustifying the sample covariance matrix of the *x*-variables and the sample

cross-covariance matrix between the *x*- and *y*-variables. For this the highly robust Stahel-Donoho estimator is used with Huber's weight function (Huber, 1964; Huber, 1973). To minimize the computational cost the sub-sampling scheme used to compute the estimator starts by drawing subsets of size $p + 2$. This means that the method cannot be applied to high-dimensional regressors ($n << p$) which is a major disadvantage. It is not possible to extend the method to PLS2 (Hubert & Vanden Branden, 2003). A robust version of SIMPLS algorithm called RSIMPLS was proposed by Hubert & Vanden Branden (2003). This algorithm is based on replacing the cross-covariance matrix $\mathbf{C}_{xy}$ and the empirical covariance matrix $\mathbf{C}_x$ by robust estimates, and by performing a robust regression method instead of MLR. This method is resistant to all types of outliers, can handle data with more variables than samples and with $q \geq 1$. The RSIMPLS method is reminiscent to the minimum covariance determinant which is known to have quite a low efficiency (Croux & Haesbroeck, 1999). Recently, Serneels et al. (2005a) proposed a method, Partial Robust *M*-Regression (PRM), for robust regression based on *GM*-estimators. PRM uses continuous weights, resulting in a gradual down-weighting of outliers according to their degree of outlyingness. The weighting is used both in the SIMPLS step of computing the PLSR scores as well as in the regression of y on these scores. The PRM method is computational possible for high dimensional data sets and can handle both types of outliers but the method is currently only derived for univariate *y* (i.e. PLSR1) and the highest possible breakdown point of all *GM*-estimators is in general not larger than 30 % and decreases as a function of the dimensionality *p* (Maronna et al., 1979; Rousseeuw & Yohai, 1984).

4.3.1 Application of robust PLSR

Fat is an important parameter handling marinated herring products. This carcass constituent both affects quality and production output. In addition to seasonal variation, herring caught at the same place, at the same time, show great variation within fat content (Larsen et al., 1997; Nielsen et al., 2005). Today, the fat content is based on a visual inspection and / or a laboratory analysis, which again is based on a pooled sample. A pooling of samples is done as time is a limiting factor during processing. This means that the true variation of fat content, within a catch, is not perfectly revealed. In a

study by Nielsen et al. (2005), evaluating the potential of different non-destructive methods for on-line fat measuring on single fish level, NIR demonstrated the most promising results, compared to Torry Fish Fat Meter and NMR. By implementing on-line fat measuring on single fish level in the production plant, it will be possible to differentiate the raw material into different products, thereby optimizing product quality and minimizing wastage.

In this section, two robust calibration methods RSIMPLS (Hubert & Vanden Branden, 2003) and PRM (Serneels et al., 2005a) will be compared to classical PLSR (NIPALS) when correlating the fat content in herring measured by Bligh and Dyer extraction to NIR measurements. For a more detailed description of RSIMPLS and PRM see Paper II. A major difference between the two robust regression methods studied is that the PRM method use continuous weights, resulting in a gradual down weighting of the outliers according to the severity of the very same, whereas RSIMPLS uses hard rejection, donating a weight of zero to all observations with residuals above a certain cut-off value and unity to all others. The breakdown point of all *GM*-estimators, the type used in PRM, is no higher than 30 %, whereas the *MCD*-estimator used in RSIMPLS can be as high as 50 %. However, the statistical efficiency was shown to be better for PRM than for RSIMPLS, when comparing various distributions of error terms, different samples sizes and dimensionality (Sernells et al., in press). This lower efficiency of RSIMPLS was due to the use of *MCD* which has a relatively low efficiency. The efficiency of *MCD* can be improved at the expense of the breakdown point. For a reweighed *MCD,* with a breakdown point of 25 %, the efficiency is nearly always above 60 % in the Gaussian case (Croux & Haesbroeck, 1999). In the present study, breakdown values of 10 % outliers were used, thereby improving the statistical efficiency of the model.

For each herring the fat concentration was measured by Bligh and Dyer, while the *x*-variables consisted of NIR absorbance spectra. The intension was to predict the fat concentration based on 821 NIR spectra, with measurements for every 2 nm from 1.000 up to 2.222 nm. For each model (RSIMPLS, PRM and classical PLSR) the Root Mean Square Error of Prediction (RMSEP) r^2 and the bias were calculated. The said data set

has previously been studied by Nielsen et al. (2005), however, that investigation did no remove any samples due to their outlying properties. It was therefore interesting to see how the robust methods would perform, when no obvious outliers were present in the data set. It has been shown that four components were sufficient to perform the PLSR analysis. The pre-processing of the data was done in the same manner as in Nielsen et al. (2005), which resulted in a data set of NIR spectra (scatter corrected) of 230 dimensions.

For each of the three methods full cross-validation was performed.
The RMSEP value is defined as

$$\text{RMSEP}_k = \sqrt{\frac{1}{n}\sum\left(\hat{y}_{-i,k} - y_i\right)} \qquad \textit{Equation } 4.4$$

where $\hat{y}_{-i,k}$ represents the predicted *y*-value for sample *i* based on *k*-components, when sample *i* was left out of the estimation of the regression parameters.
RMSEP can be interpreted as the average prediction error, expressed in the same units as the original response values.

The Bias can be interpreted as the systematic difference between predicted and measured values. The Bias is computed as the average value of the residual

$$Bias = \sum\left(\hat{y} - y\right)/n \qquad \textit{Equation } 4.5$$

The Bias is a commonly used calculation of the accuracy of a prediction model, and should be close to 0 if the model is good.

The criteria were evaluated for $k = 1, \ldots, 6$ components. The results are summarized in Table 4.1.

For all three methods tested, more than two principal components were needed to obtain a satisfactory prediction. With more than two components there were no difference

between r^2 and bias for the obtained models. Between PLSR and PRM the RMSEP is almost identical. That indicated that no extreme outliers were present in the data set. However, a somewhat better RMSEP value was obtained for RSIMPLS compared to the other two methods classical PLSR and PRM. Though, when looking at the score plots, the influence plot and the leverage values, no samples appeared to be extreme (results not shown). Therefore, the lower RMSEP value obtained with RSIMPLS could indicate, that in this case, the samples excluded as outliers are borderline samples - those samples expanding the variance within the data. By excluding these samples, the obtained model might not cover the variance in new samples and consequently weaken the precision of the prediction. An independent test might have revealed this, unfortunately that was not possible in this study. To summarize, this study illustrated that in the case of data sets with no extreme outliers at present, the advantages of employing robust methods were ineligible. Focusing on the drawbacks of the robust methods, especially the lower statistical efficiency and the time-consuming computations leaped out.

Table 4.1. RMSEP, r^2 and bias calculated for the prediction of fat content (%) based on NIR measurement when comparing the performance of three different PLSR methods. *k* = number of PCs.

		PLSR	RSIMPLS	PRM
k = 1	RMSEP	2.56	2.11	2.58
	r^2	0.78	0.79	0.78
	Bias	0.00	-0.02	0.35
k = 2	RMSEP	2.28	1.74	2.33
	r^2	0.83	0.83	0.82
	Bias	0.00	-0.08	0.23
k = 3	RMSEP	2.19	1.65	2.21
	r^2	0.84	0.84	0.84
	Bias	0.00	-0.08	0.17
k = 4	RMSEP	2.19	1.54	2.13
	r^2	0.85	0.86	0.85
	Bias	0.00	-0.09	0.03
k = 5	RMSEP	2.03	1.54	2.04
	r^2	0.86	0.86	0.86
	Bias	0.00	-0.02	0.042
k = 6	RMSEP	2.02	1.55	2.04
	r^2	0.86	0.86	0.86
	Bias	0.00	-0.03	0.04

4.4 An approach for and application of robust PARAFAC

The algorithm to compute PARAFAC (Bro, 1998; Smilde et al., 2004) is normally a least squares fitting based on the alternating least squares procedure, which is not able to withstand the presence of severe outliers.

An attempt to make PARAFAC robust was presented at the ERCIM meeting at the Royal Veterinary and Agriculture University 2005 (Engelen & Hubert, 2005a; Engelen & Hubert. 2005b). The proposal is based on unfolding the three-way array (I x J x K) so that the sample-mode is kept intact and then applying a method for robust principal components analysis ROBPCA (Hubert et al., 2005a) on the unfolded data (I x JK). The residual for each point is computed, and the h samples with the smallest residuals are stored in the initial h-subset. Classical PARAFAC is carried out on these h samples, and

a new *h*-subset is constructed by taking the *h* samples with smallest residuals with respect to the PARAFAC model. The procedure is repeated until the relative change in fit is small. The statistical efficiency of the *MCD* estimator, used in ROBPCA, can be increased by implementing a reweighing estimator (Rousseeuw & Zomaren 1990; Rousseeuw & Van Driessen, 1999).

The robust PARAFAC method, proposed by Engelen & Hubert (2005b), is intended to find outlying samples. In the two methods, proposed by Vorobyov et al., (2005), the PARAFAC is made robust towards elementwise outliers by optimizing the least absolute error (*LAE*) fitting criterion, instead of the ordinary *LS* criterion in regression. The procedures are based on efficient interpoint methods for linear programming (LP) and weighted median filtering iteration (WMF), respectively. The breakdown point of *LAE* is 50 % compared to 0 % for the *LS*, which can be seen when considering the mean estimation under *LS* and *LAE* criteria. These correspond to arithmetic mean and median operators, respectively, where the arithmetic mean can be ruined by even a single outlying sample, whereas the *LAE* will stay stable. In a simulation study, it turned out that both algorithms are computationally efficient, but the WMF iteration is particularly appealing from a simplicity point of view compared to LP (Vorobyov et al., 2005). Both methods also outperform the classical *LS* PARAFAC fitting under heavy tailed noise, and show good tendency for impending scrutiny (Vorobyov et al., 2005).

4.4.1 Application of robust PARAFAC

A common phenomenon, and problem, when fitting PARAFAC to fluorescence landscapes (excitation-emission matrix), is the light scatter effects, such as Raman and 1st and 2nd order Rayleigh scattering (Andersen & Bro, 2003; Thygesen et al., 2004). The 1st and 2nd order Rayleigh scattering are the ridges seen in the lower right and upper left part, respectively, in Figure 4.4.

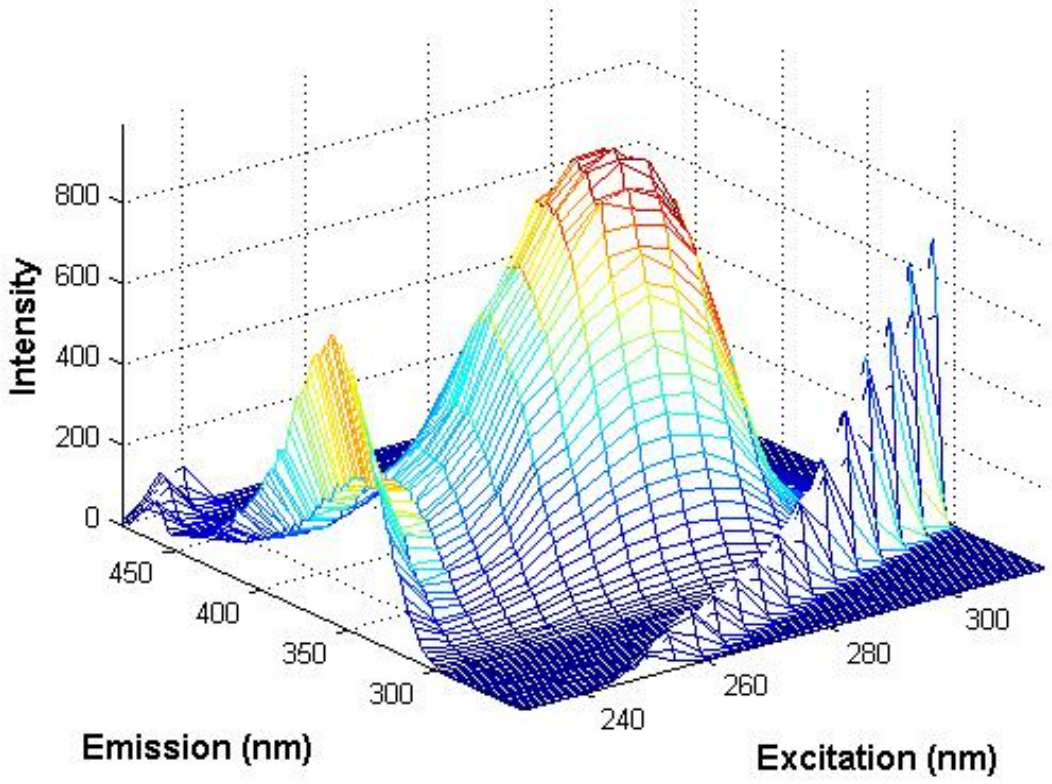

Figure 4.4. Example of a fluorescence excitation-emission landscape. The 1st and 2nd order Rayleigh scatter are the ridges seen in the lower right and upper left part, respectively.

This scatter contains no chemical information and will most possibly give a model inadequacy, influencing the estimated model parameters (Andersen & Bro, 2003) - this explains why this effect should be removed or reduced as much as possible. As such, scatter can be considered as outlying elements. Different proposals of how to handle these scatter effects can be found in the literature; subtracting a standard (Wentzell et al., 2001; McKnight et al., 2001), down weighting the scatter (Bro et al., 2002; JiJi & Booksh, 2000), inserting missing values (Bro, 1997), simply avoiding the part containing scatter (Bro, 1999), interpolating the scatter area (Zepp et al., 2004; Bahram et al., 2006) or insertion of zeros outside the data area (Thygesen et al., 2004). Unfortunately, all of the proposed methods seem to have some drawbacks, e.g. they can only be used in special cases, unacceptable decomposition of the spectra affecting the convergences of PARAFAC algorithm or they are computational cumbersome (Andersen & Bro, 2003; Thygesen et al., 2004, Rinnan & Andersen, 2005). A common

problem is the visible inspection of the data before the methods can be applied. This makes it difficult to perform all these proposed methods on several data sets at once. It even becomes harder to reduce the effect of scatter when the signal and scatter are overlapping, which is often the case.

In the following, the *LAE* criterion, proposed by Vorobyov et al. (2005), is adapted for fitting PARAFAC to fluorescence landscapes, to investigate if the elemental robust PARAFAC method can dispose of the scatter effects in the data. In the classical algorithm for fitting PARAFAC, the *LS* criterion is replaced with *LAE* in all three modes. The method was tested on different well analyzed fluorescent data. The overall impression was equal, and therefore only the results obtained with fluorescence data of mixtures of four known fluorophores (Baunsgaard, 1999; Riu & Bro, 2003), will be shown here. The four compounds are phenylanaline, 3, 4-dihydroxyphenylalanine (DOPA), 1,4-dihydroxybenzene and tryptophan. For every sample an excitation-emission matrix was obtained by measuring the emission spectra from 200 to 450 nm at 5 nm intervals, with excitation at every 5 nm from 200 to 350 nm on a Perkin-Elmer LS50 B fluorescence spectrometer. The excitation from 200 to 230 nm and the emission below 260 nm were excluded from the analysis since it is highly influenced by the condition of the xenon lamp as well as by the physical environment and mainly contained missing elements, respectively (Baunsgaard, 1999). From previous investigations (Baunsgaard, 1999; Riu & Bro, 2003), it is known that four components are appropriate and that four samples can be considered as outliers, these are therefore removed from the data set, as this analysis is aimed at testing elementwise outliers, not whole samples. The data set then consists of 23 samples, 18 excitation wavelengths and 116 emission wavelengths, and will in the following be refereed to as the full Dorrit data set.

The emission loading (second mode) from a four component *LS* PARAFAC model fitted to the Dorrit data set where scatter has been removed is shown in Figure 4.5 (left). The loadings have a reasonable shape resembling the pure spectra of the four fluorophores. This method is based on removing the Rayleigh scatter by inserting a mixture of missing values and zeroes. The emission loadings, when fitting a *LS*

PARAFAC model to the full Dorrit data set will appear as illustrated in Figure 4.5 (right). Both models are fitted with non-negativity constraints. The loadings in Figure 4.5 (left) have a reasonable shape resembling the pure spectra of the four fluorophores. When comparing the emissions loadings from the two models, it is clear that the light blue peak in the model fitted to data with Rayleigh scatter is wrong, this is caused by the scatter. This clearly indicates that the Rayleigh scatter need to be removed to obtain a good model. A problem with inserting missing values in the area covered by the Rayleigh scatter lines is that the scatter lines may be confounded with chemical information, and thus it is interesting to keep these areas. Furthermore, it might be difficult to accurately estimate the exact width of the Rayleigh peak (Rinnan & Andersen, 2005).

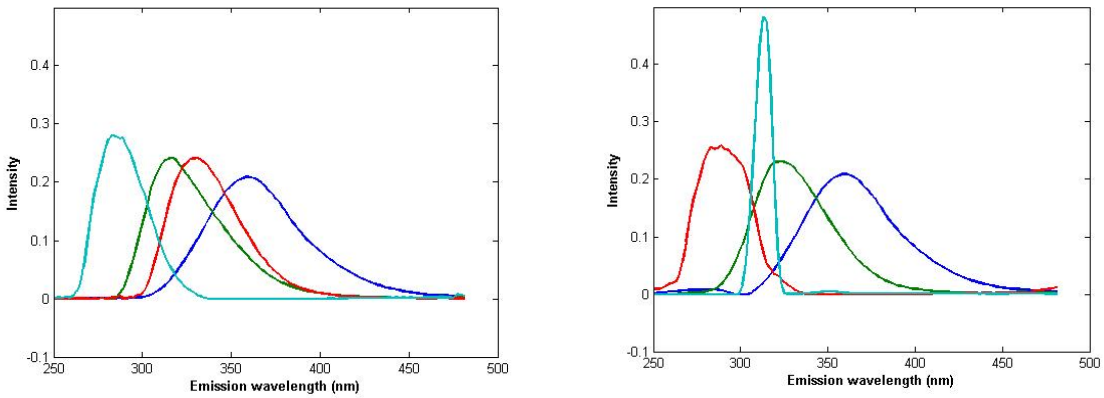

Figure 4.5. Left: Emission loadings from a four component *LS* PARAFAC model, fitted to the data set with scatter removed. Right: Emission loadings from a four component *LS* PARAFAC model, fitted to the full data set.

By applying the *LAE* PARAFAC to the full Dorrit data set, the obtained model seems almost perfect, as indicated below in Figure 4.6, showing the four emission loadings obtained. The shape of the loadings is almost identical with the pure spectra of the fluorophores as for the *LS* model with Rayleigh scatter removed.

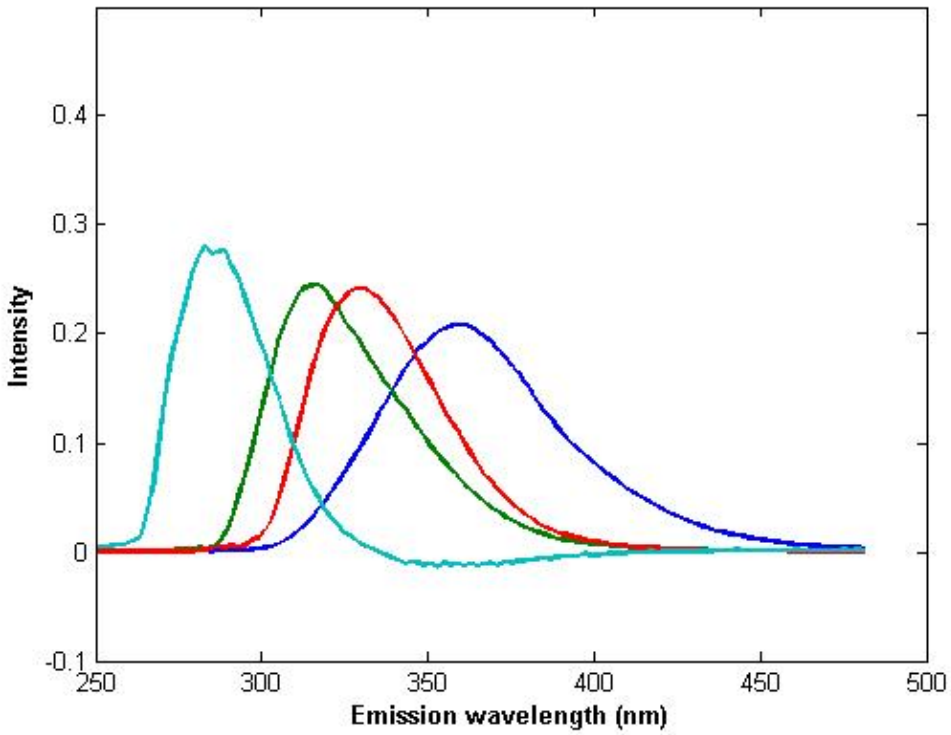

Figure 4.6. Emission loadings from a four component *LAE* PARAFAC model, fitted to the data set.

The result was encouraging, but unfortunately this will not be achieved in "reality". When different subsets of data are analyzed independently, the results vary to a great extent. Even the removal of one single sample can deteriorate the *LAE* model. In Figure 4.7 examples of the emission loadings from *LS* PARAFAC (left) and *LAE* PARAFAC (right) conducted on 12 different subsets of the Dorrit data are shown. Four of the subsets correspond to split-half analysis, and in four other subsets only one sample, randomly chosen, is removed from the full Dorrit data set. The subsets vary in sample number from 22 samples and down to 12 samples.

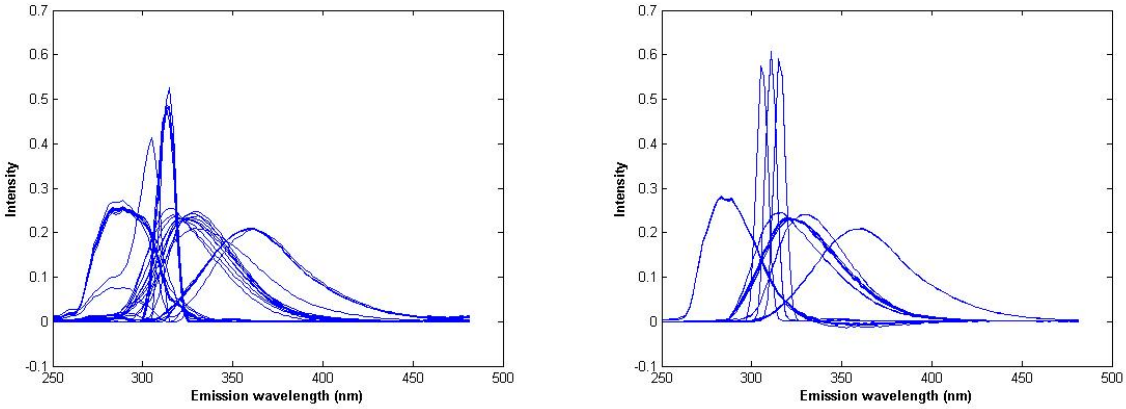

Figure 4.7. Emission loadings from four component PARAFAC models fitted to 12 different subsets of the data with the classical *LS* approach (left) and the robust *LAE* approach (right).

A problem with scattering is that it is systematic and occurs with positive values in all samples. Furthermore, some peaks containing chemical information only occur in e.g. two or three samples. This means that with *LAE*, minor real chemical peaks that only occur e.g. in two or three samples, will be downweighted as outliers, and some part of the scatter will be approximated by one or two PARAFAC components, because the scattering elements are not seen as outliers, but regarded as regular observations in the regression part of *LAE*. Examples of samples where the Rayleigh scatter is dominant compared to the relevant chemical information are shown in Figure 4.8.

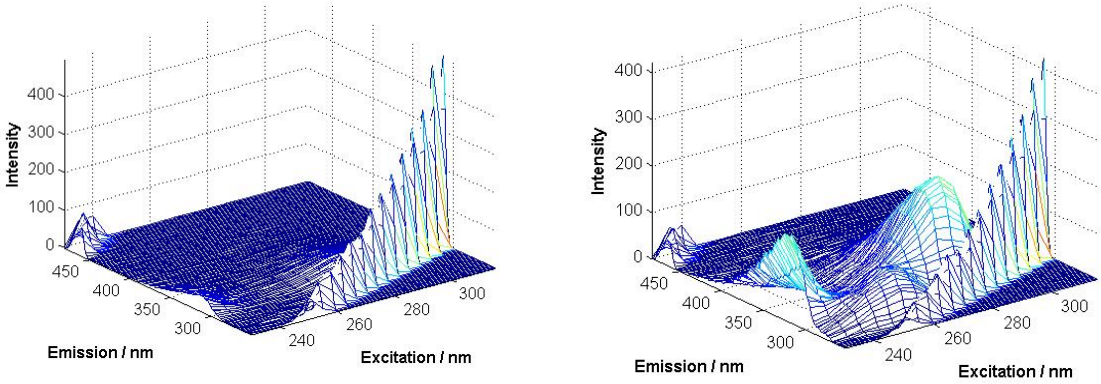

Figure 4.8. Examples of a fluorescence excitation-emission landscapes where the Rayleigh scatter is dominant compared to the relevant chemical information.

The conclusion is that *LAE* PARAFAC cannot be considered as a confident method for handling scatter as a result of the systematic nature of the scattering.

4.4.2 Automatic scatter identification

Another approach for identification of scatter was tested (Paper IV). This method is based on robust statistics and takes advantage of the systematic nature of the scatter. The method is automatic as no visual inspection of the data prior to modelling is required.

The method is based on ROBPCA (Hubert et al., 2005a). ROBPCA prevents the corruption of the principal components by outliers through a combination of robust subspace estimation (based on projection pursuit techniques) and the *MCD* estimator (Rousseeuw, 1984) for robust covariance and centre estimation. Additionally, samples are marked as regular samples or outlying samples for the concerned model making the procedure useful as outlier identification tool. For a detailed description see Hubert et al. (2005a).

ROBPCA can only be performed on two-way data matrices. Such two-way matrices can be extracted from three-way data like the EEM (Figure 4.9 A). By slicing the data along the sample mode, the scattering is situated in one or more diagonal lines in each sliced observation (see Figure 4.9 B). ROBPCA is not able to handle elementwise-outliers but only sample outliers. This means that taking each sample separately as input matrix for ROBPCA will not work well since the scattering does not correspond to a whole sample in these data, but only to a part of the sample. Therefore the proposed method starts by slicing the data $\underline{\mathbf{X}}$ along the emission and excitation mode, establishing useful two-way matrices in which the scattering is situated in columns for some of these matrices (see Figure 4.9 C and D).

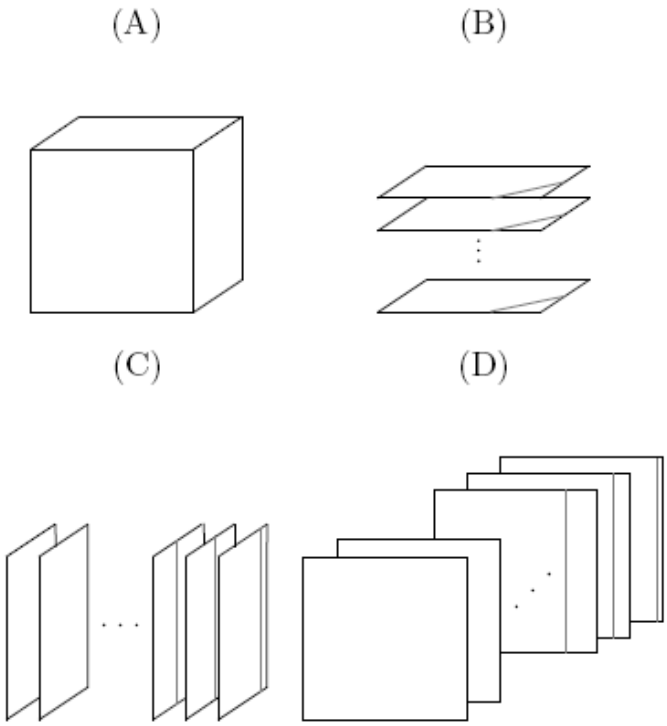

Figure 4.9. A visualization of the scattering in the three-way data (A) sliced in the sample mode (B), the second mode (C), and the third mode (D). The grey line represents the scattering.

In this way several matrices are obtained, and on the transposed of these matrices ROBPCA is applied. By applying ROBPCA on the transpose of the sliced matrices in the emission and excitation mode leads to identification of the scattering. As a result, two weights are assigned to each data element. The weight is assigned 1 to an element which is a regular point and 0 to an outlier. Merging both weights by taking the maximal value finally flags the outlying elements. For a detailed description of the method see Paper IV. The results of this automated scatter identification method can then be used as input data for PARAFAC. Since a classical PARAFAC algorithm is applied on the data after removing scatter, outlying samples will corrupt the final result. Removing of outlying samples is therefore necessary.

The proposed automatic scatter identification method was tested on different fluorescent data set with focus on how well the scatter was reduced and the signal preserved. Furthermore, the performance of the scatter identification method in combination with three different PARAFAC methods (inserting missing values, interpolate the scatter and

down-weighting the scatter regions) were evaluated. The results from the tests performed on the full Dorrit data set will be shown in the following.

In Figure 4.10 the emission profiles of sample 4 for the 18 excitation wavelengths are shown. The elements flagged as outliers by the scatter identification algorithm are marked with dots on the x-axis. The scatter corresponding to 2nd order Rayleigh scatter is clearly identified for the first 3 excitation wavelengths (3 first plots), and from excitation 5 and further on the regions according to the 1st order Rayleigh scatter are clearly identified. The successful detection of Rayleigh scatter in the remaining samples performs likewise (results not shown). From other data sets tested, it is known that the identification of Raman scatter performs likewise successfully (see Paper IV).

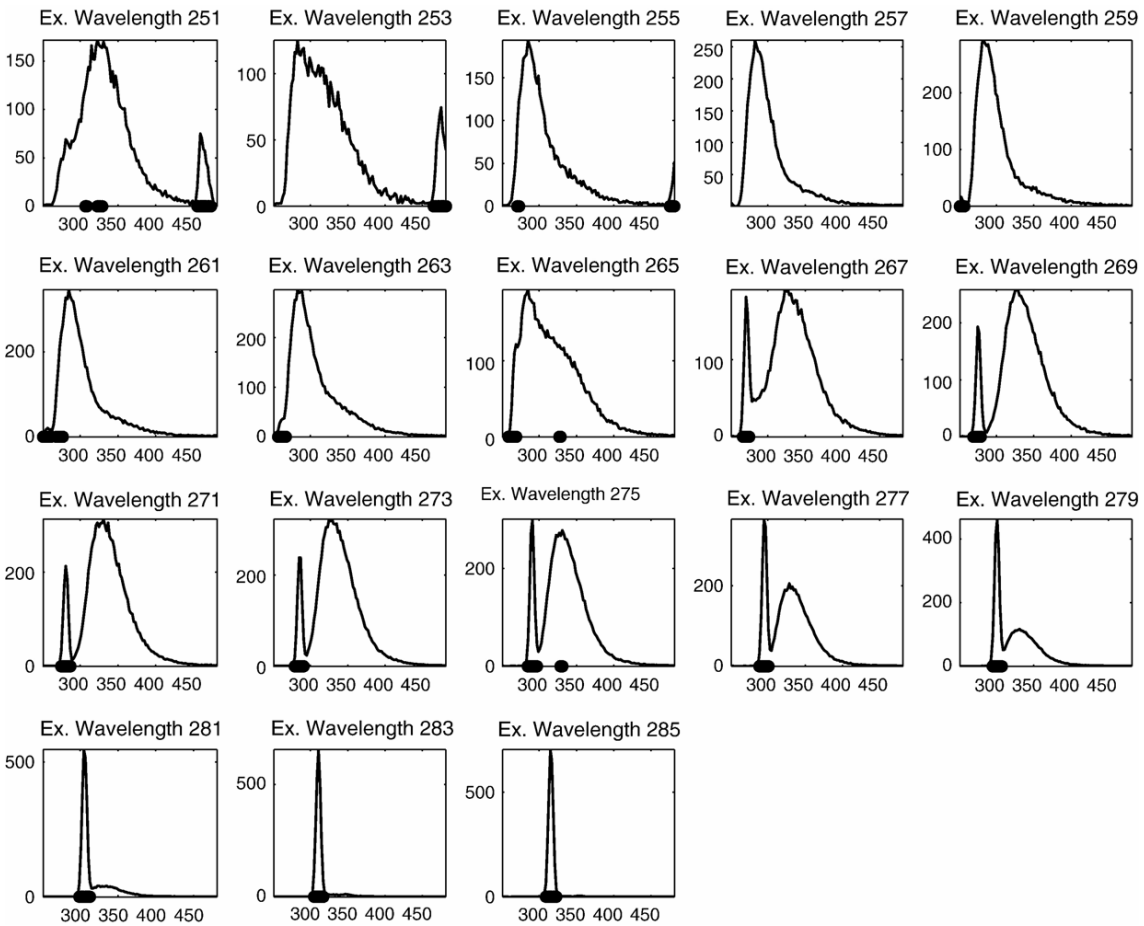

Figure 4.10. The emission profiles of the fourth sample of the full Dorrit data for the 18 excitation wavelengths. The regions identified as scatter are marked by dots.

The emission and excitation loadings obtained with the three different PARAFAC algorithms tested on the full Dorrit data in combination with the information about the scatter regions are shown in Figure 4.11. Both emission and excitation loadings for all three tested methods are almost identical with the pure spectra of the four fluorophores. This clearly indicates that this method for identifying scatter has worked well with respect to 1st and 2nd order Rayleigh scatter. For the full Dorrit data no obvious differences are observed between the three tested PARAFAC methods.

The overall evaluation of the proposed method clearly shows that the method always succeeds in finding the scatter regions both concerning Rayleigh (1st and 2nd order) and Raman scatter without marking too much of the signal as outlying due to chemicals under investigation. However, smaller parts of the scattering are sometimes hard to detect depending on the data complexity e.g. noise and overlap between scatter and chemical signal. This means that scatter might be included to a minor extent in the PARAFAC modelling step, but also smaller part of the chemical signal might be flagged as outlying and thereby excluded from the analysis.

However, this seems not to be an invincible problem for estimating the final PARAFAC estimates. The three tested PARAFAC methods after removal of the scattering work for the cases they can handle. This means that for the data with the missing values fitting problems are only encountered when the signal and scatter coincide too much, such that essential information vanishes. Secondly, classical PARAFAC applied on interpolated data also performs well, but it is most subject to the parts of the scattering that are not flagged as outlying. Finally, down-weighting the outlying elements is also a good option, provided that the scattering is in the region of the signal. For too severe scatter, this technique is not useful and actually is the least robust of the three investigated procedures.

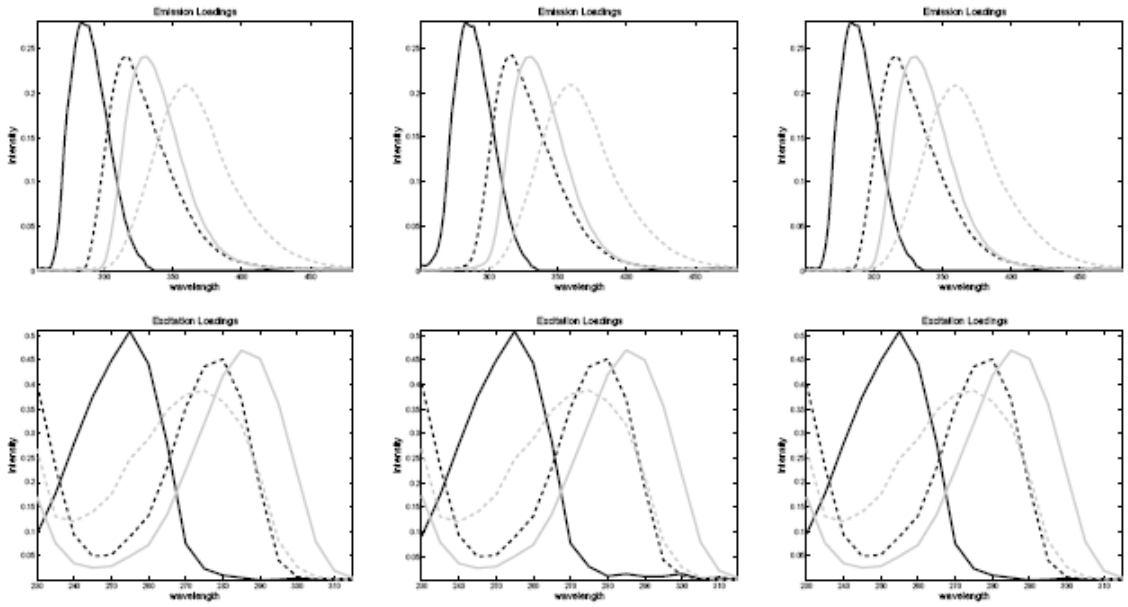

Figure 4.11. Four component PARAFAC models (left column) Missing, (middle column) Interpolation, and (right column) Weighted) fitted to the full Dorrit data where the scatter has been detected by the automated method. First row corresponds to the emission loadings and second row to the excitation loadings.

