Handling Complex Data in Food Analysis

A Chemometric Approach

PhD Dissertation by Vibeke Tølbøl Svensson

FACULTY OF LIFE SCIENCES
UNIVERSITY OF COPENHAGEN

PhD Dissertation

Vibeke Tølbøl Svensson

Handling Complex Data in Food Analysis

A Chemometric Approach

Supervisor:
Professor Rasmus Bro
Faculty of Life Science
University of Copenhagen

October 2008

Cover illustration by Sune Hansen
Title: Handling Complex Data in Food Analysis – A Chemometric Approach

Faculty of Life Sciences, University of Copenhagen,
Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark

ISBN 978-87-7611-278-3
Printed by SL grafik, Frederiksberg C, Denmark

Preface

This dissertation is the final product of my PhD work carried out at The Quality and Technology Group at The Department of Food Science at The University of Copenhagen. The work has been under the supervision of Professor Rasmus Bro and has been partly funded by FØTEK 3 (93S.2444-Å01-00100). The project has involved the industrial partners Arla Foods amba and Lykkeberg A/S. In the context of this dissertation work, I would like to show my gratitude to the following people:

Professor Age Smilde is thanked for letting me visit his group for a two month research stay at the University of Amsterdam. Dr. Sarah Rutan is greatly appreciated for letting me join her research group at The Virginia Commonwealth University during my stay in Virginia, USA. Arla Foods amba, Lykkeberg A/S, and co-authors are thanked for fruitful collaboration.

I am grateful to my supervisor Rasmus Bro for giving me the opportunity to explore the world of food and chemometrics and for the inspiration and discussions during this project. To Rasmus, I especially want to express my gratitude for the support I got when living in the US and for being patient with me. I also want to give special thanks to Frans van den Berg for his support and patience. Lisbeth T. Hansen is appreciated for her work in the laboratory and for always being a great help when needed. Flemming Hofmann Larsen is thanked for helping me in the final phase of this dissertation work and for being an incredible patient officemate.

Sarah Graham Porter is appreciated for making my stay at VCU more joyful. I am grateful to Peter I. Hansen for stepping up, at a time when I needed it the most. I would also like to thank my coworkers (former and present) at the Quality and Technology Group for making this a memorable experience. Especially I want to thank Åsmund, Jakob, Frans, Giorgio, Bonnie, Charlotte, and Elisabeth for being good friends and for making everyday life, conferences and travel – just a little bit more fun!

Finally I would like to thank my family and friends for their support. Casper, Christopher, and Victoria, to you I am deeply grateful for the experience of knowing what life is all about, and to you Carsten: Tak!

Frederiksberg, October, 2008.

Vibeke Tølbøl Svensson

Summary

The advancement within the field of analytical technology results in an increasing amount of data with multivariate dimensionality and complexity. Often multivariate data problems are approached by standard two-way methods such as PCA, PCR and PLS regression, but in certain situations specially designed chemometric methods are required, which are able to handle the data in their full dimensionality, e.g., multi-way and multi-block methods.

This research has been performed in order to provide a better understanding of the principles and the contribution of the more advanced chemometric technologies from the PAT "*toolbox*". The area of application has been food, and food products selected from the categories of vegetables, fish, and dairy products. The chemometric applications are focused on using the multi-way method, PARAFAC and two multi-block methods, i.e., multiblock PLS and LSparPLSc. The main analyses for chemical assessment have been spectroscopy, including NIR, NMR and fluorescence spectroscopy, but other types of analyses such as various chromatographic methods, physical measurements and sensory evaluation are also included. The work has been carried out in an explorative manner where the selection of the chemometric approach is based on preserving the natural structure of the data problem. Thus, a single data matrix has been analyzed using two-way PCA or PLS regression, several data matrices have been approached as individual block contribution in multi-block analysis, and finally data represented as a three-way array has been treated as a multi-dimensional problem, and PARAFAC has been applied.

The work emphasized how chemometrics in general is applicable in food science and that traditional two-way chemometrics is suited for handling straightforward two-way data. In PAPER IV a method to determine the protein content in brine was developed. The novelty of this work is that the measurements are performed on the brine from barrel salted herring instead of the herring itself. The approach shows the potential of using NIR spectroscopy and traditional two-way chemometrics for monitoring ripening characteristics in brine from barrel salted herring. A section is dedicated to the pre-processing of NIR spectra, and it illustrates the importance of performing reference measurements as they can be used to correct for instrumental disturbances which may occur during an experiment.

This dissertation work also illustrates how feasible it is to approach multivariate data in their original data structure instead of complicating matters by restructuring it into more simple structures e.g., two-way and/or one block. Applying advanced

chemometric methods does not complicate the data analysis of more complex data structures. Instead it provides more intuitive and often directly interpretable results. Multi-way data problems have been treated in PAPER III and PAPER VI and here it is illustrated how fluorescence spectroscopy measurements are capable of providing detailed information about fluorophores in food products. These fluorophores are shown to reflect the chemical composition and the environmental influences during storage of processed cheese and barrel salted herring, respectively. PAPER II is an example of how two-way ^{1}H relaxation curves can be rearranged into a three-way array and that way provide more detailed information about the water distribution in potatoes than obtained by direct analysis. Multi-block problems are treated in PAPER I and PAPER V, where the issues concerning block scaling and factor selection in multi-block analysis are discussed in a separate section in the dissertation work. Both multi-block applications illustrate how a more detailed level of information is obtained when performing multi-block analysis instead of one block two-way analysis on the concatenated blocks.

There is still a gap between the amount of existing multi-way and multi-block data problems and the actual amount of problems treated as multi-way and multi-block data questions. This dissertation work shows that spectroscopy and the chemometric methods specially designed for two-way, multi-way, and multi-block problems have great potential as PAT tools as they fulfill the primary goal of PAT which is to obtain a better process understanding in a faster and more intuitive way especially when approached in the original data structure and dimensionality.

Sammendrag

Den teknologiske udvikling indenfor analytiske metoder er i kraftig fremdrift, og det medfører en øget mængde af multivariate data af stigende data kompleksitet. Multivariate data problemer behandles typisk med standard tovejs kemometri såsom PCA, PCA og PLS regression. Men i nogle situationer vil det være mere passende at analysere data med metoder, som er specielt designet til at håndtere data i deres naturlige datastruktur såsom multivejs og multiblok struktur.

Indeværende projekt er udført for at afdække principperne bag og bidraget fra de mere avancerede kemometriske metoder, som er en del af PAT "værktøjskassen". Metoderne er anvendt indenfor fødevarer og fødevareproduktion, hvor udvalgte grøntsags-, fiske- og mejeriprodukter er evalueret. Der er fokuseret på kemometriske metoder såsom multivejsmetoden, PARAFAC og multiblok metoderne, multiblok PLS og LSparPLSc. Det er hovedsageligt de spektroskopiske metoder NIR, NMR og fluorescens spektroskopi, som er anvendt til at analysere fødevareprodukterne, men chromatografiske metoder, fysiske målinger og sensorik er også anvendt. Den kemometriske indfaldsvinkel har været eksplorativ, hvor fokus i udvælgelsen af de kemometriske analyser har været baseret på at bibeholde den oprindelige datastruktur. Dermed ment at en enkelt blok af tovejs data analyseres med de gængse tovejs analyser PCA og PLS, mens en trevejsstruktur anses for et multi-dimensionelt data problem, og derfor anvendes PARAFAC. Situationer hvor flere tovejsblokke bidrag er tilstede, vil udgøre et multi-blok problem.

Projektet viser, at kemometri er et redskab, som med fordel kan anvendes i fødevareforskning, og at traditionel tovejs kemometri er velegnet til analyse af tovejs data. I PAPER IV er udviklet en metode, der gør det muligt at bestemme proteinindholdet i sildelage fra gammeldagsmodnede sild, hvor nyhedsværdien er, at man måler på sildelagen og ikke på den hele fisk. Med kombinationen af NIR spektroskopi og kemometri er det muligt at monitorere proteinkoncentrationen i sildelagen, hvilket er et udtryk for modningsprocessen i sildene. Afhandlingen indeholder også et afsnit, som omhandler forbehandling af NIR spektre. Dette redegør for, hvorfor det er vigtigt at optage referencespektre undervejs i ens forsøg, da de selvsamme spektre vil kunne bruges til at korrigere for eventuelle instrumentelle forstyrrelser, som kan forekomme undervejs.

Ved at analysere data med multivariat dataanalyse og samtidig tage højde for den naturlige datastruktur undgår man at komplicere dataanalysen unødigt. Hvis man ændrer datastrukturen ved at konvertere multivejsdata til tovejsdata ved at udfolde data eller lægger flere blokbidrag sammen til et stort blokbidrag, kan man

besværliggøre den efterfølgende fortolkning. Ved at bibeholde datastrukturen og bruge avancerede kemometriske metoder, så opnår man mere intuitive løsninger som ofte kan give direkte fortolkninger. I PAPER II og Paper IV er det vist hvordan multivejs problemer i form af fluorescens landskaber kan give detaljerede informationer omkring de tilstedeværende fluorophorer i hhv. smelteost og sildelage, når man bruger PARAFAC. For begge metoder gælder det, at fluorophorerne afspejler de kemiske ændringer, der sker under lagring og ved udsættelse for forskellige typer af stress under lagring. PAPER II illustrerer, hvordan ^{1}H relaksationskurver, omstrukturet til trevejs data, kan give mere detaljeret information omkring vanddistributionen i kartofler end givet ved traditionel tovejs analyse. PAPER I og PAPER V er eksempler på multiblok data analyse, hvor det vises, at multiblok analyse giver en mere detaljeret information omkring data, hvilket giver bedre muligheder for data forståelse.

Der er stadig en kløft mellem, hvor mange multivejs og multiblok problemer, der eksisterer og hvor mange problemer der reelt håndteres som multivejs og multiblok problemer. Denne afhandling viser, at spektroskopi og kemometri kan håndtere tovejs, multivejs og multiblok data, og at de udgør et stort potentiale i PAT regi, idet de kan bidrage med nogle af de egenskaber, som er formålet med PAT, nemlig at opnå bedre procesforståelse på en hurtig og mere intuitiv måde.

List of Publications

PAPER I

Vibeke Tølbøl Povlsen and Connie Benfeldt: Application of Multiblock PLSR in the Dairy Industry. PLS and Related Methods, Proceedings of the PLS'01 International Symposium, V. Esposito Vinzi, C. Lauro A Morineau, M. Tenenhaus (Eds.), 371-383, 2001

PAPER II

Vibeke Tølbøl Povlsen, Åsmund Rinnan, Frans van den Berg, Henrik J. Andersen, and Anette K. Thybo: Direct decomposition of NMR relaxation profiles and prediction of sensory attributes of potato samples. Lebensmittel-Wissenschaft und Technologie – Food Science and Technology, 36 (4), 423-432, 2003

PAPER III

Jakob Christensen, Vibeke Tølbøl Povlsen and John Sørensen: Application of Fluorescence Spectroscopy and Chemometrics in the Evaluation of Processed Cheese During Storage. Journal of Dairy Science, 86 (4), 1101-1107, 2003

PAPER IV

Vibeke Tølbøl Svensson, Henrik Hauch Nielsen and Rasmus Bro: Determination of the protein content in brine from salted herring using near-infrared spectroscopy. Lebensmittel-Wissenschaft und Technologie – Food Science and Technology, 37 (7), 803-809, 2004

PAPER V

Stine Kreutzmann, Vibeke Tølbøl Svensson, Anette K. Thybo, Rasmus Bro and Mikael A. Petersen: Prediction of sensory quality in raw carrots (*Daucus Carota L.*) using multi-block LS-ParPLS, Food Quality and Preference , 19, 609-617, 2008

PAPER VI

Vibeke Tølbøl Svensson and Charlotte Møller Andersen: Characterization of Brine from Salted Herring using Fluorescence Spectroscopy. Lebensmittel-Wissenschaft und Technologie – Food Science and Technology, (*submitted*), 2008

List of Abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
ALS	Alternating Least Square
ANOVA	ANalysis Of VAriance
BSA	Bovine Serum Albumin
CANDECOMP	CANonical DECOMPosition
Corcondia	Core Consistency
CPCA	Consensus Principal Component Analysis
CPMG	Carr Purcell Meiboom Gill
DECRA	Direct Exponential Curve Resolution Algorithm
DTLD	Direct TriLinear Decompositon
EISC	Extended Inverted Scatter Correction
EEM	Excitation Emission Matrix
EMSC	Extended Multiplicative Scatter Correction
FAD	Flavin Adenine Dinucleotide
GCCA	Generalized Canonical Covariate Analysis
GC-MS	Gas chromatography-Mass Spectrometry
GPA	Generalized Procusters Analysis
GRAM	General Rank Annihilation Method
HPCA	Hierachical Principal Component Analysis
HPLC	High Pressure Liquid Chromatography
HPLS	Hierarchical Partial Least Squares
LF-NMR	Low Field Nuclear Magnetic Resonance
LS	Least Squares
LS-parPLS	Least Squares parallel Partial Least Squares
LS-parPLSc	Least Squares parallel Partial Least Squares with common loadings
NADH	Nicotinamide Adenine Dinucleotide
NIR	Near InfraRed spectroscopy
NIT	Near Infrared Transmission
NMR	Nuclear Magnetic Resonance
MB	Multi-Block
MBPLS	Multi-Block Partial Least Squares
MSC	Multiplicative Scatter Correction
MSPC	Multivariate Statistical Process Control
OLS	Ordinary Least Squares
PARAFAC	PARAllel FACtor Analysis
PAT	Process Analytical Technology
PCA	Principal Component Analysis

PLS	Partial Least Squares
R	Correlation coefficient
RMSECV	Root Mean Square Error of Cross Validation
SIS	Spectral Interference Substraction
SNV	Signal Normal Variate
S-PLS	Serial Partial Least Squares
TPA	Texture Profile Analysis

List of Notations

x	Scalar
$\mathbf{x}$	Vector
$\mathbf{X}$	Matrix
$\underline{\mathbf{X}}$	Three way array

a,A	A-score in PARAFAC
b,B	B-loading in PARAFAC
bl,BL	Block number in multi-block analysis
c,C	C-loading in PARAFAC
ex,Ex	X Residual
$\underline{\mathbf{E}}$	Three-way residual
ey,Ey	Y Residual
f,F	Factor
i,I	number of samples
j,J	number of variables in the second array
k,K	number of variables in the third array
λ_{ex}, λ_{em}	wavelengths for excitation and emission
λ	wavelengths
M_0	magnitude of the relaxation curve
$m(t_{NMR})$	Total relaxation signal
N	number of underlying pure mono-exponential relaxation curves
p,P	Loading vector and matrix in multi-block and traditional PCA, and PLS
S_0,S_1,S_2	Levels of energy
t,T	Score vector and matrix in multi-block and traditional PCA, and PLS
$\mathbf{t_T}$	Super score in multi-block PLS
T_2	LF-NMR – transverse or spin-spin relaxation time constant
τ	time between NMR pulse
t	time
w	weights
w_s	super weights
y,Y	Response variable - true value
$\hat{\mathbf{y}}, \hat{\mathbf{Y}}$	Response variable - predicted value

Table of Contents

1. Introduction

The food industry is continuously working on means to improve, optimize and gain a better understanding of the way food productions are run. The increased focus on food quality demands that the industry places a big effort into development and control of food productions. Ensuring the quality of food products requires monitoring and evaluation of every step from the raw material, to the production, to the final product, and in the distribution.

The increased focus on food quality raises the requirements of the analytical methods which are used in food production. Luckily the technological development in analytical instruments has kept up, and advanced methods are available today. This advancement in instrumental technology results in an increased amount of data with higher complexity and dimensionality, e.g., data of a multivariate nature. An illustration of an estimated overview of the existing amount of multivariate data problems represented by two-way, multi-way, and multi-block data is given in Figure 1.

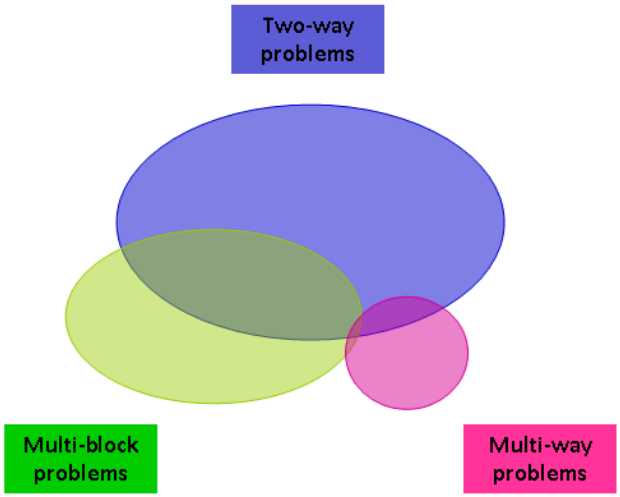


Figure 1. Approximated diagram over distribution of multivariate data divided into two-way, multi-way, and multi-block data.

The diagram shows that a large amount of two-way data exists and the regular two-way data and multi-block data do overlap. Such overlapping data problems can be two-way data consisting of several block contributions treated as a single block problem or a single block, which can be split up into several block contributions.

