The importance of data quality and traceability in data mining. Applications of robust methods for multivariate data analysis. A case-study conducting the herring industry

Frosch, Stina

Publication date: 2006

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

Link back to DTU Orbit

Citation (APA): Frosch, S. (2006). The importance of data quality and traceability in data mining. Applications of robust methods for multivariate data analysis. A case-study conducting the herring industry. Kgs. Lyngby: Danish Institute for Fisheries Research, Department of Seafood Research & The Technical University of Denmark.
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 $1^{st}$ and $2^{nd}$ 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.
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.


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.


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.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALS</td>
<td>Alternating Least Squares</td>
</tr>
<tr>
<td>CSW</td>
<td>Chilled Sea Water</td>
</tr>
<tr>
<td>ICES</td>
<td>International Council for the Exploration of the Sea</td>
</tr>
<tr>
<td>LAE</td>
<td>Least Absolute Error</td>
</tr>
<tr>
<td>LS</td>
<td>Least Squares</td>
</tr>
<tr>
<td>LTS</td>
<td>Least Trimmed Squares</td>
</tr>
<tr>
<td>MCD</td>
<td>Minimum Covariance Determinant</td>
</tr>
<tr>
<td>NIPALS</td>
<td>Nonlinear Iterative Partial Least Squares</td>
</tr>
<tr>
<td>NIR</td>
<td>Near Infrared Reflectance</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear Magnetic Resonance</td>
</tr>
<tr>
<td>PARAFAC</td>
<td>Parallel Factor Analysis</td>
</tr>
<tr>
<td>PC</td>
<td>Principal Component</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PCR</td>
<td>Principal Component Regression</td>
</tr>
<tr>
<td>PLSR</td>
<td>Partial Least Squares Regression</td>
</tr>
<tr>
<td>RMSEP</td>
<td>Root Mean Square Error of Prediction</td>
</tr>
<tr>
<td>RSW</td>
<td>Refrigerated (chilled) Sea Water</td>
</tr>
<tr>
<td>SVD</td>
<td>Singular Value Decomposition</td>
</tr>
<tr>
<td>TRU</td>
<td>Traceable Resource Unit</td>
</tr>
</tbody>
</table>
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., von Frese, J. and Bro, R.

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

2Spectroscopy 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


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

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

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

2Danish 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
  1.1 Background .................................................................................................................. 1
  1.2 Objectives ................................................................................................................... 3
  1.3 Approach .................................................................................................................... 4

2.0 Multivariate data analysis methods .............................................................................. 5
  2.1 Multivariate data analysis .......................................................................................... 5
  2.2 Principal component analysis .................................................................................... 6
    2.2.1 PARAFAC ............................................................................................................. 8
  2.3 Multivariate regression methods ............................................................................... 9
    2.3.1 Principal component regression .......................................................................... 10
    2.3.2 Partial least squares regression ............................................................................ 11

3.0 Data from a herring industry ...................................................................................... 13
  3.1 The production line for marinated herring products ................................................... 14
  3.2 Traceability ............................................................................................................... 20
  3.3 Data presentation ...................................................................................................... 21
  3.4 Data analysis ............................................................................................................. 25
    3.4.1 Overview of trips included in the data analysis ................................................... 25
    3.4.2 Data obtained in connection with the trip ............................................................ 27
    3.4.3 Quality evaluation of the raw material ................................................................ 29
    3.4.4 Evaluation of the marinated products ................................................................. 36
    3.4.5 Quality determination at topping ......................................................................... 39
  3.5 Multivariate data analysis of data from the herring industry ...................................... 42
  3.6 Additional measurements .......................................................................................... 48
  3.7 Concluding remarks ................................................................................................. 49

4.0 Applications of robust multivariate methods .............................................................. 51
  4.1 Outliers ..................................................................................................................... 52
  4.2 Robust PCA .............................................................................................................. 56
    4.2.1 Application of robust PCA ................................................................................. 57
  4.3 Robust PLSR ............................................................................................................. 63
    4.3.1 Application of robust PLSR ................................................................................. 64
4.4 An approach for and application of robust PARAFAC ..............................................68
  4.4.1 Application of robust PARAFAC .......................................................................69
  4.4.2 Automatic scatter identification ........................................................................75
  4.5 Software....................................................................................................................79
  4.6 Concluding remarks...............................................................................................80

5.0 Conclusion and perspectives ....................................................................................81

References .....................................................................................................................85
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 (Kourtì 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 the 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 \( \mathbf{X} \) is decomposed into the matrix products \( \mathbf{T} \mathbf{P}' \) and the residual matrix \( \mathbf{E} \) (Equation 2.1).
The matrix product, $\mathbf{TP}'$, 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, $\mathbf{E}$, contains the remaining information or noise in $\mathbf{X}$ that was not described by $\mathbf{TP}'$.

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, \( X(I \times J \times K) \), 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

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

\textit{Equation 2.2}

where \( a_{if}, b_{jf} \) and \( c_{kf} \) denote elements of the loading matrices, \( A(I \times F), B(J \times F) \), and \( C(K \times F) \), 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

\[
X = AD_k B + E_k, \ k = 1, \ldots, K
\]

\textit{Equation 2.3}

where, \( D_k \) is a diagonal matrix holding the \( k \)th 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, \( y \):

\[
y = Xb + e
\]

\[\text{Equation 2.4}\]

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,

$$y = Tb + e$$  \hspace{1cm} \text{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 its 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](image)

**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).
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.

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, 178/2002/EC 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.