4.5 Software

The common basic methods for robust estimation of location and scatter (i.e. *MCD*) and robust regression (i.e. *M*-, *LMS*-, *LTS*-, *S*- and *MM*-estimators) are all available within the standard statistical software packages SAS (release > 6.12) (Chen, 2002), S-Plus (S-PLUS, 2001; S-PLUS, 2002) and R (Fox, 2002). An implementation for robust PCA is also available for S-Plus (Hubert et al., 2005c). Recently, a comprehensive MATLAB toolbox, "LIBRA", for robust estimates and multivariate methods has appeared (Verboven & Hubert, 2005). Apart from *MCD* and *LTS* it also contains implementations of other methods that have been developed at the research groups at the University of Antwerp and the Katholieke Universiteit Leuven, in particular for robust PCA (ROBPCA) and PLS (RSIMPLS). The toolbox also includes many graphical tools for model checking and outlier detection. Additionally, an incorporation of several of these methods into the widely used PLS_Toolbox for Matlab is in preparation (Eigenvector Inc., Pers. Comm.). The partial robust M-regression is also available as Matlab

implementation (Serneels et al.2005b). The algorithm for robust RSVD used in section 4.2.1 (Paper III) was kindly provided by A. Belousov, Münster.

4.6 Concluding remarks

This small investigation of robust methods clearly indicates that robust methods are not the solution to the whole problem concerning outliers, but they offer a substantial improvement over standard techniques, which to a certain degree depends on the type of data and outliers (sample- or elementwise) given in the data set.
Conditions that prove the most promising employing robust methods appear to be in situations with many samples and variables, such as in the case of gas chromatographic data, as illustrated in this investigation. Furthermore, the outliers might not be systematic as illustrated with the scatter example in section 4.4. In such situations the outliers are not seen as outliers, but regarded as regular observations in the modelling. However, as illustrated with the proposed automatic scatter identification procedure, the systematic nature of the scatter can be utilized and turned to something constructive.

5.0 Conclusion and perspectives

In this project the traceability chain from fishing vessel to final product has been scrutinised and the information (data) obtained throughout the production chain has successively been analysed. The objective has been to investigate the possibilities of integrating multivariate techniques into the industrial documentation system. Furthermore, the potential of using robust multivariate methods within a data miming process has been investigated.

It is easy to generate large data sets that contain little or no information. Moreover, it is an extensive task to find significant information in large amounts of data. Therefore, two essential questions emerge: 1) how to get data that contain as much relevant information as possible, and 2) how to extract information from large and complicated data sets. With the introduction of multivariate data analysis, the problem of extracting information from vast data sets is as good as solved, leaving as the challenge how to generate data containing information relevant for the purpose under investigation, as in the case of this study. When predicting the influence of the quality of the raw material on the quality of the final product, apt measurements reflecting these qualities are necessary.

In this study, the analyses of data obtained during the production of marinated herring, indicated that the data, in the present form, were not suitable for further multivariate data analysis. The reason for that is the lack of variability and/or the lack of traceability on the needed level (in particular specification of place of catch) in a range of essential measurements/registrations, such as fat content and final product quality. In this study, many of the data reflected quality related to the process, e.g. cutting procedure and marinating procedure, rather than quality related to the raw material. In addition, the methods for final product quality determination did not reflect the true variation of the products. These data were for that reason used to point out what types of measurements were missing or needed to be improved – an informative task, in the sense of process control and process optimisation, to the herring industry.

As pointed out in Paper I some challenges for the future, in respect of process control and process optimization within the herring industry are:

- Development of an information system for usage on board the fishing vessels. Such a system should include important information about the herring. As a minimum, information about data of catch, position of catch, and the time/temperature profile for storage on board should be obtained. If the system is capable of gathering additional information, e.g. size and quality, and is capable of passing this information on to the systems on land, these crucial parameters of information could be transmitted in advance, allowing the production setup to be prearranged, thus saving production time.
- Development of a quality measuring for evaluating the quality of marinated herring. In particularly, this is important if the quality of the final product should be used as a process control parameter.
- Development of an on-line system for measuring fat content on single fish level with subsequent sorting according to determination. As a notice, promising results have been shown for applications of NIR, even though authentic research is still needed.

Hence, an upcoming challenge is to define a well designed traceability system from raw material to final product. This includes identifying and defining measuring points relevant for the process, and finding the right positions for integrating a new on-line/at-line evaluating method to achieve the optimal utilization of the raw material, beneficial to both the fish processing industry and the consumers.

As clearly demonstrated in this study, when investigating the data from the herring industry, some measurements/samples deviated strongly from the major part of the measurements/samples, as a matter of fact, this finding proved to be more the rule than the exception. Such deviating samples, called outliers, may deteriorate the common multivariate models based on a least squares estimation. Whilst huge amount of data are collected, as is often the case in the industry, visual based evaluation and screening for outliers are difficult. Furthermore, there might not be unlimited resources of time

available for analyzing production data. Implementation of robust methods therefore seems a possible alternative to the classical multivariate methods. Different methods of robust PCA, PCR and PLSR exist (Paper II). The practicability of these methods varies, and some can in advance be disqualified for application within industrial use as a result of computational costs, and the missing capability to handle situations with more variables than samples. A majority of examples shown in the literature so far, presenting the advantages of robust methods compared to the classical alternative, exploit data sets with extreme outliers. A remark to that approach is that outliers with such characteristics are also identified using classical methods, truly, a simple outlier warning system may remedy the problem. A recalculation of the model, without the outliers, might be the solution. With this in mind, there is a price to be paid for using robust methods, in particular when looking at the extreme robust methods. Apart from higher computational complexity, robust methods usually also exhibit a lower statistical efficiency and convergence rate. However, a breakdown value of 50 % will rarely be relevant within the industry – with half of the samples being outliers, something tremendous might be wrong in the production. For methods with adjustable breakdown properties, such as ROBPCA and RSIMPLS, a good compromise between robustness and efficiency ought to be obtained.

The study also revealed that robust PCA might be advantageous compared to classical PCA when analysing the entire profile of gas chromatographic data, in the case of suboptimal peak-alignment or other situations where outlying measurements occur, e.g. due to bad baselines or errors in sample amount injected (Paper III). This means that a perfect alignment of the chromatograms is not strictly required to extract useful information from the chromatograms, and thereby the time spent on perfectly aligning the chromatograms might be reduced considerably. What type of robust method, sample or elementwise to choose depends on the type of outliers present in the data set. Situations where only some part of the chromatograms are not properly aligned would benefit the best, using element-wise robust methods, e.g. RSVD. When outliers are due to a specified characteristic throughout the chromatogram, sample-wise robust methods, e.g. ROBPCA, perform the best.

When the occurrence of outliers are systematic, as in the case of Rayleigh scatter in fluorescence data, robust elementwise PARAFAC (*LAE* PARAFAC) turned out not to be a reliable and confident method of handling scatter. However, the systematic nature of scatter can be used constructively for automated scatter identification. Such a method for automatically identifying scatter in fluorescence data using robust techniques is present in Paper IV. A further challenge will be a fully robust procedure able to both identify sample outliers and scatter designed for analysing fluorescence data.

When no extreme outliers are presented in the data set, the advantages of employing robust methods were doubtful. Further research is needed to evaluate the prediction performance of robust models on independent test set. Focusing on the drawbacks of the robust methods, especially the lower statistical efficiency and the time-consuming computations, the improvement of prediction error should be convinced.

The different studies in this project clearly reveal that robust methods in some cases are a good alternative to traditional methods, such as PCA based on least squares estimation, whereas in other cases they are not the complete solution to the problem. A more systematic going through of the advantages and drawbacks of robust methods on more difficult data sets would be interesting. Furthermore, a user-friendly interface is necessary to extend the usage of robust methods, especially to individuals that do not pursue any research. In addition, the time to complete calculations needs to be condensed, before any practical utilisation will take place in the industry.

References

Ammann LP. Robust principal components. *Communications in Statistics – Simulation and Computation* 1989; 18: 857 - 874.

Andersen CM & Bro R. Review, Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. *Journal of Chemometrics* 2003; 17: 200 - 215.

Bahram M, Bro R, Stedmon C & Afkhami. Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation. *Journal of Chemometrics* 2006; 20: 99 - 105.

Baunsgaard D. Factors affecting 3-way modeling (PARAFAC) of fluorescence landscapes. Royal Veterinary and Agricultural University, Department of Dairy and Food technology, Frederiksberg, Denmark, 1999.

Bechmann IE, Jensen HS, Jessen K, Bøknes N, & Nielsen J. Prediction of chemical, physical and sensory data from process parameters for frozen cod using multivariate analysis. *Journal of the Science of the Food and Agriculture* 1998; 78: 329 - 336.

Birkeland S, Sivertsvik M, Nielsen HH, & Skåra T. Effects of brining conditions on weight gain in herring (*Clupea harengus*) fillets. *Journal of Food Science* 2005; 70: 418 - 424.

Bro R, Sidiropoulos ND, & Smilde AK. Maximum likelihood fitting using ordinary least squares algorithms. *Journal of Chemometrics* 2002; 16: 387 - 400.

Bro R. Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis. *Chemometrics and Intelligent Laboratory Systems* 1999; 46: 133 - 147.

Bro R. *Multi-way Analysis in the Food Industry*. PhD thesis, The Royal Veterinary and Agricultural University, Denmark, 1998.

Bro R. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems* 1997; 38: 149 - 171.

Campbell NA. Robust procedures in multivariate analysis I: robust covariance estimation. *Applied Statistics* 1980; 29: 231 - 237.

Carroll JD & Chang J. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika* 1970; 35: 283 - 319.

Charm SE, Learson RJ, Ronsivalli LJ, & Schwartz M. Organoleptic technique predicts refrigeration shelf life of fish. *Food Technology* 1972; 26: 65 - 68.

Chen C. Robust regression and outlier detection with the ROBUSTREG procedure. In

Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference; SAS Institute: Cary, NC, 2002.

Christensen JH, Tomasi G, & Hansen AB. Chemical fingerprinting of petroleum biomarkers using time warping and PCA. Environmental Science and Technology 2005; 39: 255 - 260.

Croux C & Haesbroeck G. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis* 1999; 71: 161 – 190.

Croux C & Haesbroeck G. Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika* 2000; 87: 603 - 618.

Croux C & Ruiz-Gazen A. A fast algorithm for robust principal components based on projection pursuit. In: Compstat: *Proceedings in Computational Statistics*, Pat A (eds) Heidelberg: Physica-Verlag, 1996; 211 - 217.

Croux C, Filzmoser P, Pison G, & Rousseeuw P J. Fitting multiplicative models by robust alternating regressions. *Statistics and Computing* 2003; 13: 23 - 36.

Cummins DJ & Andrews CW. Iteratively reweighted partial least squares: a performance analysis by Monte Carlo simulation. *Journal of Chemometrics* 1995; 9: 489 - 507.

Daigle G & Rivest LP. A robust biplot. *The Canadian Journal of Statistics* 1992; 20: 241-235.

Danish Directorates of Fisheries, Ministry of Food, Agriculture, and Fisheries. http://www.fd.dk [30.12.05].

de Jong S. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems.* 2003; 18: 251-263.

Devlin SJ, Gnanadesikan R, & Kettenring JR. Robust estimation of dispersion matrices and principal components. *Journal of American Statistical Association* 1981; 76: 354 - 362.

Donoho DL & Huber PJ. The notion of breakdown point. In *A Festschrift for Erich Lehmann*. Ed. Bickel PJ, Doksum K, Hodges Jr, Wadsworth, Belmont, CA, 1983.

Donoho DL. Breakdown Properties of Multivariate Location Estimators. *PhD Qualifying paper*, Harvard University, 1982.

EC (European Commission) 2002. Regulation No 178/2002 of the European parliament and of the council of 28 January 2002 laying down the general principles and requirements of food law, establishing the European food safety and laying down

procedures in matter of food safety. Official Journal of the European Communities No. L 31, 01.02.2002, 1-24.

Engelen S & Hubert M. *A robust PARAFAC method*. Katholiek University Leuven, Department of Mathematics, http://www.stat.jyu.fi/icors2005/icorsabstracts/engelen.pdf, 2005b [31.12.05].

Engelen S & Hubert M. *A robust version of principal component analysis: ROBPCA.* ERCIM meeting at the Royal Veterinary and Agricultural University, Denmark, 2005a.

Fox J. *An R and S-PLUS companion to applied regression.* Saga Publications: 2002.

Galpin JS & Hawkins DM. Methods of L1 estimation of a covariance matrix. *Computational Statistics and Data Analysis* 1987; 5: 305 - 319.

Geladi P & Kowalski BR. Partial least-squares regression – A tutorial. *Analytica Chimica Acta* 1986; 185: 1 - 17.

Gil JA & Romera R. On robust partial least squares (PLS) methods. *Journal of Chemometrics* 1998; 12: 365 - 378.

Gnanadesikan R. *Methods for Statistical Data Analysis of Multivariate Observations.* Wiley: New York, 1977.

Griep MI, Wakeling IN, Vankeerberghen P, & Massart DL. Comparison of semirobust and robust partial least squares procedures. *Chemometrics and Intelligent Laboratory Systems* 1995; 29: 37 - 50.

Hampel FR, Ronchetti EM, Rousseeuw PJ, & Stahel WA. *Robust Statistics: The Approach Based on Influence Functions.* Wiley: New York, 1986.

Hampel FR. A general qualitative definition of robustness. *Annals of Mathematical Statistics* 1971; 42: 1887 - 1896.

Hansen P, Ikkala P, & Bjornum M. Holding fresh fish in refrigerated sea water. *Bullettin d' Institute International de Refrigeration* 1970; 50: 299 - 309.

Harshman, RA. *Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi-modal factor analysis*, UCLA Working Papers in Phonetics, 1970; 16: 1 - 84.

Hattula T, Miettinen H, Luoma T, Arvola A, Kettunen J, & Setälä J. Effects of different on-board cooling methods on the microbiological and sensory quality of Baltic herring (*Clupea harengus* L.). *Journal of Aquatic Food Product Technology* 2002; 11; 167 - 175.

Hawkins DM, Liu L, & Young SS. Robust singular value decomposition. *National Institute of Statistical Sciences Technical Report 122,* 2001.

Herborg L. Marinering af sild. Fiskeriministeriets forsøgslaboratorium 1978. (In Danish).

Hotelling H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 1933; 24: 417 - 441, 498 - 520.

Hotelling H. The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology* 1957; 10: 69 - 79.

Huber PJ. Robust estimation of a location parameter. *Annals of Mathematical Statistics* 1964; 35: 73 - 101.

Huber PJ. Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics* 1973; 1: 799 - 821.

Hubert M & Vanden Branden K. Robust methods for partial least squares regression. *Journal of Chemometrics* 2003; 17: 537 - 549.

Hubert M, Rousseeuw PJ, & Van Aelst S. Multivariate outlier detection and robustness. In *Handbook of Statistics, Data mining and Data Visualization*, Elsevier, 2005b; 263 - 302.

Hubert M, Rousseeuw PJ, & Vanden Branden K. http://wis.kuleuven.be/stat/robust.html [8 Aug., 2005c].

Hubert M, Rousseeuw PJ, & Vanden Branden K. ROBPCA: a new approach to robust principal component analysis. *Technometrics* 2005a; 47: 64 - 79.

Hubert M, Rousseeuw PJ, & Verboven S. A fast method for robust principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems* 2002; 60: 101 - 111.

Huss HH. Quality and quality changes in fresh fish. *FAO Fisheries Technical Paper* – 348, 1995.

Iles TD. The duration of maturation stages in herring. *Journal du Conseil Permanent International pour l'Exploration de la Mer* 1964; 29; 166 - 188.

International Organization for Standardization (ISO). Quality management and quality assurance – Vocabulary, 8402, 1994.

Jackson JE. *A user's guide to principal components*. John Wiley & Sons, Inc., 1991.

Jensen AJC. Mængde og vækst af sildeyngel i de danske farvande. *Beretning til Fiskeriministeriet fra den danske biologikse station* 1949; 51: 5 - 16 (In Danish).

Jessen B. Sildehalvkonserves I: Råvare og halvfabrikata. *Fisker-Bladet* 2, 1987 (In Danish).

JiJi RD & Booksh KS. Mitigation of Rayleigh and Raman spectral interferences in multi-way calibration of excitation-emission matrix fluorescence data. *Analytical Chemistry* 2000; 72: 718 - 725.

Johannessen A & Jørgensen T. Stock structure and classification of herring (*Clupea harengus* L.) in the North Sea, Skagerrak/Kattegat and the Western Baltic based on multivariate analysis of morphometric and meristic characters. In *Proceedings of the International Herring Symposium*. Anchorage, Alaska, USA. October 23-25, 1990; 223 - 243.

Karl H & Münkner W. Quality and processing possibilities of Western Baltic Sea spring spawning herring. *Journal of Aquatic Food Product Technology* 2002; 11: 31 - 43.

Karl H, Roepstoff A, Huss HH, & Bloemsma B. Survival of *Anisakis* larvae in marinated herring fillets. *International Journal of Food Science and Technology* 1995; 29: 661 - 670.

Kendall MG. *A course in multivariate analysis*, Griffin, London, 1957.

Kim HM, Fox MS, & Gruninger M. Ontology of Quality for Enterprise Modeling. In *Proceedings of WET-ICE*, Los Albamitos, CA, USA, 1995; 105 - 106.

Kolakowska A, Czerniejewska-Surma L, Gajowiecki L, Lachowicz K, & Zienkowicz L. Effect of fishing season on shelf life of iced Baltic herring. In *Quality Assurance in the Fish Industry*, Huss HH, Jacobsen M, Liston J (eds). Elsevier Science Publishers, Amsterdam, Netherlands, 1992; 81 - 91.

Kourti T, Lee J, & MacGregor JF. Experiences with Industrial Applications of Projection Methods for Multivariate Statistical Process Control. *Computers & Chemical Engineering* 1996; 20: 745 - 750.

Larsen E, Jensen S, & Zappey H. Quality management and measurement in the fish industry. In: *Seafood from producer to consumer, integrated approach to quality. Proceedings of the international seafood conference on the 25th anniversary of the WEFTA*, Luten JB, Børresen T, Oehlenschläger, Noordwijkerhout (eds), The Netherlands, 1997; 403 - 410.

Li G & Chen Z. Projection-Pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo. *Journal of the American Statistical Association* 1985; 80: 759 - 766.

Liang YZ & Kvalheim OM. Robust methods for multivariate analysis – a tutorial review. *Chemometrics and Intelligent Laboratory systems* 1996; 32: 1 - 10.

Liu L, Hawkins DM, & Ghosh S, Young SS. Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy Sciences of the United States of America* 2003; 13167 - 13172.

Maronna RA & Yohai V. Robust estimation of multivariate location and scatter. In *Encyclopedia of statistical science*. Kotz S (eds). Wiley: New York, 1998; 589 - 596.

Maronna RA, Bustos OH, & Yohai VJ. Bias- and efficiency- robustness of general M-estimators for regression with random carriers. In *Smoothing Techniques for Curve Estimation*, Gasser T, Rosenblatt M. (eds). Springer Verlag: New York, 1979; 91-116.

Maronna RA. Robust M-estimators of multivariate location and scatter. *Annals Statistics* 1976; 4: 51 - 67.

Martens H & Næs, T. *Multivariate Calibration*. Wiley; Chichester, 1989.

McKnight DM, Boyer E, Westerhoff P, Doran P, Kulbe T & Andersen DT. Spectrofluorometric characterisation of dissolved organic matter for indication of precursor organic material and aromaticity. *Limnology and Oceanography* 2001; 46: 38 – 48.

McLay R & Pirie R (1971). Development of marinated herring. *Journal of Food Technology* 1971; 6: 29 - 38.

Michaelsen K. Personal communication, 2005.

Munck L, Nørgaard L, Engelsen SB, Bro R, & Andersson CA. Chemometrics in food science – a demonstration of the feasibility of a high exploratory, inductive evaluation strategy of fundamental scientific significance. *Chemometrics and Intelligent Laboratory Systems* 1998; 44: 31 - 60.

Nielsen D, Hyldig G, Nielsen HH, & Nielsen J. Sensory properties of marinated herring (*Clupea harengus*) – Influence of fishing ground and season. *Journal of Aquatic Food Product Technology* 2003; 13: 3 - 24.

Nielsen D, Hyldig G, Nielsen J, & Nielsen HH. Lipid content in herring (*Clupea harengus* L.) – Influence of biological factors and comparison of different methods of analyses: solvent extraction, Fatmeter, NIR and NMR. *LWT Food Science and Technology* 2005; 38: 537 - 548.

Nielsen HH, Bro R, Stefansson G, & Skåra T. Salting and ripening of herring collection and analysis of research results and industrial experience within the Nordic Countries. TemaNord 1999: 578.

Nielsen KN, Guldager HS, Sørensen BS, & Nielsen J. Sensory analysis (Quality Index Method) used as part of a Q-indicator to describe quality changes in thawed cod. Poster at the LMC congress, January 2000.

Nielsen N-PV, Carstensen JM, & Smedsgaard J. Aligning of single and multiple wavelength chromatographic profile for chemometric data analysis using correlation optimized warping. Journal of Chromatography, A 1998; 805: 17 - 35.

Pell PR. Multiple outlier detection for multivariate calibration using robust statistical techniques. *Chemometrics and Intelligent Laboratory Systems* 2000; 52: 87-104.

Podolska M & Horbowy J. Infection of Baltic herring (*Clupea harengus membras*) with *Anisakis simplex* larvae, 1992 – 1999: a statistical analysis using generalized models. *ICES Journal of Marine Science* 2003; 60: 85 - 93.

Rinnan Å & Andersen CM. Handling of first-order Rayleigh scatter in PARAFAC modeling of fluorescence excitation – emission data. *Chemometrics and Intelligent Laboratory Systems* 2005; 76: 91 - 99.

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.

Rivest E & Plante N. L'analyse en composantes principales robuste. *Revue de Statistique Appliqu'ee.* 1988; 36: 54 - 66.

Rosenberg R & Palmén LE. Composition of herring stocks in the Skagerrak-Kattegat and the relations of these stocks with those of the North Sea and adjacent waters. *Fisheries Research* 1982; 1: 83 - 104.

Rousseeuw PJ & Leroy AM. *Robust regression and outlier detection*. Wiley: New York, 1987.

Rousseeuw PJ & Van Driessen, K. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics.* 1999; 41: 212 - 223.

Rousseeuw PJ & Yohai VJ. Robust regression by means of S-estimators. In *Robust and Nonlinear Time series Analysis-(Lecture Notes in Statistics, Volume 26)*; Franke J, Härdle W, Martin RD (eds). Springer Verlag: New York, 1984; 256 - 272.

Rousseeuw PJ & Zomaren B. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* 1990; 85: 633 - 639.

Rousseeuw PJ. Least median of squares regression. *Journal of the American Statistical Association* 1984; 79: 871 - 880.

Ruymgaart FH. A robust principal component analysis. *Journal of Multivariate Analysis* 1981; 11: 485 - 497.

Ryan T. *Modern regression methods.* Wiley: New York, 1997.

Serneels S, Croux C, Filzmoser P, & Van Espen P. http://chemometrix.ua.ac.be/dl/prm.php/[8 Aug., 2005b].

Serneels S, Croux C, Filzmoser P, & Van Espen P. Partial Robust M-Regression. *Chemometrics and Intelligent Laboratory Systems,* 2005a; 79: 55 - 64.

Siegel AF. Robust regression using repeated medians. *Biometrika* 1982; 69: 242 - 244.

Slotte A. Spawning migration of Norwegian spring spawning herring (*Clupea harengus* L.) in relation to population structure. Dr. Scient. Thesis. Department of Fisheries and Marine Biology, University of Bergen, Norway, 1998.

Smilde A, Bro R, & Geladi P. *Multi-way Analysis with Applications in the Chemical Sciences*. Wiley & Sons, England, 2004

Smith JGM, Hardy R, McDonald I, & Templeton J. The storage of herring (*Clupea harengus*) in ice, refrigerated sea water and at ambient temperature. Chemical and sensory assessment. *Journal of the Science of Food and Agriculture* 1980; 31: 375 - 385.

Somers JM. Herring marinades. *Food Progress* 1975; 2: 2, 4.

S-PLUS 6 for Windows Guide to Statistics. Insightful Corporation: Seattle, 2001.

S-PLUS 6 Robust Library User's Guide. Insightful Corporation: Seattle, WA, 2002.

Stahel WA. Robust estimation: infinitesimal optimality and covariance matrix estimators. *PhD Thesis,* ETH, Zürich, 1981.

Thygesen LG, Rinnan Å, Barsberg S, & Møller 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.

Tomasi G, van den Berg F, & Andersson C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometric* 2004; 18:231 - 241.

Vanden Branden K & Hubert M. Robustness properties of a robust PLS regression method. *Analytic Chimica Acta* 2004; 515: 229 - 241.

Verboven S & Hubert M. LIBRA: a MATLAB library for robust analysis. *Chemometrics and Intelligent Laboratory Systems* 2005; 75: 127-136.

Vorobyov SA, Rong Y, Sidiropoulos ND, & Gershman AB. Robust Iterative Fitting of Multilinear Models. www.ece.mcmaster.ca/~vorobyov/tsp_01489tc.pdf [31.12.05]

Wakeling IN & Macfie HJH. A robust PLS procedure. *Journal of Chemometrics* 1992; 6: 189 - 198.

Wentzell P, Nair S & Guy R. Three-way analysis of fluorescence spectra of polycyclic aromatic hydrocarbons with quenching by nitromethane. *Analytical Chemistry* 2001; 73: 1408 – 1415.

Wold S, Esbensen K, & Geladi P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 1987; 2: 37 - 52.

Wold S, Martens H, & Wold H. The multivariate calibration-problem in chemistry solved by the Pls method. *Lecture Notes in Mathematics* 1983; 973: 286 - 293.

Xie Y, Wang J, Liang YZ, Sun L, Song X, & Yu R. Robust principal component analysis by projection pursuit. *Journal of Chemometrics* 1993; 7: 527 - 541.

Zepp R, Sheldon W & Moran M. Dissolved organic fluorophores in southeastern US costal waters: correction method for eliminating Rayleigh and Raman scattering peaks in excitation-emission matrices. *Marine Chemistry* 2004; 89: 15 – 36.

P a p e r I

How to, turn data compilation and traceability in the herring processing industry into a profitable business (viewpoint).

Frosch Møller, S.

Trends in Food Science and Technology (submitted)

(Viewpoint)

How to turn data compilation and traceability in the herring processing industry into a profitable business

Frosch Møller, S.

Danish Institute for Fisheries Research, Department of Seafood Research, The Technical University of Denmark , DK-2800 Kgs.Lyngby, Denmark

Tel.: +45 45 25 49 22. Fax: + 45 45 88 47 74. E-mail: sfr@dfu.min.dk

Abstract

Increased focus on food safety and the occurrence of food scandals has been the driving force for consumers requesting traceability for foods in general, including seafood. New food regulations are now introduced within the EU, enforced by January 2005, which includes the requirement for food traceability.

Implementation of traceability will require investments and higher product costs, but, where effective traceability systems are in place they can also bring extensive benefits to business, when used under proper conditions, for instance for; process control, optimization and better marketing.

Introduction

Increasing demands to food safety imply that from 1st of January 2005, the regulation 178/2002/EC on General Food Law will require traceability to be established at all stages of the food chain. This means that it should be possible to trace and follow a food, feed, food-producing animal or substance through all stages of production, processing and distribution (EC Regulation 178/2002). In commercial practice, traceability systems often usefully include information about what has happened to the food or feed (its processing history) as well as where it came from (backward traceability) and who it was sent to (forward traceability). Traceability is an essential aspect of quality management. Where effective traceability systems are in place they can also bring extensive benefits to business, when used under proper conditions, for instance; process control, optimization and better marketing.

Herring

Herring is important to Denmark as well as to other European countries. In 2003, 114700 tons of herring for consummation was landed in Denmark with a landing value of approx. 28 million Euros. The herring is used for marinated products, but also a significant share is exported semi-manufactured. The very competitive situation in the fish processing industry today means that there is an increased commercial interest in making the production more cost effective and raising the efficiency by rationalizations (Larsen, Jensen & Zappey, 1997). The movement from competition mainly on prices in the early industrialization of food processing to eating quality, food-safety, nutritional value and environmental aspects as very essential parameters has resulted in introduction of quality management systems such as GMP (Good Manufacturing Practice), HACCP (Hazard Analysis Critical Control Points), ISO 9000 (ISO 22000), and TQM (Total Quality Management) in cooperation with new standard measuring methods in the fish processing industries. The introduction of new methodologies for measuring and monitoring throughout the production chain (raw material, intermediate products and final products), results in increasing quantities of available data for use in quality assessment and management. This means that proper handling and analysis of complex data is also required. With the introduction of multivariate data analysis it is now possible for the industry to analyse properly such kinds of data.

Monitoring is an enabling tool for process control and thus helps in either preventing expensive rework or disposing out-of-specification products. To obtain the best possible results, collecting relevant data and a well designed traceability systems is needed. Otherwise industries risk ending up with a large number of useless data which tend to be irrelevant to the process and in the worst case, impossible to connect or trace. Additionally, a well designed traceability system in combination with well defined "target" and quality parameters can be used for process optimization e.g. minimize product variability, maximize yield or minimize the working procedure. A target is an optimal value for a property that is related to and important for the quality of the product under consideration (Næs, 1994).

As implementation of traceability will require investments and higher product costs, it is assumed that it will be the larger companies and retailers that will take the lead in the implementation of seafood traceability. The objective of this Viewpoint is to discuss how the fish processing industries can benefit from traceability imposed by the EU regulation by using the traceability system as a basis for process control and optimization. In the first section to follow, the concept of traceability is described with special emphasis on the current case study. Before one can start controlling or optimizing a process, a great number of aspects have to be considered seriously. A clear definition of a target or an optimal product quality is especially important, a topic dealt with in the next section. In the third, it is illustrated how traceability and collecting of relevant

data can be used to sort raw material with focus on end-point quality in the herring process industry. Finally a discussion of traceability and how to benefit from it is discussed and perspectives outlined.

The article is based on a case-study using data from one of the largest herring industries in Denmark.

Traceability

According to the ISO standard (ISO 8402), traceability can be defined as:

Traceability is the ability to trace the history, application or location of an entity, by recorded identifications.

Product traceability is first of all based on the ability to identify products uniquely. In practice this can be done either by physically marking or by keeping records. Unique identification means according to traceability, that no other unit can have exactly the same, or comparable, characteristics. Unique identification and traceability in any system hinges on the definition of what is the batch size or, using the terminology by Kim, Fox and Gruninger (1995), the Traceable Resource Unit (TRU). The batch size depends on what level (single fish, catch, or production day) it is possible to get specific information from. In some cases different batches are pooled which will create new TRU's.

When considering traceability, two distinct practices enclose the way to keep track of products through the production chain (Moe, 1998). Within a company or

location, the term “internal traceability” relates to the origin of materials, the processing history and finally the distribution of the product after retail delivery. On the other hand, looking at production information from one link in the chain to the next is called “chain traceability”.

In this case study, **Figure 1** illustrates the traceability chain for the production of marinated herring. As can be seen from Figure 1, the chain covers from catching through processing to the final production of semi-manufactured marinated herring. The semi-manufactured marinated herring are sold to another company for final processing before the product is ready to sell, for instance in supermarkets. These last links; company 2, transportation from company 1 to company 2, transportation from company 2 to supermarket, and the storage information from the supermarket were not available for this study but play an important role when analyzing the whole traceability chain.

Considering **Figure 1**, in theory it should be possible to track a single product (lot) back to its catching ground. Due to at least three reasons this does not hold in practice, this owing to; 1) catches from different grounds are mixed onboard the fishing vessel, 2) during offloading, fish from different holds are mixed, and 3) continuous processing means that fish from different vessels can be mixed. It is almost possible to eliminate the problem with mixing of fish from different vessels. This means that the “smallest” TRU for backward traceability in this case is the vessel. During the process the TRU will be split up in different herring sizes, different cuts (e.g. butterfly fillets or fillets without skin), different

marinating recipes, and at last, different lots when packing for sale. This means that it is possible to track a specific lot back to the vessel. The TRU for forward traceability will then be the lot. A special case will be when all catches are from the same catching ground. In such a situation it will be possible to track a lot back to the catching ground.

The amount of available information from a TRU depends on the use, and can range from only the most necessary traceability informations to all know data concerning the particular TRU. Like most other individual food industries, the fish process industries can seldomly sustain information transfer through the whole chain from fishing ground to consumer, but each link has a role to play in collecting and storing information about ingredients, products and processes under their control. In the herring processing industry, the missing information link is often between the vessel and the industry on shore. This information-gap between vessel and industry on land also turned out to be the limiting factor in this study. The industry buys the herring directly from the vessels and as is often the case today, sparse or no information at all follows the herring. The information chain is thus broken. This is a general problem since it nearly happens every time the commodity goes from one link to the next in the chain even though some measurements are rather important for almost all links in the chain e.g. time/temperature and size (kg/fish or No./kg). Weighing for instance is the most widespread measurement in the fish production chain and sometimes as

many as six weightings are done on the same material without any manufacturing being done in between (Larsen et al.,1997).

The overall fish quality continues to decrease from the moment of catch and death due to enzymatic and microbial activity. The rate of spoilage depends highly on storage time and temperature. Regarding shelf life the most important factor is the microbial activity, and the extension of shelf life due to chilling can be explained directly by the temperature influence on the growth of the fish micro flora (Huss, 1996). Data loggers can be used for monitoring the temperature and help preventing undesirable temperature fluctuations and prolonged exposure at elevated temperatures since these conditions stimulate chemical reactions (autolysis) that reduce product flavors, color and texture while allowing further bacterial growth. Data loggers use electrically temperature measurement systems and periodically report the information to a computer and a memory chip inside the logger. By knowing the time/temperature profile from the storage onboard the vessel it is possible to predict the freshness of the herring when landed provided a proper handling after catch (Doyle, 1989). Freshness is a very important parameter since it makes a major contribution to the quality of the final product.

It could be advantageous if information such as the time/temperature profile and weightings from one link in the production chain passes on to the next link, and especially if it is done in advance. This would make it possible to adjust the process according to the specific product and thereby optimize the process.

Target and data collecting

As hinted in the previous section, a well defined traceability and information system can be a valuable advantage to process control and process optimizing. But before one can start controlling or optimizing a process, there are a number of aspects that have to be considered seriously; two are especially important (Næs, 1994). First off all, it is important to have a clear definition of a "target" or an optimal product quality. Secondly, when a clear definition of target or quality is made, a measurement technique for the relevant parameter must be selected, properly installed and calibrated (Næs, 1994).

When dealing with food, the definition of quality is often a very complex task since food is related to many aspects of quality, i.e. sensory, chemical, microbial, physical and nutritional properties. Furthermore, different consumer aspects must be taken into consideration. The quality measurements and measurements in general must cover and reflect the differences in the products. Known differences in the product should be reflected in the relevant measurements, otherwise the measurements are at best misleading. Additionally, there is no meaning to measure irrelevant things, this just creates a lot of useless data and may even mess up the process manageability.

The objective of the measurement is not necessarily directly related to quality, but could also be reflecting e.g. waste and shrinkage. The herring will shrink in the

marinating process due to loss of water. To find the shrinkage percent, the measure is weight before and after marinating, after which the shrinkage has to be calculated from these measurements. Calculating the shrinkage percent can be hindered by material being rejected between the two weightings, if the amount of rejected material is not measured. When calculating values, it is important that the differences between the measurements only reflect the information of specific interest; otherwise the measurement is useless. This is one reason for measuring the amount of waste. Otherwise it is impossible to distinguish between what is due to shrinkage and what is due to waste.

The reason for rejecting material in general is also extremely important since this information can be used to improve the production and avoid the same failure to happen again in the future. In the best case it could also help the company to identify catching grounds e.g. subjected to contamination, provided that the traceability chain is complete. When analyzing data from several years, patterns might appear showing that at a specific time of the year herring are less qualified for marinating or meeting certain product specifications. These herring might thus profitably be used for other products, and the amount of discarded material reduced.

Clear identifications and definitions of target or quality points relevant for the process are necessary to set up practical specifications and guidelines that can be used for taking relevant measurements, which in combination with a well designed traceability system later can be used for process control and process optimization.

Focus so far has been on measurements for specific, dedicated purposes as this is practice today in the herring industry. Even so, it might be beneficial to combine all available data from different measurements throughout the production chain (from catch to final product) in order to extract even more relevant information from the collected data by multivariate data analysis. **Table 1** illustrates how such a data matrix might look. All information belonging to a TRU is arranged in one column, and each measurement makes up one row. When colleting data exceptional conditions should also be registered. Such information can be helpful to understand otherwise unexplainable alterations in the data, or prevent inconvenient conditions to happen again.

Sorting raw materials with focus on end-product properties

This section will focus on how traceability and collecting of relevant data can be used to sort raw material with focus on end-point quality in the herring process industry.

The objective of the sorting procedure is to identify "functional" groups in the raw material and to find corresponding optimal process conditions for each group. The number of groups is determined from how many categories it is practical to process differently and by the expected overall end-product quality. Since the process can be adjusted according to the raw material quality in each category, all

raw materials can be better utilized and both quality and stability of the end product can be improved. Processes that can be improved by sorting are typically processes with much variation in raw material quality and where raw materials of different qualities have different optimal levels of the process variables.

Today the herring are sorted according to size into e.g. four categories. Sorting into such more homogeneous categories with respect to size lead to a simpler and more stable control of the process, because it can be run on one category at a time, with optimal settings for each category.