The multi-way data problems differ from the two-way problems as illustrated by only having a small overlap. As can be seen the univariate data problems are not included in the diagram. Univariate data problems still exist but often univariate data analysis is no longer sufficient when trying to solve the increasing amount of research questions that arise due to the advancements in technology. Multivariate data analysis, also known as chemometric analysis, is designed to handle data problems of multivariate structure. Multivariate data analysis is necessary when studying food and food systems. The natural variation in the chemical composition of foods can be expected to be high since most food originates from living creatures or plants. In order to be able to study and understand this complex structure and interactions between the food constituents multivariate technologies are required.

Process Analytical Technology (PAT) has been in focus ever since the U.S. Food and Drug Administration (FDA) released the "Guidance for Industry PAT – A Framework for Innovative Pharmaceutical Development, Manufacturing and Quality Assurance" (www.fda.gov). In fact, PAT has been a focus area in many industries even years before the term was coined. FDA encourages the pharmaceutical industry to use PAT to overcome some of the drawbacks associated with the existing ways of monitoring and controlling pharmaceutical productions. PAT, as defined by FDA is "*a system for the analysis and control of manufacturing processes based on timely measurements of critical quality parameters and performance attributes of raw materials and in-process materials*". Besides ensuring the product quality by demanding that the product is produced within specifications, applying PAT to food production will give a better process understanding and provide information targeted towards the relevant specifications within a process-compatible time frame. This will ensure that quality is *built into* the product early on in the process (ideally in the development phase) and this can be continued throughout the process.

PAT is like a toolbox filled with tools to improve process understanding and process monitoring. The industry has secretly been peeking in the toolbox but finally the FDA recognize the potential of the toolbox and permission has been granted (FDA guidelines) to dig into the toolbox and explore.

Several of today's food products have existed for centuries e.g., salted herring or semi-hard ripened cheese, and production of such products is associated with great traditions and craftsmanship. Even in production up-scaling the craftsmanship is invaluable, but unfortunately also associated with a high degree of variation and uncertainty due to the fact that the outcome is dependent on the individual employee. Implementing PAT can help reduce the productions' dependency on the

employee at hand and secure a more uniform evaluation of the productions. PAT is not meant as a substitute for good craftsmanship, but as a tool that can determine focus points, speed up the analysis time giving high reproducibility. Where in the process can PAT make a difference? The processing of food can be split up in five parts as illustrated in Figure 2. The first step is the assessment of the raw material, the second step is the pre-processing step, and the third step is the processing of the food, which can be a multi-step process depending on the product at hand. The fourth step is the packing and the fifth step is the distribution. Depending on the product the distribution always has to be in a controlled environment e.g., products that are refrigerated or frozen. Applying PAT has mainly been focused on the step 1 to 3 but step 4 to 5 can also gain from PAT.

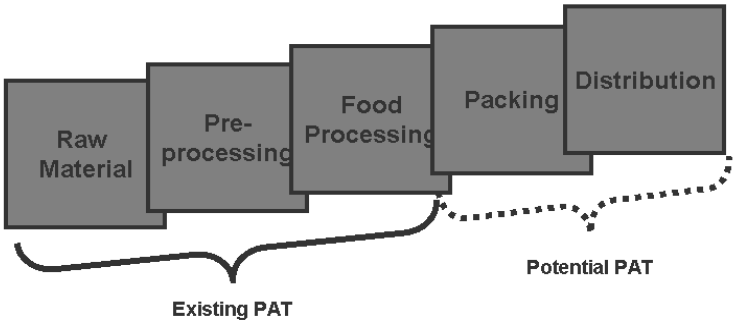


Figure 2. Food processing divided into 5 steps.

PAT is already a well known phenomenon in the food industry, where spectroscopy and chemometrics in the last decades has gained recognition due to their ability to provide fast and relevant information about the food products, especially by Near Infrared spectroscopy. In most cases the spectroscopic analysis is very fast and typically little or no sample preparation is needed. This makes it well suited for implementation on-line or in-line as an automated routine analysis. If the analysis is used at-line the limited sample preparation ensures that all personnel can carry out a measurement. In addition most instrumental software packages allow for instant data processing so the operators can react promptly and take the necessary action. The PAT initiative emphasizes the use of spectroscopy or similar multivariate sensors as they represent a fast and automated alternative to current chemical (laboratory) analysis. Spectroscopy, if implemented and used correctly, can provide highly detailed (and often highly relevant) information about the chemical composition of food products in a short period of time. Often, spectroscopic measurements can either directly, or indirectly through calibration models, provide much more relevant information than traditional quality measures. This makes it

ideal for monitoring processes that change rapidly. Furthermore most types of spectroscopic measurements are non-invasive and non-destructive. This is of great importance in situations where the sample material is very valuable or limited. If you can not control or select the quality of the raw materials you have to control the factors under which a product is produced. Adjusting the production to fit the raw material, thus selecting the right adjustments is crucial in order to perform satisfactory monitoring and assessment throughout the food production. In order to optimize the spectroscopic analysis towards a certain type of production, several issues must be considered such as selecting the appropriate spectroscopic method and the best suited data analysis.

Spectroscopy and chemometrics are being recognized as going hand in hand and are well suited for assessing the quality of food products.

The strength of spectroscopy in combination with multivariate data analysis is that it opens up a level of information (the exploratory spirit), which otherwise could be overlooked or neglected. The exploratory aspect of multivariate data analysis is a gift in process understanding since you can perform a chemometric analysis and just see where it gets you, without knowing anything about the process beforehand. In food processing the exploration can take you places you did not expect especially if you dare to look at a process with open eyes instead of restricting yourself to see only what you expect to find. Such revelations are due to the fact that chemometrics makes it possible to view and compare hundreds or thousands of chemical and/or spectral variables, something which can be quite overwhelming by univariate data analysis and direct visual inspection. Chemometric analysis processes the multivariate nature of spectroscopic measurements in a way that the "essence" of the data will be extracted as descriptive chemical factors. Chemometrics can also handle traditional precursors for process monitoring, e.g., pH, pressure, temperature, etc. But often the traditional monitoring parameters can not explain the chemical composition, the chemical changes and the covariate nature of food systems as well as spectroscopic methods can. PAT is also about combining the traditional process parameters with the multivariate sensors and multivariate data analysis, as the impact of their interconnectivity can give better process understanding and provide a targeted process monitoring. Principal component analysis (PCA) and partial least squares (PLS) regression are among the most common chemometric methods applied to spectroscopic data and have proved to be useful in many situations[1,2,3,4].

Multi-block analysis is superior to the corresponding two-way analysis for data visualization and interpretation of multiple blocks of data.

In the initial screening phase or during a monitoring process where a number of instrumental methods are applied, the intuitively straightforward way to perform data analysis will be to concatenate the data in one block and perform PCA or PLS regression. What is not so straightforward is to try and make sense of the models. An alternative way of evaluating this type of data is to apply multi-block analysis. By this approach the natural structure of data can be maintained, so concatenation of the data blocks is not necessary. Instead the data are modeled in a block structure, thus the data can be treated by their individual block contribution, but also the overall model of all blocks can be evaluated. When the data are expanded by increasing the number of dimensions in e.g., batch monitoring or fluorescence landscapes measurements, multi-way analysis is applicable such as PARAFAC and multi-way regression.

The present work is an elaboration on the conclusions drawn from Figure 1 which illustrates that there is large number of multivariate data problems that needs to be solved. But how much focus is there on using the appropriate multivariate technique? How often is two-way treated as two-way and not univariate data, how often are multi-block problems solved by using multi-block modeling and how about multi-way problems?

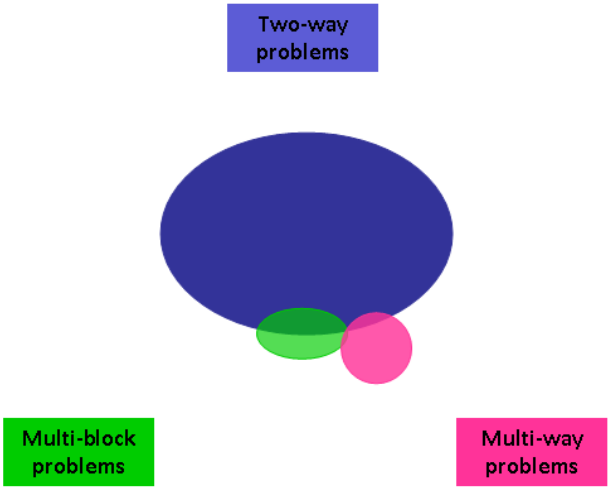


Figure 3. Estimated overview of multivariate data problems treated as such.

Figure 3 is an illustration of an estimate as to how many multivariate problems are treated in their natural data structure. From Figure 3 it clearly appears that there is a big gap between the actual amount of multivariate data problems (Figure 1) and how many are solved as multivariate problems in practice.

As described in the previous sections the "PAT toolbox" includes tools designed to handle these complex data structures. Two-way chemometrics is used more often and the recognition of multivariate sensors and multivariate data techniques expressed through the PAT framework does increase the interest and focus on these multivariate technologies. Because multivariate data analysis is still fairly new, it has taken some time for the industry to adapt to the thought of applying two-way analysis. Because it has taken a while for two-way analysis to be acknowledged there is a tendency to limit the use of multivariate analysis to only single block two-way analysis with the *excuse* that applying more advanced multivariate modeling may complicate things even more. But it is my belief that it is simpler and much more rewarding to view the data problem in its natural structure, thereby handling single block two-way data using PCA and PLS modeling, multiple block problems by multi-block PCA and PLS modeling and multiple dimensional data using multi-way analysis. With this approach there is no unnecessary data compression, no loss of information due to data compression, and interpretation can often be performed directly.

Multi-block and multi-way analyses provide an overview of complex data by processing them in their natural structure and without compressing unnecessarily.

The objective of this dissertation work is to use the tools provided in the PAT framework with the main focus on spectroscopy and advanced chemometrics such as multi-way and multi-block methods to assess food quality.

How to Read the Dissertation

The "backbone" of this dissertation is the three topics; two-way, multi-way, and multi-block chemometrics. The dissertation is divided into two parts where the first part ties the knots between the six papers (PAPER I-VI) which constitute the second half. The work evolves from the *straightforward* two-way chemometric approach into the more advanced methods, e.g., multi-way and multi-block chemometrics. In between the main subjects the analytical methods which provide the multivariate data are presented and described with the emphasis on the spectroscopic methods. Within each of the main data analysis areas applications related to the analytical measurement, food product and/or chemometric methods are given.

Chapter 2 illustrates what multivariate data structure is and introduces the assumptions made for spectral analysis in terms of linearity in order for chemometric analysis to be valid.

Chapter 3 is an illustration of how most multivariate chemometrics are applied today in the form of traditional two-way data analysis on NIR spectral data. An introduction to NIR spectroscopy and applications of NIR spectra treated by two-way chemometrics are given in relation to PAPER IV, which is performed on brine from barrel salted herring. This chapter also includes a section about spectral pre-treatment of NIR and how instrumental disturbances can be treated using such pre-treatment procedures.

Chapter 4 introduces multi-way data in the form of three-way fluorescence excitation-emission spectroscopy (landscapes) and construction of three dimensional data from two-way low field NMR (LF-NMR) relaxation curves (SLICING).

In Chapter 5 multi-way analysis is the topic and an introduction to Parallel Factor Analysis (PARAFAC) and the principles of PARAFAC modeling are given. The chapter also includes applications of PARAFAC modeling of fluorescence landscapes and LF-NMR data reorganized by SLICING (PAPER II, PAPER III, and PAPER V).

Chapter 6 is all about multi-block analysis. An introduction to four multi-block methods is given and pros and cons are discussed. An overview of reported multi-block applications in the area of food science is made (PAPER I and PAPER VI).

Chapter 7 is “the play ground chapter”. It includes the results of a number of topics that were explored for optimizing multi-block modeling in order to make more intuitive solutions and make them easier to access for the inexperienced user. Matrix correlation between blocks and genetic algorithms for optimizing regression was tested. Finally a semi-automated approach for selecting the optimal factor and block weighting combinations is also introduced.

Chapter 8 gives some concluding remarks and formulates some future perspectives on the matters involved in this dissertation work.

Figure 4 gives an overview of the six papers presented in this work with their application, spectroscopy methods and chemometric analysis.

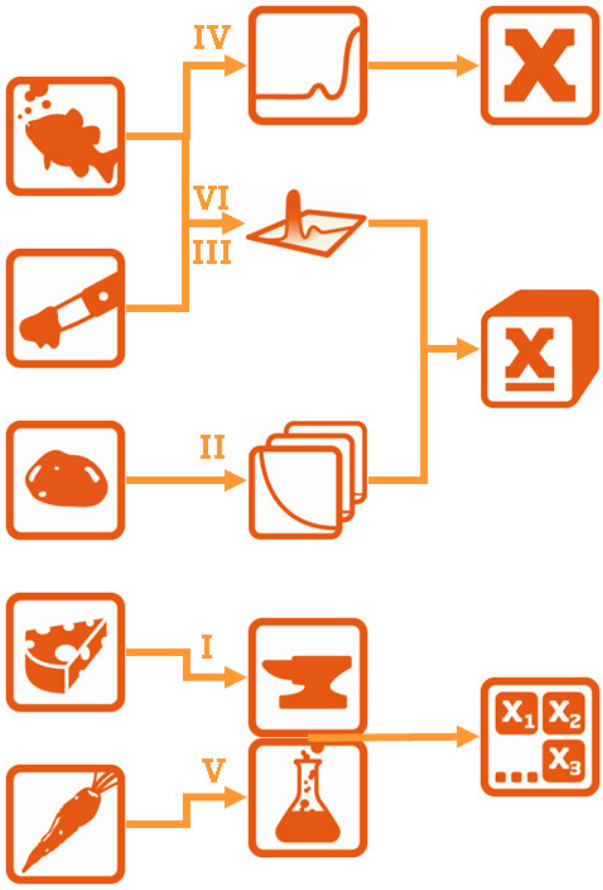


Figure 4. Overview of PAPER I to VI. Herring products, processed cheese, potatoes, semi-hard cheese, and carrots have been analyzed by; NIR spectroscopy, fluorescence spectroscopy, ^{1}H NMR relaxation, and physical/chemical measurements. Data analysis has been divided into three categories: Two-way, multi-way and multi-block.

2. Multivariate Data in Food Science

In the field of food science performing chemical and analytical measurements will result in data presented in different dimensions. A measurement giving one single observation is considered a univariate measurement, e.g., a pH analysis. A measurement resulting in several observations such as spectral measurements is denoted two-way data. Three-way data consists of a three dimension structure such as fluorescence excitation-emission landscapes for a series of samples. An illustration of the data dimensions is given in Figure 5.

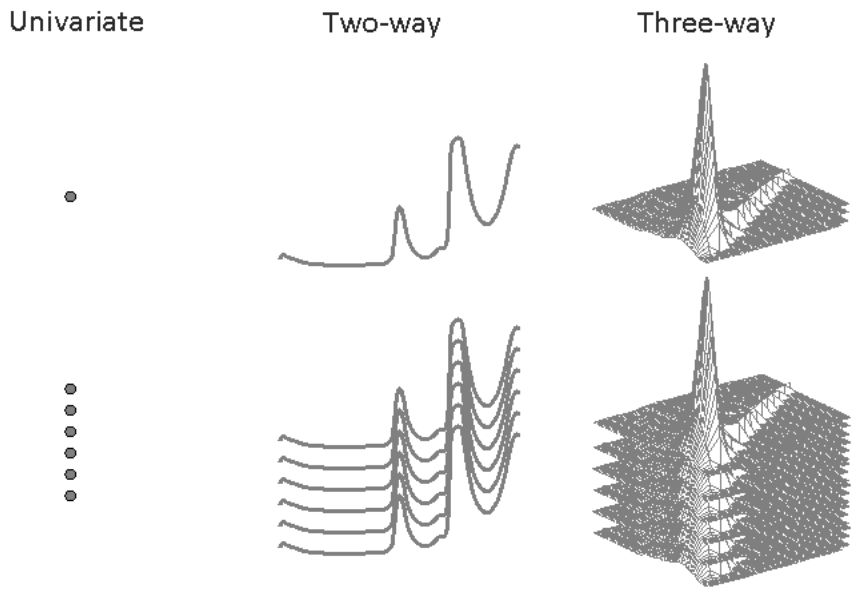


Figure 5. Univariate, two-way and three-way data. Univariate is a single number and several univariate observations result in a vector. Two-way data is a vector, here represented by a NIR spectrum and several spectral measurements give a matrix. Three-way data is illustrated by a fluorescence landscape also denoted a data matrix - several matrices stacked on top of each other result in a three-way array.

The multivariate data examples used in the figure above are spectral measurements and this is not a coincidence since spectroscopic techniques are well suited for analyzing food and food products. They can be applied for both qualitative and quantitative purposes (Box 1). Spectroscopy is one of the analytical methods which are emphasized in the PAT approach, due to high level of information. A full spectrum may contain information about several constituents

and their interactions, information which can not be obtained via a univariate approach.

Box 1. Beer 's Law

Beer's Law (Eq. 1) states that the relationship between the absorbance and concentration of an absorber of electromagnetic radiation is linear, which is essential when spectroscopy is used for quantitative purposes. With this assumption the amount of absorbed light will be proportional to the concentration of an absorbing substance in a sample.