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

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>
<thead>
<tr>
<th>Year</th>
<th>Number of trips</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>24*</td>
</tr>
<tr>
<td>2000</td>
<td>77</td>
</tr>
<tr>
<td>2001</td>
<td>90</td>
</tr>
<tr>
<td>2002</td>
<td>148</td>
</tr>
<tr>
<td>2003</td>
<td>132*</td>
</tr>
</tbody>
</table>

* 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 bromphenolic 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.

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

*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 °C). In both situations, off-flavours will develop in the herring, if the seawater is not renewed (Smith et al., 1980).
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.
No. of elements

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.

**Industrially calculated fat content, %**

![Graph](image)

**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.

<table>
<thead>
<tr>
<th>Time</th>
<th>Fat content, %</th>
<th>Batch level (Industry)**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single level (Research)*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>May</td>
<td>6.93</td>
<td>1.82</td>
</tr>
<tr>
<td>November</td>
<td>6.49</td>
<td>2.72</td>
</tr>
<tr>
<td>February</td>
<td>4.48</td>
<td>2.04</td>
</tr>
</tbody>
</table>

* 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”, “Homogeneity” and “Specifications kept” are presented in Figure 3.11.

![Figure 3.11](image-url)  
**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.

![Graph showing evaluation results](image)

**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.

![Bar chart](Figure 3.13) The results obtained from the general assessment of the final marinated products.

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.

<table>
<thead>
<tr>
<th>Month</th>
<th>2000</th>
<th>2000</th>
<th>2001</th>
</tr>
</thead>
<tbody>
<tr>
<td>August</td>
<td>11</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>September</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>October</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 \( (x_i) \) are considered as outliers, but it is equally relevant to look for outliers in variables \( (x_j) \) 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 \( n \times p \) data matrix \( X = (x_1, \ldots, x_n)' \), with \( x_i = (x_{i1}, \ldots, x_{ip})' \) the \( i \)th observation, as described in section 2.1. The common estimates for the multivariate location \( \hat{\mu}_0 \) and scatter matrix \( \hat{\Sigma}_0 \) 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 \( (x_i, y_i) \) with \( i = 1, \ldots, n \), it holds that

\[
y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + r_i
\]

\textbf{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 \( \text{LS} \) not being resistant to outliers follows from the properties of the objective function for \( \text{LS} \) procedures. The objective function to be minimized is the sum of the squared residuals:

\[
\text{minimize } \sum_{i=1}^{n} r_i^2
\]

\textbf{Equation 4.2}

in which the residuals \( r_i \) are given by

\[
r_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip}
\]

\textbf{Equation 4.3}
where \( y_i \) \((i = 1, \ldots, n)\) 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 \( LS \) 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 \textit{least median of squares} (Rousseeuw, 1984) and \textit{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 $\varepsilon^*$ (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 $\varepsilon^* = 50 \%$, 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 robustifying
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, $x_i$, 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](image)

**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 °C (△), -30 °C (○), and -80 °C (▲). Three outliers (all -30 °C 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 ($n \ll p$) 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 $C_{xy}$ and the empirical covariance matrix $C_x$ 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 (Serneels 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

$$
RMSEP_k = \sqrt{\frac{1}{n} \sum (\hat{y}_{-i,k} - y_i)}
$$

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

$$
Bias = \frac{\sum (\hat{y} - y) }{n}
$$

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. $k = \text{number of PCs.}$

<table>
<thead>
<tr>
<th>$k$</th>
<th>PLSR</th>
<th>RSIMPLS</th>
<th>PRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = 1$</td>
<td>RMSEP</td>
<td>2.56</td>
<td>2.11</td>
</tr>
<tr>
<td>$r^2$</td>
<td>0.78</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td>Bias</td>
<td>0.00</td>
<td>-0.02</td>
<td>0.35</td>
</tr>
<tr>
<td>$k = 2$</td>
<td>RMSEP</td>
<td>2.28</td>
<td>1.74</td>
</tr>
<tr>
<td>$r^2$</td>
<td>0.83</td>
<td>0.83</td>
<td>0.82</td>
</tr>
<tr>
<td>Bias</td>
<td>0.00</td>
<td>-0.08</td>
<td>0.23</td>
</tr>
<tr>
<td>$k = 3$</td>
<td>RMSEP</td>
<td>2.19</td>
<td>1.65</td>
</tr>
<tr>
<td>$r^2$</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>Bias</td>
<td>0.00</td>
<td>-0.08</td>
<td>0.17</td>
</tr>
<tr>
<td>$k = 4$</td>
<td>RMSEP</td>
<td>2.19</td>
<td>1.54</td>
</tr>
<tr>
<td>$r^2$</td>
<td>0.85</td>
<td>0.86</td>
<td>0.85</td>
</tr>
<tr>
<td>Bias</td>
<td>0.00</td>
<td>-0.09</td>
<td>0.03</td>
</tr>
<tr>
<td>$k = 5$</td>
<td>RMSEP</td>
<td>2.03</td>
<td>1.54</td>
</tr>
<tr>
<td>$r^2$</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>Bias</td>
<td>0.00</td>
<td>-0.02</td>
<td>0.042</td>
</tr>
<tr>
<td>$k = 6$</td>
<td>RMSEP</td>
<td>2.02</td>
<td>1.55</td>
</tr>
<tr>
<td>$r^2$</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>Bias</td>
<td>0.00</td>
<td>-0.03</td>
<td>0.04</td>
</tr>
</tbody>
</table>

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 ($I \times J \times K$) 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 ($I \times JK$). 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.
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](image)

**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.
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.

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 \( \mathbf{X} \) 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.](image)
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 mining 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


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.


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


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.


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


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


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


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


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.


Michaelsen K. Personal communication, 2005.


Nielsen KN, Guldager HS, Sorensen 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.

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


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.


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


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.


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