A further sorting criterion, beyond the allready existing size sorting, could be sorting according to fat content since the fat content vary considerable within a catch (**Figure 2**). This finding is in accordance with other studies (Nielsen, Hyldig, Nielsen & Nielsen, 2005; Larsen et al. 1997). Nielsen et al. (2005) showed that herring size and maturity status could not be used to sort herring according to fat content. **Figure 2** shows the results from fat measurements on four commercial catches in Denmark. For each catch the fat content is measured on approx. 50 herring. The overall variation in fat content between catches is due to season variation. Today's practice is to calculate the fat content as an average value on a pooled sample of e.g. 20 minced herring due to limited time and equipment. One value appears for each cut and the calculation is based on dry matter. Hence, the variation *within* the cut is not revealed and the fat content declared to the customer is imprecise. Instead, the optimal solution is on-line

measuring of fat content and subsequent sorting. This will enable more homogeneous products according to fat content.

If a connection between fat content and shrinkage content exists, herring with a high fat content having a higher shrinking percent than herring with lower fat contents, a sorting system would be beneficial for the industry since herring with high fat content could be sorted out for other products of higher value. Furthermore, a connection between fat content and quality of final product might also be motivating for sorting raw material according to fat content.

In all cases the perspective is the development of equipment for rapid and non-destructive measurement of fat content for whole herring or herring fillets, and a subsequent individual sorting device so that the process may be adjusted appropriately.

Discussion

The implementation of a well prepared traceability system is not only about technique, it is a time-consuming and expensive process which involves all levels of personal in the company. Before starting the process of implementing traceability, the company needs readiness, an implementation plan and above all a business plan. When implemented, optimal benefits from quality control, production control in fulfilling consumer demands et cetera should be achieved resulting in a more effective production and, confidently reduced costs. It is important to involve all kind of personal from production staff to the managers,

and benefit from their experience and knowledge. To obtain the best possible solution the implementation should be a common goal and all should benefit from the information technology introduced.

An efficient traceability system makes it possible for all links in the chain to fast and effective recall defective products. The amount of recalled material depends on the TRU size. This means that an extensive traceability level makes it possible to recall only very small quantities. In the herring industry the lowest level of traceability - the smallest TRU – is on lot level, this means that in the "best" case it is only necessary to recall a lot.

Traceability can also be used for marketing. With an identifier on the product the consumer should be able to get information about the product history straight back to catching ground to final product by entering the identifier on a computer. This opportunity will make the consumer able to choose between similar products.

Combining data from different measurements throughout the production chain (raw material measurements to final product quality) in combination with multivariate data analysis (chemometrics) of data-structures as illustrated in **Table 1** may give additional information about unknown relations or scientifically explore relations that have previously only existed as "experienced personnel knowledge". Unfortunately, lack of variability in a range of crucial

measurements/data implies that the historical data are not suitable for further multivariate data analysis in the present case study. This is often the case when using historical data since historical data probably have been collected for other and maybe now irrelevant reasons and therefore are not relevant or applicable to multivariate data analysis. Instead, analysis of the historical data can hint at which measurements are missing and which data need to be improved to be sufficiently informative.

Process control and process optimization can benefit from a well designed traceability system, under the condition that data relevant for the process are measured and that the measurements reflect the variability in the products. A quality system where all end products are giving essentially the same quality grading is useless, and may even destroy utilization of other measurements since it is not possible to say anything about e.g. how the raw material qualities affect the end product quality.

Traceability systems may be used as foundation for process control e.g. by differentiating the raw material according to size and/or fat content. Fat and lean herring are by example turned into different end products, or fat herring and lean herring require different marinating time. From the traceability system one knows which category of herring that is under process presently running when the herring enter the marinating section.

In contrast to process control where the goal is to keep the process running, process optimization optimizes the process to perform optimal with e.g. the best possible yield of a given process. Thus, the combination of a well defined quality parameter and a traceability system convey an improved approach to reach the target of a given production.

Implementation of traceability in the processing industries can be done according to **Table 2**. Phase 3 and 6 are not necessary but by adding those, a huge benefit concerning process control and optimization can be drawn from the traceability system.

Perspectives

A traceability system onboard the vessel containing important informations about the herring (e.g. date of catch and catch ground) and storage conditions (time/temperature profile) capable to communicate with the systems on land will be of great benefit for the herring industry, especially if the informations are sent to the industry in advance. This will make the industry able to prearrange the production line and thus save production time.

Furthermore a well-developed quality measurement for final marinated herring is needed, in particular if the quality of the final product should be used as process control parameter.

Hence, the challenge of the future is to define a well designed traceability system from raw material to final product, identify and define target or quality points relevant for the process and find the right integration of new on-line/at-line measuring methods to achieve the optimal use of the raw material to benefit of both the fish processing industry and the consumers.

Acknowledgement

This work was supported by the Danish Ministry of Food, Agricultural and Fisheries. Dr. Bo M. Jørgensen is gratefully acknowledged for valuable advice during preparation of the manuscript.

References

Doyle, J. P. (1989). Seafood Shelf Life as a Function of Temperature. Alaska Sea-Grant, No. 30. Marine Advisory Program, University of Alaska, Fairbanks.

EC (European Commission) (2002). Regulation No 178/2002 of the European parliament and of the council of 28 January 2002 laying down the general principles and requirements of food law, establishing the European food safety and laying down procedures in matter of food safety. *Official Journal of the European Communities No. L. 31, 01.01.2002*, pp. 1 – 24.

Huss, H. H. (1995). Quality and quality changes in fresh fish. FAO Fisheries Technical Paper, Vol. 348, pp. 1 – 195.

International Organization for Standardization (ISO) (1994). Quality management and quality assurance – Vocabulary.

Kim, H. M., Fox, M. S. & Gruninger, M. (1995). Ontology of Quality for Enterprise Modelling. In Proceedings of WET-ICE (pp. 105 – 106), Los Albamitos, CA, USA, IEEE.

Larsen, E., Jensen, S. & Zappey, H. (1997). Quality management and measurement in the fish industry. In Luten, JB, Børresen, T. & J. Oehlenschläger,

Seafood from producer to consumer, integrated approach to quality. Proceedings of the international seafood conference on the 25th anniversary of the WEFTA (pp. 403 – 410). Noordwijkerhout, The Netherlands.

Moe, T, (1998). Perspectives on traceability in food manufacture. *Trends in Food Science & Technology*, 9, pp. 211 – 214.

Nielsen, D., Hyldig, G., Nielsen, H.H., Nielsen, J. (2005): Lipid content in herring (Clupea harengus L.) – Influence of biological factors and comparison of different methods of analyses: solvent extraction, Fatmeter, NIR and NMR. *Food Science Technology on-line december 2004.*

Næs, T, (1994). Information technology and process control. *World of Ingredients*, pp. 18 – 21.

Legends to figures

Figure 1.The traceability chain for production of semi-manufactured marinated herring. The bold line illustrates the straight production line whereas the dotted arrows illustrate the complexity within the production due to various herring sizes and different customer demands. The ellipses indicate a complete production line.

Figure 2. Fat content measured in four commercial catches.

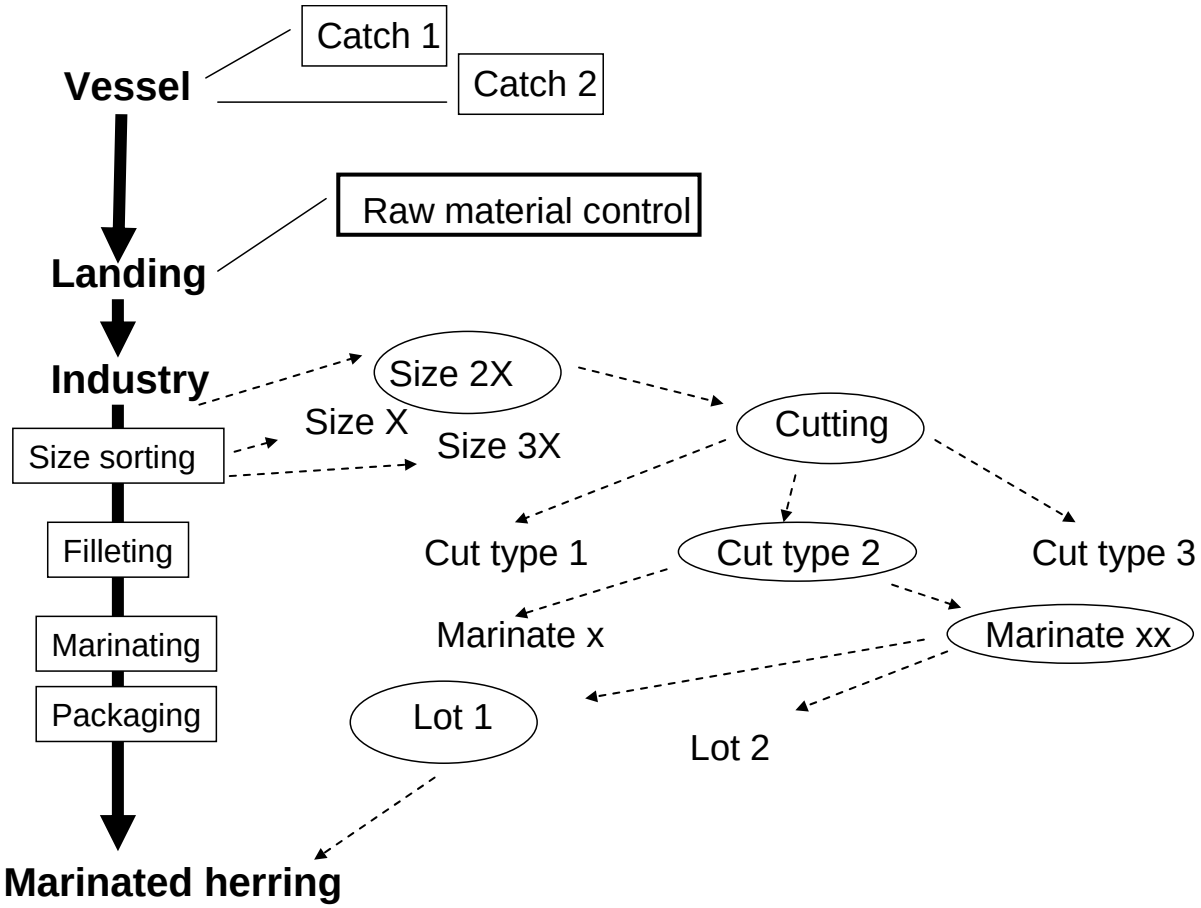
Catch 1
Catch 2
Vessel
Raw material control
Landing
Industry
Size 2X
Size X
Size 3X
Size sorting
Cutting
Filleting
Cut type 1
Cut type 2
Cut type 3
Marinating
Marinate x
Marinate xx
Packaging
Lot 1
Lot 2
Marinated herring

Figure 1.

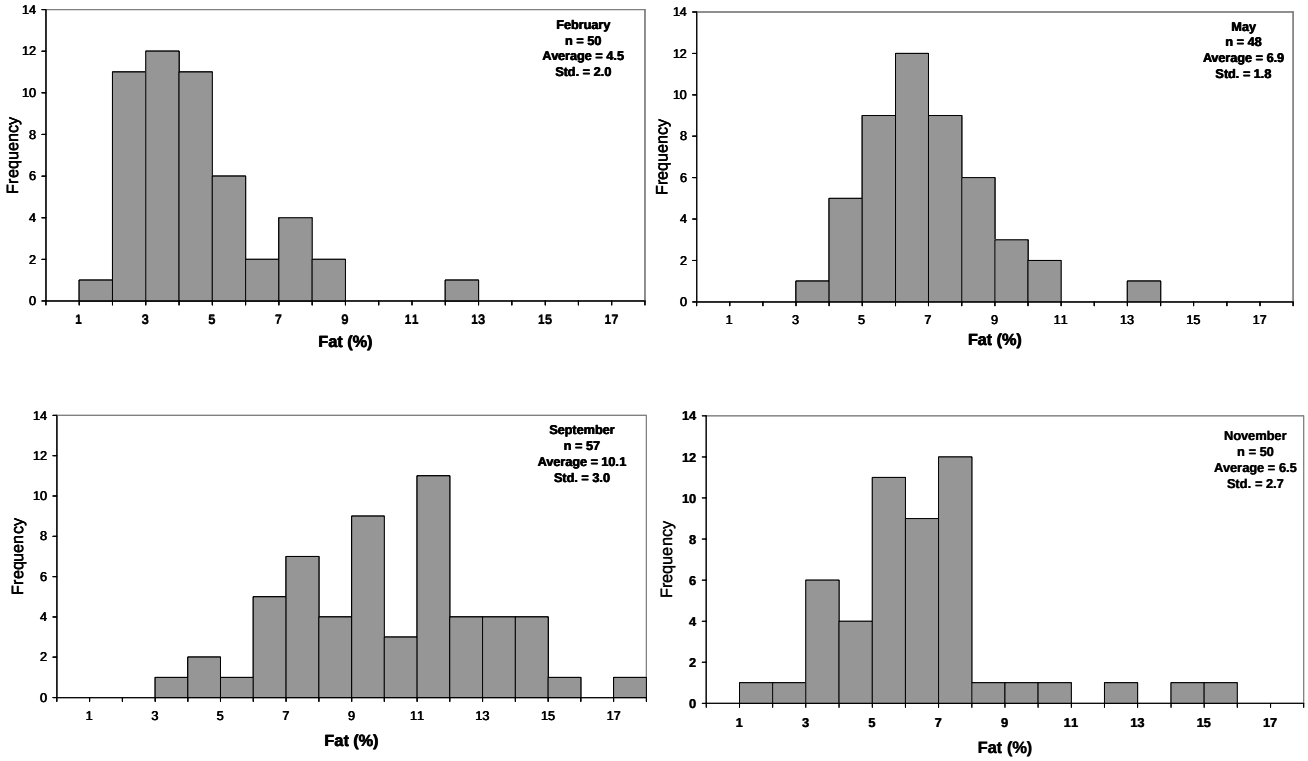
February
n = 50
Average = 4.5
Std. = 2.0
May
n = 48
Average = 6.9
Std. = 1.8
September
n = 57
Average = 10.1
Std. = 3.0
November
n = 50
Average = 6.5
Std. = 2.7
Frequency
Fat (%)

Figure 2.

Table 1. Schematic data set structure suitable for multivariate data analysis of traceability data.

Each column contains measurements on the specified TRU.

Information / measurement		**TRU number**								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Vessel:	Catching ground									
	Date of catch									
	Date of landing									
	Storage temperature									
	Comments (e.g. bad weather)									
Production:	Raw material quality									
	Size									
	Fat content									
	Nematodes									
	Cutting yield									
	Weight before marinating									
	Weight after marinating									
	Product quality									
	Comments									

Table 2. Implementation plan for a traceability system directed towards food production.

Phases shown in italic are not mandatory but very advantageous.

Phase number:	**Action:**
1	Analyzing the production chain
2	Define the traceability level
3	*Define targets and or optimal quality and be sure that measurement technique for the relevant parameter exist*
4	Programming
5	Implementation
6	*Calibration and validation of the system*
7	Maintenance of the system

P a p e r I I

Robust methods for multivariate data analysis (review article).

Frosch Møller, S., von Frese, J. and Bro, R.

Journal of Chemometrics, 19: 549 – 563, 2005

JOURNAL OF CHEMOMETRICS
J. Chemometrics 2005; **19**: 549–563
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.962

Robust methods for multivariate data analysis

S. Frosch Møller[1]*, J. von Frese[2] and R. Bro[2]

[1]Department of Seafood Research, Danish Institute for Fisheries Research, The Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
[2]Spectroscopy and Chemometrics Group, Quality and Technology, Department of Food Science, The Royal Veterinary and Agricultural University, DK-1958 Frederiksberg C, Denmark

Received 24 February 2005; Revised 06 December 2005; Accepted 27 January 2006

Outliers may hamper proper classical multivariate analysis, and lead to incorrect conclusions. To remedy the problem of outliers, robust methods are developed in statistics and chemometrics. Robust methods reduce or remove the effect of outlying data points and allow the 'good' data to primarily determine the result. This article reviews the most commonly used robust multivariate regression and exploratory methods that have appeared since 1996 in the field of chemometrics. Special emphasis is put on the robust versions of chemometric standard tools like PCA and PLS and the corresponding robust estimates of regression, location and scatter on which they are based. Copyright © 2006 John Wiley & Sons, Ltd.

KEYWORDS: outliers, robust estimation, PCA, PCR, PLS

1. INTRODUCTION

Outliers are observations that appear to break the pattern or grouping shown by the majority of the observations. Presence of outliers is more the rule than the exception for real world data. Many branches of chemometrics in both industry and research work with huge amounts of data, which makes visually based evaluation and screening for outliers difficult. The reasons for outliers are various, for example instrument failure, non-representative sampling, formatting errors and objects stemming from other populations. Usually, only complete objects ($\mathbf{x}_{i.}$) are considered as outliers, but it is equally relevant to look for outliers in variables ($\mathbf{x}_{.j}$) and even individual data elements (x_{ij}). Most conventional multivariate methods are sensitive to outliers due to the fact that they are based on least squares (LS) or similar criteria where even *one* outlier can have an arbitrarily large effect on the estimate and deteriorate the model. Therefore, it is necessary to (1) identify outliers and (2) decide whether the outliers should be accommodated or rejected in the modeling process.

The aim of any robust method is to reduce or remove the effect of outlying data points and allow the remainder to predominantly determine the results. Robust methods are helpful for both semi-automated detection of outliers by looking at the robust residuals and model building. When no outliers are present in the data set, the result from a robust method should be consistent with the result from the corresponding non-robust method. Robust methods provide a powerful methodology extending a conventional 'manual' analysis and elimination of outliers by using exploratory methods and 'conventional' outlier diagnostics.

Rousseeuw and Leroy [1] presented an overview of robust estimates in regression and outlier detection, and Maronna and Yohai [2] described recent advances in robust estimation in multivariate location and scatter estimation. Much focus has been put on making the common chemometric techniques such as principal component analysis (PCA), principal component regression (PCR) and partial least squares (PLS) regression more robust against outliers, using robust estimates to replace the non-robust LS estimate. Reference [3] holds a review of the robust methods for multivariate analysis until 1996. An overview of the recently developed methods for multivariate data analysis, based on the minimum covariance determinant and least trimmed squares estimators for location, scatter and regression, together with a detailed description of these estimators, can be found in Reference [4].

The aim of this paper is to present an overview of the most common robust chemometric methods, that is PCA, PCR and PLS, described in the literature as many new methods have emerged subsequently.

In Section 2 outliers and their effect on least squares estimation will be discussed. Section 3 introduces the robust estimates for regression, location and covariance used in the robust multivariate methods discussed in Section 4. Section 5 contains comments on software availability. Finally, Section 6 presents a discussion on the use of robust methods.

*Correspondence to: S. F. Møller, Department of Seafood Research, Danish Institute for Fisheries Research, The Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark.
E-mail: sfr@dfu.min.dk
Contract/grant sponsor: Danish Ministry of Food, Agriculture and Fisheries.

Copyright © 2006 John Wiley & Sons, Ltd.

2. OUTLIERS

Different types of outliers can be discerned. Taking regression models as an example where the independent variables are denoted as $\mathbf{X}_{n,p}$ (n stands for the number of objects ($i = 1, \ldots n$) and p for the original number of variables ($j = 1, \ldots, p$)) and the q dependent variables are denoted as $\mathbf{Y}_{n,q}$, the following categories of outliers can be considered: (1) 'Good' leverage points which are observations isolated from the major part of the observations in the data matrix **X** but following the same regression model, (2) 'Bad' leverage points which in addition to being isolated from the major part of **X** deviate strongly from the regression model defined by the other observations and (3) Outliers that are not leverage points but have large y prediction residuals in calibration and are therefore referred to as high y residual outliers or vertical outliers. In robust analysis, the good leverage points are usually not denoted as outliers as they are not detrimental to the regression model but merely reflect an 'unfortunate design'. These three types of outliers can occur both during model fitting and during predictions with a previously established model.

Figure 1 shows a scatterplot of 10 points, $(x_1, y_1), \ldots (x_{10}, y_{10})$, with no outliers presented. The LS solution fits the data very well, and for data with normally distributed random noise without outliers, the LS solution is in fact optimal in the maximum likelihood sense.

Why LS is not resistant to outliers follows from the properties of the objective function for LS procedures. The objective function to be minimized is the sum of the squared residuals

$$\underset{\hat{\beta}}{\text{Minimize}} \sum_{i=1}^{n} r_i^2 \tag{1}$$

in which the residuals r_i are given by

$$r_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip} \tag{2}$$

where y_i $(i, \ldots, n)$ are the corresponding values of the dependent variables, x_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, p)$ the values of the explanatory variables, and $\hat{\beta}_j = (j = 1, \ldots, p)$ is the LS

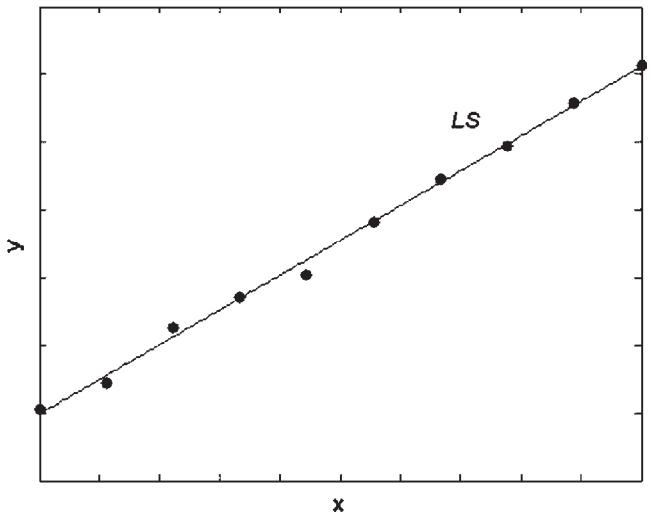

Figure 1. Scatterplot of 10 points, $(x_1, y_1), \ldots (x_{10}, y_{10})$ and the LS regression line.

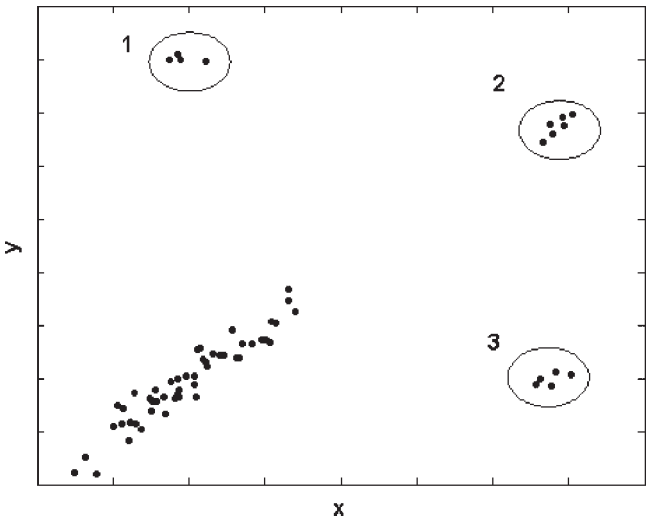

Figure 2. High y residual outliers (1) and leverage points (Good leverage points are denoted '2' and bad leverage points are denoted '3').

estimate of the parameters. This means that a large outlier will exert an inappropriately large influence on the LS-estimate as will be illustrated in the following:

Figure 2 illustrate the three outlier types where high y residual observations are marked with a '1', '2' represent good leverage points and bad leverage points are marked with '3'. Both the high y residual outliers and the bad leverage points affect the calibration model by distorting the least squares model to a certain degree and should be eliminated. From a purely experimental point of view an observation with a high y residual does not necessarily indicate a corrupted or deviating y measurement, we only know that the corresponding (x, y) pair is inconsistent with the remainder (although from a pure theoretical viewpoint the explanatory variable x is considered to be error-free). In multivariate regression models (e.g. PCR or PLS) it might be possible to assign such an outlier as originating from **X** in case it shows a strong deviation from the **X**-model. Generally, outliers are not necessarily bad measurements but could also indicate samples belonging to another group than the majority of the data.

As noted by Gnanadesikan [5], the consequence of outliers in multivariate data is intrinsically more complex than in the univariate case. A multivariate outlier can distort measures of location and scale and thereby also those of covariance structure. As a result the modeling methods may describe the shape of the majority of the data incorrectly and conclusions drawn can be misleading. An added complication is that it is much more difficult to identify multivariate outliers. A single univariate outlier may easily be detected graphically, which is not that straightforward in higher dimensions. Many multivariate methods work well for identifying single outliers, but when there are many outliers, masking and swamping effects may occur. The masking effect means that some outliers are unnoticed because the presence of other outliers masks their bad influence [6,7]. The swamping effect consists of wrongly identifying/diagnosing

Copyright © 2006 John Wiley & Sons, Ltd.

an observation as an outlier because of the presence of other outliers [8].

According to Rousseeuw and Leroy [1] outlier diagnostics and robust methods have the same goal just in opposite order: In standard or non-robust diagnostic approaches the outliers are first identified and then the remaining data analyzed with a non-robust LS criterion. In robust methods, models are fitted to the majority of the data and outliers are identified as those observations with large residuals from the robust fit. A survey of diagnostic techniques can be found in Reference [1].

3. BASIC ROBUST STATISTICS

To enable the comparison of different robust methods in various situations, measures of performance are necessary. One such performance measure for robust methods is the breakdown point ε^* [9] which can loosely be defined [10] as the smallest fraction of samples (with respect to n) that can render the estimator useless.[1] This might be given depending on the sample number (e.g. $1/n$) or as a limiting value for $n \to \infty$ (e.g. 0%). A breakdown point of zero for an estimator means that the presence of even a single outlier can completely distort the model. One such example is the LS function whose breakdown point is zero. Breakdown points vary considerably for different classes of estimators with 50% as the highest possible for the equivariant estimators discussed in this review. Conceptually, it becomes impossible to distinguish between the good and the bad parts of the data if the fraction of outliers becomes larger than 50%. Estimators with $\varepsilon^* = 50\%$ are called high breakdown point estimators. Another essential performance measure is the influence function introduced by Hampel *et al*. [8]. The influence function tries to quantify the influence an infinitesimal outlier has on the estimate. Thus, in principle this allows for a more detailed quantitative comparison of different robust methods under a single outlier. A fundamental question here is if the influence function is bounded, that is if already a single outlier can lead to a breakdown of the estimator. For assessing the influence function, distributional assumptions for the data have to be made. This often renders the analysis more intricate and might necessitate empirical comparisons with unknown general validity in particular for $n \ll p$ (e.g. [11]).

Efficiency is another important concept for the discussion of robust estimates. The relative statistical efficiency is the ratio of the mean square error from a robust estimator to the mean square error from an ordinary LS estimator when applied to data from an uncontaminated distribution, for example with normally distributed errors [6].

Equivariance properties are also important for understanding estimators. Equivariance means that a systematic transformation of the data will cause a corresponding transformation of the estimator [1]. Three types of equivariance exist for regression estimators; (1) regression-, (2) scale and (3) affine equivariance. Regression equivariance means that any (additional) linear dependence $\mathbf{Y} \to \mathbf{Y} + \mathbf{Xv}$ should be reflected in the regression vector accordingly $\mathbf{b} \to \mathbf{b} + \mathbf{v}$. This corresponds to translation equivariance in the case of location estimation. Scale invariance implies that the fit is essentially independent of the choice of measurement units for the response variable $\mathbf{y}$ and for any scaling $\mathbf{y} \to c\mathbf{y}$ the regression vector scales appropriately $\mathbf{b} \to c\mathbf{b}$. Affine equivariance means that for linearly transformed data $\mathbf{X} \to \mathbf{XA}$ the estimate of the regression vector transforms correspondingly $\mathbf{b} \to \mathbf{A}^{-1}\mathbf{b}$, such that the predicted values and residuals are invariant under this transformation.

A weaker condition than affine equivariance is the orthogonal equivariance, which means that any orthogonal transformation of the data (rotation and reflection) transforms the estimator properly. Orthogonal equivariance is sufficient in the context of PCA or PLS since even the classical procedures are only orthogonally equivariant [12]. For a detailed description of the different equivariance criteria the reader is referred to Reference [1].

Multiple linear regression, as well as estimation of sample mean and covariance are the cornerstones of multivariate data analysis methods such as PCA, PCR and PLS [1,2]. The former underlying techniques are not resistant to outliers as they are based on LS techniques and robustifying them is often the basis for obtaining robust versions of the latter multivariate data analysis methods. The focus throughout this section is restricted to robust multivariate regression estimators and robust estimates of multivariate location and covariance used in the multivariate methods described afterwards in this paper.

Overviews of the important robust multidimensional estimators for regression and location and scatter are listed in Tables I and II, respectively.

3.1. Robust multivariate regression estimates

3.1.1. *Multiple linear regression*

In multiple linear regression (MLR), the response variable y_i is regressed on p explanatory variables $(x_{i1}, \ldots, x_{ip})$ in the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{3}$$

with errors ε_i. As in univariate regression, the LS estimator of $\beta_0, \beta_1, \ldots, \beta_p$, which corresponds to minimizing the squared residuals (Equation (1), is quite sensitive to the presence of outlying points. A robust alternative to LS estimator is therefore needed.

3.1.2. *M-estimates*

The methods collected under the term M-estimators (maximum likelihood type estimators), first introduced by Huber [13,14], replace the squared residuals in Equation (1) by another function of the residuals

$$\underset{\beta}{\text{Minimize}} \sum_{i=1}^{n} \rho\left(r_i/s\right) \tag{4}$$

The function ρ is symmetric (i.e. $\rho(-t) = \rho(t)$) for all t with a unique minimum at zero, r_i is the residual of the ith observation and s is a suitable estimate of the scale obtained from the residuals. For $\rho(t) = t^2$ and $s = 1$ one obtains the LS estimator. Different choices of $\rho(t)$ correspond to assuming

[1] Robust estimates are marked by asterisk (*) throughout the article.

Copyright © 2006 John Wiley & Sons, Ltd.

Table I. Overview of the most important regression estimators in robust multivariate data analysis (*n*, sample size; *p*, number of regressors)

Estimator	ε^*	Comments	Affine equivariant	Reference
M	0%	Not robust against **X**-outliers	Yes	[13,14]
GM	$\leq$30%, decreases as p increases	Robust with respect to outliers in **X** as well as outliers in **Y**	Yes	[1]
Siegels repeated median	50%		No	[15]
LMS[a]	50%		Yes	[16]
LTS	50%		Yes	[16]
S	50%		Yes	[17]
MM	50%		Yes	[18,19]

Abbreviations are explained in the text.
[a] slow convergence ($\propto n^{-1/3}$).

Table II. Overview of the most important robust estimators for multivariate location and scatter in robust multivariate data analysis (*n*, sample size; *p*, number of regressors)

Estimator	ε^*	Comments	Equivariance properties	Reference
M	$\leq 1/(p+1)$	Not all M-estimates are affine equivariant, for example L_1 Limitation: $n > p$	Method dependent	[20]
MVT	50%	Converge quickly Limitation: $n > p$ after trimming	Affine	[21,22]
Stahel–Donoho	50% for n larger than $2p+1$	High computational cost	Affine	[23,24]
MVE	50%	Converges slowly ($n^{-1/3}$) Limitation: $n > p$	Affine	[16]
MCD	50%	Limitation: $n > p$ A fast algorithm exists (FAST–MCD)	Affine	[16]
S	50%	Limitation: $n > p$	Affine	[25,26]
MM	50%	Limitation: $n > p$	Affine	[27]

Abbreviations are explained in the text.

specific distributions for the errors (e.g. $\rho(t) = |t|$ for double exponentially distributed errors) [28].

M-estimator calculations can be computed by means of iteratively reweighted least squares (IRLS). The principle of IRLS is to obtain a weight for each observation depending on the size of the regression residual. Such weights make it possible to bound the effect of outliers on the final model.

The scale factor is necessary to achieve affine equivariance for the estimators. A possible scale estimate for the residuals would be the standard deviation. However, the standard deviation is not robust to outliers and therefore not suitable in this context. The median absolute deviation (MAD) is a commonly used robust alternative to the standard deviation [1]. Rousseeuw and Croux [29] proposed the Q_n scale estimate as an alternative to the MAD. The Q_n estimator, motivated by the Hodges–Lehman estimator [30], is for a univariate data set $(z_1, \ldots, z_n)$ defined as the first quartile of the pairwise differences between the data $Q_n = 2.2219 \cdot d \cdot \{|z_i - z_j|; i \leq j\}$ where d is a small sample correction factor (approaching 1 for increasing n) [29]. When comparing the efficiency of several robust scale estimators Rousseeuw and Croux [29] found that the Q_n estimator yielded better results than the MAD. Unfortunately, the M-estimators can be strongly influenced by any high leverage point and thus have a break down point ε^* of 0% [1].

3.1.3. GM-estimates

Generalized M-estimators (GM-estimators) frequently referred to as 'bounded influence estimators' were developed to overcome the x-outlier problem of M-estimators [1] and thereby improve the breakdown point.

The basic purpose of these methods is to bound the influence of outlying x_i by means of some weights w_i that give full influence to observations assumed to come from the main body of the data, but reduced weight or influence to outlying observations. Starting with a sensible estimate the iterative procedure will continue until the sequence of estimates has converged to within the desired accuracy. The ε^* of all GM-estimators can be no better than a certain value, in general not larger than 30%, decreasing as a function of the dimension p [17,31].

A number of weights have been proposed, for example Tukey's biweight [32], Huber [13,14], Hampel [33] and Andrew's wave [33,34]. These weights are not restricted to GM-estimators but can be used for all kind of estimates requiring a weight function. Plots of selected weight functions are illustrated in Figure 3.

3.1.4. Siegel's repeated median

The first high breakdown point regression estimator was the repeated median (RM) proposed by Siegel [15]. It is based on

Copyright © 2006 John Wiley & Sons, Ltd.

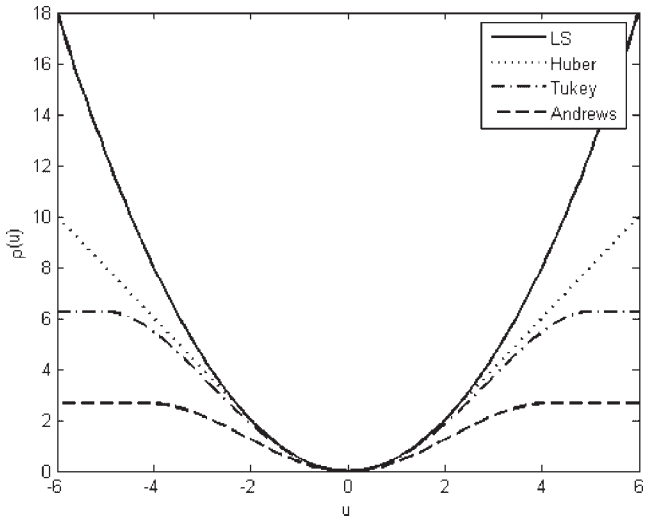

Figure 3. Objective function of the LS, Huber weights, Tukey's biweight and Andrew's wave.

calculating perfectly fitting models for all possible data subsets of size p and obtaining a final regression model through a nested coordinatewise median calculation (see Reference [15]).

As the explicit calculation of the RM would involve the consideration of all possible subsets of p points, where p is the number of variables, the resulting computational complexity of n^p [15] would mean a prohibitive amount of calculation time even for moderate p [1]. Additionally, this estimator is not affine equivariant for linear transformations of $\mathbf{x}_i$ [1].

3.1.5. *Least median of squares*

Replacing the sum of the squared residuals in Equation (1) with the robust median yields one of the most well-known instances of a high breakdown point estimator, the least median of squares (LMS) method of Hampel and Rousseeuw [16], defined by

$$\underset{\hat{\beta}}{\text{Minimize}} \quad \underset{i=1,\ldots,n}{\text{median}}\ r_i^2 \tag{5}$$

The LMS estimator has an ε^* of 50%, and is robust with respect to outliers in $\mathbf{y}$ as well as outliers in $\mathbf{X}$. Unfortunately, the LMS has a very low efficiency (converges like $n^{-1/3}$) [16].

3.1.6. *Least Trimmed Squares*

To overcome the poor convergence rate for LMS, Rousseeuw [16] proposed the least trimmed squares (LTS) estimator

$$\underset{\hat{\beta}}{\text{Minimize}} \sum_{i=1}^{h} (r^2)_{i:n} \tag{6}$$

where $(r^2)_{1:n} \leq \ldots \leq (r^2)_{n:n}$ are the ordered squared residuals that is the sum of squared residuals is formed over a suitable lower quantile h/n of the residuals. The fraction of included samples can be as low $h \approx {}^{n}/_{2}$. The LTS objective is equivalent to LS, with the exception that the largest residuals are not used in the sum, so that outliers are disregarded. Taking $h = n$ yields the LS estimator. The highest possible breakdown value (50%) for LTS is attained when $h \approx n/2$. For general h, the breakdown is $(n - h + 1)/n$. The LTS converges like $n^{-1/2}$ and behaves satisfactorily with respect to asymptotic efficiency [1]. The disadvantage of LTS is the sorting of the squared residuals, which blows up an already significant computation time [16]. A fast algorithm, FAST–LTS, for computing the LTS estimators was developed by Rousseeuw and van Driessen [35]. For small data sets the exact LTS is found whereas for larger data set the new algorithm gives more accurate estimates than existing LTS algorithms. A general problem of the FAST–LTS and other approximate algorithms with a given number of starting trial sets consists in a lack of consistency for larger and larger training data sets as pointed out by Hawkins and Olive [36]. As possible solution a specific prior clustering has been suggested in [36] and FAST–LTS already makes uses of a more sophisticated empirical subsampling scheme for larger datasets but further research is needed [37].