Eq. 1 $$A = \frac{P_0}{P} = abc$$

Where: *A* is the absorbance
c is the concentration
b is the path length through solution (cm)
a is the absortivity coefficient
P_0 is the incident light source
P is the exiting light, which is not absorbed

The spectroscopic methods can, for the majority, be considered as non-invasive, non-destructive, rapid, environmentally friendly and relatively easy to perform. Furthermore they can be used in a variety of locations such as laboratories (off-line), in productions such as in-line, on-line, or at-line measurements and in field work where it can be used as portable instruments. Spectroscopic methods used within the food industry include ultraviolet and visual spectroscopy, fluorescence spectroscopy, nuclear magnetic resonance, microwave absorption, ultrasound transmission, and infrared techniques such as IR and NIR, and Raman spectroscopy - covering most regions of the electromagnetic spectrum (Figure 6). In the present work the following spectroscopic techniques have been applied; NIR (PAPER IV), fluorescence spectroscopy (PAPER III, PAPER VI) and low field NMR (PAPER II).

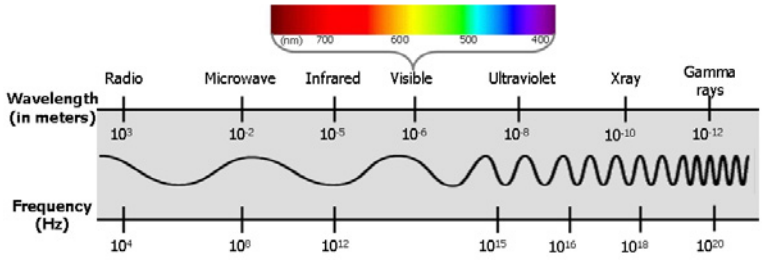


Figure 6. Overview of the electromagnetic spectrum.

2.1 Chemometrics for Handling Multivariate Data

"*The true exploratory potential of the chemometric technology is to identify the first unknown principles with a minimum of a priori bias by measuring first and hypothesizing afterwards.*"

Lars Munck, 2005[5]

The explorative perspective of chemometrics makes it well suited for analyzing data from monitoring food systems and food processing, due to food products' complex physical and chemical composition. Chemometrics is part of the PAT toolbox, because it can help to obtain a better understanding on how food and food production work. Chemometric modeling can be used to study patterns of chemical parameters in food products and it can detect samples with properties that deviate from the remaining samples. A chemometric model separates structural information from un-structured information (residual or noise), where the structured part is the chemical information. The structured information is expressed as underlying latent structures which are represented by factors in accordance with how much of the data variation they explain. Pattern recognition of the physio-chemical composition is made possible by mapping the descriptive parameters (principal components or factors) in an intuitive meaningful graphical representation. This form of data visualization makes data interpretation and data handling much easier as it becomes possible to overview and analyze large data material. Hidden patterns, which under normal circumstances would have been overlooked due to high covariance and correlation between variables, can be revealed. One of the advantages of chemometric modeling is that it can be performed with or without prior knowledge about the food product.

A great interest lies in developing and optimizing chemometric methods which provide better interpretability and intuitive solutions for as many data scenarios as possible. A typical data set up is chemical and spectral measurements represented by a data matrix (**X**) (Figure 7). Such data can be analyzed using two-way analysis e.g., PCA and PLS regression. In cases where several analyses are performed on the same material multi-block analysis is an option (Figure 7). Multi-block analysis treats several blocks [$\mathbf{X_1}$ $\mathbf{X_2}$,…,$\mathbf{X_{BL}}$] in one model and provides information about the individual block contribution combined with the overall model of all block contributions. If the data material exceeds two dimensions such as three-dimensional data ($\underline{\mathbf{X}}$) e.g., fluorescence excitation-emission landscapes and reshaped low field nuclear magnetic resonance (LF-NMR) data, multi-way analysis is appropriate (PARAFAC) (Figure 7).

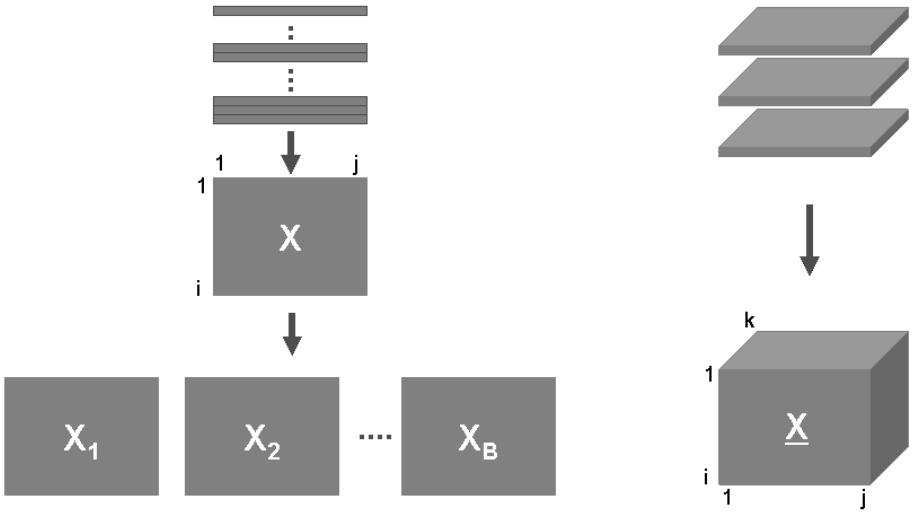


Figure 7. Examples of data structures suited for PCA and PLS analysis, the data matrix (**X)**. Multi-block analysis, data blocks [$\mathbf{X_1X_2}$,…,$\mathbf{X_{BL}}$], and multi-way array $\underline{\mathbf{X}}$.

In the PAT framework, the role of chemometrics is to build a link between the data collection and data analysis and in doing so a number of considerations must be made in order to get valid results. Important issues in PAT analysis for monitoring food processes are to define the problem at hand and try to understand the problem. This might seem obvious, but quite often in practice the problem has not been defined beforehand. In the step designing the experiment, one must remember to handle issues concerning sampling, e.g., how is sample material removed from the product and how should the samples be analyzed? Is it by chemical analysis and/or sensor technology? How are data collected and stored and what type of data analysis should be performed? And when performing the data analysis is scaling and signal pre-processing necessary? What is the best way to visualize and interpret data? All these topics are crucial to consider before and when applying PAT (Figure 8).

The forthcoming chapters will illustrate how traditional two-way chemometrics can be performed and then move on to the more advanced chemometric methods where the focus will be on multi-way analysis methods for analyzing fluorescence landscapes and LF-NMR relaxation curves on multi-block regression analysis for handling several blocks of data. The multi-block and multi-way methods are extensions based on the concepts from PCA and PLS regression. In PAPER I, PAPER II, PAPER III, PAPER V, PAPER VI, PCA and PLS regression are used for preliminary

data analysis followed by multi-way or multi-block modeling (results not always included). PAPER IV is based solely on the basic PCA and PLS regression of NIR measurement on brine. PCA and PLS regression are the most frequently used chemometric methods within the area of food science; hence it is assumed that the reader is familiar with terms and expressions connected to PCA and PLS regression analysis, and therefore it will not be explained further - for more information about basic two-way analysis the reader is referred to Martens and Næs[6] and Martens and Martens[7].

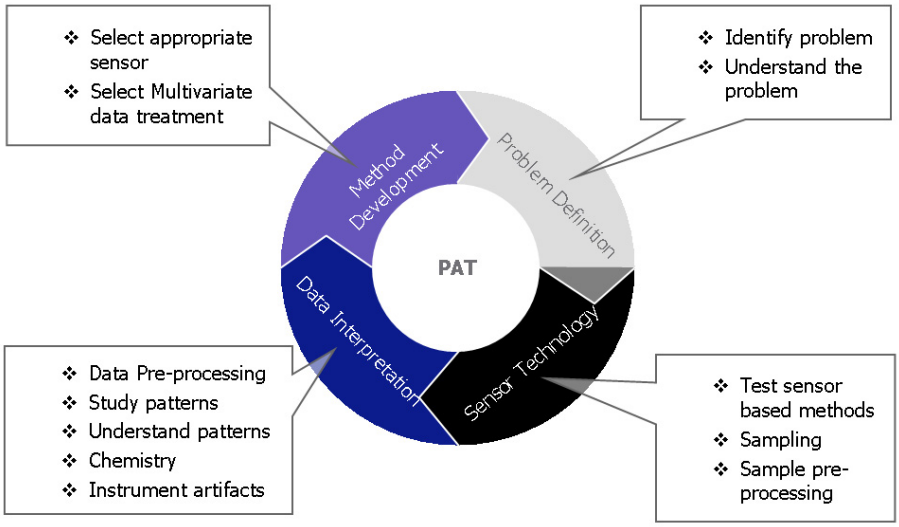


Figure 8. What to be aware of when implementing PAT in food science.

3. Traditional Two-way Chemometrics of NIR Data

The discovery of infrared light was discovered in 1800 by Sir Frederick William Herschel (1738-1822) when he discovered a light radiation beyond the visible spectrum when passing sunlight through a prism. This discovery laid the ground work for the vibrational spectroscopy methods such as Raman, IR and NIR spectroscopy. NIR spectroscopy is the most commonly used spectroscopy for assessing food products and food processes and is therefore of great relevance in the PAT framework for evaluating food quality. Analyzing NIR spectra using chemometrics such as PCA or PLS regression can extract information based on the entire spectra. This is a huge advantage when looking for patterns, correlations and in general when trying to obtain a better understanding of NIR data when limited or no prior information exists.

Box 2: Principles of NIR Spectroscopy

The NIR region covers the range from 4.000 to 13.000 cm^{-1} (780 to 2500 nm). The NIR spectra reflect the overtones and combination bands from molecules with a small energy difference in their vibrational and rotational state. The vibrations of a molecule can be divided into stretching and bending of covalent bonds (scissoring, rocking, wagging and twisting) as illustrated in Figure 9.

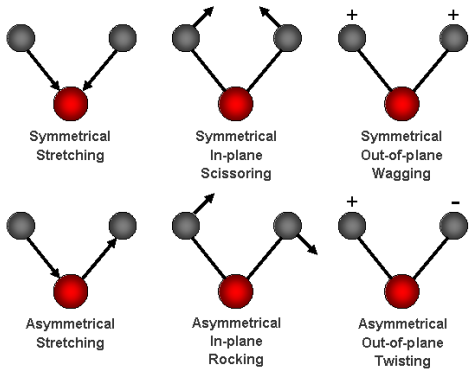


Figure 9. Overview of the molecular vibrations illustrated by H_2O. O: red and H: grey.

The NIR absorption bands are typically broad, overlapping and 10-100 times weaker than their corresponding fundamental mid-IR absorption bands. The rather weak absorption bands express mainly the functional groups which have a hydrogen atom attached to a carbon (CH), nitrogen (NH), and oxygen (OH).

NIR spectroscopy measurements can provide information about the chemical composition, since some of the main constituents in food, namely water, fat, carbohydrates, and proteins absorb in the near infrared region. Because NIR analysis is fast and non destructive and can handle solid samples as well as liquid samples, it is well suited for food evaluations. In Table 1 an overview of the most relevant molecules in food evaluation are listed.

Table 1. Overview of relevant wavelengths for major food components in the near infrared region (modified from Li-chan and coworkers[8]).

Food constituent	Wavelength (nm)	Assignment
Protein	910	C-H stretch 3rd overtone
	1020	2*N-H stretch + 2*amide I
	1510	N-H stretch 2nd overtone
	1980	N-H asymmetric stretch + amide II
	2050	N-H- symmetric stretch + amide II
	2180	1*amide I + amide III
Fat	928	C-H stretch 3rd overtone
	1037	2*C-H stretch + 2*C-H deformation + $(CH_2)_n$
	1200	C-H stretch 2nd overtone (CH_2 groups)
	1734	C-H stretch 1st overtone (intramolecul. H bond)
	1765	C-H stretch 1st overtone (intramolecul. H bond)
Starch	990	O-H stretch 2nd overtone
	1440 (Sucrose)	O-H stretch 1st overtone
	1528	O-H stretch 1st overtone (intramolecul. H bond)
	1540	O-H stretch 1st overtone (intramolecul. H bond)
	1580 (Glucose)	O-H stretch 1st overtone (intramolecul. H bond)
	1900	O-H stretch + 2*C-O stretch
	2100	2*O-H deformation + 2*C-O stretch
	2252	O-H stretch + O-H deformation
	2276	O-H stretch + C-C stretch
	2461	C-H stretch + C-C stretch
	2488	C-H stretch + C-C stretch
	2500	C-H stretch + C-C stretch
Moisture	970	O-H stretch 3rd overtone
	1450 (Starch)	O-H stretch 1st overtone
	1940	O-H stretch + O-H deformation

The potential of NIR spectroscopy as a fast method to assess agricultural and food products was first introduced in 1961 by Norris and coworkers[9] and is to be considered the primary vibrational spectroscopy used for quantitative analysis of the major food components. The number of food related NIR publications has doubled within the last 10 years[8]. The latest trends are the pursuit to use NIR real time process monitoring and control. A recent review outlines the major trends within NIR on-line/in-line monitoring of food and beverages[10]. It shows that the food industry has acknowledged NIR's potential for monitoring process quality, even long before the release of the PAT guidelines. In the recent years the majority

of NIR data are analyzed using some kind of multivariate data analysis where PCA, PCR and PLS are the most dominant.

3.1 NIR Spectroscopy Applied to Fish Products

Assessing fish and fish products using NIR spectroscopy and chemometrics are well known in the field of aquaculture. It turns out that rainbow trout is the most studied fish by NIR spectroscopy, but also salmon, cod and tuna are assessed frequently by NIR. The majority of the performed studies have been used to determine one or more of the major chemical components in fish such as fat, protein, salt or water content. The first NIR application for assessing fish dates back to 1987 where Gjerde and Martens[11] studied the chemical composition in form of the main constituents, e.g., water, fat and protein of freeze dried rainbow trout. Their study was followed by a similar study on freeze dried trout and arctic charr by Mathias and coworkers[12]. Since then the major milestones in this area have been to measure NIR directly on the meat from rainbow trout with no prior treatment besides freezing [13]. In 1992, a NIR measurement was performed on a whole intact rainbow trout and studied in the short wave region (700-1100 nm) by Lee and coworkers[14]. NIR spectroscopy used for predicting sensory profiles of fish were reported in 1997 by Jørgensen and Jensen[15], where sensory assessment and water holding capacity of cod were correlated to NIR spectroscopy. In 2002 Uddin and coworkers[16] applied NIR spectroscopy to detect the end-point temperature and water holding capacity of three types of fish and shellfish. In 2003, Solberg and coworkers[17] performed a study on live fish under sedation to evaluate the fat content in farmed Atlantic salmons. Sollid and Solberg[18] narrowed down the region from 850-1050 nm to assess the fat content in salmon. This short wave NIR region (700-1100 nm) has since been used to asses the major chemical components in rainbow trout[14], salmon and salmon related products (fat/fat and protein)[17,18,19,20,21,22,23], halibut (fat, protein, drymatter) [24], tuna [25,26], mackerel [27], surimi [28,29]. A few studies comparing the performance of NIR to other instrumental methods for assessing fish have been performed. A study comparing NIR to the Torry Fatmeter, a microwave based method and the fexIKA (Commercial modification of the Soxhlet procedure), where the microwave based method and NIR performed equally well when predicting the fat content in herring[30]. Another study compared NIR to the Fatmeter and NMR[31] spectroscopy, and it concluded that NIR was suited as a method for sorting whole herring or fillets in a production line. Since texture and the chemical composition are closely related, NIR spectroscopy has also been studied for classification of salmon based on the texture profiles by Isaksson and coworkers[32]. The study showed that NIR spectroscopy correlated fairly well with the textural shear force measurement. NIR has also been studied as a method to express the freshness of fish in terms of

sensory quality. This has been studied in both untreated fish[15,33,34,35] and in processed fish products[34] with promising results.

NIR Spectroscopy for the Assessment of Herring

In PAPER IV NIR spectroscopy was proven to be an excellent and fast method for determining the protein in brine for barrel salted herring. In general applications for assessing herring by NIR spectroscopy are very limited, thus only a few studies of NIR and herring have been reported. Recent work includes a sensory study where NIR spectroscopy on raw fish species including herring was tested as a predictor of the sensory quality of cooked fish[36]. Herring was also used in the two studies previously mentioned, where NIR spectroscopy was compared to other instrumental methods to determine the fat content[30,31]. PAPER IV presents a standard two-way analysis (PCA and PLS regression) approach for analyzing NIR data in order to determine the protein content in brine of barrel salted herring. The method is an indirect way to monitor the quality of whole salted herring. The approach of measuring the protein content in the surrounding brine instead of sampling the whole herring is a big advantage as collecting brine samples and measuring them is much easier than handling the issues of homogenous sampling of whole fish carcasses due to the in-homogeniety of the chemical composition in fish[37,38]. The results show that performing PLS regression on the NIR spectra can predict protein content in brine with a correlation of r=0.93 and an RMSECV of 0.25 g/100g when selecting spectral regions. PLS regression on the entire NIR region resulted in predictions of r= 0.87 and an RMSECV of 0.35 g/100g. This corresponds well with other studies predicting protein in other fish species. No reported studies have been performed to determine the protein content in herring or herring products, but protein assessment based on the NIR region from 1100-2500 nm has been performed by Mathias and coworkers[12], where freeze dried rainbow trout and arctic charr were analyzed using two different NIR instruments for the assessment of protein and resulted in correlations of r= 0.88 and 0.97. A study by Isaksson and coworkers[39] did not give the same satisfactory results when predicting protein in whole fillets of salmon, but instead promising results were obtained for ground salmon. Other studies of protein show that a correlation of r= 0.85 can be reached when assessing fishmeal[40]. The most recent study by Khadabux and coworkers[41] predicted the protein content of two tuna species, Skipjack and yellow fin tuna with a high correlation, r= 0.99. In whole sea bass fillets the prediction of crude protein failed, but r=0.68 was reported in freeze dried fillets[42].