The ε^* for both LTS and LMS is independent of p, the number of variables and they also satisfy all three equivariance properties mentioned earlier [1].

Rousseeuw and Yohai [17] generalized the LMS and LTS estimators to S-estimators. The class of S-estimators corresponds to replacing the scale of the residuals in Equation (4) by a robust measure that minimizes the dispersion of the residuals. S-estimators are regression-, scale- and affine equivariant and possess a convergence rate $n^{-1/2}$. The ε^* can attain 50% with a suitable choice of the constants involved, and in contrast to GM-estimators, S-estimators have a high breakdown point for any dimensionality. But S-estimators cannot simultaneously achieve high efficiency and a high breakdown point [1,17]. If a 50% breakdown point is imposed, the asymptotic Gaussian efficiency of S-estimators is at most 33% [38].

Computing the exact S-estimator is often not feasible, and may present difficult problems due to the existence of many local minima [18]. The computational cost of methods based on subsampling increases exponentially with p, and makes these estimates very costly for high dimensions.

To circumvent the efficiency problem of S-estimators Yohai introduced MM-estimators [18], which are basically efficient M-estimators using the result of an S-estimator as starting value and for obtaining a robust auxiliary scale estimate and thereby obtaining improved robustness with an ε^* up to 50%. Thus, MM-estimators combine a high breakdown point with a high efficiency, requiring the same (high) computation time as S-estimators and are for example the 'official recommendation' for robust regression in the statistics software S-Plus [39].

The solution to the multivariate regression problem can also be reformulated in terms of the joint location $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_X, \hat{\boldsymbol{\mu}}_Y)$ and scatter matrix $\hat{\mathbf{C}} = \begin{pmatrix} \hat{\mathbf{C}}_{XX} & \hat{\mathbf{C}}_{XY} \\ \hat{\mathbf{C}}_{YX} & \hat{\mathbf{C}}_{YY} \end{pmatrix}$ of the explanatory and dependent variables since the LS estimators of intercept and slope can be written as functions of the joint location and scatter matrix [40]. Thus, a robust regression estimate can also be obtained by applying robust estimators of location and scatter instead of the LS estimates. Following this approach, Rousseeuw *et al*. [40] recently derived the MCD regression as robust regression method using the MCD estimator of multivariate location and scatter (see Section 3.2. below). The resulting robust regression estimator has the appropriate equivariance properties, a bounded influence function, and inherits the breakdown value of the MCD

Copyright © 2006 John Wiley & Sons, Ltd.

estimator. In order to improve the rather low efficiency a reweighting scheme was proposed [40].

3.2. Robust estimates of multivariate location and covariance (scatter)

For a data set $\mathbf{X}$ of n points in p dimensions the most well-known estimator of the multivariate location is the arithmetic mean

$$T(\mathbf{X}) = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \tag{7}$$

which can also be viewed as an LS estimator because it minimizes

$$\sum_{i=1}^{n}\|\mathbf{x}_i - T\|^2 \tag{8}$$

where $\|\ldots\|$ is the L_2 norm. The breakdown point is 0%. The maximum likelihood estimator for the population covariance matrix $\mathbf{C}$ is defined as

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - T)'(\mathbf{x}_i - T) \tag{9}$$

for $\mathbf{x}_i$ the ith row of the $(n \times p)$ data matrix $\mathbf{X}$, and T the arithmetic $(1 \times p)$ mean vector. Robust estimates for multivariate location and covariance will be described in the following.

3.2.1. *M-estimates of location and covariance*

A generalization of M-estimators to multivariate location is given by

$$\underset{T}{\text{Minimize}}\sum_{i=1}^{n}\rho(\|\mathbf{x}_i - T\|) \tag{10}$$

where T can be regarded as the location estimate [1]. These estimates are not necessarily affine equivariant as the example of the L_1 location estimator shows

$$\underset{T}{\text{Minimize}}\sum_{i=1}^{n}\|\mathbf{x}_i - T\| \tag{11}$$

The L_1 location estimator also known as the 'spatial median' or 'median center' is a generalization of the univariate median and its breakdown point is 50% [13,41]. The L_1 estimator only satisfies the weaker condition of orthogonal equivariance.

Affine equivariant M-estimates of multivariate location and scatter were formally proposed by Maronna [20]. A major drawback of these is that the breakdown point of affine equivariant M-estimators is at most $1/(p+1)$, that is relatively low for even a moderately high number of variables [23,42,43]. Furthermore, Devlin *et al*. [44] found that M-estimators in practice could tolerate even fewer outliers than indicated by this upper bound. In addition Wisnowski *et al*. [45] found empirically that the usefulness of M-estimates of covariance for detecting multiple outliers is limited to low-dimensional, low-density scenarios.

3.2.2. *Stahel–Donoho Estimator*

The first affine equivariant multivariate location and scatter estimator with a high breakdown point was the Stahel–Donoho estimator (SDE) or 'outlyingness-weighted median' [23,24]. For each $\mathbf{x}_i$ in $\mathbf{X}$, one looks for a one-dimensional projection in which $\mathbf{x}_i$ is most outlying in the sense defined as

$$r_i = \sup_{\|\mathbf{v}\|=1}\frac{\left|\mathbf{x}_i\mathbf{v}^t - \underset{j}{\text{med}}(\mathbf{x}_i\mathbf{v}^t)\right|}{\underset{k}{\text{med}}\left|\mathbf{x}_k\mathbf{v}^t - \underset{j}{\text{med}}(\mathbf{x}_j\mathbf{v}^t)\right|} \tag{12}$$

Where $\underset{j}{\text{med}}(\mathbf{x}_j\mathbf{v}')$ is the median of projections of all data points $\mathbf{x}_j$ on the direction of the vector $\mathbf{v}$. The location and scatter are then estimated by the weighted mean and the weighted covariance matrix with weights of the form $w(r)$ where w is a strictly positive and decreasing weight function of $r \geq 0$. The estimator is related to projection pursuit since one principally searches over all possible projections $\mathbf{v}$ with r_i as projection index. Due to the high computational cost, when calculating the SDE in its exact form, approximate methods with subsampling procedures are commonly used [46]. The breakdown properties of the MAD (i.e. the denominator in Equation (13) basically determine the breakdown of the SDE and corresponding modifications have been suggested to obtained further improvements in this respect [47,48].

3.2.3. *Multivariate trimming*

Ellipsoidal multivariate trimming (MVT) was proposed by Gnanadesikan and Kettenring [21] and Devlin *et al*. [22]. In each step of this iterative procedure the squared Mahalanobis distance (d_i^2) of the observation vectors $\mathbf{x}_i$ from the current robust estimate of location $\bar{\mathbf{x}}^*$, are measured in the metric of $\mathbf{C}^*$, the current robust estimate of the covariance matrix of the $\mathbf{X}$ data. A specified percentage (the trimming percentage) of the most extreme observations (i.e. objects with the largest d_i^2) is temporarily set aside (max. 50% of the observations) and the remaining observations are used to compute $\bar{\mathbf{x}}^*$ and $\mathbf{C}^*$ exactly as $\bar{\mathbf{x}}$ and $\mathbf{C}$, the sample mean vector and covariance matrix. A number of samples with highest d_i^2 corresponding to the trimming percentage is again set aside and the process is repeated with the remaining samples. The iterative process terminates when both $\bar{\mathbf{x}}^*$ and $\mathbf{C}^*$ converge. The effect of trimming is that observations with large distances do not contribute to the calculations for the remaining observations. In most approaches the starting values for $\bar{\mathbf{x}}^*$ and $\mathbf{C}^*$ are taken to be $\bar{\mathbf{x}}$ and $\mathbf{C}$, even though robust starting values may appear as natural choices for $\bar{\mathbf{x}}$ and $\mathbf{C}$. Empirically, MVT has been found to converge quickly, usually in two or three steps [44,49]. Devlin *et al*. [44] claimed that the ε^* of MVT was the same as its trimming percentage (max. 50%), and does not decrease with the number of variables. However, Donoho [24] argued that the ε^* of MVT is at most about $1/p$, thus rendering this method less attractive due to its low breakdown point. In its original version the MVT procedure can be applied only if the number of objects in the $\mathbf{X}$ matrix after trimming exceeds the number of variables. This limitation can be avoided by applying it to the score matrix $\mathbf{T}$ from a PCA model of the data matrix $\mathbf{X}$ [49], but in this case the result also depends on the robustness properties of the applied method for obtaining the PCA model and furthermore the affine equivariance of a conventional covariance estimate would be lost.

Copyright © 2006 John Wiley & Sons, Ltd.

3.2.4. Minimum volume estimator

Another affine equivariant high breakdown point estimator of multivariate location and covariance is the minimum volume estimator (MVE) [16,50]. The objective of the MVE is

$$T(\mathbf{X}) = \text{Center of the minimum volume ellipsoid covering (at least) a fraction of } h \text{ points of } \mathbf{X} \tag{13}$$

where h can be taken as low as $(n/2)+1$. The location estimator is then given by the center of the ellipsoid. The corresponding covariance estimator is defined as the covariance matrix of the ellipsoid multiplied by a suitable correction factor to obtain consistency with the multivariate normal distribution. The ε^* can be the highest possible namely 50% as $n \to \infty$. The algorithm will have a slow convergence rate similar to the LMS estimate, $n^{-1/3}$ [1]. Furthermore, the algorithms suffer from inefficiency and high computational complexity, making it impractical for use with large data sets [51].

3.2.5. Minimum covariance determinant

The objective of the minimum covariance determinant (MCD) estimator of multivariate location and scatter is [16,50]

$$T(\mathbf{X}) = \text{Mean of the } h \text{ points of } \mathbf{X} \text{ for which the determinant of the covariance matrix is minimal} \tag{14}$$

The MCD seeks the h points out of the whole data set (n objects) for which the classical tolerance ellipsoid (for a given level) has a minimum volume among all possible subsets of size h. Then the location and scatter estimates are given by the mean and covariance matrix for this optimal subset h_n. The covariance matrix has to be scaled by a consistency factor c_δ, in order to obtain a consistent estimator for the multivariate Gaussian distribution. In the univariate case this corresponds to the least trimmed squares estimator where each data point receive a weight of one if it belongs to the robust confidence interval and zero otherwise.

The limitation of this method is that the number of observations should be larger than the number of variables, if $p > h$ the covariance matrix of any h-subset has a zero determinant and thus cannot be minimized. Therefore, high dimensional data sets should first be reduced by variable selection or by using principal components. The breakdown point is the highest possible when $h = 0.5n$. For a better compromise between efficiency and breakdown value h should be $\approx 0.75n$ ($\varepsilon^* \approx 25\%$) [52]. A fast algorithm for calculating the MCD estimator (FAST–MCD) has been derived [52]. For small data sets it finds the exact MCD, whereas for larger data sets it is claimed to give more accurate results than alternative methods at the time of development [52].

If the outlier contamination can be estimated *a priori*, h can be conveniently reformulated in terms of the trimming percentage α (where $0 < \alpha \leq ½$): one can replace $h = [n(1-\alpha)]+1$ in Equation (14) and (15). The breakdown point of these estimators is equal to α. For $\alpha \to 0$ the MVE yields the center of the smallest ellipsoid covering all the data, whereas the MCD objective tends to the arithmetic mean [50].

Butler *et al*. [53] show that MCD has better statistical properties than MVE since MCD is asymptotically normal and further that MVE has a slower convergence rate ($n^{-1/3}$) [54]. Other authors have also noted the theoretical superiority of MCD to MVE [52,55]. The FAST–MCD is also an order of magnitude better than all MVE algorithms in terms of computational complexity [45].

The statistical efficiency of the MCD estimator can be increased by implementing a reweighting estimator [52,56]. After obtaining the raw MCD estimators of location ($\hat{\boldsymbol{\mu}}_{\text{MCD}}$) and scatter ($\hat{\mathbf{C}}_{\text{MCD}}$) each observation receives a weight $\mathbf{w}_i$; which is zero if its robust distance, $d(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_{\text{MCD}}, \hat{\mathbf{C}}_{\text{MCD}}) = \sqrt{(\mathbf{x_i} - \hat{\boldsymbol{\mu}}_{\text{MCD}})' \hat{\mathbf{C}}_{\text{MCD}} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{MCD}})}$ exceeds $\sqrt{\chi^2_{n,\,0.975}}$.

The reweighted MCD estimator is then defined as the weighted mean and covariance matrix.

3.2.6. S-estimators of multivariate location and scatter

Davies [25] and Lopuhaä [26] extended multivariate regression S-estimates to multivariate location and covariance estimates. The S-estimator of multivariate location and scatter is defined as that vector T and positive definite symmetric matrix $\mathbf{C}$ which minimize det $|\mathbf{C}|$ subject to a limit on the magnitudes of the corresponding Mahalanobis distances, $d_i = \sqrt{(\mathbf{x}_i - T)' \mathbf{C}^{-1} (\mathbf{x}_i - T)}$

$$\frac{1}{n} \sum_{i=1}^{n} \rho(d_i) = b_0 \tag{15}$$

where $\rho(\cdot)$ is symmetric and $\rho(0) = 0$. The constant b_0 is often taken as the expected value of $\rho(d_i)$ assuming a multivariate Gaussian distribution [57].

These S-estimators are affine equivariant, asymptotically normal, and for well-chosen $\rho(\cdot)$ their breakdown points can be as high as 50% [1].

4. ROBUST MULTIVARIATE MODELS

Almost all robust multivariate models in the literature and described in the following, work under the assumption that outliers are samples (rows in the data matrix). That is all data of one sample is treated as *one* observation. Hence, these models aim at identifying and minimizing the influence of individual outlying samples. The situation where, for example an individual element in the data matrix x_{ij} is considered as outlying has not gained much attention in the literature. First approaches for this situation have been provided by Hawkins *et al*. [58], Liu *et al*. [59] and Croux *et al*. [60].

4.1. Robust principal component analysis

Principal component analysis (PCA) is often the first step of the data analysis, typically followed by a more quantitative analysis using PCR, PLS, discriminant analysis or other multivariate techniques. Classical PCA involves computing the eigenvectors and eigenvalues of the sample covariance or correlation matrix. Simple but powerful algorithms have been developed for finding the principal components with the singular value decomposition (SVD) as the most widespread [61].

Copyright © 2006 John Wiley & Sons, Ltd.

In the context of PCA, an outlier can be defined as an observation/object that either lies far away from the subspace spanned by the correct k eigenvectors, and/or for which the projection into the model lies far from the remainder of the data within the subspace [62].

Several ways of robustifying principal components have been proposed. They can be grouped as follows:

(1) Techniques that replace the classical covariance matrix by a robust covariance via robust estimators of location and shape such as the MCD. Calculating the eigenvalues and eigenvectors of this robust covariance matrix provides eigenvectors that are robust to sample outliers. These approaches are limited to relatively low-dimensional data.
(2) Projection pursuit (PP) searches for structure in high dimensional data by projecting these data into a lower-dimensional space which maximizes a robust measure of spread called the projection index, $\rho(\cdot)$, for example using the *MAD*. PP methods obtain robust eigenvector estimates by explicitly solving the maximization (or minimization) problem: find the direction (eigenvector) $\mathbf{v}_p$ that maximizes $\rho(\mathbf{Xv})$ [63]. In subsequent steps, each new direction is constrained to be orthogonal to all previous directions. The procedure results in robust principal components and a robust covariance matrix. Classical PCA, which uses the variance as projection index is a special case of the PP algorithm. Since the principal components are computed sequentially, this approach can handle high dimensional data, $n < p$.
(3) A Combination of (1) and (2). This approach can handle high-dimensional data.
(4) Adjustments to the internal computations of the SVD algorithm by replacing the LS criterion with a robust estimate. This approach can handle high-dimensional data [60] and elemental outliers [58].

Table III lists some of the robust PCA methods belonging to group (2) and (3) discussed in this paper.

Replacement of the classical covariance or correlation matrix by one of the abovementioned or other robust estimators is perhaps the simplest and also an intuitively appealing approach.

Applications using GM-estimators of means and covariances can be found in Campbell [34], Rivest and Plante [70] and Daigle and Rivest [71]. These methods are less suitable for high dimensionalities due to the fact that ε^* decreases towards zero with increasing dimensionality.

Many simulation studies, starting with Devlin *et al*. [44] have been carried out to find out which robust estimator should be used for estimating a covariance/correlation matrix and its principal components. More recently, Croux and Haesbroeck [72] have shown that the one-step reweighted MCD method [50] and the S-estimator of location and shape are well suited for robustifying the estimate of the covariance matrix. The theoretical results as well as the simulations favor the use of S-estimators, since they combine a high efficiency with appealing robustness properties. The results are more robust than results obtained with GM-estimators, but are unfortunately limited to small dimensions [72]. This limitation is a severe restriction in chemometrics where it is usually important to have robust PCA methods for situations with $p > n$. A second problem is the computation of these robust estimators in high dimensions. Todays fastest algorithms [52,73] can handle up to about 100 variables, which might pose a substantial practical limitation in many chemometric applications.

Other approaches to robustify PCA, based on PP, have been considered by Ruymgaart [74], Li and Chen [64], Ammann [75], Galpin and Hawkins [7], Xie *et al*. [65], Croux and Ruiz-Gazen [66], Hubert *et al*. [67] and Hubert *et al*. [68].

The projection index in the method proposed of Li and Chen [64] was chosen as Huber's M-estimator of scale, whereas Galpin and Hawkins [7] and Xie *et al*. [65] replace the variance norm by a robust measure of the spread based on optimizing the L_1 norm (e.g. MAD). Xie *et al*. [65] introduced the generalized simulated annealing algorithm as an optimization procedure in the process of PP calculation to pursue the global minimum. Li and Chen [64] proved that the resulting method inherits the breakdown value of the robust scale estimator and are qualitative robust. These methods can handle high dimensional data, where the number of variables exceeds the number of samples [67].

Unfortunately, the algorithm of Li and Chen [64] has a high computational cost [67]. In Croux and Ruiz-Gazen [66] a computationally more advantageous method was presented (C-R algorithm) with the L_1—median estimator as the center of the data and the Q_n scale estimator [29] as projection index. To speed up the computation the directions to be investigated have been restricted to all directions that pass through the center and a data point. Unfortunately, the

Table III. Robust high-dimensional PCA methods discussed in this paper

Name	Robust estimate	References	Comments
PP-PCA	Weighted mean (center of the data) Hubers M-estimate (projection index)	[64]	High computational complexity
PP-PCA	L_1 (center of the data) MAD (projection index)	[65]	
PP-PCA	L_1 (center of the data) Q_n (projection index)	[12,66]	
RA-PCA	L_1 (center of the data) Q_n (projection index)	[67]	Fast to compute
ROBPCA	Combined PP with reweighted MCD	[68]	A fast alternative algorithm exsist, ROBPCA-$k_{\max}$[69]

Outliers are considered as whole samples.

Copyright © 2006 John Wiley & Sons, Ltd.

algorithm has a numerical accuracy problem in high dimensions (large p) due to round-off errors and is still computationally costly [61]. Hubert *et al*. [67] proposed an improvement of the C-R algorithm, a modified two-step version, called reflection-based algorithm for robust principal component analysis (RAPCA). According to the authors the method is more stable numerically and computationally much faster (when PCA compression to the rank of the data is performed as a first step), and can deal with both low and high dimensional data. The compression step in RAPCA can also be built in the C-R algorithm and thereby improve the speed. By doing so no difference with respect to computation speed were observed for up to 15 components [76]. An improved version of the C-R algorithm, which should not suffer from the numerical problem in high dimensions can be found in Croux and Ruiz-Gazen [12].

The most attractive advantages of the PP-procedures are that they yield covariance matrices and PCs that are of both orthogonal equivariance and show a high ε^*, which can be as high as 50%, and does not depend on the dimensionality [64].

A different type of approach is the robust singular value decomposition method described by Ammann [51]. In the calculation of the SVD successive transposed QR decompositions are used where the corresponding regression steps have been replaced by their robust alternatives. Mallow's GM-estimator [1] together with an iterative update of location by M-estimation has been suggested for the computations. Final estimates are obtained by an ordinary SVD on the weighted covariance matrix, where weights have been obtained within the previous QR decompositions.

Hubert *et al*. [68] introduce a new method for robust PCA called ROBPCA where PP ideas are combined with robust location and covariance estimation in lower dimensions. After preprocessing the data, PP is used for initial dimension reduction ($k \ll p$), and the reweighted MCD estimator is then applied to this lower dimensional data space to find a robust centre and a robust covariance estimator of the projected samples. Dimension reduction is necessary because the MCD estimator is only applicable for $p < n$ due to singularity of the covariance matrix when $h < p$. Finally these estimates are back transformed to the original space and a robust estimate of the location of $\mathbf{X}$ and of its scatter are obtained. This method can handle both low and high dimensional data, and according to the authors, it produces more accurate estimates for non-contaminated data and more robust estimates for contaminated data. The ROBPCA method is orthogonal equivariant [68]. Unfortunately, this algorithm is not very fast when the results for several principal components ($k = 1, \ldots, k_{\max}$) are required as it needs to run the whole procedure for each component separately [69]. A faster alternative algorithm, ROBPCA-$k_{\max}$, for moderate data sets of sizes up to 100 and $k_{\max} = 10$ is proposed by Engelen *et al*. [69]. This algorithm is less precise and more time consuming if $k_{\max}$ is chosen too large because it computes the MCD estimator for $k_{\max}$ dimensions in one run. When comparing the ROBPCA and ROBPCA-$k_{\max}$ algorithms Engelen *et al*. [69] found that ROBPCA-$k_{\max}$ performed almost as well as the original ROBPCA. Only for very particular contaminations the ROBPCA-$k_{\max}$ estimator was unfavorable compared to ROBPCA.

All methods discussed so far consider entire samples, $\mathbf{x}_i$ as outliers but there are many examples where individual data elements, x_{ij}, in otherwise good rows, $\mathbf{x}_i$ are corrupted, for example in fluorescence, microarray or proteomics data and image analysis. Methods that can handle this kind of outliers have been proposed by Hawkins *et al*. [58], Liu *et al*. [59] and Croux *et al*. [60]. These methods are based on the alternating least-squares algorithm proposed by Gabriel and Zamir [77]. In this algorithm the minimization problem is solved with criss-cross regressions, which involve iteratively computing dyadic (rank 1) fits using LS (similar to NIPALS). The original Gabriel-Zamir SVD algorithm is then rendered robust by using robust regression estimates such as L_1[58], LTS [59] or weighted L_1 (weights based on MVE) [60] instead of the ordinary LS regression in the SVD algorithm.

4.1.1. *Multiway methods*

For the N-way methods Pravdova *et al*. [78] proposed a robust version of the three-way Tucker3 model or multimode PCA. The three-way data cube, $\underline{\mathbf{X}}$, is decomposed into a number of components as in PCA but as opposed to PCA the number of components can be different for the three modes (i.e. the dimensions or directions). The method is based on robust initialization of the Tucker3 algorithm using MVT or MCD. The three-way data are unfolded to two-way matrices and the loadings are obtained by SVD. In this version the outliers are identified in the first mode only ($\mathbf{A}$), but as all modes are treated symmetrically, one can look for outliers in any mode. First a so-called clean subset is constructed using MVT or MCD. A clean subset means that the data set contains no detected outliers. In each iteration of the ALS subroutine, the loadings $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ for the different modes are calculated for the clean subset of objects only. The loadings $\mathbf{A}$ are then predicted for all objects and the data is reconstructed with the predefined number of factors. Residuals between the initial data and the reconstructed data are calculated and sorted, and 51% of objects with the smallest residuals are selected to form the clean subset for the next ALS iteration. The objective function to be minimized is the sum of squared residuals for the h clean objects from the first mode. Once the robust Tucker3 model is constructed the outliers are identified by a robust estimate of the standardized residuals based on the scale estimator MAD. The final Tucker3 model is constructed as the least squares model for the data after outlier elimination. Empirically studies show that for 20% outliers in the dataset, the robust Tucker3 model converged to a good solution but for 40% outliers, the algorithm had problems for some type of outliers. According to authors MCD is a better algorithm for finding the clean subset than MVT [78], which can be understood from the deteriorating breakdown properties of MVT with increasing dimensionality [24].

4.2. Robust principal component regression

Principal component regression (PCR) can be used to build regression models for rank deficient regressors $\mathbf{X}$ [62]. The principal components are obtained by decomposing $\mathbf{X}$ via PCA and subsequently the response(s) is regressed onto the score, $\mathbf{T}$.

Both steps of the PCR procedure can be affected by outliers because regression as well as PCA is based on least squares

Copyright © 2006 John Wiley & Sons, Ltd.

fitting. The bad leverage points of the PCA model are important as they deteriorate the estimation of the **T** matrix, influencing in this way also the second step of PCR. The regression step of the PCR can additionally be hampered by outliers in **y**. It is thus important to robustify PCR both with respect to the PCA step as well as the subsequent multiple linear regression. Such robust PCR methods which make both steps robust have been presented by Walczak and Massart [49], Pell [79], Filzmoser [80], Hubert and Verboven [81] and Zhang *et al*. [82].

Table IV contains an overview of the different robust PCR methods discussed in this paper.

The robust PCR (RPCR) method proposed by Walczak and Massart [49] uses robust PCA based on a robust covariance matrix (MVT), followed by LMS regression applied on all samples. The robust covariance matrix and thus the robust principal components provide the background to reveal outliers, either good or bad leverage points, in the **X** data while standardized residuals from the robust model are applied to detect outliers in both **X** and **y**. The breakdown point of MVT is the same as its trimming percentage (max. 50%) and does not decrease with the number of variables. However, Donoho [24] argued that the ε^* of MVT is at most about $1/p$, thus rendering this method less attractive due to its low breakdown point. The breakdown point of LMS is the highest possible, namely 50%. The lack of efficiency for LMS can be regarded as a possible disadvantage of this method, as the LMS minimizes the median of the residuals (which is equivalent with consideration of only 50% of the data) instead of the sum of the squared residual. Therefore Walczak and Massart [49] recommend that the model should be used as an outlier detection tool and not the final building method for the model.

Filzmoser [80] introduced a robust method for PCR based on the idea of PP proposed by Li and Chen [64] to obtain robust principal components, combined with LTS regression for prediction. The algorithm is the C-R algorithm [66] with the median absolute deviation from the median (MAD) estimator as projection index. Due to the fact that that first principal components (PC) with the largest variance are not necessarily the best predictors, a step-wise selection of PC's is performed for subset ($k < p$) selection of PC's for prediction. The selection starts with the PC resulting in the best prediction of the response variables (according to an appropriate association measure). The selection process continues until the quality of prediction cannot be increased significantly.

In the method of Hubert and Verboven [81] first a robust PCA method is applied on the regressors. For low-dimensional data ($p < n$), the MCD estimator [16] is applied as a robust estimator of the covariance matrix of **X**, and for high-dimensional data ($p > n$) the ROBPCA method [68] is used. Next a robust regression method is applied. If there is only one response variable the LTS regression estimator [16] is preferred, otherwise the MCD regression estimator [40] is utilized. All the robust estimates in this method are based on a hard thresholding, that is binary weights 0/1 are applied.

The rather empirical approach of Zhang *et al*. [82] relies on robust regression diagnostics of the predictions for identifying outliers before building an ordinary LS-PCR model with the good data. For regression diagnostics so called 'principal sensitive vectors' (RPPSV) [83] are used, that is an analysis of how the predictions for a sample change when each training sample is omitted in turn. In addition, samples with significantly high prediction residuals are also omitted. The estimates computed by the procedure are affine equivariant and the convergence rate is $n^{-1/2}$ [83].

In the method proposed by Pell [79], 'resampling by half-means' (RHM) [84] is used to detect outliers in the response matrix, **X**, in order to remove those samples from consideration before the PCA step. RHM is based on repeatedly estimating the Euclidean distance of each sample to the mean of a randomly drawn subset of 50% of the data. In case **X** consists of variables on different scales, autoscaling is first applied, using the standard deviations from the respective 50% subset. The frequency with which each sample is among the 5% most distant observations is used as the criterion for outlier detection. The PCA decomposition is performed without those extreme samples but the extreme samples are projected onto the PCA space and the scores from the extreme samples are used with the rest of the calibration samples in the regression step. Simulation studies [84] showed slightly better breakdown properties for the RHM than for the MCD 20–45% and 20–30% respectively, depending on the fractions of outliers and their multivariate distance from the rest of the data. In addition, the method can handle rank deficient data. In the regression step (step 2) the LTS estimator is used with multiple trimming parameters

Table IV. Overview of different robust PCR methods discussed in this paper

Name	Principle	Comments	References
RPCR	PCA: MVT Regression: LMS	Only suggested as outlier detection tool due to lack of efficiency of LMS	[49]
RPCR	PCA: PP Regression: LTS		[80]
RHM-PCR	PCA: RHM Regression: LTS	Not orthogonal equivariant	[79]
RPCR	PCA: MCD (low dim.) or ROBPCA Regression: reweighted LTS (one response var.) or MCD regression		[81]
RPPSV	Principal sensitivity vectors	Empirical	[82]

Copyright © 2006 John Wiley & Sons, Ltd.

(50, 40, 30, 20 and 10%) as suggested by Ryan [6]. Unfortunately, the RHM estimator does not possess orthogonal and translation equivariance [81].

4.3. Robust partial least squares regression

PLS is a linear regression technique developed to deal with high-dimensional regressors and one, $\mathbf{y}_{n\times 1}$ (PLS1) or several response variables, $\mathbf{Y}_{n\times q}$ (PLS2). PLS makes use of ordinary least squares regression steps in the calculation of weights, loadings, scores and regression coefficients. Therefore outliers, either in the **X** or in the **y** or **Y** variables, pose a serious problem to PLS regression. The most common algorithms for PLS in the chemometric field are the NIPALS algorithm and SIMPLS algorithm. In cases with only one response variable ($q=1$) and without missing values, SIMPLS and PLS1 (NIPALS) are equivalent.

In Table V the different robust PLS methods discussed in this paper are listed.

Gil and Romera [88], Wakeling and Macfie [85] and Griep *et al*. [86] claimed that substituting all steps of the ordinary LS regression steps in the NIPALS PLS algorithm by a robust regression procedure will make the model completely resistant to outliers, but there are several prices to be paid: Not just a higher computational demand, but also a lower efficiency of the robust steps. Thus, an alternative approach is to replace one or two selected regression steps instead of all steps together which could still show a good performance in terms of handling outliers. Such procedures are called semi-robust [86,88].

A first robust PLS2 algorithm, (RPLS) [85] has been developed by replacing the non-robust regression steps for **w** (weight vector for **X**) and **c** (loading vector for **Y**) in the PLS2 algorithm by the robust biweight method [32]. This makes these estimates resistant to outliers by down-weighting cases with high residuals. Final values of **t** (score vector for **X**) and **b** (score vector for **Y**) are formed from the un-weighted data as in the conventional PLS2 algorithm. The price to be paid consists of a lack of orthogonality on successive **X** weight vectors, **w**. The RPLS algorithm is designed to compensate independently for outliers in both **X** and **Y**.

Following the idea of Wakeling and Macfie [85], Griep *et al*. [86] carried out a comparison among three different methods of robust regression and studied their incorporation into the PLS1 algorithm. In their study Griep *et al*. [86] replaced the regression step for the weight vector **w** with three different methods of robust regression: LMS, Siegel's RM and IRLS. Their empirical results indicate that the best option is to use IRLS compared to LMS and Siegels RM.

According to Gil and Romera [88] this way of making the PLS models robust, by substituting some or all regressions steps with a robust alternative, does not necessarily catch 'multivariate' outliers. This is due to the fact that the first step of PLS is not just one regression, but is formed from the individual regressions of each variable $\mathbf{x}_j$ on **y**.

Thus, the application of, for example IRLS in the first step consists of the application of IRLS to each simple regression. Therefore outliers in the projections of the data onto planes $[\mathbf{x}_{.1},\mathbf{y}]$, $[\mathbf{x}_{.2},\ \mathbf{y}],\ldots\ldots,[\mathbf{x}_{.j},\ \mathbf{y}]$ are taken into account, but the multivariate nature of the $\mathbf{X}'\mathbf{Y}$ data is not considered.

Another version of the IRLS algorithm for PLS has been obtained by Cummins and Andrews [87] and Pell [79] called IRPLS. The idea is the same as in ordinary IRLS, but in these cases using the cross validated residuals of the PLS regression in the sample weight function. The sample weights are initially set to one and are updated after each iteration. Both Cummins and Andrews [87] and Pell [79] tested four weight functions including bisquare, Cauchy, Fair and Huber. The only difference between the algorithms proposed by Pell [79] and Cummins and Andrews [87] is that in the algorithms of Pell [79] the number of components is fixed until the sample weights converge whereas in the algorithm proposed by Cummins and Andrews [87] the number of components is allowed to change as the sample weights change. Gil and Romera [88] claimed that a problem with the method proposed by Cummins and Andrews [87] is that the residuals for each sample would depend strongly on the number of components calculated in PLS because different criteria for choosing the number of components could cause different weights for each sample.

The breakdown point for IRPLS depends on the chosen weight function but was around 44% for the tested weight functions and moderate outliers [87].

These algorithms were only derived for a one-dimensional response variable (i.e. PLS1) and are not resistant to high

Table V. Overview of different robust PLS methods discussed in this paper

Name	Principle	Basic algorithm	Comments	References
RPLS	Biweight-IRLS	NIPALS	Semi-robust Derived for PLS2	[85]
Robust PLS1	IRLS	NIPALS	Semi-robust	[86]
IRPLS	IRLS	NIPALS	Derived for PLS1 Not resistant to leverage points	[87]
IRPLS	IRLS	NIPALS	Derived for PLS1 Not resistant to leverage points Tendency towards overfitting	[79]
PLSMR	Stahel–Donoho estimator	NIPALS	Derived for PLS1 Only valid for $n>p$	[88]
RSIMCD	ROBPCA for robust scores MCD regression for regression	SIMPLS	Derived for both PLS1 and PLS2	[89]
RSIMPLS	ROBPCA for robust scores Information from ROBPCA for regression	SIMPLS	Derived for both PLS1 and PLS2 Fast (twice as fast as RSIMCD)	[89]
PRM	GM-estimators	SIMPLS	Derived for PLS1	[90]

Copyright © 2006 John Wiley & Sons, Ltd.

leverage points since the weights only depends on the residuals after each step [89]. Different weight functions as well as different tuning constant for the same weight function can give different results, which may make the methods less attractive [79,87]. Furthermore, Pell [79] obtained in some cases better prediction results with the robust methods than the estimated reference error, which may be indicative for an overfitting problem.

In Gil and Romera [88] a robust PLS1 method, PLS matrix robust (PLSMR), is obtained by robustifying the sample covariance matrix of the x-variables and the sample cross-covariance matrix between the x- and y-variables. For this the highly robust Stahel–Donoho estimator [10,23,24] is used with Huber's weight function. To minimize the computational cost the sub-sampling scheme used to compute the estimator starts by drawing subsets of size $p+2$. This means that the PLSMR method cannot be applied to high-dimensional regressors ($n \ll p$) which is a major disadvantage. To tackle this problem Gil and Romera [88] propose to initiate with variable selection before the robust regression. It is not possible to extend the method to PLS2 [89].

In SIMPLS [91], the PLS weights are obtained as eigenvectors of the $\mathbf{X}'\mathbf{Y}$ cross covariance matrix after successive deflation steps. Since SIMPLS is based on the sample cross-covariance $\mathbf{C}_{xy}$, the empirical covariance matrix ($\mathbf{C}_x$) of the x-variables and on linear LS regression the results are affected by abnormal observations in the data set. Hubert and Vanden Branden [89] introduced two robustified versions of the SIMPLS algorithm called RSIMCD and RSIMPLS respectively, based on replacing the cross-covariance matrix $\mathbf{C}_{xy}$ and the empirical covariance matrix $\mathbf{C}_x$ by robust estimates, and by performing a robust regression method instead of MLR. The robust algorithms are built on two main stages. First, robust scores $\mathbf{t}_i$ are constructed based on a robust criterion (ROBPCA) on $\mathbf{Z}_{nm} = (\mathbf{X}_{np}, \mathbf{Y}_{nq})$, and secondly a robust linear regression is performed based on the robust scores from the ROBPCA. The first stage is similar for both methods but the regression step differs: RSIMCD is based on MCD regression [40] while RSIMPLS uses additional information from the previous ROBPCA step for a reweighted MLR. The proposed algorithms are fast compared with previously developed robust methods and can handle cases where $n \ll p$ and $q = 1$ [11]. Hubert and Vanden Branden [89] recommend RSIMPLS because it is roughly twice as fast as RSIMCD. The breakdown value for RSIMPLS is roughly $1-\alpha$ (where α is the assumed minimal fraction of regular observations) as for the MCD estimator [11]. Both RSIMPLS and RSIMCD are equivariant for translation and orthogonal transformations in x and y [89].

Recently, Serneels *et al*. [90] proposed a method, partial robust M-regression (PRM), for robust regression based on GM-estimators. PRM uses continuous weights, resulting in a gradual down-weighting of outliers according to their degree of outlyingness. This is in contrast to, for example RSIMPLS where a weight of zero is given to all observations with residuals above a certain cut-off value and unity to all others. As GM-estimator the weights are computed from both the residuals as well as the leverage for a sample. This can be performed in a way such that the orthogonal and scale equivariance of the PLS estimator is retained. The weighting is used both in the SIMPLS step of computing the PLS scores as well as in the regression of y on these scores. For $p \gg n$ the computation time is sped up by carrying out a prior SVD on $\mathbf{X}$ $(n \times p)$, $\mathbf{X}' = \mathrm{VSU}$. The iteration procedure is then applied to the reduced data matrix, $\hat{\mathbf{X}} = \mathbf{US}$ $(n \times n)$ and the resulting PRM regression estimate $\tilde{\beta}$ needs then to back transformed into $\hat{\beta} = \mathbf{V}\tilde{\beta}$.

A simulation study testing various distributions of the error terms, different samples sizes and dimensionalities showed that in terms of statistical efficiency, PRM generally outperformed PLS and RSIMPLS. Only for the normal error terms PLS was more efficient than PRM. In situations with 10% bad leverage points PRM and RSIMPLS performed almost similar and clearly outperformed PLS. When comparing the computation time for PRM and SIMPLS it turned out that the computation time for RSIMPLS was consistently substantially higher than for PRM both for increasing number of observations and increasing number of predictor variables [90].

The PRM method is computational possible for high dimensional data sets and thus provides an appealing alternative to RSIMPLS in particular when a gradual outlier behavior can be expected. But the method is currently only derived for univariate y (i.e. PLS1) and the highest possible breakdown point of all GM-estimators is in general not larger than 30% and decreases as a function of the dimensionality p [17,31].

5. SOFTWARE

The common basic methods for robust estimation of location and scatter (i.e. MCD) and robust regression (i.e. M-, LMS-, LTS-, S- and MM-estimators) are all available within the standard statistical software packages SAS (release > 6.12) [92], S-Plus [39,93] and R [94]. An implementation for robust PCA is also available for S-Plus [95].

Recently, a comprehensive MATLAB toolbox, 'LIBRA', for robust estimates and multivariate methods has appeared [96]. Apart from MCD and LTS it also contains implementations of other methods that have been developed at the research groups at the University of Antwerp and the Katholieke Universiteit Leuven, in particular for robust PCA (ROBPCA) and PLS (RSIMPLS). The toolbox also includes many graphical tools for model checking and outlier detection. Additionally, an incorporation of several of these methods into the widely used PLS_Toolbox for Matlab is in preparation (Eigenvector, Inc., Pers. Comm.). The partial robust M-regression is also available as Matlab implementation [97].

6. DISCUSSION

Real world data is often contaminated with gross outliers, which can lead conventional data analysis methods based on least squares completely astray. Thus, outlier detection and/or the use of robust methods are of paramount importance for applied multivariate data analysis in general and chemometrics in particular.

Although simple robust methods such as the median, median absolute deviation or interquartile range are easy to

Copyright © 2006 John Wiley & Sons, Ltd.

compute and have been applied for a long time, the use of robust methods for multiple regression or latent variable models has been computationally prohibitive for multivariate datasets of even moderate size. This is because theoretically rigorous methods usually require the consideration of all possible sample subsets of a given size, which means that their computationally complexity shows a combinatorial growth. Thus, although the exponentially increasing computational power (Moore's law) has contributed a lot to the advancement of computationally demanding data analysis methods, the progress in the applicability of robust methods has largely been achieved by improved fast approximate methods.

When considering robust methods for applied multivariate data analysis, practical considerations might deviate from a pure theoretical viewpoint. Data sets with more than 10–20% outliers would probably be rejected completely and the generation of new data with a higher quality required. Hence, a theoretical breakdown point of 50% might be helpful for initially assessing the data quality and detecting outliers. But on the other hand it should be kept in mind that there is a price to pay for such an extreme robustness. Apart from the higher computational complexity, robust methods usually also show a lower statistical efficiency and convergence rate. For example the median only shows 64% asymptotic efficiency for normally distributed data in comparison to the mean [98]. Thus, although robust methods are able to obtain reasonable estimates in the presence of gross outliers in the original data, their estimates for the 'good' data are usually subject to a higher uncertainty and they require more training samples than conventional least squares methods. For methods with adjustable breakdown properties, therefore a good compromise between robustness and efficiency should be aimed for.

In addition it might be remarked that breakdown points are derived for arbitrarily large, that is 'infinite' outliers, whereas in practice outliers are usually bounded due to a finite measurement range. Extreme outliers might also be easy to detect individually by applying appropriate outlier diagnostics.

With state-of-the-art robust methods, for example based on projection pursuit or fast MCD estimators, efficient tools for robust multivariate data analysis are available and convenient software implementations, for example in Matlab exist.

Acknowledgements

First author acknowledges support provided by the Danish Ministry of Food, Agriculture and Fisheries. We thank the two referees for their helpful comments.

REFERENCES

1. Rousseeuw PJ, Leroy AM. *Robust Regression and Outlier Detection*. Wiley: New York, 1987.
2. Maronna RA, Yohai VJ. Robust estimation of multivariate location and scatter. In *Encyclopedia of Statistical Sciences*, Kotz S (ed.). Wiley: New York, 1998; 589–596.
3. Liang YZ, Kvalheim OM. Robust methods for multivariate analysis—A tutorial review. *Chemometrics Intell. Lab. Syst*. 1996; **32**: 1–10.
4. Hubert M, Rousseeuw PJ, Van Aelst S. Multivariate outlier detection and robustness. In *Handbook of Statistics, data mining and Data Visualization*, Rao CR, Wegman EJ, Solka JL (eds). Elsevier: North Holland, 2005; 263–302.
5. Gnanadesikan R. *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley: New York, 1977.
6. Ryan T. *Modern Regression Methods*. Wiley: New York, 1997.
7. Galpin JS, Hawkins DM. Methods of L1 estimation of a covariance matrix. *Comput. Statist. Data Anal*. 1987; **5**: 305–319.
8. Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. *Robust Statistics: The Approach Based on Influence Functions*. Wiley: New York, 1986.
9. Hampel FR. A general qualitative definition of robustness. *Ann. Math. Statist*. 1971; **42**: 1887–1896.
10. Donoho DL, Huber PJ. The notion of breakdown point. In *A Festschrift for Erich Lehmann*, Bickel PJ, Doksum K, Hodges Jr (eds). Wadsworth: Belmont, CA, 1983.
11. Vanden Branden K, Hubert M. Robustness properties of a robust PLS regression method. *Anal. Chim. Acta*. 2004; **515**: 229–241.
12. Croux C, Ruiz-Gazen A. High breakdown estimators for principal components: The projection-pursuit approach revisited. *J. Multivariate Anal*. 2005; **95**: 206–226.
13. Huber PJ. Robust estimation of a location parameter. *Ann. Math. Statist*. 1964; **35**: 73–101.
14. Huber PJ. Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist*. 1973; **1**: 799–821.
15. Siegel AF. Robust regression using repeated medians. *Biometrika* 1982; **69**: 242–244.
16. Rousseeuw PJ. Least median of squares regression. *J. Am. Statist. Assoc*. 1984; **79**: 871–880.
17. Rousseeuw PJ, Yohai VJ. Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis-(Lecture Notes in Statistics, Volume 26)*, Franke J, Härdle W, Martin RD (eds). Springer Verlag: New York, 1984; 256–272.
18. Yohai VJ. High breakdown point and high efficiency robust estimates for regression. *Ann. Statist*. 1987; **15**: 642–656.
19. Yohai VJ. A procedure for robust estimation and inference in linear regression. In *Directions in Robust Statistics and Diagnostics, Part II*, Stahel WA, Weisberg SW (eds). Springer Verlag: New York, 1991.
20. Maronna RA. Robust M-estimators of multivariate location and scatter. *Ann. Statist*. 1976; **4**: 51–67.
21. Gnanadesikan R, Kettenring JR. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* 1972; **28**: 81–124.
22. Devlin SJ, Gnanadesikan R, Kettenring JR. Robust estimation and outlier detection with correlation coefficients. *Biometrika* 1975; **62**: 531–545.
23. Stahel WA. Robust estimation: Infinitesimal optimality and covariance matrix estimators. PhD Thesis, ETH, Zürich, 1981.
24. Donoho DL. Breakdown properties of multivariate location estimators. PhD Qualifying paper, Harvard University, 1982.
25. Davies PL. Asymptotic behavior of S-estimators of multivariate location and dispersion matrices. *Ann. Statist*. 1987; **15**: 1269–1292.
26. Lopuhaä HP. On the relation between S-estimators and M-estimators of multivariate location and covariance. *Ann. Statist*. 1989; **17**: 1662–1683.
27. Tatsuoka KS, Tyler DE. The uniqueness of S and M-functionals under non-elliptical distributions. *Ann. Statist*. 2000; **28**: 1219–1243.
28. Draper NR, Smith H. *Applied Regression Analysis*. 3 ed., Wiley-Interscience: New York, 1966.

Copyright © 2006 John Wiley & Sons, Ltd.

29. Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. *J. Am. Statist. Assoc*. 1993; **88**: 1273–1283.
30. Hodges JL, Lehmann EL. Estimates of location based on rank tests. *Ann. Math. Statist*. 1963; **34**: 598–611.
31. Maronna RA, Bustos OH, Yohai VJ. Bias- and efficiency-robustness of general M-estimators for regression with random carriers. In *Smoothing Techniques for Curve Estimation*, Gasser T, Rosenblatt M (eds). Springer Verlag: New York, 1979; 91–116.
32. Beaton AE, Tukey JW. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics* 1974; **16**: 147–185.
33. Andrews DF, Bickel PJ, Hampel FR, Huber PJ, Rogers WH, Tukey JW. *Robust Estimates of Location: Survey and Advances*. Princeton University Press: Princeton, 1972.
34. Campbell NA. Robust procedures in multivariate analysis I: Robust covariance estimation. *App. Statist*. 1980; **29**: 231–237.
35. Rousseeuw PJ, van Driessen K. An algorithm for positive-breakdown regression based on concentration steps. In *Data Analysis: Scientific Modeling and Practical Application*, Gaul W, Opitz O, Schader M (eds). Springer Verlag: New York, 2000; 335–346.
36. Hawkins DM, Olive DJ. Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm. *J. Am. Statist. Assoc*. 2002; **97**: 136–159.
37. Hubert M, Rousseeuw PJ, van Aelst S. Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm—Comment. *J. Am. Statist. Assoc*. 2002; 97.
38. Hössjer O. On the optimality of S-estimators. *Statist. Prob. Lett*. 1992; **14**: 413–419.
39. *S-PLUS 6 for Windows Guide to Statistics*. Insightful Corporation: Seattle, 2001.
40. Rousseeuw PJ, van Aelst S, van Driessen K, Agulló J. Robust multivariate regression. *Technometrics* 2004; **46**: 293–305.
41. Hogg RV. Statistical robustness: One view of its use in application today. *Am. Stat*. 1979; **33**: 108–115.
42. Huber PJ. *Robust Statistics*. Wiley: New York, 1981.
43. Huber PJ. Robust covariances. In *Statistical Decision Theory and Related Topics*, Gupta SS, Moore DS (eds). Academic Press: New York, 1977; 165–191.
44. Devlin SJ, Gnanadesikan R, Kettenring JR. Robust estimation of dispersion matrices and principal components. *J. Am. Statist. Assoc*. 1981; **76**: 354–362.
45. Wisnowski JW, Simpson JR, Montgomery DC. A performance study for multivariate location and shape estimators. *Qual. Reliab. Engng. Int*. 2002; **18**: 117–129.
46. Maronna RA, Yohai VJ. The behaviour of the Stahel-Donoho robust multivariate estimator. *J. Am. Statist. Assoc*. 1995; **90**: 330–341.
47. Zuo Y. A note on finite sample breakdown points of projection based multivariate location and scatter statistics. *Metrika* 2000; **51**: 259–265.
48. Zuo Y. Projection-based affine equivariant multivariate location estimators with the best possible finite sample breakdown point. *Statistica Sinica*. 2004; **14**: 1199–1208.
49. Walczak B, Massart DL. Robust principal components regression as a detection tool for outliers. *Chemometrics Intell. Lab. Syst*. 1995; **27**: 41–54.
50. Rousseeuw PJ. *Multivariate Estimation With High Breakdown Point*, Grossmann W, Pflug G, Vincze I, Wertz W (eds). Bad Tatzmannsdorf: Austrai, 1983; 283–297.
51. Ammann LP. Robust singular value decompositions: A new approach to projection pursuit. *J. Am. Statist. Assoc*. 1993; **88**: 505–514.
52. Rousseeuw PJ, van Driessen K. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 1999; **41**: 212–223.
53. Butler RW, Davies PL, Jhun M. Asymptotics for the minimum covariance determinant estimator. *Ann. Statist*. 1993; **21**: 1385–1400.
54. Davies PL. The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *Ann. Statist*. 1992; **20**: 1828–1843.
55. Rocke DM, Woodruff DL. Identification of outliers in multivariate data. *J. Am. Statist. Assoc*. 1996; **91**: 1047–1061.
56. Rousseeuw PJ, Zomaren BC. Unmasking multivariate outliers and leverage points. *J. Am. Statist. Assoc*. 1990; **85**: 633–639.
57. Campbell NA, Lopuhaä HP, Rousseeuw PJ. On the calculation of a robust S-estimator of a covariance matrix. *Statist. Med*. 1998; **17**: 2685–2695.
58. Hawkins DM, Liu L, Young SS. Robust singular value decomposition. *National Institute of Statitical Sciences Technical Report 122*, 2001.
59. Liu L, Hawkins DM, Ghosh S, Young SS. Robust singular value decomposition analysis of microarray data. *P. Natl. Acad. Sci. USA*. 2003; 13167–13172.
60. Croux C, Filzmoser P, Pison G, Rousseeuw PJ. Fitting multiplicative models by robust alternating regressions. *Stat. Comput*. 2003; **13**: 23–36.
61. Golub GH, van Loan CF. *Matrix computations*. The John Hopkins University Press: Baltimore, 1989.
62. Martens H, Næs T. *Multivariate Calibration*. Wiley: Chichester, 1989.
63. Huber PJ. Projection pursuit. *Ann. Statist*. 1985; **13**: 435–475.
64. Li G, Chen Z. Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo. *J. Am. Statist. Assoc*. 1985; **80**: 759–766.
65. Xie Y, Wang J, Liang YZ, Sun L, Song X, Yu R. Robust principal component analysis by projection pursuit. *J. Chemometrics* 1993; **7**: 527–541.
66. Croux C, Ruiz-Gazen A. *A fast algorithm for robust principal components based on projection pursuit*. COMPSTAT, Physica-Verlag, 1996; 211–216.
67. Hubert M, Rousseeuw PJ, Verboven S. A fast method for robust principal components with applications to chemometrics. *Chemometrics Intell. Lab. Syst*. 2002; **60**: 101–111.
68. Hubert M, Rousseeuw PJ, Vanden Branden K. ROBPCA: A new approach to robust principal component analysis. *Technometrics* 2005; **47**: 64–79.
69. Engelen S, Hubert M, Vanden Branden K. A Comparison of three procedures for robust PCA in high dimensions. In *Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling*, 2004; 11–17.
70. Rivest E, Plante N. L'analyse en composantes principales robuste. *Rev. Stat. Appl*. 1988; **36**: 54–66.
71. Daigle G, Rivest LP. A robust biplot. *Can. J. Stat*. 1992; **20**: 235–241.
72. Croux C, Haesbroeck G. Principal components analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika* 2000; **87**: 603–618.
73. Woodruff DL, Rocke DM. Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *J. Am. Statist. Assoc*. 1994; **89**: 888–896.
74. Ruymgaart FH. A robust principal component analysis. *J. Multivariate Anal*. 1981; **11**: 485–497.
75. Ammann LP. Robust principal components. *Commun. Statist. Simulat*. 1989; **18**: 857–874.

Copyright © 2006 John Wiley & Sons, Ltd.

76. Stanimirova I, Walczak B, Massart DL, Simeonov V. A comparison between two robust PCA algorithms. *Chemometrics Intell. Lab. Syst*. 2004; **71**: 83–95.
77. Gabriel KR, Zamir S. Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics* 1979; **21**: 489–497.
78. Pravdova V, Estienne F, Walczak B, Massart DL. A robust version of the Tucker3 model. *Chemometrics Intell. Lab. Syst*. 2001; **59**: 75–88.
79. Pell PR. Multiple outlier detection for multivariate calibration using robust statistical techniques. *Chemometrics Intell. Lab. Syst*. 2000; **52**: 87–104.
80. Filzmoser P. Robust principal component regression. Aivazian S, Kharin Y, Rider L (eds). *Proceedings of the Sixth International Conference on Computer Data Analysis and Modeling*, Minsk, Belarusia, 2001; **1**: 132–137.
81. Hubert M, Verboven S. A robust PCR method for high-dimensional regressors. *J. Chemometrics*. 2003; **17**: 1–15.
82. Zhang MH, Xu QS, Massart DL. Robust principal components regression based on principal sensitivity vectors. *Chemometrics Intell. Lab. Syst*. 2003; **67**: 175–185.
83. Penã D, Yohai VJ. A fast procedure for outlier diagnostics in large regression problems. *J. Am. Statist. Assoc*. 1999; **94**: 434–445.
84. Egan WJ, Morgan SL. Outlier detection in multivariate analytical chemical data. *Anal. Chem*. 1998; **70**: 2372–2379.
85. Wakeling IN, Macfie HJH. A robust PLS procedure. *J. Chemometrics* 1992; **6**: 189–198.
86. Griep MI, Wakeling IN, Vankeerberghen P, Massart DL. Comparison of semirobust and robust partial least squares procedures. *Chemometrics Intell. Lab. Syst*. 1995; **29**: 37–50.
87. Cummins DJ, Andrews CW. Iteratively reweighted partial least squares: A performance analysis by Monte Carlo simulation. *J. Chemometrics* 1995; **9**: 489–507.
88. Gil JA, Romera R. On robust partial least squares (PLS) methods. *J. Chemometrics* 1998; **12**: 365–378.
89. Hubert M, Vanden Branden K. Robust methods for partial least squares regression. *J. Chemometrics* 2003; **17**: 537–549.
90. Serneels S, Croux C, Filzmoser P, Van Espen P. Partial robust M-regression. *Chemometrics Intell. Lab. Syst*. 2005; In press.
91. de Jong S. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics Intell. Lab. Syst*. 2003; **18**: 251–263.
92. Chen C. Robust regression and outlier detection with the ROBUSTREG procedure. In *Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference*; SAS Institute: Cary, NC, 2002.
93. *S-PLUS 6 Robust Library User's Guide*. Insightful Corporation: Seattle, WA, 2002.
94. Fox J. *An R and S-PLUS Companion to Applied Regression*. Sage Publications: Thousand Oaks, 2002.
95. Hubert M, Rousseeuw PJ, Vanden Branden K. http://wis.kuleuven.be/stat/robust.html [8 Aug., 2005].
96. Verboven S, Hubert M. LIBRA: A MATLAB library for robust analysis. *Chemometrics Intell. Lab. Syst*. 2005; **75**: 127–136.
97. Serneels S, Croux C, Filzmoser P, Van Espen P. http://chemometrix.ua.ac.be/dl/prm.php/[8 Aug., 2005].
98. Kenny JF, Keeping ES. Relative merits of mean, median, and mode. In *Mathematics of Statistics*, Van Nostrans NJ (ed). 1962; 211–212.

Copyright © 2006 John Wiley & Sons, Ltd.

P a p e r I I I

Peak alignment and robust principal component analysis of gas chromatograms of fatty acid methyl esters and volatiles.

Frosch Møller, S. and M. Jørgensen, B.

Journal of Chromatographic Science, 45; 169 - 228, 2007

Peak Alignment and Robust Principal Component Analysis of Gas Chromatograms of Fatty Acid Methyl Esters and Volatiles

Stina Frosch Møller and **Bo M. Jørgensen***
Danish Institute for Fisheries Research, DTU Build. 221, DK-2800 Kgs. Lyngby, Denmark

Abstract

Gas chromatograms of fatty acid methyl esters and of volatile lipid oxidation products from fish lipid extracts are analyzed by multivariate data analysis [principal component analysis (PCA)]. Peak alignment is necessary in order to include all sampled points of the chromatograms in the data set. The ability of robust algorithms to deal with outlier problems, including both sample-wise and element-wise outliers, and the advantages and drawbacks of two robust PCA methods, robust PCA (ROBPCA) and robust singular value decomposition when analysing these GC data were investigated. The results show that the usage of ROPCA is advantageous, compared with traditional PCA, when analysing the entire profile of chromatographic data in cases of sub-optimally aligned data. It also demonstrates how choosing the most robust PCA (sample or element-wise) depends on the type of outliers present in the data set.

Introduction

Chemometric tools, such as principal component analysis (PCA) for visualisation and data mining, are frequently used to analyse chromatographic data. In most cases, chromatographic data are transformed to peak areas, which are then used for further analysis. The method relies on subjective peak selection and peak identification and on integration parameters, which if not properly set, may cause great errors in the calculated peak areas. Implications of the data extraction method, thus, are incorporated in the PCA analysis. The disadvantages concerned with peak area analysis, such as loss of information due to the selection of a subset of peaks and to erroneous peak areas, can be avoided by using the entire chromatographic profile *per se* when analysing the data. In addition, peak shapes and information about the absence or presence of peaks are automatically included in the data analysis.

* Author to whom correspondence should be addressed: email boj@difres.dk.

Unavoidable retention time shifts from one run to another obscure differences due to chemical variations between samples. Because multivariate data analysis requires uniform presentation of data [i.e., all data vectors have to be of the same length with corresponding elements (variables) representing similar phenomena in all samples], an appropriate pre-processing technique to align the chromatograms is needed. Variations, thus, are not dominated by shifts between variables but by different levels of the variables as they should.

Several retention time alignment algorithms have been reported in the literature (1–3). In the present study, the correlation optimization warping (COW) algorithm (2), originally developed as a data pre-processing step in multivariate modelling of chromatographic data. The COW algorithm has been successfully employed to align chromatograms from gas chromatography (GC)–flame ionization detection (FID) (3,4) and GC–mass spectrometry (5) measurements. According to Tomasi et al. (4), COW is less flexible than other warping methods, thus giving fewer artefacts and improving the quality of the alignment when applied to complex chromatographic data. COW allows aligning complex chromatograms with different number of peaks, peak intensities, and peak widths. Furthermore, it corrects peak shifts in both directions and aligns many chromatograms simultaneously, without any knowledge or identification of peaks.

PCA, like most other common chemometric methods, is based on the less robust least squares estimation. This means that the presence of even one single outlier in the data set can hamper the analysis and lead to incorrect conclusions. Outliers are measurements that do not fit into the pattern or grouping shown by the majority of measurements in a properly designed experiment. The most common outlier types are complete sample measurements (data vectors), but also individual "strange" data elements in the chromatogram may be considered as outliers.

The outlier problem can be solved in two ways: (*i*) by diagnostics or (*ii*) by robust estimators (6). In the first approach, outliers are identified and expelled from the data set prior to making the chemometric model. A complication is that it may be difficult to identify outliers, even when multivariate data are available, and

Reproduction (photocopying) of editorial content of this journal is prohibited without publisher's permission.

the task gets harder and more time-consuming when the amount of data is huge. In the second approach, which is used in this paper, robust estimators are used instead of the ordinary non-robust least squares estimator. Robust methods reduce or remove the effect of outlying data points, allowing the remainder to predominantly determine the model.

In this study, the advantage of using all collected data points from the GC in the chemometric analysis combined with COW pre-processing is illustraited. Because of the outlier problem, concerning both sample-wise and element-wise outliers, the advantages and drawbacks of two robust PCA (ROBPCA) methods, ROBPCA (7) and robust singular value decomposition (RSVD) (8), are also investigated for the analysis of GC data. Opposite to the methods that rely on subjective peak selection and peak areas, the PCA analysis is able to identify relevant peaks and use all information contained in the chromatograms.

The analyses are performed on two data sets differing in quality. The first is GC–FID data from fatty acid methyl esters (FAME), which are "well behaved" in the sense that outliers are expected to be due to insufficient peak alignment only. The second data set consists of GC–FID data of volatile lipid oxidation productions (ATD), which have a relatively higher risk of artefacts and with larger sample differences and peak shifts.

Materials and Methods

Data sets

Gas chromatograms of FAMEs and of ATDs collected by dynamic head-space were kindly provided by the lipid group of the authors' institute. An FID was used for both types of chromatograms. The data from gas chromatograms of FAMEs show the fatty acid composition of triglycerides or phospholipids. In the present case, samples of fish oil from farmed rainbow trout fed two different diets were included. The data from gas chromatograms of ATDs show volatile lipid oxidation products (mostly aldehydes, ketones, and short-chain fatty acids). The samples included were from farmed rainbow trout kept frozen at –20°C, –30°C, or –80°C for 0–24 months. Detailed results concerning the experiments and the chemical findings are under preparation for publication.

The chromatograms were imported from the instrumental result files (ASCII text format) into MatLab 7.0.4 (The MathWorks) where the pre-processing (normalization, baseline correction, and alignment) and multivariate data analyses were performed. Each chromatogram was loaded into a MatLab workspace as a vector composed of the FID-signal collected over the duration of the GC run. The chromatograms were appended into a matrix where each row was the chromatogram from a single sample. The algorithms for COW and ROBPCA were downloaded from the literature (9,10). The algorithm for RSVD was kindly provided by A. Belousov (11).

Pre-processing of data

Pre-processing of the chromatograms prior to PCA is necessary to remove variations unrelated to chemical compositions. The pre-processing consists of baseline correction, normalization, and peak alignment using COW.

Baseline shift removal

Because baseline shifts affect both the warping and the normalization, a baseline correction is necessary. Furthermore, PCA cannot separate variance due to peak misalignment from variance due to baseline shifts. Hence, the baseline correction was essential. All chromatograms were individually baseline-corrected by subtracting the average signal for the last 1300 s and first 150 s, respectively, from the full chromatogram.

Normalization

Normalization to a constant area was used to compensate for differences in the amount of injected sample for Data set 1 (gas chromatograms of FAMEs), taking advantage of the unspecificity of the FID. Data set 2 (gas chromatograms of ATDs) was normalized by dividing each chromatogram by the injected amount of sample, giving informational value to the total amount of volatiles produced. In both cases, normalization was necessarily applied after baseline adjustment in order to give meaningful results.

Chromatographic alignment by COW

The aim of COW was to align two chromatographic profiles by piecewise linear stretching and compression, also known as warping, of the time axis of one of the profiles relative to the other. The chromatograms are subdivided into segments that were iteratively stretched and compressed by interpolation. The optimal alignment is the solution that maximizes the correlation between corresponding segments in the sample and the reference chromatogram. The number of data points each segment is allowed to change (maximal warping) is determined by the so-called slack parameter and depends on the peak shift to correct. According to Nielsen et al. (2) the optimal alignment will be achieved when the segment length is in the region of the number of data points making up the sharpest peak in the chromatogram.

The optimal chromatographic alignment settings in this study were selected as the segment length and slack that maximizes the first singular value as proposed by Christensen et al. (5). Combinations of segment lengths from 10 to 60 data points, and increments of 5 and slacks between 1 and 5 were tested to find the best settings. The optimal settings were based on the evaluation of the whole data set for the data from gas chromatograms of FAMEs and of 30 randomly selected samples for the data from gas chromatograms of ATDs.

PCA

The classical PCA method is not robust against outliers because of the least squares criterion. This means that even one single outlier in the data set can have an arbitrarily large effect on the model and lead to wrong interpretation and conclusions.

Different approaches have been proposed for making a robust version of PCA. They can be grouped as follows: (*i*) techniques that replace the classical covariance matrix by a robust covariance estimator (6,12,13) as the minimum covariance determinant (MCD) (14). Unfortunately, these approaches are limited to relatively low-dimensional data and are computational costly. (*ii*) Another group is methods that use projection pursuit (PP) techniques (15–20). PP searches for structure in high dimensional

data by projecting these data into a lower-dimensional space that maximizes a robust measure of spread called the projection index. These methods can handle situations where the number of variables exceeds the number of samples. (*iii*) A combination of (*i*) and (*ii*) called ROBPCA (7) is used, which should yield more accurate estimates than the raw PP algorithm. The final group (*iv*) involves adjustments to the internal computations of the singular value decomposition (SVD) algorithm by replacing the least squares criterion with a robust estimate (8,21,22). These RSVD methods can handle high-dimensional data and element-wise outliers. Element-wise outliers exist where one or several individual data elements in otherwise good rows are corrupted.

In this study, the classical least square PCA will be compared with the two robust versions, ROBPCA and RSVD. Both robust methods can handle situations with more variables (columns) than samples (rows), are computationally feasible, and have shown good performance in other studies (17,23).

ROBPCA

The ROBPCA approach combines PP with robust covariance estimation in lower dimensions (7). The ROBPCA method can be divided into three major steps. First, the data, stored in an $n \cdot p$ data matrix X, were pre-processed by reducing their data space to the affine sub-space spanned by the n observations. This was performed by SVD of the column mean-centred X, without loss of information. In the next step of the ROBPCA algorithm, PP was used for initial dimension reduction ($k << p$). A measure of "outlyingness" was computed for each data point. The h data points with smallest outlyingness were then retained, the covariance matrix of this h-subset computed, and the number of principal components to retain (k) selected. In the last step of the ROBPCA algorithm, the re-weighted MCD estimator is then applied to this lower dimensional data space to find a robust center and covariance estimator of the projected samples. Finally, these estimates were back-transformed to the original space, and a robust estimate of the location of X and of its scatter were obtained.

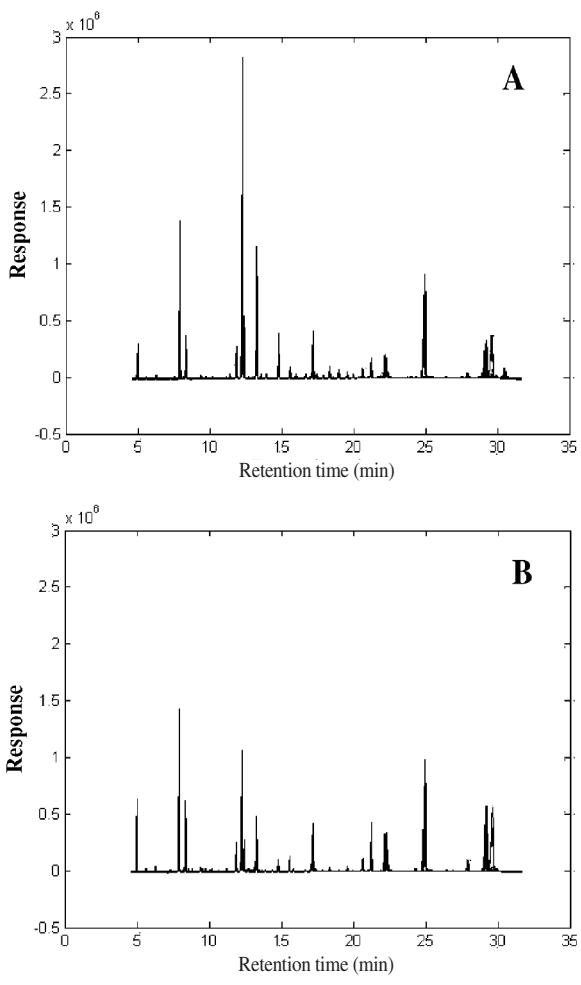

Figure 1. Chromatograms (GC–FID of FAMEs), after alignment using COW with a segment length of 15 data points and a slack of 3 points, of samples from fish fed on diets containing vegetable oil (A) or fish oil (B).

Robust singular value decomposition

This method, called RSVD (8), was based on the alternating least squares algorithm for SVD proposed by Gabriel (24). In this algorithm, the minimization problem was solved with criss-cross regressions, which involves iteratively computing dyadic (rank 1) fits using least squares regression. The original Gabriel–Zamir SVD algorithm is then rendered robust by substituting the non-robust least squares regression with a robust estimator, which in this case, was the alternating L1-norm (the sum of absolute residuals).

Results and Discussion

Data set 1 (GC–FID of FAMEs)

Optimal warping parameters

Figure 1 shows the aligned chromatograms appearing from fish whose feed contained mostly vegetable oil or pure fish oil, respectively. In all chromatograms, the same fatty acids appear, but with different concentrations, reflecting the different feed types. Fish fed vegetable oil contained higher amounts of 18:1 (n-9), 18:2 (n-6), and 18:3 (n-3) than did fish fed fish oil. On the other hand, fish fed fish oil contained the highest amount of 14:0, 16:0, 16:1 (n-7), 18:4 (n-3), 20:4 (n-3), 20:5 (n-3), 20:1 (n-9), 22:1 (n-11), 22:5 (n-3), and 22:6 (n-3). The relatively high amount of long chain polyunsaturated fatty acids in the fish fed vegetable oil is due to small amounts of fish meal in the feed.

The peak identified around 24.7 to 27.3 min in the un-warped data is due to an internal standard in some of the samples. This peak is isolated from the other peaks, and for that reason, it is possible to exclude the part of the chromatogram from the data analysis allowing samples both with and without an internal added standard to be included in the data matrix. If the part containing the standard was retained, severe artefacts in both the normalization step and in the following PCA modelling would occur.

The warping parameters segment length and slack were considered optimal when maximizing the first principal component from a PCA model fitted to the warped data. Combinations of segment lengths of 10 to 60 data points with increments of 5 data points and slacks between 1 and 5 were tested. Furthermore, the mean relative difference together with the maximal decrease and

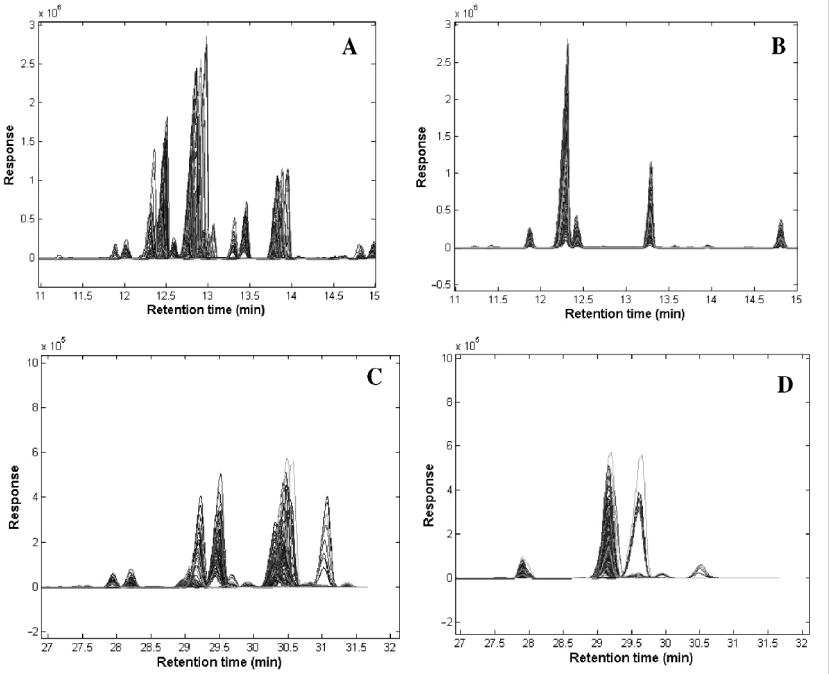

Figure 2. The effect of warping on two selected regions of the chromatograms (GC–FID of FAMEs): before warping (A and C); and after warping (B and D). For warping, COW was used with a segment length of 15 data points and a slack of 3 points.

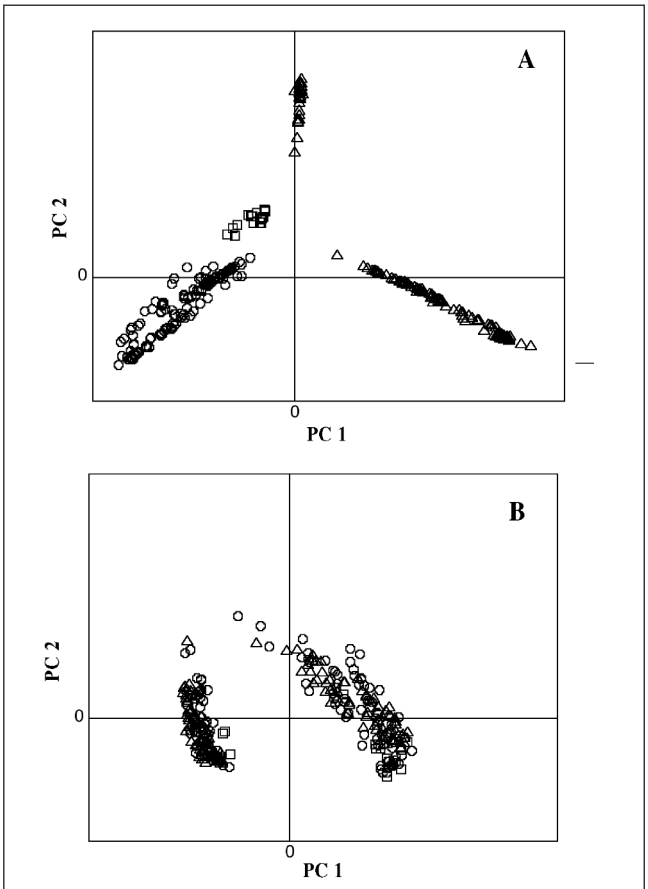

Figure 3. PCA scores: PC2 versus PC1, without warping (A) and with warping (B). The samples are marked according to frozen-storage time: 0 months (n), 4 months (1) and 24 months (s).

increase in area difference between the unwarped and the warped chromatograms were calculated for all tested settings to evaluate the warping effect on the chromatogram profiles.