3.2. Pre-processing of NIR Spectra

The ideal situation in spectral analysis would be if the absorption band of every analyte could be presented as isolated absorption bands - but in NIR spectroscopy unfortunately this is usually not the case. NIR spectra of food samples are often broad and diffuse and furthermore they can be influenced by noise. These disturbances can be due to light scattering, chemical shifts, sample composition and molecular interactions. In Figure 10 the raw NIR spectra from brine samples are shown.

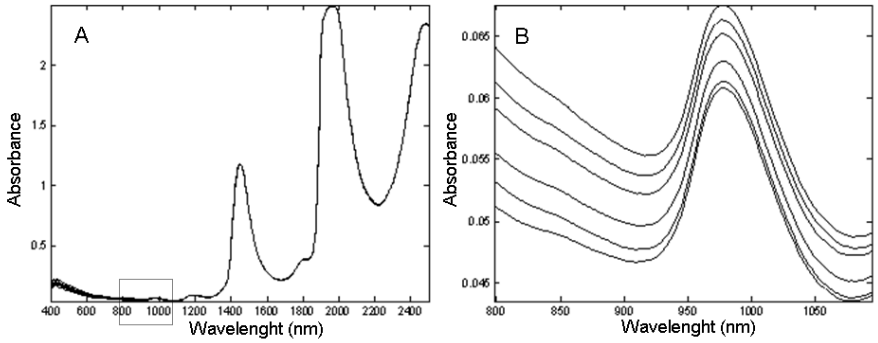


Figure 10. A). The raw NIR spectra from 400 to 2400 nm. B). A zoom of the region from 800 to 1094 nm from brine sample.

In order to be able to extract the chemical information from the NIR spectra with spectral artifacts pre-processing of the spectra is required. A number of spectral preprocessing methods exist. The most used methods are the reference based methods which can be divided into scatter correction methods and smoothing methods (derivative). The present work has used the scatter correction methods; multiplicative scatter correction (MSC), extended inverted scatter correction (EISC) and signal normal variate (SNV) and the derivative method Savitzky-Golay. Other methods exist, but these will not be mentioned in this work; they can be found in the reference by Rinnan and coworkers[43], where more detailed information about the presented pre-processing methods can be found.

Multivariate scatter correction, originally designed to handle multiplicative and additive effects in spectra due to light scattering, can be used to handle other spectral interferences[44]. The principle of the MSC method is to use a correction coefficient to correct the spectra for non-linearties (Figure 11A). A reference

spectrum (typically the average spectrum) is used when calculating the correction coefficient.

An extended version of the MSC (EMSC), including wavelength corrections and corrections based on prior spectral information (spectral interference subtraction (SIS) has been purposed[45]. The EISC method is a further development of the extended MSC methods and is suited in situations where the reference spectrum is noisier than the measured spectrum[46]. The standard normal variate[47] approach also uses a correction coefficient, but in contradiction to MSC it does not require a reference spectrum.

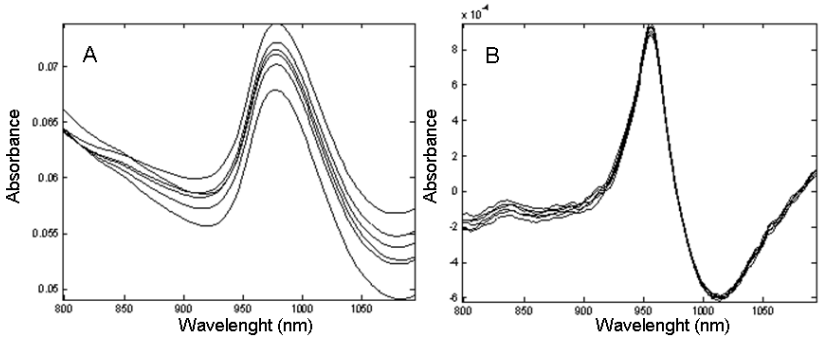


Figure 11. A). MSC corrected NIR spectra and B). 1st derivative Savitzky-Golay corrected NIR spectra. The illustrations are a zoom from the entire NIR spectra from 400-2500 nm where spectral preprocessing has been performed.

The Savitzky-Golay derivative[48] can be performed as a 1st or 2nd derivative and is normally used for smoothing and noise reduction purposes (Figure 11B). The principle of the Savitzky-Golay derivative is to smooth over a number of points of the spectra by fitting a predefined polynomial against the spectral points. A number of factors have to be determined when performing Savitzky-Golay, where the size of the smoothing window and the order of the polynomial to be fitted against have to be selected.

3.3. Handling Instrumental Artifacts by Spectral Pre-processing

The following section focuses on how to manage the type of instrumental disturbances which occurred when performing NIR spectroscopy measurements on

the brine samples from PAPER VI. A total of 176 barrels (batches) of salted herring had samples taken six times during the ripening period.

All samples were evaluated by Near-infrared transmission (NIT) spectroscopy using an Infraprover II fourier transform spectrometer (Bran and Luebbe, Germany) over a period of 10 days and every day a Bovine Serum Albumin (BSA) standard was measured. After the fourth day of measurements the NIR instrument broke down, the remaining samples being measured on a similar, second NIR instrument.

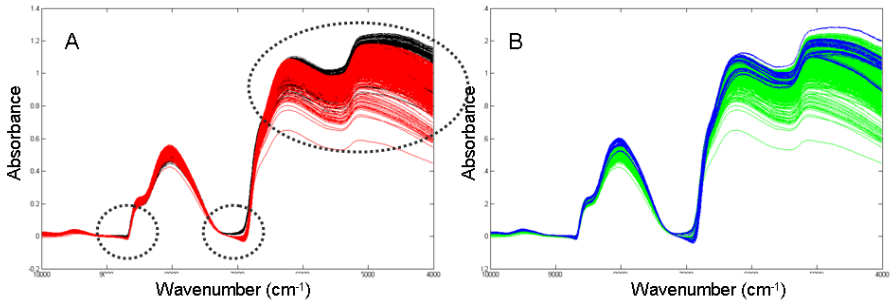


Figure 12. A). Raw NIR spectra from brine samples measured on two instruments. Black – Instrument I and red – Instrument II. Circles indicate the areas where there is a difference in spectra between instrument I and II. B). Blue -BSA standard NIR spectra and green NIR spectra from brine samples.

Changing the NIR instrument during the measurement trial resulted in spectral disturbances which could be assigned to NIR instrument I and instrument II as illustrated in Figure 12A and Figure 13. Performing spectral pre-processing in the form of MSC and differentiation by Savitzky-Golay can not remove information due to instrument I and II. Instead correcting using the standard BSA spectra measured every day of sampling seems to remove the instrument variation. Comparing the BSA spectra from the two instruments it is obvious that they express the same spectral curvatures unique for each of the two instruments and it makes sense to assume that subtracting the reference spectra can correct for the instrument variation (Figure 12B).

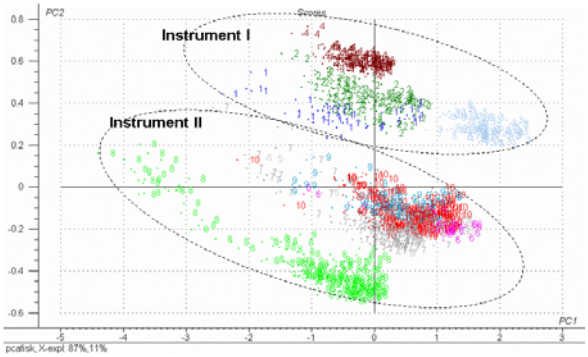


Figure 13. PCA score plot of the brine samples based on raw NIR spectra. A separation of samples according to sampling day and instrument can be observed.

After the BSA standard correction day to day variations are still present and in order to correct for this MSC, 1st and 2nd derivative (Savitzky-Golay) were tested (results not shown). EISC pre-processing was able to remove the remaining variation caused by sampling days (Figure 14).

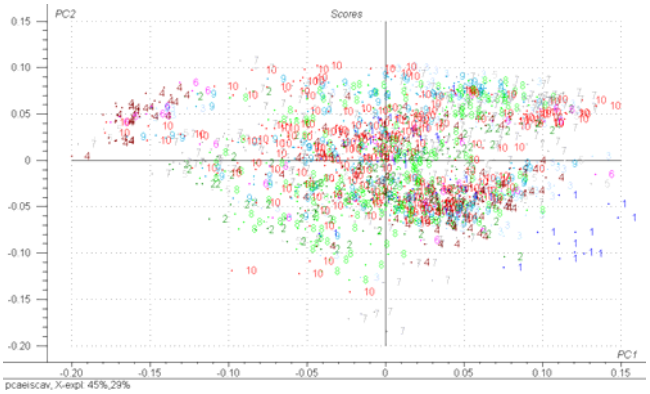


Figure 14. PCA plot from NIR measurement on brine samples after correction with BSA standards and pre-processed with EISC.

4. Beyond Traditional Two-way Chemometrics

Higher data dimensionality and an increased number of data blocks are a result of the advancements in the instrument technology. Standard two-way chemometrics is still the most used data analysis when analyzing multi-way and multi-block problems.

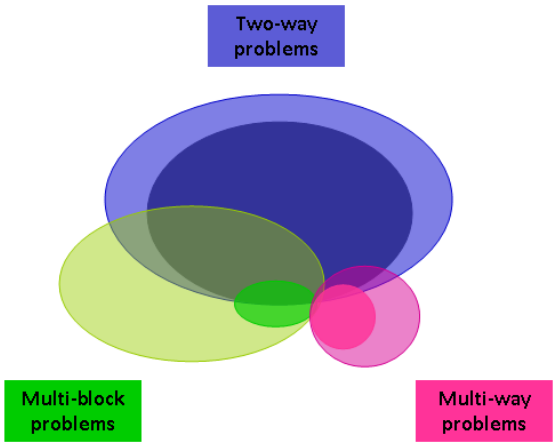


Figure 15. A view of the world of multivariate data. The large circles represent the existing data problems and the smaller circles the actual problem treated as such.

In order to extract the right information in a more intuitive way the data problems should be addressed by multivariate methods designed especially to handle multi-way and multi-block problems. But in order to even out the difference between the existing number of two-way, multi-block and multi-way problems and the actual number of problems handled as such (Figure 15), the focus and accessibility of these more advanced methods must be emphasized. Figure 15 is an illustration of Figure 1 and Figure 3 merged in order to give a better visual comparison.

4.1. Fluorescence Landscapes – A Three-way Data Source

The first observation of fluorescence was made by Sir John Frederick William Herschel (1792-1871) in 1845. Similar to his father's discovery 45 years earlier, sunlight was the light source for his discovery of fluorescence from a quinine solution[49]. The advantage of fluorescence spectroscopy compared to other spectroscopic methods is its sensitivity and high specificity. Compounds that absorb at the same wavelength in other spectral regions may be differentiated due to their

different fluorescence properties and this is a great advantage when analyzing food samples.

Box 3: Principles of Fluorescence spectroscopy

Fluorescence spectroscopy is of special interest in the UV/Vis region ranging from 200-700 nm. The UV region is colorless for the human eye, but in the visible regions some of the wavelengths have characteristic colors. Photons in these regions possess energies which promote electronic transitions in certain molecules that lead to emission of light. Molecular luminescence spectroscopy describes emission of light from molecules in electronically excited states, where usually only the ground state (s_0), the first excited state (s_1), and the second excited state (s_2) are involved in fluorescence. The emission appears 10^{-11} – 10^{-7} seconds after excitation and can arise from excitation by way of absorption of light (photoluminescence) or by way of a chemical reaction (chemiluminiscence). Photoluminescence can be divided into fluorescence and phosphorescence, where fluorescence is the focus of this study. In Figure 16 an illustration of a Jablonski diagram is given.

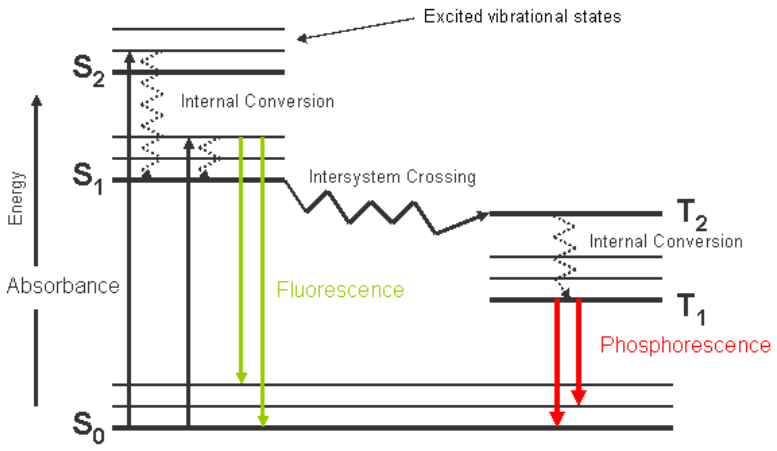


Figure 16. Jablonski diagram. The principles of fluorescence spectroscopy (modified after Lakowitz[49].

It can be observed that the wavelength of the emitted energy will be lower than the energy during absorption. This is due to loss of energy in the excited state. The absorption frequency is usually greater than the fluorescence frequency because the absorption process puts the molecule in an excited vibrational level of the excited electron state. Rapid decay to the lowest vibrational level (s_1) then occurs before emission (internal conversion) [49 50].

Performing so-called excitation-emission fluorescence spectroscopy consists of a set of spectra recorded in a range of wavelengths (λ_{ex}) and their corresponding emission spectra. This representation of an excitation-emission matrix (EEM) is also known as a fluorescence landscape. The three-way structure of fluorescence landscapes makes them an ideal data source for multi-way modeling. The advantages of EEM's are the ability to providing a higher level of information compared to e.g., uni- or unfolded two-way fluorescence measurements. It is very powerful when analyzing more complex mixtures such as food products, and the combination of EEM with multi-way data analysis (PARAFAC) has made the EEM analysis much more assessable and much easier to interpret due to its more intuitive solutions. This approach is used in the steady-state fluorescence[a] applications presented in PAPER III and PAPER VI.

In food related studies especially the intrinsic fluorescence from aromatic amino acids is relevant, where the protein fluorescence is due to tryptophan, tyrosine and phenylalanine. Other natural occurring fluorophores are the cofactor enzyme nicotinamide adenine dinucleotide (NADH), pyridozyl phosphate and the flavins (adenine dinucleotide (FAD)). Vitamin fluorescence is represented by the water soluble vitamins B_2, B_6, and fat soluble vitamins are vitamin A and vitamin E. Pigment fluorescence in food is often due to chlorophyll and hematoporphyrin[49,51]. The fluorescence properties of the fluorophores are highly dependent on several environmental factors such as pH, temperature and solvent, and can affect their fluorescence. The molecular structure also influences how the fluorescence will be executed. If the fluorescent compound is hidden in the molecular structure or if it resides close to the surface[49,52]. When analyzing biological material by fluorescence spectroscopy both intrinsic and instrumental factors can give spectral disturbances which may result in deviations from Beer's Law. Light scattering from in-homogenious sample material, containing larger molecules or aggregation of matter can occur. Raman scatter is caused by scatter and is explained in the following section. If the optical density of the sample is too high it may hinder emitted light from a fluorophore to reach the detector. This will result in non-linear absorption measurements which will deviate from Beer's Law (Box 1) and thereby linear modeling is no longer appropriate.

Scatter Effects

When performing fluorescence landscapes measurements the phenomena of Raman and Rayleigh scattering is almost impossible to avoid. Rayleigh scatter refers to the scattering of light by particles and molecules smaller than the wavelength of

[a] Steady-state fluorescence is performed with constant illumination when the emission is recorded. The opposite of steady-state is time-resolved fluorescence.

the light. Rayleigh is so-called elastic scatter which means that no energy loss is involved and therefore the scattered light occurs at the same wavelength as the incident light. In praxis Rayleigh scatter mainly influences fluorophores with a small Stoke's shift. This is because the fluorescence signal of this type of fluorophores will be excited and emit in the area close to the Rayleigh scatter. Rayleigh scatter can be present in 1^{st} order Rayleigh and 2^{nd} order Rayleigh scatter. 2^{nd} order Rayleigh appears at emission wavelengths twice the given wavelength. Raman scatter is inelastic scatter. A constant energy loss will appear for Raman scatter which results in scattered light having a higher wavelength than excited light with a constant difference in wave numbers (energy). Raman scatter will be a diagonal with a systematic increasing deviation from the Rayleigh scatter. Scatter may interfere with the data analysis of EEM's, but can be dealt with simply by removing the data points with scatter [53,54].

4.2. ^{1}H NMR Relaxation Curves - A Three-way Data Source

In contrast to the well-established vibrational spectroscopy and fluorescence spectroscopy, nuclear magnetic resonance (NMR) has not existed for more than 60 years. The first successful NMR study using bulk magnetization was performed in the mid 1940s by two independent research groups at Harvard [55] and Stanford [56]. Despite the fact, that it is one of the youngest spectroscopic methods, NMR has shown remarkable growth[57] and has been widely used in the analysis of foods[58].

Performing CPMG (Carr-Purcell-Meiboom-Gill) experiments on foods can lead to information about the physical properties of the molecules in foods by studying the relaxation rates of the protons. Box 4 introduces the basic CPMG measurement terminology in relation to determination of T_2 relaxation times for protons. For basic NMR theory Harris[59] and Eads and Davis[60] are recommend and for more advanced theory Abragam[61] is suggested.

PAPER II focuses on the principles of ^{1}H low field NMR CPMG experiments[62,63] with emphasis on determination of the transverse relaxation time (T_2) to study water distributions in potatoes. The relaxation curve consists of the maximum time points of each of the spin echoes illustrated in Figure 17.

Box 4: The Principles of NMR Relaxation

NMR is performed on nuclei whereas the previously described spectroscopic methods are performed on electrons. Only nuclei with a non-zero spin quantum number (I) will be NMR active. An example of a nucleus with I = ½ is hydrogen (^{1}H), The proton (^{1}H) can provide information about the water distribution and is therefore of great interest when studying food products and other biological systems[64]. Other popular nuclei are ^{13}C, ^{15}N and ^{31}P all with the spin quantum number of I=½. When exposed to a strong external magnetic field (B_0) the nuclei will be distributed in (2I+1) energy levels. ΔE (= γB_0) is the difference in energy between the two energy levels, where the spin can be parallel or anti-parallel to B_0.

The CPMG experiment

The following is based on ^{1}H which is the most abundant NMR nucleus in food systems. In the rotating frame (coordinate system rotating with the Larmor frequency[61]), the CPMG experiment consists of a 90˚ ×-pulse to turn the equilibrium z-magnetization along the y–axis, then a delay τ, a 180˚ y-pulse (refocusing pulse), a delay τ and then the first spin-echo is formed. After the second τ-delay the magnetization is directed along the y–axis as it was just after the initial 90˚ ×-pulse. That is the reason for the term "echo". The echo maximum is the first data point. Repeating the "delay τ, 180˚ y-pulse, delay τ"-block N times results in creation of N additional echoes and thereby acquisition of N additional data points separated in time by 2τ. In Figure 17 the CPMG pulse sequence is illustrated.

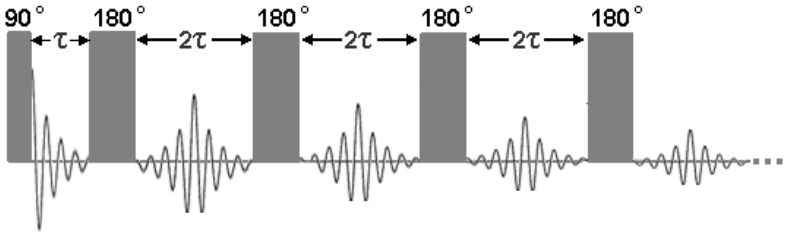


Figure 17. CPMG sequence.

The relaxation curve consists of a sum of exponentials and can be described by Eq. 2.

Eq. 2. $$m(t_{NMR}) = \sum_{n=1}^{N} M_{0,n} \cdot \exp\left(-\frac{t_{NMR}}{T_{2,n}}\right) + Ex$$

Where: $m(t_{NMR})$ is the total relaxation signal
N is the number of underlying pure mono-exponential relaxation curves
M_0 is the magnitude of the relaxation curve
t_{NMR} is the time
T_2 is the transverse relaxation time
Ex is the residual

A short T_2 indicates low molecular mobility (e.g., crystalline or bound water) whereas high molecular mobility (e.g., free water) results in a long T_2. The number of exponentials present in a relaxation curve reflects the number of underlying mono-exponentials and equals the number of ^{1}H sites with different relaxation properties. A number of ways to estimate the number of mono-exponentials present in the relaxation curves can be used. The most familiar approaches are bi-exponential fitting and distribution analysis.

4.3. The SLICING Approach

The SLICING method was developed by Pedersen and coworkers[65] to restructure CPMG LF-NMR data into a three-way array of relaxation curves. This procedure is then followed by decomposition of the tri-linear structured data by multi-way analysis into individual exponentials corresponding to the number of relaxation compounds. The method is an alternative to the existing fitting methods, e.g., such as bi-exponential fit, distributed exponential fit, discrete exponential fit. The method originates from the direct exponential curve resolution algorithm (DECRA) introduced by Windig and Antalek[66]. SLICING is a method that reorganizes two-way relaxation curves into a pseudo-three-way structure and analyzes the data using PARAFAC modeling. The third dimension is retrieved by reorganizing the data by forming “slices”, as the name implies.

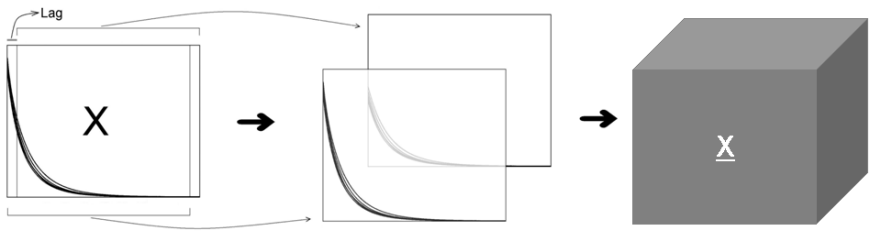


Figure 18. Principles of SLICING lag and slab.

A relaxation curve expresses the number of relaxation times (T_2's) present in data and the number of T_2's equals the number of exponentials. Because the relaxation curve can be described as a sum of exponentials as given in Eq. 2, it is possible to rearrange data by stacking the data behind each other (slab) introducing a slight displacement (lag) for each slab. This is illustrated in Figure 18. The first slab will be equal to the original data and will include relaxation components represented in the first part of the data, e.g., the fast relaxation components, whereas the second slab is dominated by the relaxation components with a slower T_2. The dimension of the rearranged data structure will be a three way array, where the first mode is the number of samples, second mode is the relaxation profiles as a function of time, and the third mode is the number of slabs (X_{ijk}=[object (*i*) × time (*j*) × slab (*k*)]). Parameters, which can be changed when performing SLICING are the slab size, lags and number of components. This type of data has a tri-linear structure and is thus suited for PARAFAC modeling, where the A-scores will be proportional to the magnitude of the relaxation curve (M_0), B-loadings are the estimated relaxation curves and the C-loadings have no practical function. SLICING as proposed in the work from Pedersen et al[65] has been applied in PAPER II. In a later publication Engelsen and Bro[67] addressed some of the disadvantages of the SLICING approach and proposed a faster and more accurate approach (PowerSlicing). In the original DECRA[66] algorithm the general rank annihilation method (GRAM)[68] was used which limits the size of data resulting in a maximum number of two slabs. The SLICING method has overcome this problem as it is based on the direct tri-linear decomposition (DTLD)[69] which can handle more than two dimensions in the slab directions. The advantage of increasing the number of slabs is that it gives more accurate estimates of the relaxation times (T_2). The fast relaxing components will be resolved based on the first slab, whereas the slower relaxing components will be dominating in the following slabs, and by including more slabs a more accurate T_2 will be obtained. The disadvantage of expanding the number of slabs is that there are an (almost infinite) number of possibilities for selecting the optimal slab and lag combination. And the drawback is that determining the optimal number of slabs

can be rather time consuming. One way to overcome this problem has been proposed in the PowerSlicing method[67]. Both SLICING and PowerSlicing suffer from the requirement of a minimum of two decaying curves, but this is no longer an issue since just recently a new method called DoubleSlicing[70] was introduced. The DoubleSlicing method differs from the previous slicing methods as it is designed to generate a three-dimensional data matrix from a single decay curve.

5. Multi-way Analysis

Handling data of higher dimensions than two are referred to as multi-way analysis. This particular work has dealt with analyzing data with three dimensions. Usually two-way data analysis is performed on a data matrix, **X** ($I \times J$) whereas multi-way analysis like three-way data analysis will be performed on a three-way array, **X** ($I \times J \times K$). When referring to the dimensions in multi-way analysis the term "mode" refers to the array direction e.g., a three-way array will have the first, second, and third mode. In the previous chapter excitation-emission fluorescence spectroscopy as a multi-way data source was described and when evaluating fluorescence landscapes Parallel Factor Analysis (PARAFAC) is used. PARAFAC is also a part of the second multi-way approach, SLICING where an "artificial" three-way array from two-way CPMG relaxation curves is created and then analyzed by PARAFAC in order to extract the relaxation NMR profiles.

5.1. Parallel Factor Analysis - PARAFAC

PARAFAC[71] originates from the field of psychometrics where two independent research groups in the 1970's published the method under two different names, CANDECOMP (Canonical decomposition)[72] and PARAFAC[71]. Bro[73] has elaborated on PARAFAC and exemplified the use of PARAFAC within the area of food science.

PARAFAC analysis is performed on data of three or higher dimensions. In the following the principles of PARAFAC modeling will be illustrated by a three-way data example (fluorescence landscapes). The three-way data is organized in a tensor **X**, with the dimensions $I \times I \times K$, where $i=1,\ldots,I$ is the number of samples, $j=1,\ldots,J$ is the excitation wavelengths and $k=1,\ldots,K$ is the emission wavelengths. PARAFAC tries to minimize the sum of squares of the residuals (e_{ijk}) for the three-way array, **X** with factors, $f=1,\ldots,F$. The number of factors (F) or the rank of **X** equals the minimum number of factors that sum up to **X**[74]. The PARAFAC decomposition is represented by the triple product of vectors in Eq. 3 and a graphical view is given in Figure 19.

Eq. 3 $$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \qquad (i=1,\ldots,I;\ j=1,\ldots,J;\ k=1,\ldots,K;\ f=1,\ldots,F)$$

The PARAFAC model can be considered an extension of the bilinear PCA model, thus one or more dimensions are added. PARAFAC decomposes three-way data (**X**) into F number of tri-linear contributions. Each factor consists of a score vector (a)

and two loading vectors (b and c) and is represented by the loading matrices **A** *(I × F)*, **B** (*J × F*) and **C** *(K × F)*, respectively.

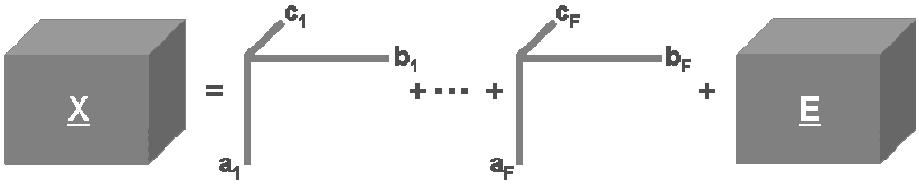


Figure 19. PARAFAC decomposition **X** of a three way into factors f=(*1,…F*) represented by the loading vectors a, b and c. **E** represents the residual.

A great advantage of PARAFAC is that it is unique in the sense that a correct estimated PARAFAC model can not be rotated without loosing fit. For this reason the estimated factors in a PARAFAC model, under the right circumstances, will be equal to the "true" underlying factors or in the case of spectral data, the "true" underlying spectra. In order to draw a direct conclusion from the extracted factors, it is *a must* that the data fulfills the requirements of tri-linearity and that the PARAFA model has been performed under the correct conditions such as selecting the right number of factors and choosing the right model parameters for the model. Estimating the PARAFAC model is usually done by alternating least squares (ALS). ALS is an iterative algorithm which will converge when the relative changes in fit is small. The drawback of ALS is that it is time consuming when a high number of components are calculated or if the data matrix is large[74]. An additional factor that adds to the time consumption is that the PARAFAC model is not nested like the PCA is[75]. This requires calculation of an individual PARAFAC model for each number of factors needed to be tested.

In order to select the correct PARAFAC model parameters some kind of validation is needed. The criteria for selecting the right number of factors can be determined by a number of parameters[73,74,75,76].

- Visual inspection in terms of evaluating scores and loadings in order to see if they make chemical sense.

- Residuals, for spectral data residuals can be interpreted by plotting the residuals after each factor to see that all spectral information is extracted.

- Split half test. The data set is split into a number of subsets and analyzed individually. Because PARAFAC is unique each PARAFAC solution should give the same results under the condition that each subset contains the same chemical information. The split half test results can be evaluated by visual inspection of the loadings.

- Core consistency (corcondia). A high core consistency (max. 100%) will indicate that the data does have the rank in question, whereas a low core consistency indicates over-fitting. The pitfall using the core consistency is that under-fitting will also give a high core consistency. Thus it is recommended to over-fit when using the core consistency so that a drop in core consistency is observed in order to exclude under-fitting.

Constraints in PARAFAC

Mathematics and reality do not always add up to the same solution. Mathematically there are many possibilities on what can be calculated. PARAFAC-ALS is not different meaning that no assumptions about the reality of the systems that are analyzed are made. Therefore it is up to the data analyst to make these assumptions and to execute them in terms of constraints by limiting the PARAFAC solution to make chemical sense. Constraints can also secure that the right solutions are reached and thereby making sure that the solutions are stable and reproducible.

The way the model is fitted will always be a trade off compared to an unconstrained model, thus a constrained model will have a lower fit, resulting in a higher sum-of-squares error e_{ijk} in Eq. 3[75]. But constraining a model can still be justified in terms of interpretability and more chemically correct solutions. Typical constraints to be applied using PARAFAC:

- Orthogonality constraints can be applied on each of the data arrays if independent factor interpretations are required. Orthogonality constraints used for spectral data are not very common as it will disturb the direct interpretation of the loadings. Orthogonality is not an issue in PCA and PLS modeling as the factors in PCA and PLS are orthogonal per definition, but in multi-way analysis orthogonal factors can not be assumed.

- Unimodality constraints. In chromatography profiles and for some spectroscopic data e.g., fluorescence spectroscopy the phenomena of unimodel profiles are present. Imposing unimodality can provide solutions where only one analyte is present in each spectral profile and thereby in each factor. The unimodality constraint does require some kind of a prior knowledge about the problem otherwise a misinterpretation can be made[77].

- Non-negativity constraints are applied when the scores and loadings of the PARAFAC model is known or assumed to have no negative values. Typical data targeted for non-negativity constraints are measurements involving physical properties such as concentrations or spectral measurements. They are by nature non-negative and in such cases applying non-negativity makes scientific sense[78].

Managing missing data in the data set can also be considered a constraint, but has not been included in this work. For further reading on how to handle missing data in PARAFAC modeling the following references are recommended[74,79].

5.2. EEM Fluorescence and PARAFAC

This section gives examples of applications where PARAFAC analysis has been used to evaluated excitation-emission fluorescence landscapes from dairy products and fish products. A review written by Bro[80] about multi-way analysis in the area of chemistry in the years from 2000-2005 states that fluorescence is the most frequently used data type for multi-way data analysis.

The concept of fluorescence landscapes was first introduced in 1975. The Video Fluorometer was capable of measuring 241 fluorescence spectra excited at 241 different wavelengths in 16.7 milliseconds and was developed by Warner and coworkers[81]. In the following years several proposals on how to analyze the fluorescence landscapes by multi-way data analysis was given e.g., a least squares based method[82], non-negative least squares[83], factor analysis[84], rank annihilation[85]. In 1981, Appellof[86] applied multi-way analysis to resolve fluorescent compounds by HPLC/EEM and principal component factor analysis[87], a method coined by themselves, which however is identical to PARAFAC.

The first to suggest PARAFAC in the form known today was Ross and coworkers[88], who used PARAFAC to resolve fluorescent spectra from plant pigments complexes. Since then PARAFAC and fluorescence have been applied in a number of areas, e.g., environmental studies[89,90], soil[91], pharmaceutical[92,93], and food. In food science, the

first PARAFAC application was a five-way data array example consisting of five constituents influencing enzymatic browning in vegetables[94]. PARAFAC and ANOVA analysis were compared and PARAFAC provided a much more interpretable solution than given by the ANOVA. The first mentioning of EEM fluorescence and PARAFAC in the food area is to my knowledge, by Nørgaard[95] who used PARAFAC to resolve fluorescence spectra of sugar samples in order to study slit width of fluorescence instruments. PARAFAC was able to differentiate the effect of both excitation and emission slit width. In 1997, Bro[73] published the PARAFAC tutorial including examples of PARAFAC and EEM. Munck and coworkers[96], laid the ground work for a more extensive study of sugar beets and sugar beet thick juice by resolving EEM landscapes using PARAFAC[97,98,99,100]. Other areas include meat, where the quality of dry cured Parma ham[101] was assessed, olive oil was studied by Guimet and coworkers[102,103,103,104] and benzoic acid studied in a number of non alcoholic beverages[105]. A review by Christensen and coworkers[106] shows that the fluorescence spectroscopy has been applied in a number of food areas with dairy products being the most dominant application. In contrast only a few studies on fish and fish products have been reported. One study determines the dioxin content in fish oil[107]. This study found that PARAFAC scores gave spectral fingerprints describing variations that could be identified and used for calibration purposes. It revealed that the higher spectral regions with peaks $\lambda_{ex}/\lambda_{em}$ maxima at 435/545 nm and 420/675 nm were of special interest. In PAPER VI brine from barrel salted herring was analyzed by EEM fluorescence spectroscopy. The area of interest turned out to be the proteins and vitamin fluorescence with $\lambda_{ex}/\lambda_{em}$ in the range 280-450/320-530 nm. Four fluorophores were identified by PARAFAC modeling and are listed in an overview of fluorophores found in fish and dairy products (PAPER VI).