The explained variance for a one-component model increased from 30.6% (un-warped and uncentred data) to 87.2%, attained with a segment length of 15 data points and a slack of 3. The absolute area of the chromatograms after warping was changed, on average, with 4.4% compared with the original chromatograms with a maximal decrease in the area of 12.0%, and a maximal increase in area of 6.0%. These changes in area were due to interpolation when warping the data. Four of the samples experienced a decrease in area of more than 10% compared with the original chromatograms. When comparing the raw data of these samples with the standard chromatogram, it appeared that they had large shifts in retention times, resulting in the maximum warping allowed. In Figures 2A and 2B, the effect of warping was illustrated on a selected region of the chromatograms where the improvement by warping was pronounced. However, in the last part of the chromatograms, the improvement was not that good (Figures 2C and 2D). This misalignment was caused by larger shifts in retention time in the last part of the chromatograms and might be addressed by modifying the COW algorithm. The chromatograms might be split into several segments along the retention time axis and different warping parameters used for each of these segments.

Alternatively, misalignment may be dealt with by using RSVD, a method that only excludes outlying elements. This means that it was not necessary to exclude whole samples because of misalignment in some part of the chromatograms because the properly aligned parts of the chromatograms are still available for analysis.

Principal component analysis

To investigate the effect of warping on the results obtained from PCA modelling, PCA was first applied to the mean-centered un-aligned data set. The score plot of PC1 versus PC2, from the model fitted to the un-aligned data, is presented in Figure 3A. Four distinct groups appear: three groups matching the storage period and time of analysis and one group where all samples belong to the same storage period, measured on the same day. Because of the experimental design, a confounding effect between storage period and time of analysis was unavoidable; it was, therefore, difficult to conclude if the grouping was due to storage period or time of analysis. When looking at the un-aligned chromatograms from samples stored for 24 months, a clear shift in retention time between the two groups appear, indicating that the clustering seen in Figure 3A was due to shifts in retention time, rather than chemical differences between the samples. Similar results were obtained when comparing the chromatograms for two groups separated along PC1. In Figure 3B the corresponding score plot of PC1 versus PC2, for a PCA

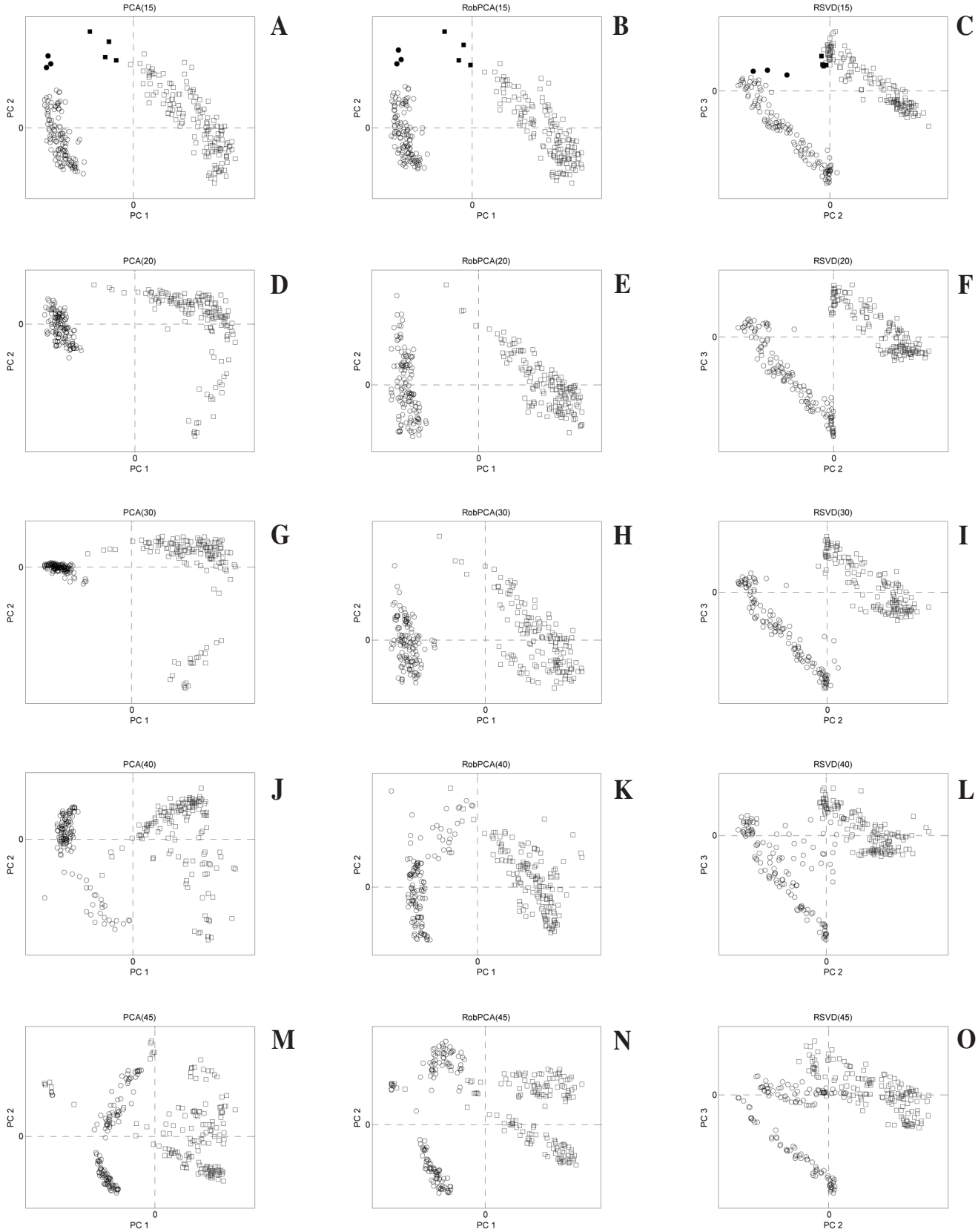

Figure 4. PCA scores: PC2 versus PC1 for classical PCA (column 1: A, D, G, J, and M) and for ROBPCA (column 2: B, E, H, K, N), and PC3 versus PC2 for RSVD (column 3: C, F, I, L, O). The chromatograms (GC-FID of FAMEs) were aligned by warping with the slack kept constant at 3 and varying segment lengths: 15 (A–C), 20 (D–F), 30 (G–I), 40 (J–L), and 45 (M–O) data points. The samples are marked according to oil type in the feed: vegetable oil (1) and fish oil (n). A few "extreme" samples are marked with filled symbols (A–C).

model fitted to the warped data, shows two groups only. These groups cannot be ascribed to a storage period or time of analysis, but they correlate to changes in the fatty acids profile caused by the different oils in the feed.

Furthermore, the loading plots for PC1 (37.6%) and PC2 (14.8%), from the un-warped data, showed complicated patterns, with many regions resembling the first derivative. This is typical for data distorted to a high degree by shifts in retention time (1). The shifts in retention time not only affect the first PC but also the subsequent components.

Thus, it was concluded that the pattern for the unaligned data was due to misalignments in the un-warped chromatograms rather than to chemical differences of the samples. For a reliable interpretation of the PCA model, alignment of the chromatograms is essential.

As illustrated in Figure 3, warping the chromatograms clearly improves a PCA, but it may be difficult to obtain optimal warping for all samples, especially in unsupervised situations. In that case, using robust PCA methods on the warped data may be helpful and provide better results than does the traditional PCA method, based, as it was, on least squares estimates. Moreover, even with perfectly aligned data, outliers may occur because of instrumental instability, etc. In this situation, the use of a robust methods was also of advantage.

In the score plot of PC1 versus PC2, both from traditional PCA and ROBPCA, clusters for each of the two treatments (vegetable or fish oil) were observed (Figure 4, first row). For both methods, PC1 scores discriminated between fish oil and vegetable oil, whereas PC2 scores displayed the variance between individuals in each group. Samples of fish fed vegetable oil were characterized by a high concentrations of 18:1 (n-9), 18:2 (n-6), and 18:3 (n-3), as their peaks in the chromatogram were positively loaded in PC1, and lower concentrations of 14:0, 16:0, 16:1 (n-7), 20:4 (n-3), 20:5 (n-3), 22:1 (n-11), and 22:6 (n-3), with peaks highly negatively loaded in PC1. The opposite results were obtained for samples of fish feed with fish oil.

In neither of the two models (traditional PCA and ROBPCA) was PC2 correlated to the experimental design, but this was primarily due to biological variation within the groups and to artefacts, such as a suboptimal baseline correction. No other groupings where found in higher order PCs. The difference in baseline was especially pronounced for the extreme samples with high score values in PC2 in both traditionally PCA and ROBPCA (filled symbols).

An even better class separation was obtained with elementwise robust PCA (Figure 4). No centering of the data was built in this RSVD algorithm, as was the case for ROBPCA, meaning that the first PC explained the centering of the data and was, for that reason, not interesting. PC2 and PC3 explained 60.0% and 22.1%, respectively, of the variance when PC1 was excluded, and these PCs are both relevant for the clustering. The same fatty acids, as found from the two previous models, were responsible for the clustering in Figure 4.

The explained variance in the first PC increases with ROBPCA 77.8%, compared with traditional PCA, 69.7%. The cumulative variance of the two components from RSVD, associated with the clustering due to different oils in the feed, was estimated to 82.1%. The variance was concentrated in the robust models, as a result of excluding outlying samples or outlying elements from the modelling step leading to increased class separation and reduced within-class variation.

In the former paragraphs it was illustrated that for well warped data, the results obtained with traditional PCA and ROBPCA were fairly good, even though the result can be improved by using the robust SVD method. Now, it will now be interesting to compare the PCA methods with decreasing data quality to investigate how well the data need to be aligned in order to yield acceptable results according to clustering. The data quality was based on the explained variance for the different warping parameters tested, fitting a one component model (PCA) to the normalized, but un-centred data (5). The slack was kept constant at 3, and the segment length was increased from 15 to 50 data points. The explained variance for a one component model when evaluating the warping parameters was: segment 20, 86.0%; segment 30, 84.4%; segment 40, 79.6%; segment 45, 72.4%; and segment 50, 67.0%.

The score plots in Figure 4 illustrate the effect of reduced data quality on the three different principal component analysis procedures. Results obtained for data warped with a segment length of 50 data points are not displayed, as they were similar to the results obtained with data warped with a segment length of 45 data points. A clustering according to different types of oil in the feed was observed for all three methods for data of high quality, although the clearest clustering was obtained with the two robust methods. With decreasing data quality (i.e., 79.6% explained variance and below in this case) the plot gets more unclear regardless of which PCA method was used to analyze the warped data. This clearly illustrates that data, and thereby the warping, need to be of a certain quality to obtain reliable results. The robust methods can not remedy problems with large shifts in retention time.

Data set 2 (GC–FID of ATDs)

Optimal warping parameters

Figure 5 shows aligned chromatograms for samples stored at –20°C and –80°C. The profiles and the total amount of oxidation products depend strongly on the storage temperature, as would be expected. The number of peaks and their areas are much higher for samples stored at –20°C than for those stored at –80°C (The storage time was 24 months in both cases).

The highest obtained explained variance for a one component un-centred PCA model was 80.1%, attained with a segment length of 20 data points and a slack of 3. In comparison, the explained variance for a one component model of un-warped and un-centred data was only 65.8%.

Principal component analysis

The score values of PC1 and PC2 from both traditional PCA and ROBPCA, as well as of PC2 and PC3 from RSVD, are shown in Figure 6. The samples are marked according to their storage temperature. For all three models, PC1 scores (PC2 for RSVD) turned out to be reasonable in storage temperature. The scores went from one sign to the other related to storage temperatures from –80°C or –30°C to –20°C. The clearest grouping according to storage temperature, –80°C or –30°C versus –20°C was observed with RSVD. No big difference in PC1 scores was

observed between classical PCA and ROBPCA. Three outlying samples were separated from the other samples along PC2 (PC3 for RSVD). With ROBPCA, the three outliers are excluded from the modelling step and are placed closer to the other samples. Additionally, the variation accounted for by PC2 scores (PC3 for RSVD) was due to variation within each storage time, reflecting the biological variation. It was not possible to identify other patterns in the data by plotting other combinations of principal components.

The explained variance for PC1 and PC2 was 62.1% and 19.4%, respectively, for classical PCA and 76.3% and 11.%, respectively, for ROBPCA, resulting in a slightly higher explained variance for a two component model when applying ROBPCA. For RSVD, the explained variance for PC2 and PC3 was 6.0% and 22.1%, respectively. The low explained variance was a result of the presence of the outlying samples; only the first principal component was associated with a common variation between all samples, whereas the following components were primarily associated with the outlying samples. PC5 from RSVD accounted for 23.5% of the explained variance and was only caused by the three outlying samples (results not shown).

The chromatographic profiles of the three outliers were almost identical. A comparison of the chromatograms from the three outliers with the other samples stored at –30°C showed that the profile from the outliers were outstanding from the other chromatograms, with some peaks reaching higher or lower intensities, whereas other peaks were missing or only found for the three outliers. The full data vectors of these samples may, therefore, be regarded as outliers. This can also explain why the robust SVD method was not able to handle these outliers efficiently. All elements from the sample ought to be excluded, but the method can "only" handle up to 50% outlying elements in each data vector. The data set was not perfectly warped, meaning that all peaks are not perfectly warped and outlying elements exists. This is why different groupings are observed

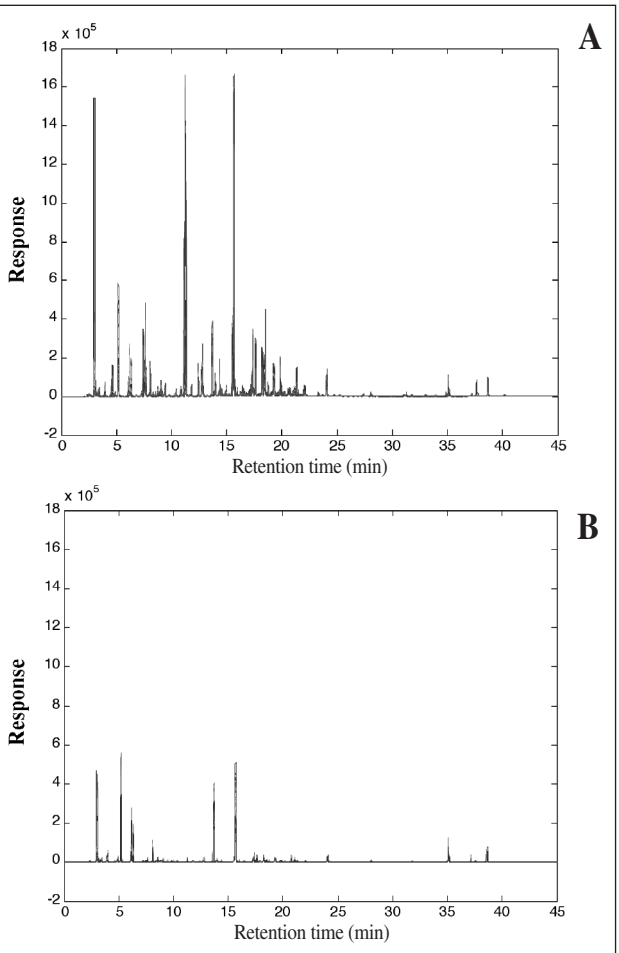

Figure 5. Chromatograms (GC–FID of ATDs) of samples stored at –20°C (A) and –80°C (B).

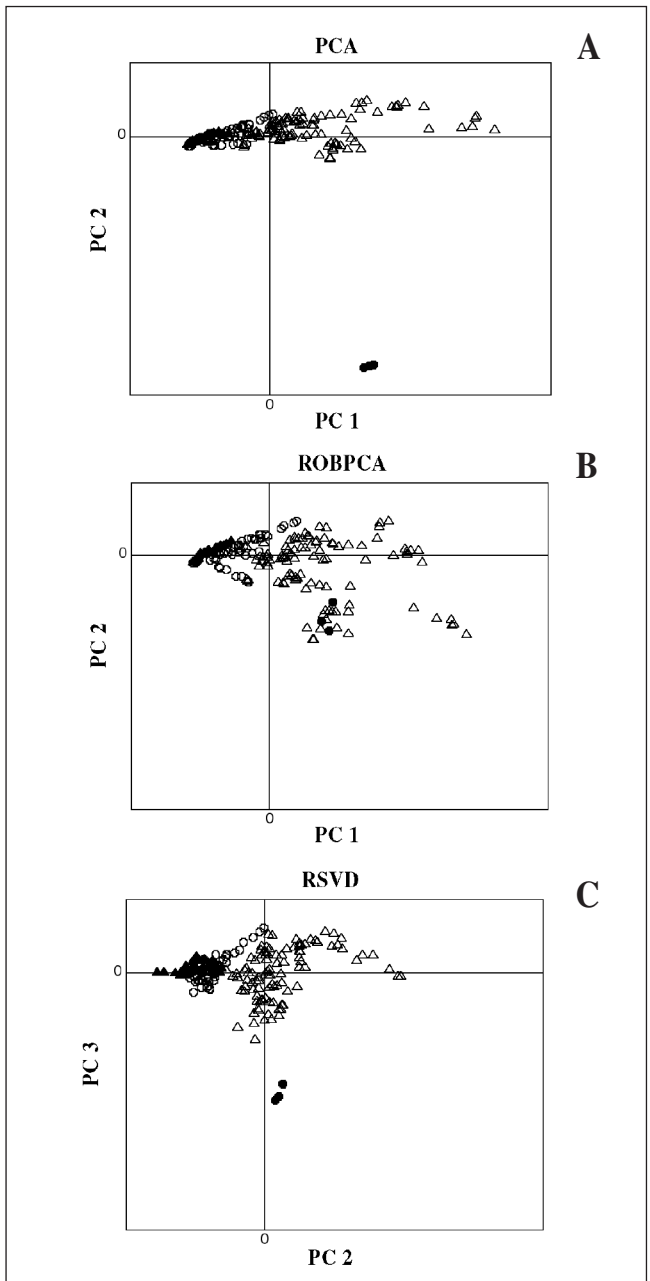

Figure 6. PCA scores: PC2 versus PC1 for classical PCA (A), ROBPCA (B), and PC3 versus PC2 for RSVD (C). The chromatograms (GC-FID of ATDs) were aligned by warping with a slack of 3 and a segment length of 20 data points. The samples are marked according to storage temperature: –20°C (s), –30°C (1), and –80°C (s). Three outliers (all –30°C samples) are marked with filled circles.

between ROBPCA and RSVD in the actual situation: in ROBPCA, the entire sample is excluded from the modelling step, leaving out the three outliers completely and thereby assigning PC2 to another, perhaps more interesting, variation.

Conclusion

In designed experiments where one looks at a whole set of chromatograms at a time, multivariate data analysis is a useful alternative to classical peak selection and area calculation procedures. Alignment of the chromatograms is necessary and may, to a large extent, be done by automatic procedures. In situations where only suboptimal alignment is obtained, or other situations where outlying measurements occur (e.g., because of bad baselines or errors in sample amount injected) robust algorithms are to be preferred in order to keep the outliers from severely interfering with the multivariate models. Situations where only some part of the chromatograms are not properly aligned are best dealt with by using element-wise robust methods (e.g., RSVD). When the outliers are due to features throughout the chromatogram, sample-wise robust methods (e.g., ROBPCA) perform the best.

Acknowledgment

This work was supported by a grant to SFM from the Danish Ministry of Food, Agriculture, and Fisheries. Gas chromatograms of FAMEs and of ATDs collected by dynamic headspace were kindly provided by the lipid group (C. Jacobsen) of the authors' institute.

References

1. G. Malmquist and R. Danielsson. Alignment of chromatographic profiles for principal component analysis: a prerequisite for fingerprinting methods. *J. Chromatogr. A* **687:** 71–88 (1994).
2. N.-P.V. Nielsen, J.M. Carstensen, and J. Smedsgaard. Aligning of single and multiple wavelength chromatographic profile for chemometric data analysis using correlation optimized warping. *J. Chromatogr. A* **805:** 17–35 (1998).
3. K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, and R.E. Synovec. Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis. *J. Chromatog. A* in press (2007).
4. G. Tomasi, F. van den Berg, and C. Andersson. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J. Chemomet.* **18:** 231–41 (2004).
5. J.H. Christensen, G. Tomasi, and A.B. Hansen. Chemical fingerprinting of petroleum biomarkers using time warping and PCA. *Environ. Sci. Technol.* **39:** 255–60 (2005).
6. P. Rousseuw and A.M. Leroy. *Robust Regression and Outlier Detection.* John Wiley & Sons, New York, NY, 1987.
7. M. Hubert, P.J. Rousseuw, and K. Vanden Branden. ROBPCA: a new approach to robust principal component analysis. *Technometrics* **47:** 64–79 (2005).
8. D.M. Hawkins, L. Liu, and G.S. Young. "Robust singular value decomposition", National Institute of Statistical Sciences, Technical Report 122. Research Triangle Park, NC, 2001.
9. http://www.models.kvl.dk. Date accessed (January 2006).
10. http://www.wis.kuleuven.ac.be/stat/robust.html. Date accessed (January 2006).
11. A. Belousov. Westfalischen Wilhelms University. Personal Communication (2006).
12. P.L. Davies. Asymptotic behavior of S-estimators of multivariate location and dispersion matrices. *Ann. Statist.* **15:** 1269–92 (1987).
13. C. Croux and G. Haesbroeck. Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika* **87:** 603–18 (2000).
14. P. Rousseeuw. Least median of squares regression. *J. Am. Statist. Assoc.* **79:** 871–80 (1984).
15. G. Li and Z. Chen. Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo. *J. Am. Statist. Assoc.* **80:** 759–66 (1985).
16. J.S. Galpin and D.M. Hawkins. Methods of L1 estimation of a covariance matrix. *Comput. Statist. Data Anal.* **5:** 305–19 (1987).
17. Y. Xie, J. Wang, Y. Liang, L. Sun, X. Song, and R. Yu. Robust principal component analysis by projection pursuit. *J. Chemometrics* **7:** 527–41 (1993).
18. C. Croux and A. Ruiz-Gazen. A fast algorithm for robust principal components based on projection pursuit. COMPSTAT, Physica-Verlag, Heidelberg, Germany, 1996, pp. 211 – 216.
19. M. Hubert, P.J. Rousseeuw, and S. Verboven. A fast method for robust principal components with applications to chemometrics. *Chemometrics Intell. Lab. Syst.* **60:** 101–11 (2002).
20. C. Croux and A. Ruiz-Gazen. High breakdown estimators for principal components: the projection-pursuit approach revisited. *J. Multivariate Anal.* **95:** 206–26 (2005).
21. L. Liu, D. Hawkins, S. Ghosh, and S. Young. Robust singular value decomposition analysis of microarray data. *P. Natl. Acad. Sci. USA* **11:** 13167–72 (2003).
22. C. Croux, P. Filzmoser, G. Pison, and P.J. Rousseeuw. Fitting multiplicative models by robust alternating regressions. *Stat. Comput.* **13:** 23–36 (2003).
23. M. Hubert and S. Englen. Robust PCA and classification in bioscience. *Bioinformatics* **20:** 1728–36 (2004).
24. K.R. Gabriel and S. Zamir. Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics* **21:** 489–97 (1979).

Manuscript received January 13, 2006.

P a p e r IV

Automatically identifying scatter in fluorescence data using robust techniques.

Engelen, S., Frosch Møller, S. and Hubert, M.

Chemometrics and Intelligent Laboratory Systems, 86: 35 – 51, 2007

Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 86 (2007) 35–51
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Automatically identifying scatter in fluorescence data using robust techniques

Sanne Engelen [a,*], Stina Frosch Møller [b], Mia Hubert [a]

[a] Katholieke Universiteit Leuven, Department of Mathematics, W. De Croylaan 54, 3001 Leuven, Belgium
[b] Department of Seafood Research, Danish Institute for Fisheries Research, The Technical University of Denmark, Søltofts Plads, Building 221, 2800 Kgs. Lyngby, Denmark

Received 1 April 2006; received in revised form 4 August 2006; accepted 8 August 2006
Available online 20 September 2006

Abstract

First and second order Rayleigh and Raman scatter is a common problem when fitting Parallel Factor Analysis (PARAFAC) to fluorescence excitation–emission data (EEM). The scatter does not contain any relevant chemical information and does not conform to the low-rank trilinear model. The scatter complicates the analysis instead and contributes to model inadequacy. As such, scatter can be considered as an example of element-wise outliers. However, no straightforward method for identifying the scatter region can be found in the literature. In this paper an automatic scatter identification method is developed based on robust statistical methods. The method does not demand any visual inspection of the data prior to modeling, and can handle first and second order Rayleigh scatter as well as Raman scatter in various types of EEM data. The results of the automated scatter identification method were used as input data for three different PARAFAC methods. Firstly inserting missing values in the scatter regions are tested, secondly an interpolation of the scatter regions is performed and finally the scatter regions are down-weighted. These results show that the PARAFAC method to choose after scatter identification clearly depends on the data, for example signal to noise ratio and overlap between signal and scatter.
© 2006 Elsevier B.V. All rights reserved.

Keywords: Raman and Rayleigh scatter; Automated method; Robustness; ROBPCA; PARAFAC; Fluorescence

1. Introduction

Fluorescence spectroscopy is a fast, non-destructive technique with high sensitivity and specificity for providing information (quantitative and qualitative) about fluorescent molecules and their environment in a wide variety of biological materials. In fluorescence excitation–emission spectroscopy, each sample is measured by the excitation of the sample at several wavelengths and measuring the emitted light at several wavelengths. The result of such a measurement is an excitation–emission matrix (EEM). When several samples (I) are measured the data can be arranged in a three-way array, X ($I \times J \times K$), where $j=1, \ldots, J$ and $k=1, \ldots, K$ represent the emission and excitation mode respectively. Parallel factor analysis (PARAFAC) [1,2] is a widespread method for modeling such fluorescence excitation–emission landscapes (see e.g. [3–8]). PARAFAC decomposes the fluorescence data into trilinear components according to the number of fluorophores (F) present in the samples. The structural model can be described as

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk}$$

where x_{ijk} is the intensity of sample i at emission wavelength j and excitation wavelength k, and where a_{if}, b_{jf} and c_{kf} are parameters describing the importance of the samples/variables to each component f. The residual e_{ijk} contains the variation not captured by the PARAFAC model [9]. For approximate low-rank trilinear data the relative concentrations of analyte f and pure analyte spectra from fluorescence measurements of chemical analytes in mixtures can be extracted, when the correct number of components, equal to the number of fluorophores present in the data, is used.

* Corresponding author.
E-mail addresses: sanne.engelen@wis.kuleuven.be (S. Engelen), sfr@dfu.min.dk (S.F. Møller), mia.hubert@wis.kuleuven.be (M. Hubert).

0169-7439/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2006.08.001

A common phenomenon, and problem, when fitting PARAFAC to an excitation–emission matrix, is the light scatter effects, such as Raman and first and second order Rayleigh scattering [10–12]. In an EEM-landscape the scatter can be typically found as depicted in Fig. 1. This scatter is due to a physical process, which happens when light passes through some kind of medium, like e.g. water. As such, the scatter contains no chemical information and does not conform to the low-rank trilinear model. Therefore it will probably give a model inadequacy, influencing the estimated model parameters [10,12]. Different proposals of how to handle these scatter effects, can be found in the literature; subtracting a standard [13,14], down-weighting the scatter [15–17], inserting missing values [9], avoiding the part containing scatter [4], inserting zeros outside the data area [11] or interpolating the scatter area [18,19]. Unfortunately, all of the proposed methods seem to have some drawbacks. Some of them can only be used in special cases. Others give rise to unacceptable decomposition of the spectra, affect the convergence of the PARAFAC algorithm or are computational cumbersome [10–12]. A common problem is the visible inspection of the data before the methods can be applied. This makes it difficult to perform all these methods on several data sets at once. It even becomes harder to reduce the effect of scatter when the signal and scatter are overlapping, which is often the case.

In this chapter, we present an automated scatter identification tool using robust statistical methods. Robust statistics overcome badly modeled data due to outliers, i.e. samples that deviate much from the majority of the data points. It is well known that estimates based on a least squares condition are corrupted by outliers in a sense that the models explain the outliers very well, but fit the majority of the data poorly. A lot of research has been done the last decades to adapt known algorithms or to create new ones that can cope with such anomalous observations. For instance in the context of principal components analysis (PCA), the least squares model can be heavily influenced by already one single outlier. Therefore, different robust PCA methods are developed. Among them are the Reflection based Algorithm for PCA (RAPCA) [20] and the Robust PCA (ROBPCA) procedure [21].

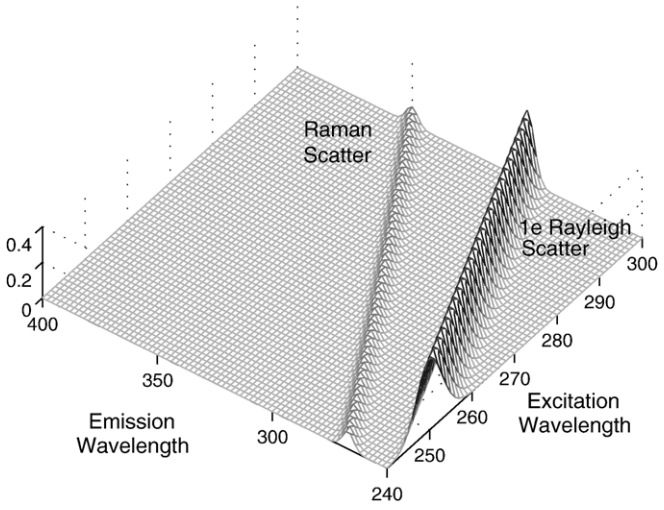

Fig. 1. Raman and Rayleigh scatter in an EEM landscape.

The output of these robust multivariate methods is two-fold. Firstly, the provided model fits the majority of the data and is stable in the presence of anomalous points. Secondly, each sample is marked as a regular observation or an outlying point for the concerned model, making all these robust procedures useful as outlier identification methods.

An outlying sample can have abnormal values for one or several variables, or it might be deviating from the majority of the samples for almost all of its variables. In three-way data, the latter situation implies that the whole sample landscape is highly different from the others. But outlying values (elements) can also occur in many or all samples. This element-wise contamination often occurs in multi-way data. A typical example is scattering which affects all samples, and which gives rise to many unexpectedly high values, certainly compared to the other values in the neighborhood.

The correction towards both types of outliers is highly recommended for the PARAFAC model, as an alternating least squares algorithm is used to estimate the scores and loadings [9]. For that reason, the algorithm breaks down when the three-way data contain outlying samples or/and outlying elements. However, a traditional approach, such as the element-wise $L1$-approach, suggested by [22], for handling outlying elements in combination with scattering will not work well (results not shown). In this $L1$-approach, PARAFAC estimates are found by minimizing the $L1$-norm of the residuals instead of the Frobenius norm. This approach would even often lead to discarding chemical information, while keeping the scatter in. Also the robust PARAFAC method for outlying samples proposed in Ref. [23] cannot handle three-way data with scattering. The major problem with scattering is that it is a systematic corruption of the data within a sample and situated for all the samples in more or less the same area. Randomly placed outlying elements would be easier to handle, but for data sets with systematic deviating parts, it is even not trivial to find for instance robust initial loadings. Nevertheless, robust techniques can still be used in a less conventional way as outlier detection tools to establish an automated identification of the scattering. We focus on ROBPCA, because it can handle high-dimensional two-way data and it is an excellent tool for outlier detection. Moreover, in Ref. [21] it is shown that ROBPCA outperforms several other robust PCA methods, such as projection pursuit techniques (e.g. [24,20]) and spherical and elliptical PCA [25].

In the following section we elaborate on this ROBPCA algorithm together with the automated search-engine for element-wise outliers in the form of scattering. Then, we assess the proposed procedure using two laboratory-made data sets in Section 3. The first one is a well-known standard data set, whereas the second one contains highly overlapping components and impurities. Finally, in Section 4 we apply the technique on two real-life examples, where in the first case the algorithm is evaluated for really noisy data. The second data set also is challenging, as the scattering and the signal are very difficult to separate.

All used programs are written in MATLAB. Most of them concerning the robustness are available in the LIBRA toolbox [26], which can be downloaded from http://www.wis.kuleuven.be/stat/robust.html. The programs handling multi-way data are available in the PLS-toolbox [27].

2. The automated scatter procedure

The proposed method for automated scatter identification is based on ROBPCA [21]. ROBPCA prevents the corruption of the principal components by outliers through a combination of robust subspace estimation (based on projection pursuit techniques) and the Minimum Covariance Determinant (MCD) estimator [28] for robust covariance and center estimation. A crucial step in ROBPCA and in the MCD procedure is the search for an outlier-free subset of size h, which will then be used for parameter estimation. The value of h lies between half the number of samples and the total number of samples, n. The higher h, the more accurate, but the less robust the algorithm will be, and vice versa. The default value of h is equal to $\lfloor 0.75n \rfloor$, which often ensures a good compromise between robustness and efficiency.

In the first step of ROBPCA, a preliminary PCA has been performed, such that all the data points are projected in their own space. This means a large dimension reduction for high-dimensional data sets. Secondly, a measure for how far a data point lies from the majority of the other samples, called the outlyingness, is defined for all samples. A further dimension reduction is then obtained by representing all the observations in the space spanned by the d dominant eigenvectors of the h points with smallest outlyingness, with d being the number of principal components to retain. In the next step a reweighted MCD procedure is performed, which provides a robust center and covariance matrix of the d-dimensional data. The principal components are determined as the eigenvectors belonging to the d largest eigenvalues of this covariance matrix. Finally, they are back-transformed to the original data space.

Outlier identification with ROBPCA is obtained by considering two distances for each observation. The orthogonal distance of an observation is defined as the distance between the point and the subspace spanned by the principal components. The second distance, the score distance, can be obtained by computing a robust, Mahalanobis-type distance in the space spanned by the principal components of an observation to the center of the data. If one of these two distances exceeds a certain cut-off value, a sample is flagged as an outlier and receives a zero weight. Other observations obtain a weight equal to 1. The cut-off value for the score distance is based on the assumption that the projected data are normally distributed. As such, the score distances are approximately χ^2-distributed and the 97.5% quantile of this distribution is taken as cut-off value for the score distances. On the other hand, it can be proven that the orthogonal distances to the power 2/3 are approximately normally distributed ([29,30]). The 97.5% quantile of the normal distribution to the power 3/2 is therefore taken as cut-off value for the orthogonal distances. For more information on the cut-off values, we refer to Ref. [21]. Hence, a weight vector is given as an extra output when applying ROBPCA, which determines whether a point is an outlier or not.

To elaborate on the construction of the automated scatter identification method, we first remark that the ROBPCA method can only be performed on two-way data matrices, which should be extracted from three-way data like EEM. In Fig. 2 the considered two-way matrices are illustrated. If there is scattering present in the three-way data, it can be found in each observation as shown for example in Fig. 3. By slicing these data along the sample mode, the scattering is situated in one or more diagonal lines in each sliced observation (see Fig. 2B). This is a problem for ROBPCA, as the scattering might corrupt all or at least a large majority of the variables (element-wise corruption).

On the other hand, slicing the three-way data along the B- or C-mode establishes useful two-way matrices, in which the

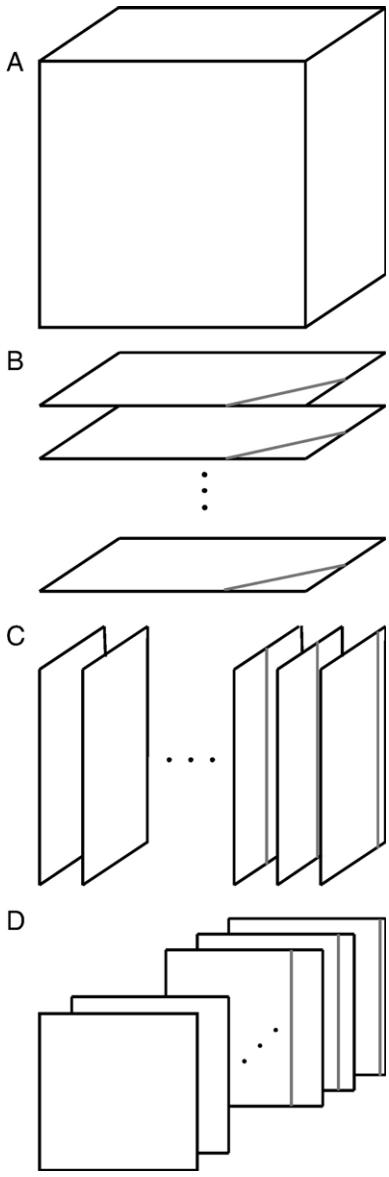

Fig. 2. A visualization of the scattering in three-way data (A) sliced in (B) the first modes, (C) the second mode and (D) the third mode. The grey line represents the scattering.