In the field of dairy, milk was analyzed to study the effect of heat treatment and acidification[108]. In this study three components were identified and according to Christensen and coworkers[109] yoghurt is likewise consisting of three components. Milk and yoghurt both contain tryptophan and riboflavin but in yoghurt, the compound lumichrom was detected whereas vitamin A was identified in milk. Vitamin A was also found in processed cheese when studying the effects of induced light and temperature stress (PAPER III). In processed cheese, the compound riboflavin was not found, as it was in milk treated at different temperatures. Instead the third component in processed cheese was believed to be due to a product of Maillard reactions. Riboflavin was also identified when characterizing active photosensitizers in butter[110], and since this study focused on the higher spectral regions of the spectra the remaining 5 components in butter were not similar to the compounds mentioned in the above studies of dairy products.

The findings in PAPER III and PAPER VI indicate that fish (brine) and dairy products do share some similarities in their composition of fluorophores when studying changes in the chemical composition during storage/ripening. Table 2 gives an overview of the reported fluorophores identified in fish and dairy products using EEM fluorescence and PARAFAC.

Table 2. Overview of fluorophores identified in fish and dairy products by EEM fluorescence and PARAFAC modeling.

$\lambda ex_{max}/\lambda em_{max}$	Tryptophan	Vit. A	Vit. B_3	Maillard reaction product	Lumi-chrom	Riboflavin
Milk	292/342	323/423				460/520
Yoghurt	290/340				260/445	380,460/520
Butter						530
Proc. Cheese*	280/339 300/347	320/411		360/431		
Brine	290/241 300/350		330/396			390,440/520

*Processed cheese

5.4. Applying SLICING for Food Quality Assessment

Only a few applications performing SLICING of CPMG relaxation curves for assessing food quality have been recorded. I managed to find three dedicated food SLICING applications which include PAPER II. In addition to the three food related applications, three papers purposing or optimizing the SLICING methods including practical application exist[65,67,70].

In the paper by Pedersen and coworkers[65] introducing the SLICING approach an example for the assessment of fish was given. Two independent studies using SLICING on relaxation curves of fish have been reported. In the first study by Andersen and Rinnan[37] SLICING revealed two water populations in fresh cod fillets. The water population with the shortest relaxation time was present near the head and the water with the longest relaxation time was present near the tail. The second study studied herring when caught under different conditions e.g., season of catch, fishing grounds, fishing vessel and biological variation[111]. Two other SLICING applications outside the area of food have been published. Both studies are performed by group of Manetti. The first study tests the POWERSLICING[67] method versus the Marquardt-Levenberg algorithm on a single relaxation curve obtained by performing a controlled experiment using copper sulphate solutions[112]. POWERSLICING performed satisfactorily as the second study followed up by using POWERSLICING to study the effect of tears on contact lenses using artificial tear solution on hydrogel contact lenses[113]. This study also compares PCA versus the scores from the SLICING method and it is shown that the SLICING scores provide

better and more direct information about the dynamics of the system. The same conclusions were drawn in PAPER II which also shows that the PARAFAC scores from the SLICING approach carry descriptive information about drymatter, cultivars and storage times of potatoes as illustrated in Figure 20. The same level of information was not present in the relaxation profiles extracted from distributed or exponential fitting.

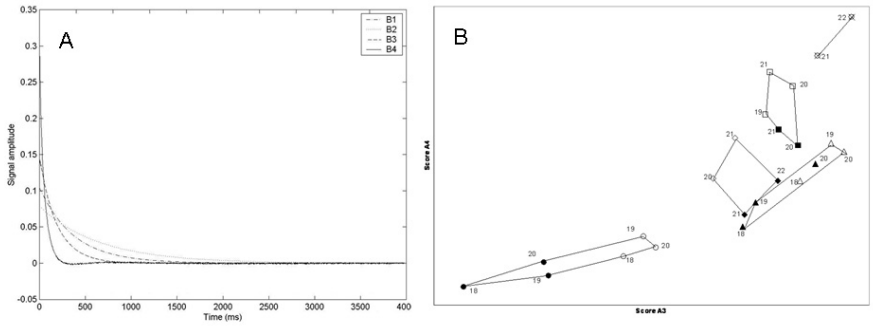


Figure 20. A. The extracted relaxation loading profiles from the SLICING of the CPMG relaxation profiles of potatoes. B. The score plot of PARAFAC scores 3 and 4. Potato varieties, Ditta (◇), Sava (Δ), Bintje - low dry matter (□), Bintje - high dry matter (⊗), and Berber (○) for storage 1999 (filled) and 2000 (open). The numbers from 18 to 22 represent the dry matter bins (PAPER II, Figure 4).

Four water populations were found to be represented in the relaxation profiles from the potatoes. In the second half of the research the CPMG relaxation profiles are correlated with six of the texture related sensory attributes comparing different data by PARAFAC scores from the SLICING approach, bi-exponential fit parameters, distribution analysis parameters and the raw CPMG relaxation profiles. It was concluded that PARAFAC scores from SLICING data did not improve the prediction of the sensory attributes as the compared methods performed equally well with the exception of the relaxation profiles from distributed fitting, which gave the poorest prediction.

6. Multi-block Analysis

In food analysis it is often the case that more than one analysis is performed at the same time on the same product. This will result in several measurements (data blocks) that need to be compared and analyzed. The reason for performing several analyses varies, but often the reason is a lack of understanding in what the measurement can provide information about. It is often the case in such situations that only one or two analyses are needed to provide sufficient information about the sample. Other situations where several analyses are performed are screening studies. Here the aim is to compare several methods in order to find the method(s) best suited. In both cases the initial step is to explore the capabilities of the different methods. To perform explorative studies chemometric methods which are designed to study several measurements (blocks) are needed. So-called multi-block methods attempt to handle data with several block contributions. Multi-block methods are designed to look for possible correlation and covariation between blocks and for unique information from the individual block.

So-called multi-block methods attempt to handle several blocks of data at once by keeping the block structure. It is shown that overall multi-block models give the same results as the regular PCA or PLS model[114,115], but multi-block analysis adds the twist of giving you the possibility to “dissect” the model into individual block contributions. This additional level of information can provide detailed information about correlations and unique block information which can ease the interpretation and help in the understanding of a product or process. The local models (block level) can be studied by their block scores and loadings and the overall model (super level) by its super scores and loadings. The impact of each block and each variable can also be studied in terms of the block loadings, and this makes it possible to identify important parameters/variables and trace them back to the raw data, if necessary.

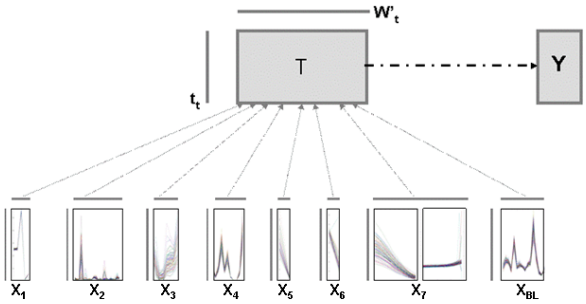


Figure 21. Example of multi-block data consisting of different chemical and physical measurements performed on the same sample (semi-hard cheese). T is the super scores and represents the concatenated block scores from each of the block contributions

Figure 21 is an illustration of multi-block data, where each type of analysis is kept in separate blocks. The upper level is referred to as the super level and describes the overall information given in the data and consists of the super scores ($\mathbf{t}_T$) and the super loadings ($\mathbf{p}_T$) and the super weights ($\mathbf{w}_T$). The **T** matrix represents the concatenated block scores from the individual block. For most multi-block methods one restriction applies and that is that one dimension (mode) of each block contribution will have to be in common (consensus between blocks).

Multi-block methods are based on the principles from PCA and PLS regression and are well suited for handling data which can be divided into conceptually meaningful blocks. Multi-block PCA or multi-block PLS regression can be used when a typical block structure of several blocks is present. This can be a situation where multiple analyses have been performed on the same product (PAPER I) or where a single analysis provides a result that can be split up into blocks, e.g., spectra[116]. The present study presents applications of multi-block PLS regression (PAPER I, PAPER V), thus multi-block PLS regression will be the main focus of this section.

The Development in Multi-block Analysis

Multi-block modeling originates from the area of path modeling[117,118,119], and throughout the years several multi-block PCA and multi-block PLS methods have been proposed. In the area of multi-block PCA, the best known multi-block PCA methods are the consensus PCA[120] (CPCA) and hierarchical PCA[121] (HPCA), but other approaches have also been suggested[122,123,124].

The pioneers in performing regression on several blocks of data simultaneously were Frank and coworkers [118,119]. But it was Wold and coworkers[120] who laid the

grounds for the multi-block methods, as they are applied in the fields of chemistry and food today, by their presentation of the consensus PCA (CPCA) and hierarchical PLS (HPLS). Shortly after the introduction of the HPLS, Wangen and Kowalski[125] proposed the multiblock PLS (MBPLS). The main difference between HPLS and the MBPLS is the level on which the PLS regression is performed. The MBPLS performs a PLS regression at the block level to obtain the block scores whereas the HPLS performs the PLS regression on the block scores at the super level obtaining the super weights and new updated super scores[114,126].

Since the first publication on the MBPLS by Wangen and Kowalski[125], Westerhuis and coworkers have given two suggestions as how to optimize the deflation[2] step instead of the way it was done in the original approach where the block scores were used to deflate **X**. The first alteration from Westerhuis and coworkers[114] showed that this approach will give misleading predictions as information in X is removed before it is used to predict Y, and instead they suggested using the super scores when deflating X. The second change given by Westerhuis and Smilde[127] was to only deflate Y using the super scores and use **X** without deflation[127]. The HPLS has also been the subject for further development. An extension to the HPLS was made by Berglund and Wold[128], who came up with the serial PLS (S-PLS). In this approach the blocks were handled as a series of blocks where Y variation is searched for in the first block and what is not explained of Y is searched for in block two etc.

Optimization studies on how to handle missing values in multi-block modeling and how to handle data structures deviating from the normal parallel block structure were performed by Muteki and coworkers[129,130].

An important issue in multi-block modeling is scaling as this it crucial for the outcome of result. When dividing data into blocks it may result in differently sized blocks, i.e. where the number of variables in each block is different. If no prior knowledge about the product or process is present it is not unusual in normal one block modeling to perform auto scaling to unit variance. In this way you do not favor any of the variables over others and spectral data treatment offsets and scatter effects can be handled to a certain extent. In multi-block analysis, performing auto scaling to unit variance without considering the data structure can give misleading results. When performing auto scaling to unit variance the variance of each block will equal the number of variables in the block. The impact of blocks

[2] Deflation of X and/or Y is performed in the PLSR algorithm after each component is calculated. Subtracting the calculated component from X gives the residual E and subtracting the component from Y gives the residual, F. When additional components are calculated the new residuals E and F will be used instead of X and Y, respectively.

with a few variables will then have less influence on a multi-block model and the opposite is the case with blocks with many variables. Individual block weighting is a solution, where it is possible to downscale large blocks and upscale small blocks if needed. Such weighting require prior knowledge about the process or product at hand in order to perform the proper scaling. On the other hand this may also force the solution in a direction where the exploratory aspect is suppressed. As of now there are no set rules on how to perform scaling and weighting of blocks in multi-block modeling and therefore weighting of blocks may be considered a drawback in situations where no prior knowledge is present about the variables. The LS-PLS is a multi-block method that can overcome this issue and was introduced by Jørgensen and coworkers[131]. The LS-PLS is based on ordinary least squares (OLS) and PLS regression and is invariant to scaling. The method is a stepwise approach where one layer of information is being extracted at a time. The first step is to subtract prior knowledge e.g., experiment design setup. Subtracting this type of information may help promote the revelation of underlying factors relevant for prediction and interpretation. Måge[132] elaborated work on the LS-PLS lead to two additions, the LS-parallel-PLS (LS-parPLS) and the LS-parPLS with common components (LS-parPLSc). The principle of the LS-parPLS is that after the initial subtraction of the design information, the unique variation from each block is found. In the LS-parPLSc approach an additional search for common information is included. Måge's dissertation work[132] includes an in-depth study comparing the predictability and interpretability of the two LS-parPLS methods, least squares PCA (LS-PCA), and a least squares MBPLS (LS-MBPLS). The LS-PCA performs a PCA on each block contribution instead of the original PLS step and in the LS-MBPLS the PLS step is replaced by MBPLS regression[127]. The issue of scaling in the LS-MBPLS is addressed by performing scaling of the blocks to equal sum of squares. The comparison is based on two case studies both including simulated data and real data sets. Both case studies include design variables and parallel blocks of spectroscopic measurements.

The overall conclusion of the study is that there is no significant difference in the predictive performance. The advantage is to be found in the interpretation, where both LS-parPLS methods were superior to the LS-PCA and LS-MBPLS. The difference between the two LS-parPLS methods is that the LS-parPLSc is recommended in situations where overlapping variation between blocks may be present. This is due to the additional search for common structures. This way a distinction between common and unique variations are made and this makes the interpretation easier. On the other hand the LS-parPLS method is best suited for data blocks without common variation by keeping it simple without complicating matters by searching for non-existing common variation. It should also be mentioned that other multi-block PLS regressions have been proposed in the recent years. Eriksson and

coworkers[133] suggested an approach which combines O2-PLS[134] with hierarchical modeling. Vivien[122] introduced the GOMCIA and GOMCIA PLS and Skov[135] a method based on variance partitioning.

The predictive properties of most multi-block PLS methods are more or less the same as the ordinary PLS regression and also the same diagnostics are used in multi-block modeling as regular PLS regression. The advantage of the multi-block regression is the additional descriptor block level. In Figure 21 an example of multi-block data is given. The additional level of information is given in the data and consists of the super scores ($\mathbf{t}_T$) and the super loadings ($\mathbf{p}_T$) and the super weights ($\mathbf{w}_T$). The super block ($\mathbf{T}$) represents the concatenated block scores from the individual block. For most multi-block methods one restriction applies and that is that one dimension (mode) of each block contribution will have to be in common (consensus between blocks). This makes it possible to go back and look at the data and identify precursors responsible for the outcome of the model.

As mentioned earlier a number of multi-block PLS algorithms exist, but until now the most familiar algorithms are the HPLS, the S-PLS and the MBPLS. The LS-parPLS method is a fairly new addition to multi-block modeling, but has been applied in the present work and will be elaborated on together with the three other multi-block algorithms.

The multi-block literature can be quite confusing to read, as it can be difficult to make the distinction between which multi-block methods are used. The term MBPLS is most often used inconsistently. The original acronym MBPLS is referring to the algorithm by Wangen and Kowalski[125], but is used for more or less all multi-block regression methods. In this particular work the acronym MBPLS is only assigned to the original multiblock PLS and its revisions.

In the following the principles of the four multi-block methods, HPLS, S-PLS, MBPLS, and LS-parPLSc algorithms are exemplified with block structures in Figure 22, Figure 23, Figure 24, Figure 25, respectively. The examples are outlined by two **X** block contributions and one **Y** block or a single **y** variable.

6.1. The Principles of the Hierarchical-PLS

The original HPLS algorithm[120] has been altered a couple of times. The first suggestion of change was given by Slama who according to Westerhuis and coworkers[114] performed normalization on the super scores ($\mathbf{t_T}$) instead of the super weights ($\mathbf{w_T}$). The same approach was used by Wold[121] with the addition of orthogonalising the super scores. The alterations are presumed to be performed in order to improve the interpretation, as normalizing the super scores may center them in the model, and orthogonalization of the scores prevents them from being correlated. There is no overall difference in the result, therefore the H-PLS presented in Westerhuis and coworkers[114] will be illustrated below.

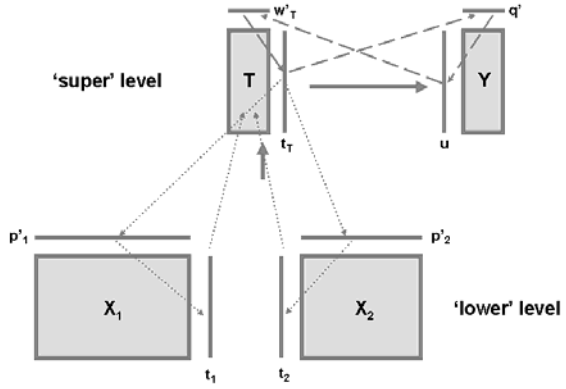


Figure 22. The principles of HPLS[114]. Lower level: (⋯▶) The super score t_T is regressed onto $\mathbf{X_{bl}}$, separately. The block loadings ($\mathbf{p_{bl}}$) and block scores ($\mathbf{t_{bl}}$) are obtained. The $\mathbf{t_{bl}}$ are collected in the super block (**T**). (– – ▶) Super level: A PLS regression is performed between **T** and **Y** to obtain the super weights $\mathbf{w_T}^T$ and a new super score $\mathbf{t_T}$. Each arrow scheme represents the regression at the lower level and the super level.

The HPLS algorithm consists of two overall steps. The first step is a CPCA performed on the block level. The second step is a PLS regression between the concatenated block scores (T) and the response variable (Y). In Figure 22 the two loops in the HPLS algorithm are illustrated by Westerhuis and coworkers[114]

CPCA cycle between $\mathbf{t_T}$ and $\mathbf{X=[X_1,\ldots,X_{BL}]}$

1. Select a super score ($\mathbf{t_T}$) (eigenvector of **X'X**, where **X** is all the blocks)
2. Regress $\mathbf{t_T}$ onto the blocks $\mathbf{X_{bl}}$ to give the block loadings, $\mathbf{P_{bl}=X_{bl}'t_T}$
3. The block scores $\mathbf{t_{bl}}$ are obtained by multiplying $\mathbf{t_{bl} = X_{bl}p_{bl}}$
4. The block scores are combined into the super block, $\mathbf{T}=[\mathbf{t}_1,\ldots,\mathbf{t}_{BL}]$

A PLS is performed between **T** and **Y** at the super level

5. The **Y** weights (**q**), $\mathbf{q}=\mathbf{Y't_T}/\mathbf{t_T't_T}$ are obtained
6. The **Y** scores (**u**), $\mathbf{u}=\mathbf{Yq}/\mathbf{q'q}$ are calculated
7. Calculated the super weights, $\mathbf{w_T}=\mathbf{T'u}/\mathbf{u'u}$
8. The super scores ($\mathbf{t_T}$), $\mathbf{t_T}=\mathbf{Tw_T}/\mathbf{w_T'w_T}$ are calculated
9. The super scores are normalized to length one, $\|\mathbf{t_T}\|=1$

Return to step 2, until convergence of $\mathbf{t_T}$

10. Deflation of $\mathbf{X_{bl}}$ using $\mathbf{t_T}$, $\mathbf{E_{bl}}=\mathbf{X_{bl}}-\mathbf{t_Tp'_{bl}}$
11. Deflation of **Y** using $\mathbf{t_T}$, $\mathbf{F}=\mathbf{Y}-\mathbf{t_Tq'_{bl}}$

If more components are needed, the calculations are repeated using $\mathbf{X}=[\mathbf{E_1},\ldots,\mathbf{E_{BL}}]$ and **Y**=**F** instead of **X** and **Y**, respectively.

6.2. The Principles of the Serial-PLS

In 1999 Berglund and Wold[128] suggested another approach which was an extension of the HPLS and referred to as serial-PLS (S-PLS). The principle of the S-PLS is to find predictive Y information in the first block. Then subtract the information from the first block from Y. A new PLS regression is performed between the second block and the residual Y. This way the second block will only explain information not given in the first block. In this approach the blocks are handled in the order they are presented, hence the Y variance not explained in the first block will be left for the second block to explain and so forth. The final model from this approach will clearly depend on the order of the blocks. It is kind of the first come first served principle, and depending on who comes first the model will turn out accordingly. Therefore the SPLS is best suited when the order of the blocks are known.

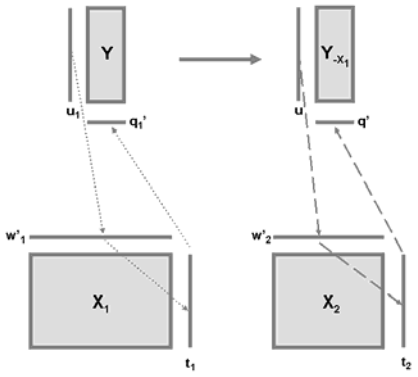


Figure 23. Illustration of the S-PLS[128].(.....▶) PLS regression between X_1 and Y (– – ▶) the residual from the first regression model (Y_{-X1}) is used to perform a new PLS with X_2.

The output of the S-PLS is a one level model where no super level exists, since it is already certain that each block contribution describe unique Y variation. The S-PLS algorithm is illustrated as a block structure in Figure 23.

The S-PLS is outlined as presented in Berglund and Wold[128]. The algorithm is given as a two block example, but if the number of blocks exceed two then $X=[X_1,\ldots,X_{BL}]$ and PLS is performed on the remaining blocks using $Y=[F_1,\ldots F_{BL}]$.

Perform PLS of Y on X_1

1. Calculate a PLS model between X_1 and Y
2. The residual F_1 is calculated, $F_1 = Y\text{-}t_1q_1'$

Perform PLS of F_1 on X_2

3. Calculate a PLS between X_2 and $Y_{\text{-}X1}$
4. The residual F_2 is calculated $F_2 = Y\text{-}t_2q_2'$

If no convergence repeat the loop

6.4. The Principles of the Multiblock PLS

The main principle in the MBPLS is to find the maximum covariance between the super block (**T**) and the response variables (**Y**) in one big loop and by focusing on explaining as much of the **Y**-variation presence in **X** as possible, where $\mathbf{X}=[\mathbf{X_1},\ldots,\mathbf{X_{BL}}]$. The relationship between the response block Y and the descriptor block T is established in the 'super' level, where the T block is a function of the original descriptor blocks ($\mathbf{X}_{bl}$) at the 'lower' level[127]. The MBPLS is outlined below and is based on the version given by Westerhuis and Smilde[127] and a block model is given in Figure 24.

PLS regression on the block level

1. Select a score **u** (a column from Y)
2. Regress **u** onto the blocks $\mathbf{X}_{bl}=[\mathbf{X_1}, \mathbf{X_2}]$ – obtaining $\mathbf{w}_1=\mathbf{X_1}'\mathbf{u}$, $\mathbf{w}_2=\mathbf{X_2}'\mathbf{u}$
3. The block weights are normalized to length one, $\|\mathbf{w_{bl}}\| = \mathbf{1}$
4. The block scores t_{bl} are obtained by multiplying,
   $\mathbf{X_{bl}w_{bl}} = \mathbf{t}_{bl}$ ($\mathbf{t_1=X_1w}_1$, $\mathbf{t_2=X_2w}_2$)
5. The blocks scores, $\mathbf{T_{bl}}=[\mathbf{t}_{1b},\ldots,\mathbf{t}_{fb}]$ are combined into the super block **T,** $\mathbf{T}=[\mathbf{T}_1\ \mathbf{T}_2]$

PLS regression on the super level

6. Regress **u** onto **T** - obtaining $\mathbf{w}_T=\mathbf{T'u}$
7. The super weights are normalized to length one, $\|\mathbf{w_T}\| = 1$
8. Calculate the super scores ($\mathbf{t}_T$), $\mathbf{t_T=Tw}_T$
9. Calculate the Y loadings (**q**), $\mathbf{q}=\mathbf{Y}^T\mathbf{t}_T / \mathbf{t}_T'\mathbf{t}_T$
10. Calculate the Y scores (**u**), $\mathbf{u} = \mathbf{Yq}/ \mathbf{q'q}$

Return to step 2, until convergence of $\mathbf{t_T}$

11. Deflation step of **Y**, $\mathbf{F= Y-t_t q'}$, where $\mathbf{t_t= Tw_t}$.

If more components need to be extracted **Y** = F and X is unchanged.

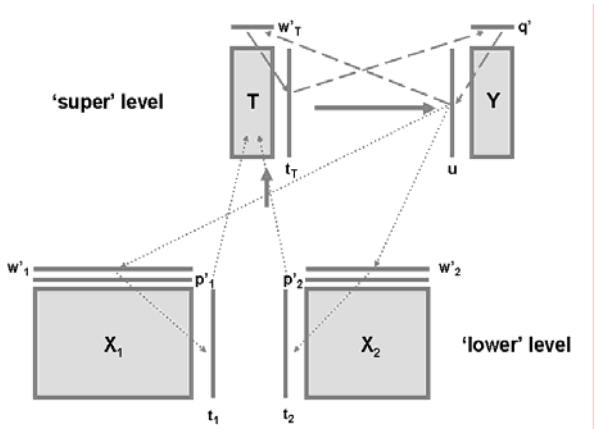


Figure 24. Illustration of the MB-PLS. First loop (·····▶) at the lower level and the second loop (– – ▶) at the super level.

6.5. The Principle of the LS-par-PLS

The principle in the LSparPLS with common loadings (LS-parPLSc) is to split up the information into 3 individual contributions: the unique information from the each individual blocks, common information shared between blocks, and the design of the experiment. Below the LSparPLSc is described as performed by Måge[132] and illustrated in Figure 25.

Extract information regarding design of the experiment

1. Fit **Y** to the design (**D**) by LS (the residual $\mathbf{Y}_{-D}$ is obtained)
2. Extract the design (**D**) from the blocks $\mathbf{X}_{bl}=[\mathbf{X}_1,\mathbf{X}_2]$, $\mathbf{X}_{bl,orth}= \mathbf{X}_{bl} \perp \mathbf{D}$.

Extract common information between blocks

3. fit $\mathbf{Y}_{-D}$ to $\mathbf{X}_{bl,orth}$ by PLS
4. find common information ($\mathbf{T}_c$) by CCA between $\mathbf{t}_{1c}$ and $\mathbf{t}_{2c}$

Extract unique information from each block

5. fit **Y** to $[\mathbf{D}\ \mathbf{T_c}]$ by LS (obtain $\mathbf{Y}_{-DTc}$)
6. fit $\mathbf{Y}_{-DTc}$ to $\mathbf{X}_{bl,orth}$ by PLS to obtain $\mathbf{T}_{bu}=[\mathbf{T}_{1u}\ \mathbf{T}_{2u}]$

Combine contributions to perform regression

7. fit **Y** to $[\mathbf{D}\ \mathbf{T}_c\ \mathbf{T}_{1u}\ \mathbf{T}_{2u}]$ by LS

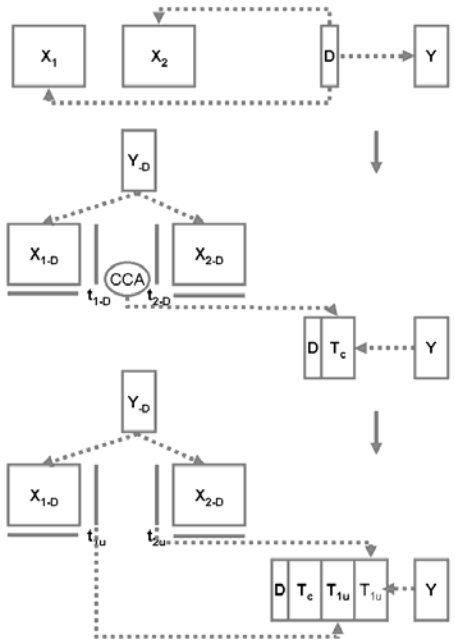


Figure 25. The principles of LS-parPLSc. First step: Extract variation due to design of the experiment. Second step: Find common structures between blocks. Third step: Extract unique block contributions.

6.6. Summary of the Four Main Multi-block Methods

Multi-block analysis makes it possible to handle several blocks of data without concatenating the blocks into one big block contribution. The multi-block approach maintains the block structure and this improved the data interpretation. Maintaining the block structure gives you the opportunity to go back and take a closer look at the individual blocks and identify the important parameters. This enhances the possibility of obtaining a better understanding of a product or process and provides means for selecting the right parameters to monitor and optimize. Four different multi-block methods have been illustrated and even though they have the same overall aim, they provide different views for data interpretation. The serial PLS is only well suited for data where a certain order is given. The serial principle of having the first block capture as much of the Y-variance as possible and then what is not explained by the first block of the Y-variance will be the input for the second block etc. This emphasizes that the order

of the blocks can not be random. Data well suited for the serial PLS will be process monitoring where the time line will be natural order of the blocks. The hierarchical PLS is a combination of the consensus PCA performed on the lower level and the PLS regression between the super block (T) and the response variable (Y) on the super level. This approach will give satisfactory predictions of Y, if the consensus PCA captures the relevant variation in **X** to predict Y.

The multiblock PLS on the other hand focuses on explaining as much of Y as possible in both the lower level and the super level. The principle of performing a PLS regression on the lower level is to let the block score capture as much Y-variance as possible. The PLS step on the super level between the super block (T) and Y provides information of which block carry the most relevant Y-variation. This approach does not take the order of the blocks into consideration and is therefore well suited for data screening or data where several analyses have been performed on the same sample. The disadvantage of the above mentioned methods is that some kind of scaling has to be performed. If no prior knowledge about the data is present it can be difficult to know how to give each block the correct weight (influence on the model).

The LS-parPLS does not experience the scaling issue and this is one of the advantages of this method. This approach is also ideal for designed data as the method can separate the Y-variance explained by the design and prevent you from making false interpretations based on design instead of relevant variation. In general, the separation of data, due to their variance contributions makes the model much easier to interpret and the LS-parPLSc (with common loadings) provides a fast view of overlapping information.

In the present work the optimized multiblock PLS version[127] (PAPER I) and the LS-PLS regression approach with the addition of searching for common structure between blocks the LS-parPLSc[132] (PAPER V) have been found to be best suited for the purpose of the two presented papers.

6.7. Multi-block Applications in Food Quality Assessment

Multi-block analysis has been applied in a number of fields where batch monitoring has been one of the dominant areas[136,137,138,139]. A more general discussion on multi-block analysis in multivariate statistical process control (MSPC) purposes is given by Kourti[140] and Kohonen[141]. Other areas where multi-block analysis has been applied is in the petro-chemistry where NIR and mid-infrared (MIR) have been evaluated using MB-PLS and S-PLS for the prediction of three quality parameters in gas and oil[142]. Real time monitoring of a petrol refining process using multi-block

PCA has also been reported[124]. Other multi-block applications worth mentioning are the areas of pharmaceuticals[143], metabonomics[144], and environmental science[145]. The first application in the food area was by Frank and Kowalski in 1984[118], who studied chemical composition of pinot noir wines related to sensory data. Since then Vivien[146] looked into the relation sensory profiles and NIR spectroscopy, when splitting the NIR spectra up in block contributions. Tenenhaus and coworkers performed two food sensory studies. The first study used sensory data from a wine tasting to demonstrate a multi-blocks PLS approach combining PLS regression with generalized canonical covariate analysis (GCCA) or generalized Procrusters analysis (GPA)[147]. The second study demonstrates the use of Path modeling and HPLS when relating sensory and chemical-physical measurements on orange juice[148]. Soybean flour[149] and peas[116] have been studied by spectroscopic methods and analyzed with multi-block analysis proving that multi-block gives improved interpretation possibilities. In the present work PAPER I and PAPER V deal with the prediction of sensory attributes related to semi-hard ripened cheese and carrots, respectively.

A couple of considerations have to be made before performing the multi-block analysis. Depending on the multi-block analysis method, decisions about data complexity (factor selection), block decomposition, and weighting will have to be addressed. There are no set of rules on how to make these decisions. You will have to find ways to perform qualified decisions. In PAPER V LS-parPLSc is used to assess a two block system where dry matter and non-volatiles (HPLC analysis) represent block I, and volatiles (GC-MS analysis) block II. In order to deal with the issue of factors selection, a test to help select the optimal factor combination for the two block contributions was performed. LS-parPLSc models were calculated for all factor combinations between 1 to 10 factors for block I and block II. In Figure 26 the RMSECV's for each block and factors 1 to 10 are plotted and result in a RMSECV landscape. The optimal factor combination is the solution with the lowest RMSECV (pointed out by the arrow).

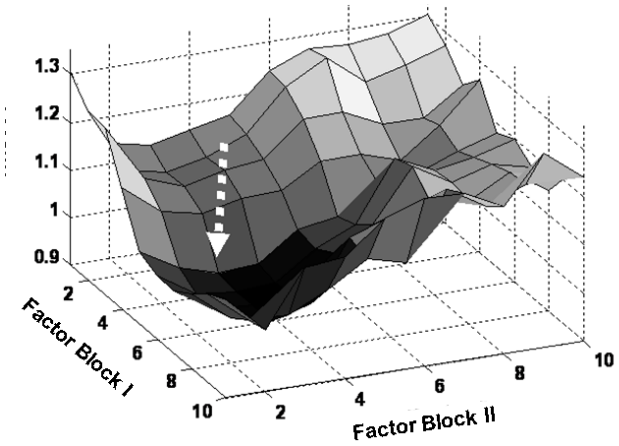


Figure 26. RMSECV versus the number of components for each of the two block contributions for the prediction of green flavour. The arrow shows the selected factor combination (block I, #3 and block II, #5, PAPER V).

PAPER I deals with data which can be divided into blocks in a number of ways. The way the experimental setup was made, two factors could determine how to decompose the data. The first thing to consider was the type of analysis, as it could be categorized as chemical and physical measurements. The second factor was the storage time, where sampling was performed three times during the storage period of 16 weeks. The study takes both measurement category and sampling time into consideration as it illustrates how a whole decomposition scheme can be tested and viewed and how it may influence the regression when information about which blocks/variables contain **Y** related information is searched for. In Figure 27 the decomposition scheme of storage data of semi-hard cheese is illustrated.