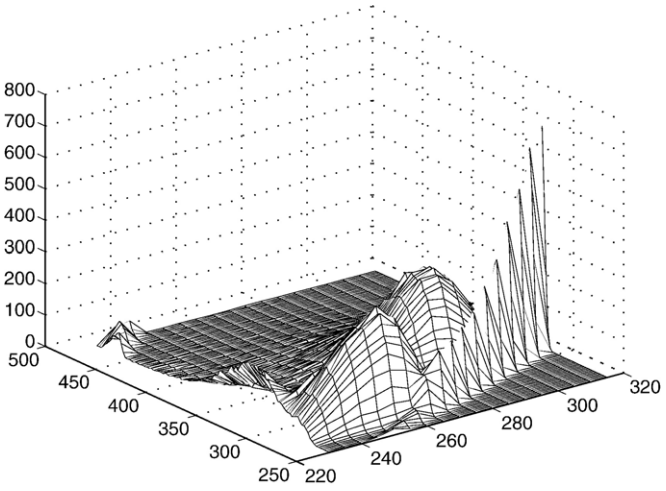

Fig. 3. Example (Sample 4) of a full excitation–emission landscape obtained from the fluorescence measurement. The 1st and 2nd order Rayleigh ridges are clearly seen as diagonal ridges. The 1st order Rayleigh scatter ridge to the right is situated at the diagonal where excitation and emission wavelengths are equal.

scattering is situated in columns for some of these matrices. So, by applying ROBPCA on the transpose of the sliced matrices in the *B*- and *C*-mode leads to the identification of the scattering, because the scattering is now manifesting as an outlying row in the considered two-way matrices (see Fig. 2C and D).

Thus, in the first step of the identification algorithm, ROBPCA is applied on the transpose of each matrix obtained by slicing the data along the emission mode, noted by $X(:, j, :)$ for the jth slice. So for each $j=1, \ldots, J$, a weight vector $\boldsymbol{w}_{B,j}$ is created which assigns 1 to a column of $X(:, j, :)$ that is a regular point and 0 to an outlier. All the weight vectors are stored in a $(K \times J)$ weight matrix $\mathbf{W}_B$.

In the second step of the algorithm, the same is done for the matrices obtained by slicing the data along the excitation mode, which is $X(:, :, k)$ for the kth excitation wavelength. Again, a weight vector $\boldsymbol{w}_{C,k}$ is obtained for each $k=1, \ldots, K$ analogously as for $\boldsymbol{w}_{B,j}$ and the $(J \times K)$ weight matrix $\mathbf{W}_C$ is constructed.

In the next step, both weight matrices $\mathbf{W}_B$ and $\mathbf{W}_C$ are converted to $(I \times J \times K)$ weight arrays $\mathbf{W}_B$ and $\mathbf{W}_C$, by repeating $\mathbf{W}_B$ and $\mathbf{W}_C$ for each sample and permuting the dimensions of both arrays. Taking the same weight matrices for each observation is justified because the scattering is present in the same area for all observations.

Now, we have two weights $w_{B,ijk}$ and $w_{C,ijk}$ from $\mathbf{W}_B$ and $\mathbf{W}_C$ respectively for each data element x_{ijk}. The next step consists of merging both weights such that each data element has only one corresponding weight w_{ijk}. This weight w_{ijk} finally defines whether the data element is outlying or not. We take the maximum of both weights $w_{B,ijk}$ and $w_{C,ijk}$ to obtain the final weight $w_{ijk}=\max(w_{B,ijk}, w_{C,ijk})$. This means that a weight w_{ijk} is still assigned a value of 1 or 0. Other weighting schemes have been tested, by substituting the minimum instead of the maximum and a smoother version, where values between 0 and 1 are allowed. But the minimum weights nor the smoother weights work well (the results are not included). Mostly they succeed in identifying the scattering, but too much of the signal is also omitted, which leads to inaccurately estimated PARAFAC parameters. The reason that the maximum weight works well to identify scattering, is that scattered elements are outliers that appear in both modes. By taking the maximum, points that are outliers in both modes are only marked as deviating samples for the whole data. So the maximum operator gives the best balance between finding the scattering, without indicating too much of the signal as being outlying.

A final step in the algorithm is turning isolated weights that are assigned a value of 0, i.e. weights that are not surrounded by other zero weights, back to 1, as these are not indicating scattering, but parts of the signal.

Note that when applying ROBPCA for each $j=1, \ldots, J$ and each $k=1, \ldots, K$, it should be determined how many components d are retained. In principle, this should be done $J+K$ times by employing common tools such as the scree-plot (see e.g. [31]) or the robust PRESS-curve ([32]). However, this is not advisable here, as this would require many user inputs and hence would result in a highly non-automated method. We thus advice to choose a fixed value d. This has the additional advantage that the sliced data sets $X(:, j, :)'$ and $X(:, :, k)'$ are all investigated on outliers towards a principal components space of the same dimension. From our experience, this optimal dimension d lies between 3 and 10. When decreasing d below 3, too many information in the data can be lost, which can lead to not-identified scatter areas. This should be avoided at any time. A too large value of d on the other hand, results in flagging smaller parts of the signal as outlying. This is not a major problem, but it also leads to a computationally more cumbersome ROBPCA algorithm. In our examples of Section 3 and Section 4, we have compared the marked scatter areas for d ranging from 1 to 10. All the results were comparable from $d=3$ to 10 components for large enough data sets. However, when J or K are really small, d should be taken large enough, such that the signal and the scatter can still be separated by ROBPCA. Taking all these results into consideration, we have set $d=10$ by default.

To summarize, the proposed method, for which a schematic overview can be found in Table 1, flags areas in the data that are considered as outliers in an automated way, i.e. without visual

Table 1
A schematic overview of the scatter identification algorithm

1. For the data sliced along the *B*-mode:
   - For each $j=1, \ldots, J$:
     - Perform ROBPCA on $X(:, j, :)'$
     - Store the weights $\boldsymbol{w}_{B,j}$ $(1 \times K)$
     - Create the jth row of $\mathbf{W}_B$: $\mathbf{W}_B(j, :)=\boldsymbol{w}_{B,j}$
   - Convert $\mathbf{W}_B$ to $\mathbf{W}_B$: $\mathbf{W}_B(i, :, :)=\mathbf{W}_B$ for each $i=1, \ldots, I$
2. For the data sliced along the *C*-mode :
   - For each $k=1, \ldots, K$:
     - Perform ROBPCA on $X(:, :, k)'$
     - Store the weights $\boldsymbol{w}_{C,k}$ $(1 \times J)$
     - Create the kth row of $\mathbf{W}_C$: $\mathbf{W}_C(k, :)=\boldsymbol{w}_{C,k}$
   - Convert $\mathbf{W}_C$ to $\mathbf{W}_C$: $\mathbf{W}_C(i, :, :)=\mathbf{W}_C'$ for each $i=1, \ldots, I$
3. Define the final weights $w_{i,jk}=\max(\mathbf{W}_B(i, j, k), \mathbf{W}_C(i, j, k))$ for each i, j and k.
4. Turn isolated zero-weights back to 1.

inspection of the data. To find the final parameter estimates by PARAFAC, the approaches mentioned in the introduction can be performed. One of them is inserting missing values in the areas that obtained a weight $w_{ijk}=0$. Another option is to estimate the values in the outlier area by means of interpolation [19]. A third possibility to estimate the PARAFAC loadings can be found in fitting the weighted PARAFAC model of Refs. [17,15], where the weights are equal to w_{ijk}. As we are working with fluorescence data which should be strictly positive, non-negativity constraints are used in all modes during this study.

To assess the proposed scatter identification method, we apply it on different kinds of data. We focus on how well the scattering is reduced from the data and how well the signal is preserved with the automated method. Moreover, we investigate the performance of the automated technique in combination with the missing values, the interpolation and the weighted PARAFAC option. The laboratory-made data sets are treated in Section 3 and the environmental data sets are analyzed in Section 4.

3. The analysis of laboratory-made data

3.1. Dorrit data

The method was tested on fluorescence data, containing mixtures of four known fluorophores [33,34]. The four compounds are phenylanaline, 3, 4-dihydroxyphenylalanine (DOPA), 1, 4-dihydroxybenzene and tryptophan. For every sample an excitation–emission matrix was obtained by measuring the emission spectra from 250 to 482 nm at 2 nm intervals, with excitation at every 5 nm from 200 to 315 nm on a Perkin-Elmer LS50 B fluorescence spectrometer.

For both excitation and emission the scan speed was 1500 nm/min. The excitation from 200 to 230 nm and the emission below 250 nm was excluded from the analysis since it is highly influenced by the condition of the xenon lamp as well as by the physical environment and mainly contained missing elements [33]. From previous investigations [33,34], we know that four components are appropriate and that four EEM

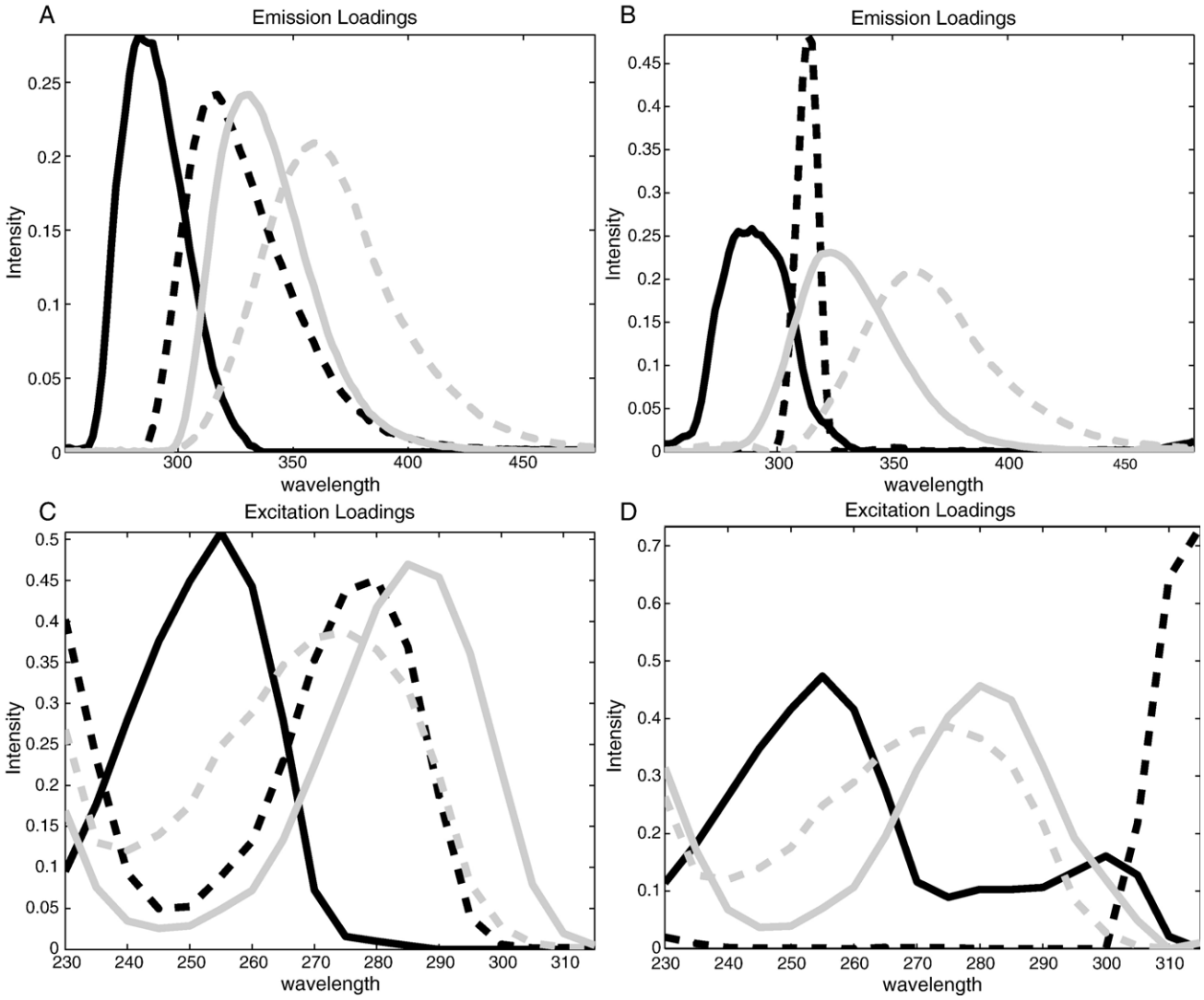

Fig. 4. Left: emission loadings from a four component PARAFAC model, fitted to the Dorrit data set where scatter has been manually removed. Right: emission loadings from a four component PARAFAC model, fitted to the full Dorrit data set.

landscapes can be considered as outliers. Since a classical PARAFAC algorithm is applied on the data after removing the scatter, outlying samples will corrupt the final results. Moreover, the focus in this paper is to put on testing the removal of scatter, considered as being in the form of element-wise outliers, not whole samples. In Ref. [23], an algorithm to deal with outlying observations is proposed, but this method is not able to withstand the effects of scattering. Handling both types of outliers together, is a challenge for future research. For now, these four observations are therefore removed from the data set. The data set then consists of 23 samples, 18 excitation wavelengths and 116 emission wavelengths, and will in the following be referred to as the Dorrit data.

An example of a full excitation–emission landscape obtained from the fluorescence measurements is illustrated in Fig. 3. The Rayleigh scatter can clearly be seen as diagonal ridges. The scatter seems to be well separated from the chemical signal and no Raman scatter is observed. This is the case for all samples in the Dorrit data set. Therefore this well-known data set appears to be perfect for illustrating the properties of the proposed method for automatic scatter removal.

The emission and excitation loadings from a four component PARAFAC model, fitted to the data set where scatter has been manually removed is shown in Fig. 4 (left). This method is based on removing the Rayleigh scatter by inserting a mixture of missing values and zeros. The loadings have a reasonable shape resembling the pure spectra of the four fluorophores. The emission and excitation loadings for the original data set will appear as illustrated in Fig. 4 (right). When comparing the emission loadings from the two models, it is clear that the highest peak in the model fitted to the data with Rayleigh scatter is wrong and caused by the scatter. Also the excitation loadings are not fitted accurately. This clearly indicates that the Rayleigh scatter needs to be removed to obtain a good model.

The identification of the scatter by our automated method performs very well as illustrated in Fig. 5, where the emission profiles of sample 4 for the 18 excitation wavelengths are shown. The elements flagged as outliers by the algorithm are marked with dots on the *X*-axis. The scatter corresponding to 2nd order Rayleigh is clearly identified for the first 3 excitation wavelengths (3 first plots) and from excitation wavelength 259 nm on the regions according to the 1st order Rayleigh scatter are clearly identified. The successful detection of Rayleigh scatter in the remaining samples performs likewise (results not shown).

The three different PARAFAC algorithms; replacing with missing values, interpolation and weighting, were then applied

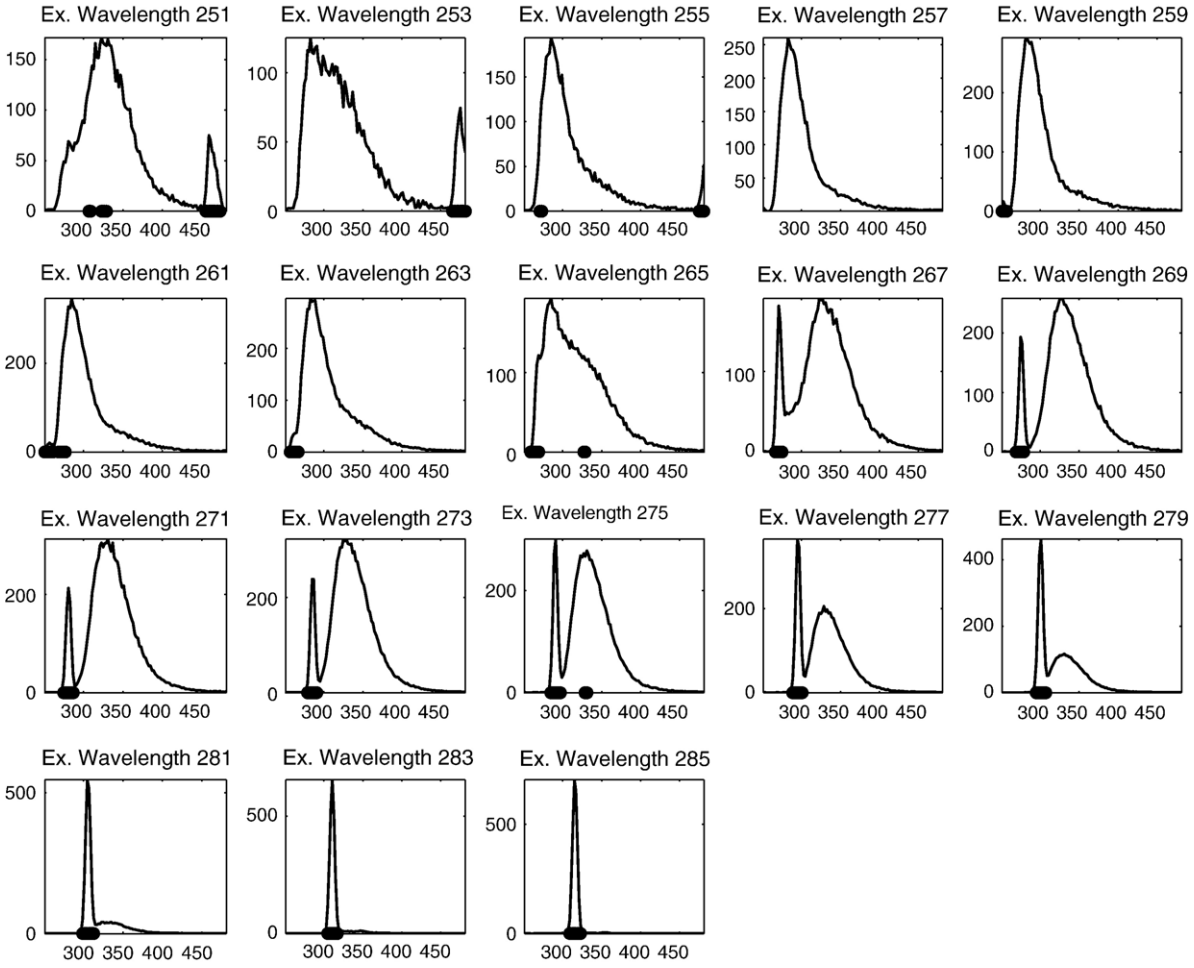

Fig. 5. The emission profiles of the fourth sample of the Dorrit data for the 18 excitation wavelengths. The regions identified as scatter are marked by dots.

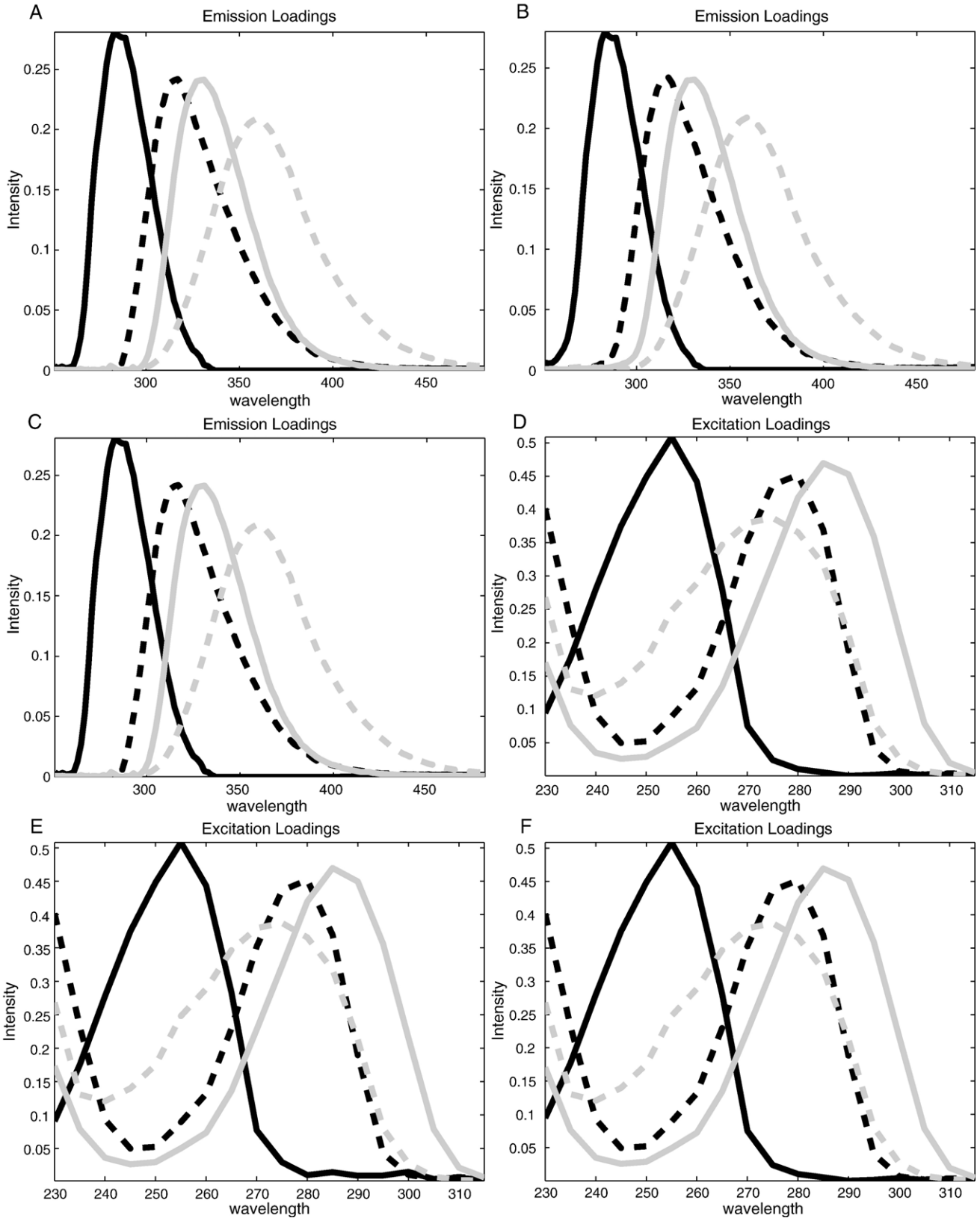

Fig. 6. Four component PARAFAC models ((above) missing, (middle) interpolation, and (below) weighted) fitted to the Dorrit data where the scatter has been detected by the automated method. The left column correspond to emission mode loading and the right column to excitation mode loading.

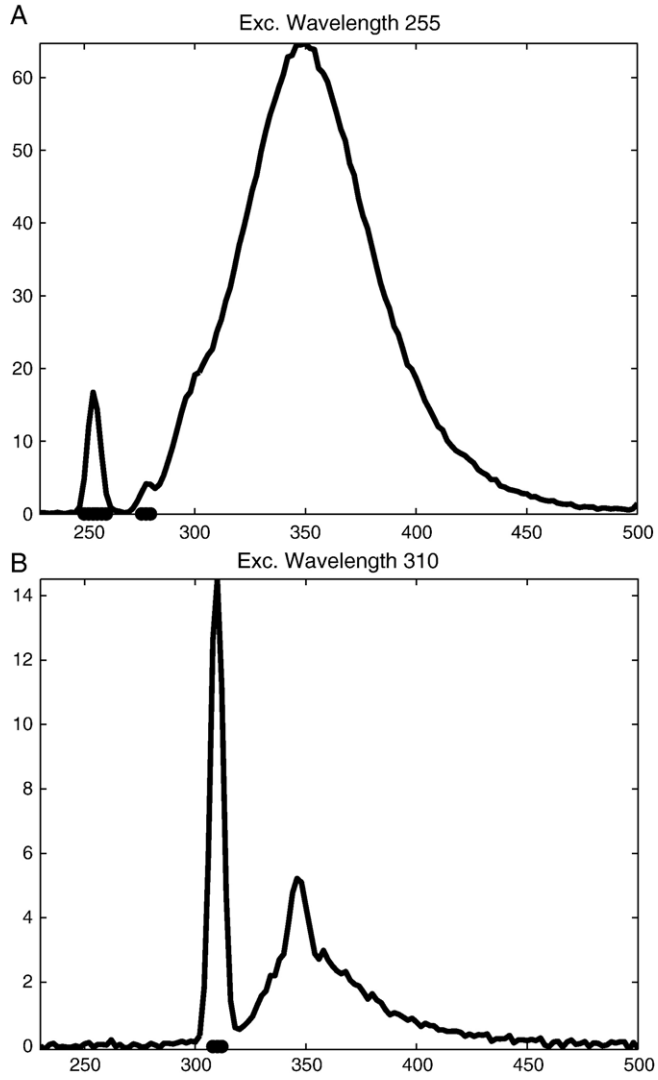

Fig. 7. The emission profiles of sample no. 20 from the fluorescence data at excitation wavelength 255 nm where the scatter region is correctly identified (left) and the emission profile at excitation wavelength 310 nm where not the whole scatter region is identified (right).

to the data set in combination with the information about outlying elements.

The emission and excitation loadings obtained with the three different PARAFAC algorithms are shown in Fig. 6. The three tested algorithms have in common that both emission and excitation loadings are almost identical with the pure spectra of the four fluorophores. This clearly indicates that the automated method for identifying scatter has worked perfectly in marking both 1st and 2nd order Rayleigh scatter, which results in fairly good PARAFAC models. No obvious differences are observed between the three tested PARAFAC algorithms.

3.2. Fluorescence data

This data set, called Fluorescence data, consists of 35 samples of a larger data set consisting of 405 samples, built by experimental design [35,36]. The Fluorescence data is made from five known analytes; catechol, hydroquinone, indole, tryptophane and tyrosine, with two to five analytes present in each sample varying in concentration. These data were chosen on the basis of closeness to the 1st order Rayleigh scatter line and their overlap in both emission and excitation spectra. The prepared samples were measured on a Varian Eclipse Fluorescence Spectrometer with slit widths 5 nm (for both emission and excitation), emission wavelengths 230–500 nm (recorded every 2 nm) and excitation wavelengths 230–320 nm (recorded every 5 nm), and a scan rate at 1920 nm/min. The sample was left in the instrument and was scanned five consecutive times but only the first measurement is included in this analysis. From the experimental set-up it is known that Catechol contains some impurities, and thus gives rise to an extra component in the data sets. This means that the PARAFAC model should be fitted with 6 components. Furthermore, overlapping emission and excitation profiles [36] might make the modelling part hard and not as simple as for the Dorrit data in the previous section.

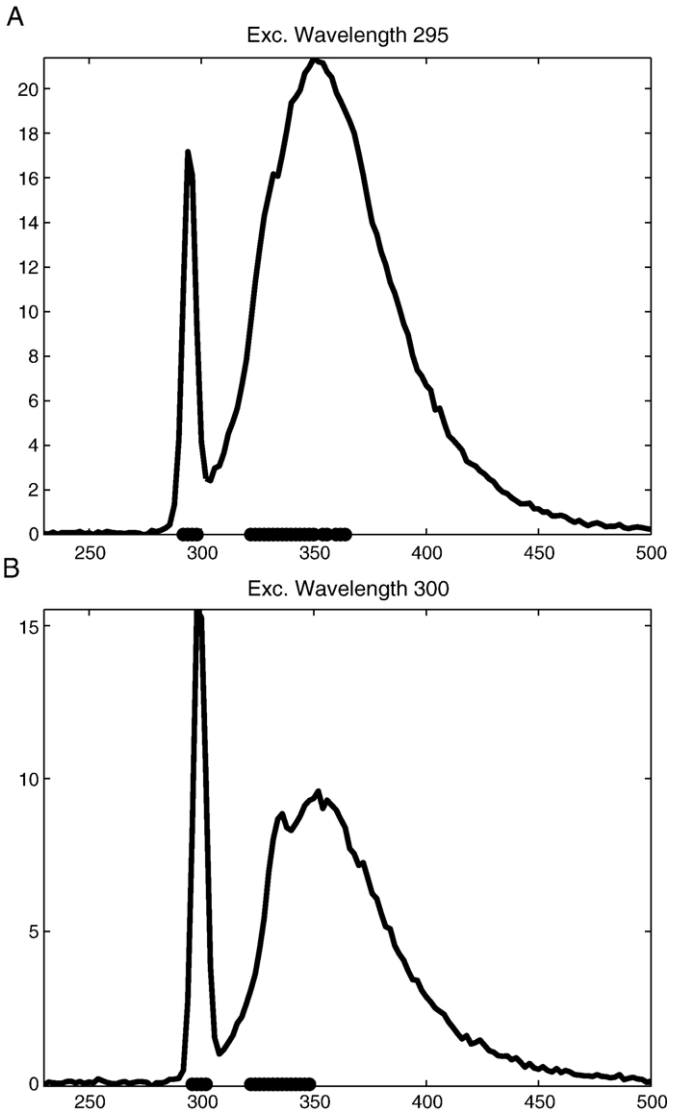

Fig. 8. The emission profiles of sample 20 from the fluorescence data excitation wavelength 295 nm (left) and 300 nm (right), respectively, showing wrongly identification of the chemical signal as outlying.

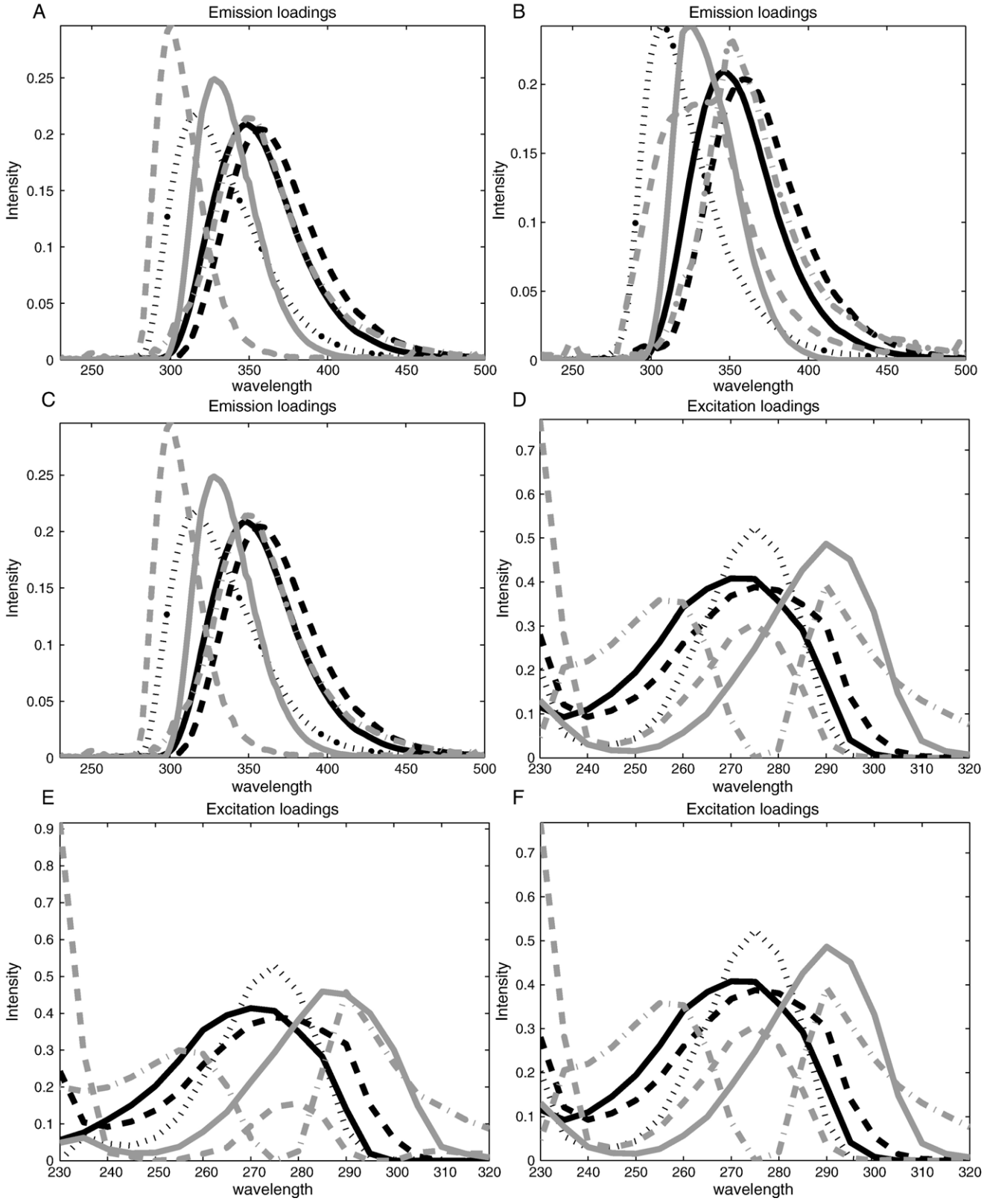

Fig. 9. The emission (left) and excitation (right) loadings for the fluorescence data from 6 components PARAFAC model for the missing algorithm (above), the interpolation algorithm (middle), and the weighted algorithm (below).

A standard, containing only the solvent, exists for this data set. By subtracting this from all other samples the Raman scatter line and possibly the Rayleigh scatter line can be removed or at least reduced [12]. This is not done here since the purpose of this study is to test the possibilities of removing all kinds of scatter by the automated scatter removing method proposed within.

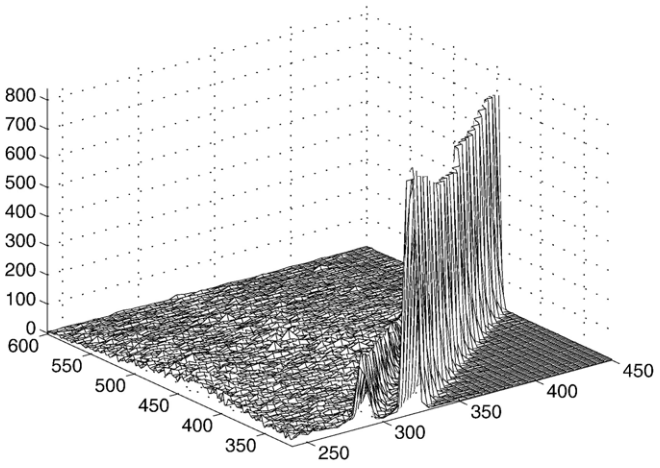

Fig. 10. The 10th sample of the North Sea data with very severe Raman and first order Rayleigh scattering. The highest peak corresponds to the Rayleigh scattering, the smallest to the Raman scattering.

The automated scatter identification method with 6 components was applied on the data set. The scatter region is almost always indicated by the method (Fig. 7, left) but sometimes not the whole scatter region is left out (Fig. 7, right). This means that a small part of the scatter is still left in the data set, and consequently included in further computations. This failure of not finding the edges of the scatter region is due to the maximum operator to obtain one weight w_{ijk}. For the data sliced along the *C*-mode, ROBPCA flags the whole scatter peak as being outlying together with a large part of the signal. But for the sliced data in the *B*-mode, the scattering appears rather small compared to the signal. The scatter is still found as an outlier by ROBPCA, but the edges are not deviating any more and thus not identified. Taking the maximum over the weights $w_{ijk,B}$ and $w_{ijk,C}$ finally results in a weight $w_{ijk}=1$ for the edges of the scatter area. The problem will not be solved by taking another operator than the maximum, like e.g. the minimum or smoother weights, as too much of the signal will be omitted then. A better alternative is to enlarge the indicated region with a certain number of zeros. However, we have not done this in this example, because the included scattering is rather small and adding zeros comprises also a greater loss of the signal.

Furthermore, another problem turned up here. A part of the chemical signal is sometimes wrongly identified as scattering as shown in Fig. 8. This is caused by the general property of robust methods that the majority of the data determine the final estimates. In this example, for some wavelengths the investigated data contain more than 50% low-value profiles, i.e. profiles that contain no signal, nor scatter. This means that the scatter and also the signal is seen as highly deviating. But, this is only happening for wavelengths 230 nm, 235 nm, 290 nm, 295 nm and 300 nm. Thus only a small part of the chemical signal is deleted. As such, enough signal is left to estimate the loadings correctly and again no changes to the algorithm are made to circumvent this problem.

The estimated excitation and emission loadings when fitting 6 components PARAFAC models in combination with the information from the automated scatter identification method to the Fluorescence data set are illustrated in Fig. 9.

For the missing and weighted algorithms the estimated excitation and emission profiles are in accordance with profiles of the pure spectra (Fig. 9 (first and last row)). The interpolation algorithm has some problems in both the excitation and emission mode (Fig. 9, second row). The component indicated by the grey dashed–dotted line is due to the impurities in the samples containing catechol. The results obtained with these data show that even small inaccuracies in identifying the scatter regions, will establish a good PARAFAC model at the end, when using missing values or the weighted PARAFAC version. The interpolated PARAFAC on the other hand has some problems which is due to the not completely removed scatter, that still causes a small peak in the interpolated data.

4. Examples

Data created in a laboratory provide an excellent platform for testing newly developed methods, because estimates can be compared to a priori information. But it is also interesting to

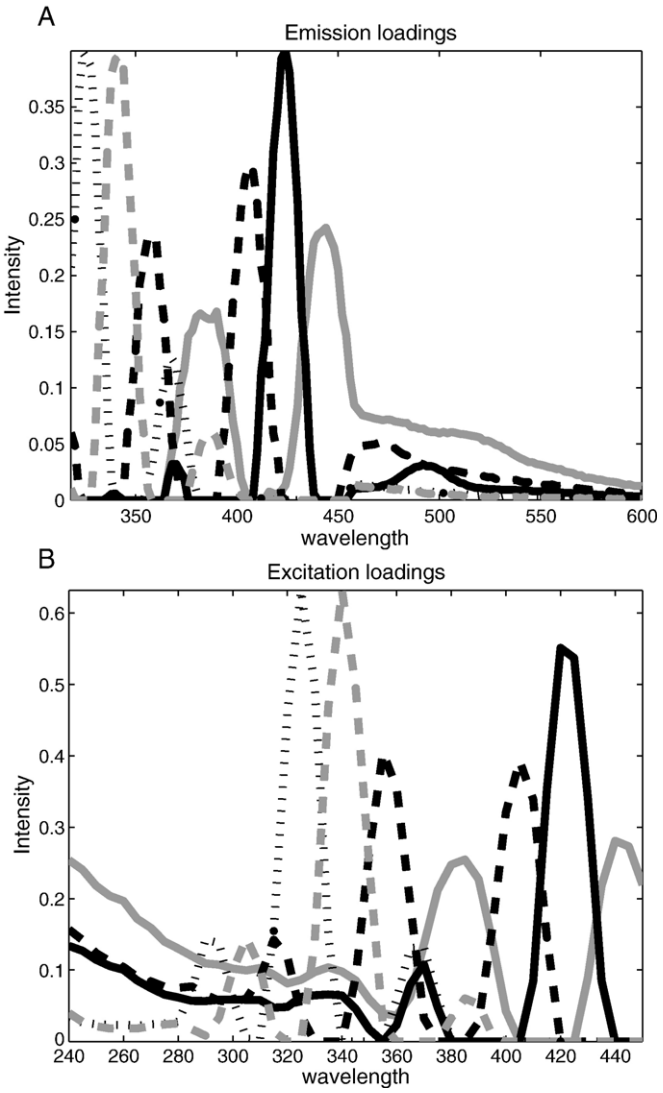

Fig. 11. The emission (left) and excitation (right) loadings of the North Sea data obtained by the classical PARAFAC algorithm.

assess new techniques on environmental data, to find out how well the method can cope with extra difficulties on top of the scattering problem typical for real-life data, such as e.g. noise. Therefore the automated scatter identification algorithm is carried out on the following two examples.

4.1. North Sea data

The 37 samples of the North Sea data, which are kindly provided by Colin A. Stedmon (personal communication), reflect the fluorescence of dissolved organic matter (DOM) of water of the Dogger bank in the North Sea. Measurements were taken with a Varian Eclipse fluorescence spectrophotometer from 2 vertical profiles at 5 m depth intervals. The excitation wavelengths range from 240–450 nm every 5 nm and the emission wavelengths from 240–600 nm each 2 nm. This results in a $(37\times181\times43)$ data cube. No pre-treatment on the data has been carried out, besides a correction for instrument specific effects and a Raman calibration (see [7]). As no blank is subtracted from the samples, severe Raman and Rayleigh scattering are present in all the fluorescence measurements. Moreover, some artifacts could be distinguished in the first 39 emission wavelengths. As we focus on removing scattering effects, we delete these artifacts before analysis. This leads to a $(37\times142\times43)$ data array. It is also known that the signal to noise ratio of the measurements is very low, which means that we have to deal with really noisy data. An EEM landscape is shown in Fig. 10. The scattering is so strong that the relevant signal cannot be distinguished. A classical PARAFAC analysis therefore fails in estimating useful loadings (see Fig. 11).

A split half and residual analysis on the data, obtained after subtracting a blank and after removing manually the scatter, was performed by Colin Stedmon (pers. com.) and indicated that 5 components were suitable for modeling the data. No outlying samples were present in the data.

In Fig. 12, the emission profiles of sample 9 for the first 20 excitation wavelengths are shown. In the first 9 plots, there is no scattering present and only small parts of the signal are marked as being outlying. For the other graphics, the scattering is left out in all cases together with minor parts of the signal. So, the identification of the scattering is performed really well by the automated method.

We use the information about the scatter region and perform the PARAFAC algorithm on the data with missing values instead of outlying elements. We have decreased the stop criterium to 10^{-10} to obtain stable PARAFAC estimates in this very noisy data set. The resulting excitation and emission loadings are depicted in Fig. 13 (first row). The emission loadings seem

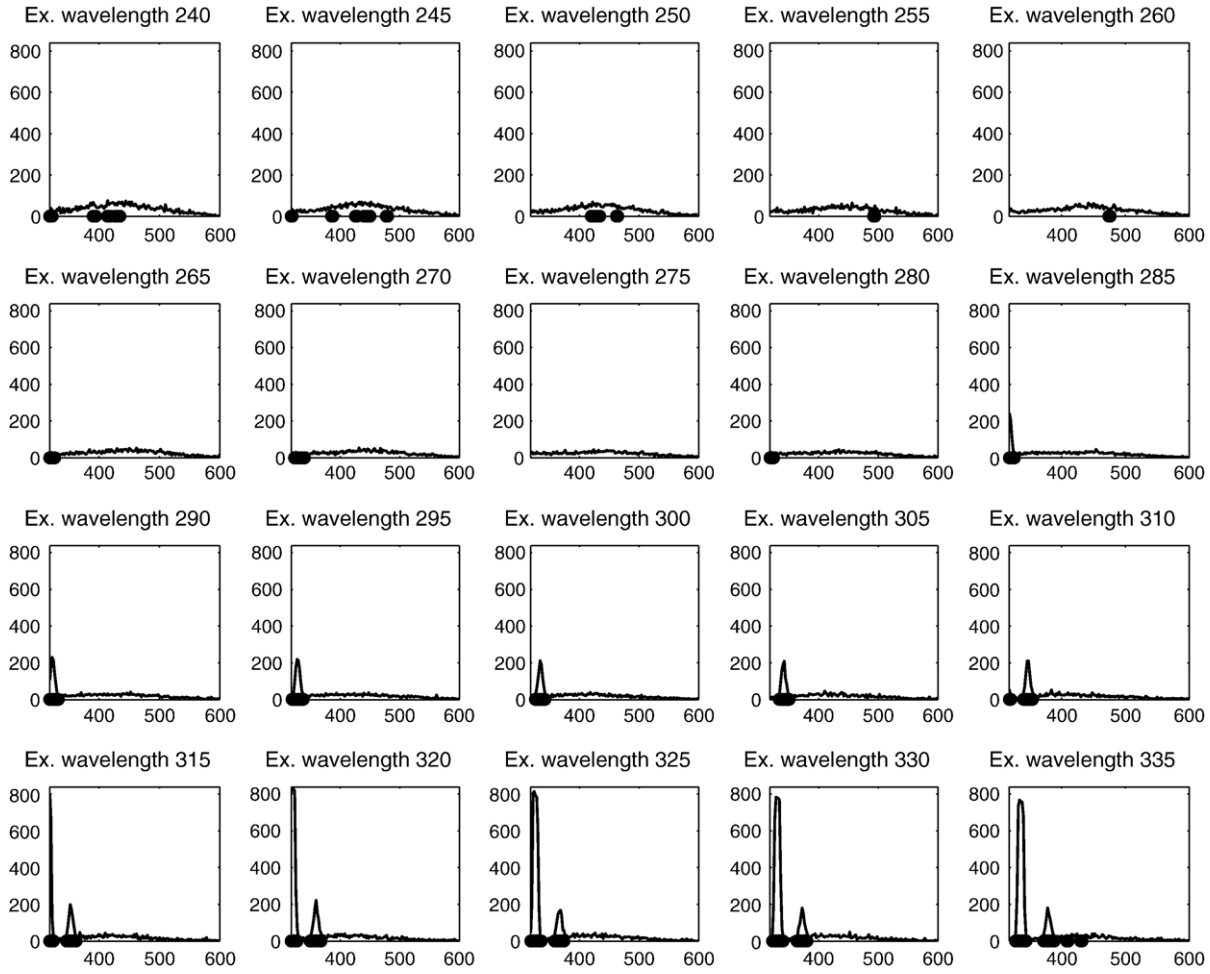

Fig. 12. The emission profile of the ninth observation of the North Sea data for the first 20 excitation wavelengths.

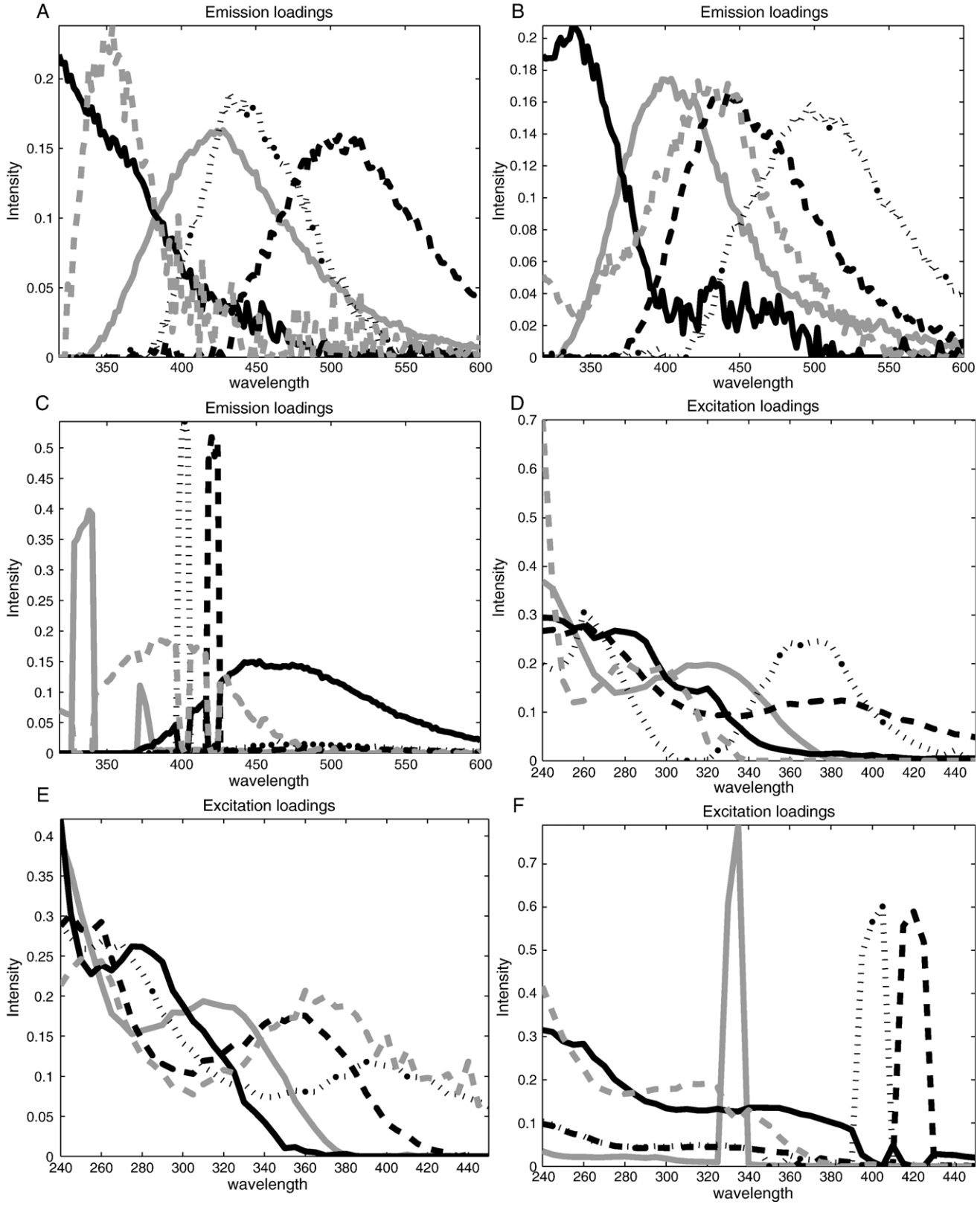

Fig. 13. The emission (left) and excitation (right) loadings for the North Sea data using missing values (above), interpolation (middle) and weighted data (below).

chemically relevant, except perhaps the dashed grey loading, that is quite noisy. This is the effect of a high noise level in the data rather than of the scattering.

Furthermore, we also applied the classical PARAFAC algorithm on the interpolated data with the same constraints as for the missing values. We end up with emission and excitation

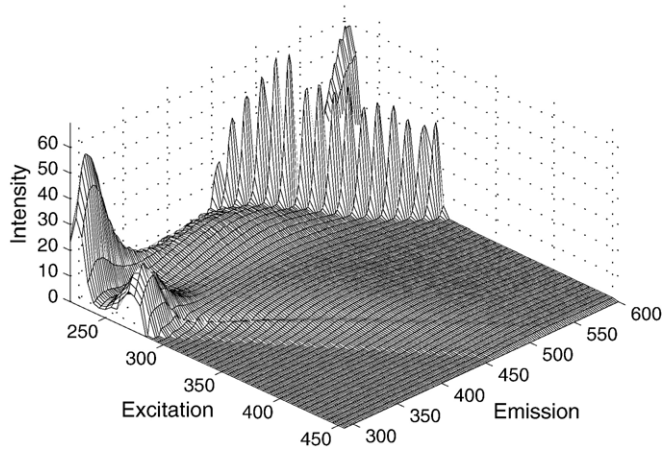

Fig. 14. The third observation of the Kauai data, before the artifacts are removed.

loadings of Fig. 13 (second row). The difference between these loadings and the one estimated by inserting missing values, is in contradiction with the unique property of PARAFAC, which states that a unique solution is provided under mild conditions [37]. An explanation for this discrepancy can be given by the different approximations of the final data on which the PARAFAC model is applied. While missing values discard certain issues present in the data, the interpolation technique fits an approximate value. This leads to quite different data elements at certain areas in the final data set and hence at different estimated loadings. On the other hand, the obtained loadings with the interpolation technique are comparable to the one depicted in Ref. [19], which is not surprising as in Ref. [19] a similar interpolation technique is used. Because this is real-life data, it is not known a priori which should be the correct loadings. So, both solutions are possible, and we cannot favor one above the other.

Finally, the weighted PARAFAC is carried out and the resulting loadings can be found in Fig. 13 (last row). It is obvious that this procedure has broken down because of the scattering. The three loadings with a narrow, but high peak, are fitting the scattering instead of chemically relevant information. The reason why this method failed is a combination of a non-robust initialization of the loadings and too heavy scattering, such that the starting point of the iterative loops in the alternating least squares PARAFAC algorithm, is taken too far from a possible solution.

4.2. Kauai data

The Kauai data, which was kindly provided by the Smithsonian Environmental Research Center in Maryland, USA, consists of 130 seawater EEMs. Of these, 53 were obtained from the ballast tanks of the container ship MV Kauai and 77 were obtained from the Pacific Ocean during the vessel's cruise between Oakland (CA)–Honolulu (HI) and Seattle (WA) in June 2003 [38]. CDOM analysis by excitation–emission matrix spectroscopy (excitation, 240–455 nm in 5-nm intervals; emission, 290–600 nm in 2-nm intervals; 5-nm bandwidths on excitation and emission modes) was performed using a spectrofluorometer at the University of Maine, USA, using a SPEX FluoroMax-2.

Each EEM from the Kauai Dataset consisted of 156 emission wavelengths and 48 excitation wavelengths. As for the Dorrit dataset, the first 4 excitation wavelengths were deleted prior to modeling. In addition, data from the last 26 emission wavelengths (Em>518 nm) were ignored since these were dominated by a spurious signal propagated from intense protein-like fluorescence near λex/λem=270/300 nm. This resulted in a signal at identical excitation but twice the emission wavelength of the actual fluorescence (λex/λem=275/600 nm; see Fig. 14). Thus modeling was performed upon a (130×130×44) element data array.

First order Rayleigh scatter and first and second order Raman scattering are clearly evident in the data (Fig. 14). Fig. 15 shows the emission profile of the third observation for the 275 nm and 290 nm excitation wavelength. For the 275 nm excitation wavelength, the right peak is due to the Rayleigh scattering,

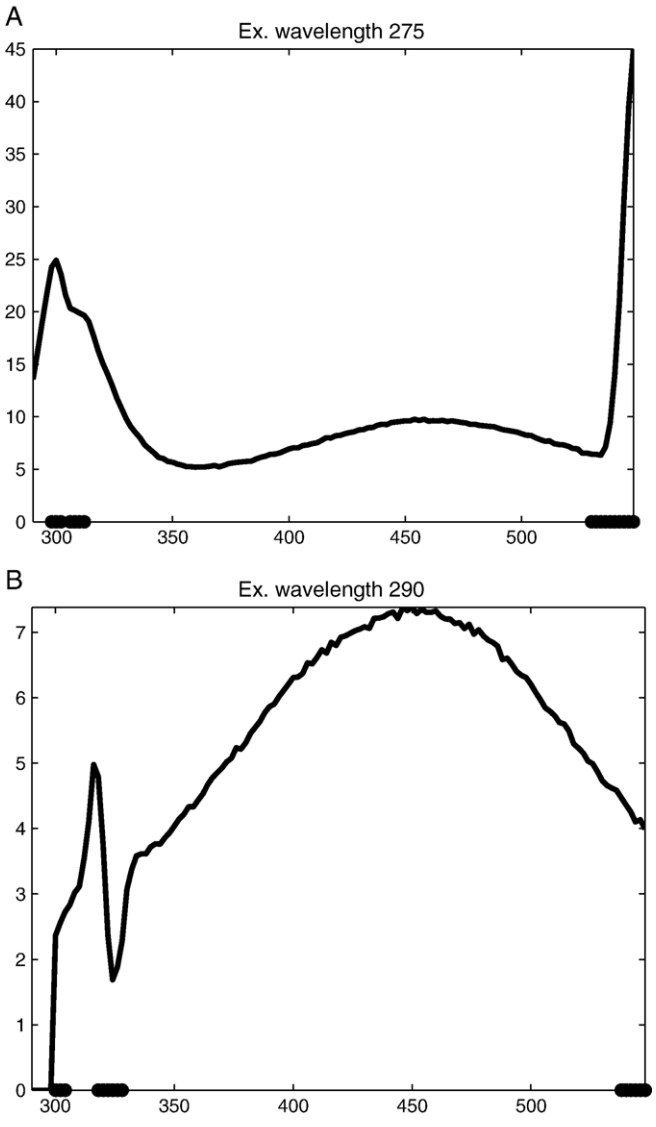

Fig. 15. The emission profile of the third observation of the Kauai data for excitation wavelength 275 nm (left) and 290 nm (right).

whereas the peak on the left consists partially of the signal and partially of the Raman scattering. The same occurs for the 290 nm excitation wavelength and most of the other wavelengths. Consequently, it is difficult to determine the extent of the scatter region from visual inspection of the data.

From a split half analysis on the data after removing scatter with the interpolation technique of Ref. [18], 6 components should be used in the PARAFAC model. Moreover, no outliers could be distinguished in the data.

Preliminary identification of scatter regions before further analysis of the data, was therefore conducted with 6 components. In Fig. 16 the results of the automated scatter identification algorithm can be seen for the first 12 emission profiles of observation 8. The remaining emission profiles of observation 8 gave similar results and are therefore not included.

We see from all these figures that the Rayleigh scattering is captured in the set of the outlying elements. The Raman scattering is indicated for each excitation wavelength, although it is not completely removed, which is caused by the presence of more than 30% zero elements in the Kauai data. The input data sets for the ROBPCA procedure in the automated algorithm therefore contain sometimes more than 50% zero rows, which makes it impossible to identify the scatter accurately. This problem is inherent for the Kauai data and cannot be solved by changing the proposed automated technique. Since, at least all the scatter area have partially been detected, we have decided to enlarge the marked outlier regions with two elements at both sides, to be sure that the scattering will not corrupt the final PARAFAC estimates. Small parts of the signal will consequently also be removed. We have depicted results for observation 8 in Fig. 16, as it was one of the samples for which the remaining scatter area was the largest. For most of the other samples, this effect is more reduced, where even for some of them, the scattering was almost totally eliminated before the enlargement of the outlier area.

Finally, the PARAFAC model with non-negativity constraints is built for the data with missing, interpolated and downweighted values. The emission (left) and excitation (right) loadings are placed in Fig. 17. It seems that imputing missing values where outliers are marked, did not perform well, because of the strange emission loading (the full grey one) and the excitation loading with two sharp peaks (the full grey one). The profile of the excitation loadings is due to the missing area going straight through the first large signal peak, which causes a split of 1 signal peak in 2 sharp peaks. For the emission loadings, the area for large emission wavelength containing only missing values or almost zero values, is to blame. Whatever is estimated as loading values instead of the missing values, it will always have no effect in the final model, as it is multiplied by very low values. This results in the narrow, but high peak at the end of the concerned emission loading. Although the model estimates are thus not correct, it is not caused by the scattering. The missing algorithm fails because of omitting crucial data parts.

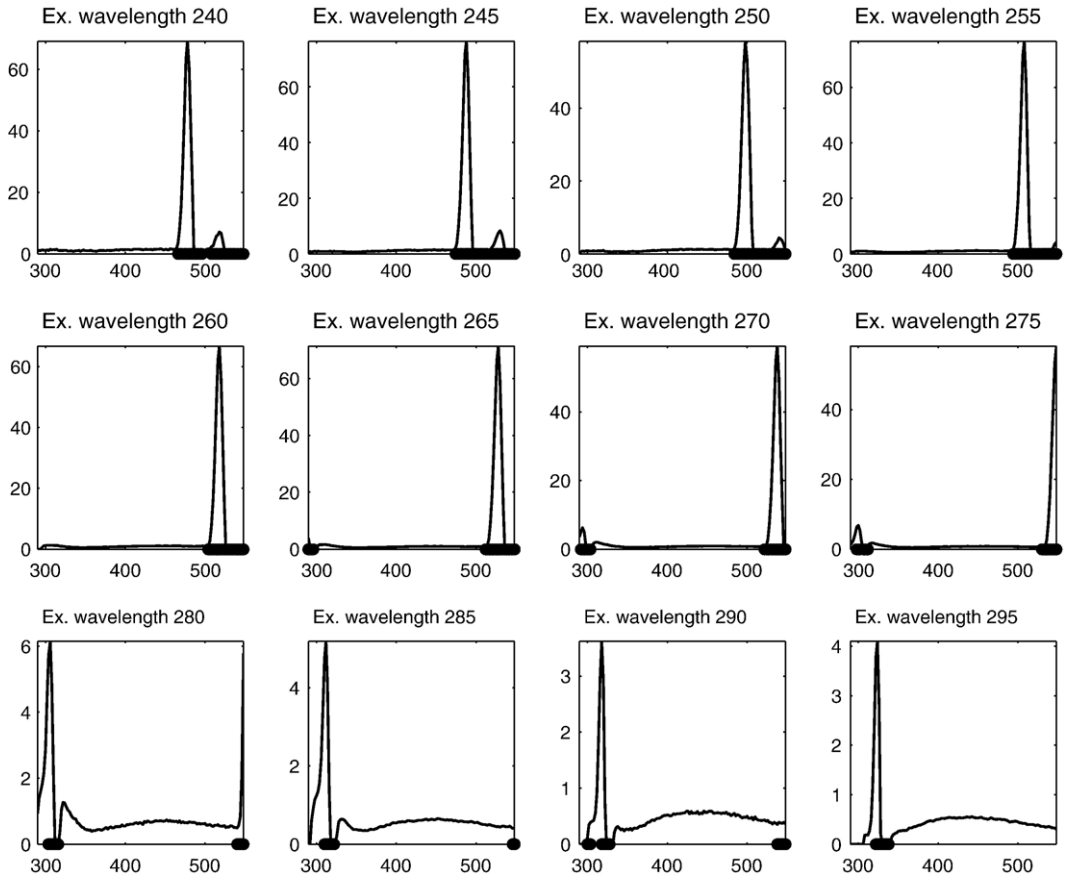

Fig. 16. The emission profiles of the observation 8 of the Kauai data for the first 12 excitation wavelengths.

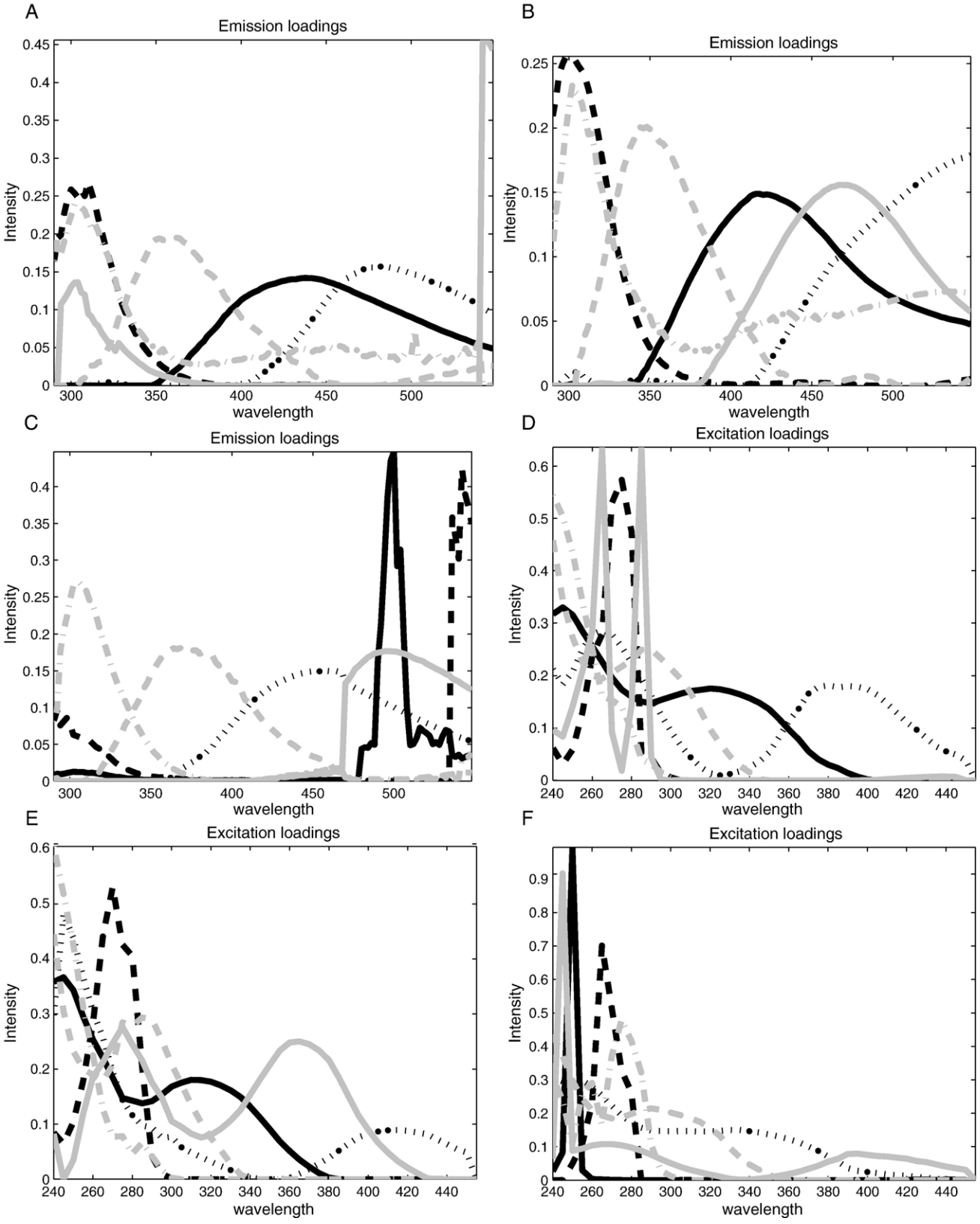

Fig. 17. The emission (left) and excitation (right) loadings of the Kauai data for a PARAFAC model with 6 components using missing values (above), interpolation (middle) and weighted data (below).

For the interpolated Kauai data, the estimated loadings for a 6 component PARAFAC model are depicted in Fig. 17 (second row). Here, the scattering is out of the model. However, the estimated loadings are not exactly the same as the one obtained in Ref. [38]. A reason can be found in a different approximation of the data, due to other interpolation techniques applied on

other sets of data elements, as the scatter areas are marked using two different approaches.

The results of the weighted PARAFAC are shown in Fig. 17 (last row). The loadings are not good, they try to fit the scattering and not only the chemical information. The same reasons as for the North Sea data cause this break down of the model and again confirm that the weighted PARAFAC model is not optimal to use for highly scattered data.

5. Discussion and conclusions

Despite different existing methods for excluding scattering effects when modeling fluorescence data by PARAFAC, no method for disregarding scattering automatically can be found in the literature. In this article we have established an automated scatter identification method which is based on ROBPCA. The method does not demand any visual inspection of the data. The evaluation of the proposed method clearly shows that the method always succeeds in finding the scatter regions both concerning Rayleigh (1st and 2nd order) and Raman scatter, without marking too much of the signal, due to chemicals under investigation, as outlying. However, smaller parts of the scattering are sometimes hard to detect depending on the data complexity e.g. noise and overlap between scatter and chemical signal. This means that scatter might be included to a minor extent in the PARAFAC modeling step, but also smaller part of the chemical signal might be flagged as outlying and thereby excluded from the analysis.

Nevertheless, this seems not an invincible problem for estimating the final PARAFAC estimates. The three tested PARAFAC methods after removal of the scattering work for the cases they can handle. This means that for the data with the missing values a fitting problem is only encountered when the signal and the scatter coincide too much, such that essential information vanishes. Secondly, classical PARAFAC applied on interpolated data also performs well, but it is the most subject to the parts of the scattering that are not flagged as outlying. Finally, down-weighting the outlying elements is also a good option, provided that the scattering is in the region of the signal. For too severe scatter, this technique is not useful and actually is the least robust of the three investigated procedures.

We only have considered data sets where the number of components has been known before analysis throughout this paper, because the identification of scatter was the major concern. However, when the optimal value for F is not known, which is mostly the case for real-life data, the following approach could be followed to determine the scatter and F. Start with an initial guess for the number of components. In the next step, identify and remove the scatter in the data. Then, use existing techniques (like e.g. a split half analysis) to define F on the data without the scatter. Finally, identify again for the known value of F the scatter automatically. This approach is justified, because the scatter is not highly dependent on F, as we have already discussed earlier in Section 2.

Remark that the proposed method cannot cope with outlying samples yet. A fully robust procedure handling both sample outlier identification and scatter identification is under development.

References

[1] J. Carroll, J. Chang, Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart–Young decomposition, Psychometrika 35 (1970) 283–319.
[2] R. Harshman, Foundations on the PARAFAC procedure: model and conditions for an explanatory multimode factor analysis, UCLA Working Papers in Phonetics 16 (1970) 1–84.
[3] R. Ross, C. Lee, C. Davis, B. Ezzeddine, E. Fayyad, S. Leurgans, Resolution of fluorescence spectra of plant-pigment complexes using trilinear models, Biochimica et Biophysica Acta 1056 (1991) 317–320.
[4] R. Bro, Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis, Chemometrics and Intelligent Laboratory Systems 46 (1999) 133–147.
[5] R. Jiji, G. Andersson, K. Booksh, Application of PARAFAC for calibration with excitation–emission matrix fluorescence spectra of three classes of environmental pollutants, Journal of Chemometrics 14 (2000) 171–185.
[6] D. Pedersen, L. Munck, S. Engelsen, Screening for dioxin in fish oil by PARAFAC and N-PLSR analysis of fluorescence landscapes, Journal of Chemometrics 16 (2002) 451–460.
[7] C. Stedmon, S. Markager, R. Bro, Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy, Marine Chemistry 82 (2003) 239–254.
[8] J. Christensen, E. Miquel Becker, C. Frederiksen, Fluorescence spectroscopy and PARAFAC in the analysis of yoghurt, Chemometrics and Intelligent Laboratory Systems 75 (2005) 201–205.
[9] R. Bro, PARAFAC, Tutorial and applications, Chemometrics and Intelligent Laboratory Systems 38 (1997) 149–171.
[10] C. Andersen, R. Bro, Practical aspects of PARAFAC modelling of fluorescence excitation–emission data, Journal of Chemometrics 17 (2003) 200–215.
[11] L. Thygesen, A. Rinnan, S. Barsberg, J. Møller, Stabilizing the PARAFAC decomposition of fluorescence spectra by insertion of zeros outside the data area, Chemometrics and Intelligent Laboratory Systems 71 (2004) 97–106.
[12] A. Rinnan, C. Andersen, Handling of first-order Rayleigh scatter in PARAFAC modelling of fluorescence excitation–emission data, Chemometrics and Intelligent Laboratory Systems 76 (2005) 91–99.
[13] P. Wentzell, S. Nair, R. Guy, Three-way analysis of fluorescence spectra of polycyclic aromatic hydrocarbons with quenching by nitromethane, Analytical Chemistry 73 (2001) 1408–1415.
[14] D.M. McKnight, E. Boyer, P. Westerhoff, P. Doran, T. Kulbe, D.T. Andersen, Spectrofluorometric characterisation of dissolved organic matter for indication of precursor organic material and aromaticity, Limnology and Oceanography 46 (2001) 38–48.
[15] R. Bro, N. Sidiropoulos, A. Smilde, Maximum likelihood fitting using ordinary least squares algorithms, Journal of Chemometrics 16 (2002) 387–400.
[16] R. JiJi, K. Booksh, Mitigation of Rayleigh and Raman spectral interferences in multi-way calibration of excitation–emission matrix fluorescence data, Analytical Chemistry 72 (2000) 718–725.
[17] G. Andersson, B. Dable, K. Booksh, Weighted parallel factor analysis for calibration of HPLC–UV/Vis spectrometers in the presence of Beer's law deviations, Chemometrics and Intelligent Laboratory Systems 49 (1999) 195–213.
[18] R. Zepp, W. Sheldon, M. Moran, Dissolved organic fluorophores in southeastern US coastal waters: correction method for eliminating Raleigh and Raman scattering peaks in excitation–emission matrices, Marine Chemistry 89 (2004) 15–36.
[19] M. Bahram, R. Bro, C. Stedmon, A. Afkhami, Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation (submitted for publication).
[20] M. Hubert, P. Rousseeuw, S. Verboven, A fast robust method for principal components with applications to chemometrics, Chemometrics and Intelligent Laboratory Systems 60 (2002) 101–111.
[21] M. Hubert, P. Rousseeuw, K. Vanden Branden, ROBPCA: a new approach to robust principal components analysis, Technometrics 47 (2005) 64–79.

[22] S. Vorobyov, Y. Rong, N. Sidiropoulos, A. Gershman, Robust iterative fitting of multilinear models, IEEE Transactions on Signal Processing 53 (2005) 2678–2689.
[23] S. Engelen, M. Hubert, Detecting outlying samples in a PARAFAC model (submitted for publication).
[24] G. Li, Z. Chen, Projection–pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo, Journal of the American Statistical Association 80 (1985) 759–766.
[25] N. Locantore, J. Marron, D. Simpson, N. Tripoli, J. Zhang, K. Cohen, Robust principal component analysis for functional data, Test 8 (1999) 1–73.
[26] S. Verboven, M. Hubert, LIBRA: a Matlab library for robust analysis, Chemometrics and Intelligent Laboratory Systems 75 (2005) 127–136.
[27] B. Wise, N. Gallagher, R. Bro, J. Shaver, W. Windig, R. Koch, PLS Toolbox 3.5 for use with MATLAB, software, Eigenvector Research, Inc., August 2004 (2004). URL http://software.eigenvector.com/.
[28] P. Rousseeuw, Least median of squares regression, Journal of the American Statistical Association 79 (1984) 871–880.
[29] G. Box, Some theorems on quadratic forms applied in the study of analysis of variance problems: effect of inequality of variance in one-way classification, The Annals of Mathematical Statistics 25 (1954) 33–51.
[30] P. Nomikos, J. MacGregor, Multivariate SPC charts for monitoring batch processers, Technometrics 37 (1995) 41–59.
[31] I. Jolliffe, Principal Component Analysis, Springer, New York, 1986.
[32] M. Hubert, S. Engelen, Fast cross-validation for high-breakdown resampling algorithms for PCA (submitted for publication).
[33] D. Baunsgaard, Factors affecting 3-way modelling (PARAFAC) of fluorescence landscapes, Ph.D. thesis, Royal Veterinary and Agricultural University, Department of Dairy and Food technology, Frederiksberg, Denmark (1999).
[34] J. Riu, R. Bro, Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models, Chemometrics and Intelligent Laboratory Systems 65 (2003) 35–49.
[35] A. Rinnan, Application of PARAFAC on spectral data, Ph.D. thesis, Royal Veterinary and Agricultural University, Denmark (2004).
[36] R. Bro, A. Rinnan, N. Faber, Standard error of prediction for multilinear PLS2. Practical implementation in fluorescence spectroscopy, Chemometrics and Intelligent Laboratory Systems 75 (2005) 69–76.
[37] A. Smilde, R. Bro, P. Geladi, Multi-way Analysis with Applications in the Chemical Sciences, Wiley, England, 2004.
[38] K. Murphy, G.M. Ruiz, W.T.M. Dunsmuir, T.D. Waite, Optimized parameters for rapid fluorescence-based verification of ballast water exchange by ships., Environmental Science and Technology 40, (in press).