Both multi-block applications in PAPER I and PAPER V show that the multi-block regression models can be used for prediction. Furthermore they give the same results as the regular PLS regression. In both cases the prediction is not the main focus, but the main goal is to utilize the improved interpretation and visualizing possibilities for assessing each block contribution. This is a major advantage as food studies are often too complex to overview and from a PAT perspective, methods for handling large data sets are crucial in order to ensure and improve the quality of food

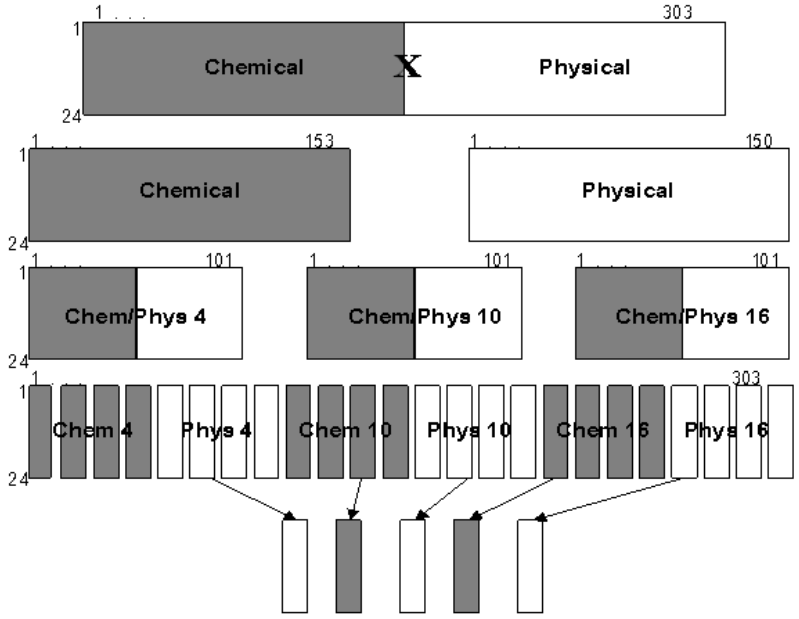


Figure 27. Decomposition scheme for chemical and physical of semi-hard smear ripened cheese sampled during a time frame of 16 weeks. The data consists of 24 samples and 303 variables. Step 1. Combined physical and chemical data in one matrix. Step 2. Decomposition into chemical and physical block contributions. Step 3. Decomposition due to storage time/sampling time. Week 4, week 10, and week 16. Step 4. Decomposition into individual physical and chemical measurements. Step 5. Block selection based on contribution and importance in the models from step 1 to 4 (Figure 3, PAPER I).

.

7. A Multi-block "Playground"

The benefits of being able to evaluate multiple blocks of data simultaneously, while keeping the structural information, are many. It is well-known that the existing multi-block methods do not give better prediction, but they are superior when it comes to providing means for data interpretation of multiple sources. The methods described previously in this section are well documented and they work satisfactorily under the right conditions such as appropriate block weighting and correct selection of the number of factors. But, as of now there are no set of rules on how to make these decisions and this may be part of the reason why multi-block analysis has not been applied more often. During my time working with multi-block analysis a couple of more or less "wild" ideas have been tested and this section is dedicated to sharing some thought on how to handle multiple blocks of data. One of the overall aims of the multi-block work was to focus on improving the graphical representation when modeling multiple data blocks. Grasping the concept of handling several blocks simultaneously can be difficult, so in order to help understand the results there is a need for tools which provide more intuitive interpretation, i.e. good visualization of results.

7.1. Matrix Correlation

The idea behind the matrix correlation was to develop a tool which could be used to compare several blocks of data and if possible provide information on how to weight the data blocks. Matrix correlation can be used to investigate the relations between the matrices. In the present work we suggest to use the RV-coefficient as an indicator of the importance of the individual blocks in relation to each other. The RV-coefficient is based on the association matrices of the block, $\mathbf{W}_i=\mathbf{X}_i\mathbf{X}_i'$ where $\mathbf{X}_i$ is column mean centered. In Eq. 4 the RV-coefficient is given[150].

Eq. 4 $$r_{RV} = \frac{trace(W_1 W_2)}{\sqrt{trace(W_1 W_1)}\sqrt{trace(W_2 W_2)}}$$

The diagonal element of the association matrix is the distance of the objects to the origin. It can be proven that two association $\mathbf{W}_i$ and $\mathbf{W}_j$ matrices are similar if and only if they can be matched by a rigid rotation. The principle behind the RV-coefficient (or normalized association index) is that it is zero if and only if $\mathbf{W}_i$ and $\mathbf{W}_j$ are found in the orthogonal subspaces, and is equal to 1 if and only if they can be matched by multiplication with a rotation matrix.

The RV-coefficient thus ranges from 0 to 1. Where 0 indicates that the matrices are not correlated and 1 indicates that it is highly correlated. Below are two examples of how the matrix correlation could be used.

Example I:

The first example is based on data from Nielsen and coworkers[151] and consists of seven blocks of NIR spectra (exposed to different kinds of spectral pre-treatment), laser-diffraction particle size curves and chemical composition: Raw NIR spectra ($\mathbf{X}_{NIR}$), SNV corrected NIR spectra ($\mathbf{X}_{SNVNIR}$), MSC NIR spectra ($\mathbf{X}_{MSCNIR}$), Second derivative NIR spectra ($\mathbf{X}_{2ndNIR}$), laser size distribution ($\mathbf{X}_{laser}$), five selected moments from the laser size distribution($\mathbf{X}_{momlaser}$), and chemical ($\mathbf{X}_{chem}$).

The results of the calculated RV-coefficient between blocks are presented in a 7x7 bar-plot (Figure 28). The RV-correlation matrix is a symmetric matrix where the diagonal will be each block contribution correlation with itself and therefore the diagonal element is equal to 1. Thus, only the part below the diagonal is of interest in the interpretation of the results.

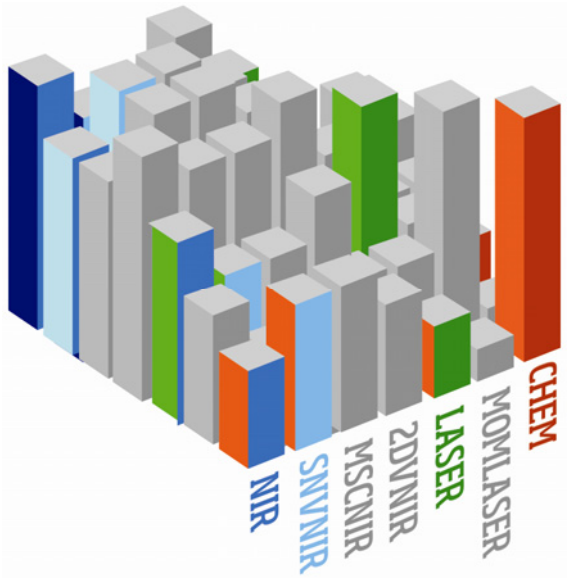


Figure 28. RV-correlation matrix (7×7) presented as a bar-plot.

The correlation matrix shows a high correlation - almost equal to 1 - between the 2nd derivative NIR spectra (2DVNIR) and the raw NIR spectra (NIR). The SNV (SNVNIR) and the MSC pre-treated NIR spectra (MSCNIR) have a similar correlation. A general observation is that the NIR spectra raw or pre-treated are highly correlated, as expected. The laser size distribution shows a reasonable correlation with the 2nd derivative NIR and the raw NIR since these spectra contain the scatter information. The chemical measurements show modest correlation with the SNV and MSC treated NIR spectra. This makes sense since the pre-treatment of the NIR spectra remove the scatter information in order to emphasize the chemical information in the NIR spectra.

This approach provides information about the correlation between the different pre-treated NIR spectra and the laser size distribution and chemical measurements. NIR measurements are typically used as an alternative, substituting measurement and it is therefore of special interest to find out if NIR correlates with the laser size distribution and chemical measurements. By performing matrix correlation it is shown that the NIR spectra are correlated with the two measurements. That is the SNV pre-treated NIR spectra and MSC pre-treated NIR spectra. They show similar behaviour where a slight difference in the correlation pattern by the raw NIR spectra is observed.

Example II:

This example uses the data from processed cheese (PAPER I). There are eight different measurements (chemical analysis, aroma analysis, casein, TPA, compression, stretch, oscillation, and peptide). All analyses have been performed three times throughout a storage period of 16 weeks (week 4, 10, and 16). All in all a total of 24 blocks can be arranged as illustrated in Figure 29. In this example the RV-coefficients are illustrated as a matrix plot where the size of the circle indicates the level of correlation between the blocks. Small dots equal no correlation and large dots corresponds a high correlation. The correlation matrix has been cut of at the diagonal as it makes interpretation easier. From the matrix plot is can be observed that the physical measurements (TPA, compression, stretch, oscillation) seem to be correlated, both individually and between sampling times. This example showed how matrix correlation can be used to find relations between sequences of blocks which are repeated but in a different time domain. The matrix correlation shows that some data blocks might contribute with more or less the same

information and this can e.g., be used as an indicator for performing data reduction.

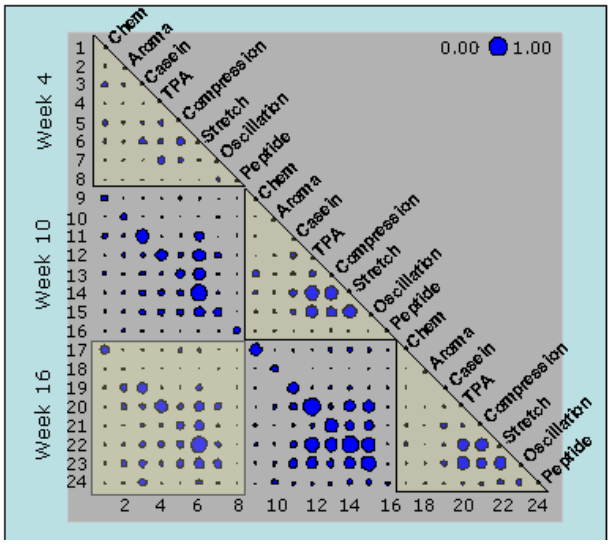


Figure 29. RV-coefficients presented in a matrix plot for 24 blocks of data. Big dots represent a high correlation whereas small dots indicate a low correlation.

7.2. Genetic Algorithm for Regression of Multiple Blocks

An approach for optimizing the predictive performance in a multivariate regression model through a Genetic Algorithm based selection mechanism was tested. Besides the purpose of improving the predictive performance, appropriate graphical representation of the model diagnostics was searched for.

The principle of a genetic algorithm in block selection is illustrated in Figure 30. In the genetic algorithm the survival of the fittest is the rule and the aim is to find the strongest combination of blocks ("genes") using crossover and mutations. This approach starts by 10 vectors with 24 block contributions which can be included (1) or excluded (0). Every block is scaled to sum-of-squares 1 which secures equal influence on the model. For each vector a PLS model using a fixed number of factors is calculated using cross validation. Minimal prediction error (RMSECV) is used as the target point. The vectors are ordered by the RMSECV vectors with the lowest RMSECV as being the best model. A new generation of ten vectors (children) is created using the five vectors (parents) giving the lowest RMSECV. One vector with the lowest RMSECV is kept (to guarantee a monotonic non-increasing

improvement over generations), four vectors are results of crossover by the five winning vectors, and five vectors by crossover and random mutations between the five parent vectors with a mutation probability of 12.5 %. This procedure is repeated 100 times and the vector with the lowest RMSECV after these runs is the "fittest parent" and the best block-set for predicting the reference value in question.

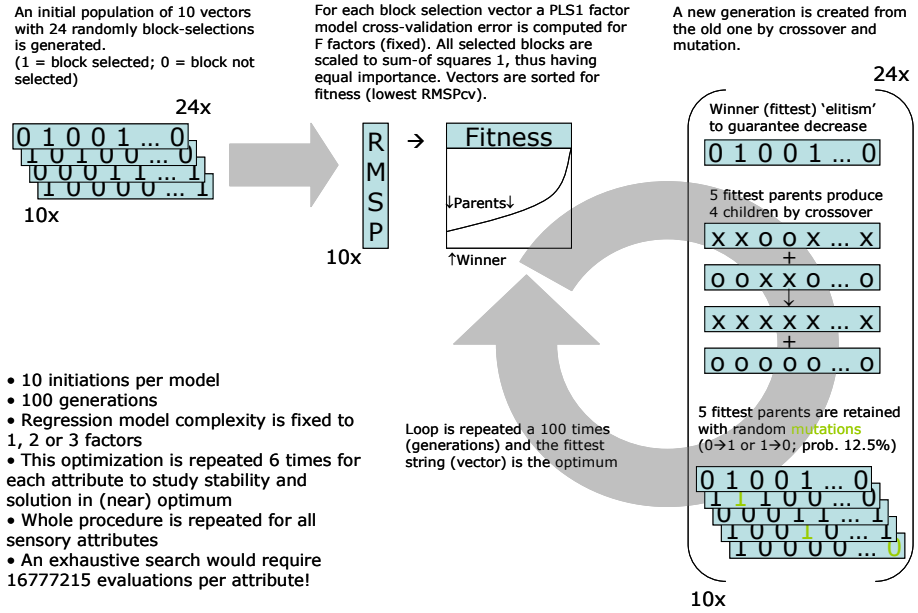


Figure 30. Regression by a genetic algorithm approach.

An example of the genetic algorithm is given using the data from the semi-hard ripened cheese (PAPER I) and in example II in the matrix correlation section. Sensory evaluation on the final cheese product was performed and in order to indentify which chemical and physical measurements (a total of 24 blocks) a 3 PLS factor genetic algorithm approach was applied. Twelve sensory attributes were evaluated and they are presented in a PCA loading plot determined on sensory attributes in Figure 31. In addition to the regular score-plot a table of 3x24 is placed below the attributes. The number of rows in the table equals the number of PLS factors and the number of columns equals the number of blocks. A filled square indicates that the block was selected by the genetic algorithm and thus has a high impact on the prediction. An example on how to interpret the sensory attribute 'cutable' will follow and a zoom on the table is also shown in Figure 31 where the right block assignments are listed in Figure 29. The horizontal bar plot shows the

RMSECV for each factor, and it can be observed that there is a decrease in the RMSECV from factor 1 to 3. The "cutable" attribute is mainly described by the physical measurements e.g., compression, texture profile analysis (TPA), and stretch. But the chemical measurements, the aroma analysis and the peptide analysis also seem to carry some information in the early weeks of storage (week 4). It is further observed that different building blocks are selected for different PLS model complexities (different number of factors). This indicates that attribute "Cutable" has only a weak link with the physical measurements. The "jumpy" behavior also gives a visible indicator that cross-validation prediction errors for small data-sets with many variables are not always reliable. For comparison, e.g., the attribute "Sticky" shows a much more consistent pattern going from 1 to 3 PLS factors.

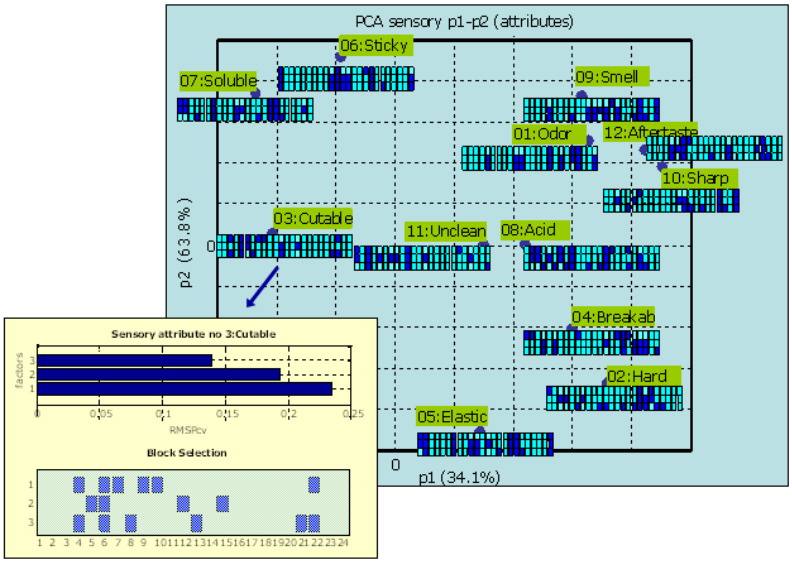


Figure 31. PCA score-plot of the sensory variables. The three horizontal bars the RMSECV from a 3 factor PLS model, where each bar represents factor 1, 2, and 3.

Another observation in Figure 31 is the connection between sensory attributes (as expressed by PCA loading values) and the block structure. E.g., the cluster "Smell", "Odor", "Aftertaste" and "Sharp" (upper-left quadrant) shows preference for building blocks with higher index numbers in regression modeling (collected in week 16).

7.3. Data Dimensionality and Building-Block Weights

This section deals with finding a semi-automated approach to determine the optimal factor and weighting scheme for the multi-block modeling. In all factor models (such as PLS regression) the bias-variance trade-off plays an important role, while for the multi-block models an additional feature besides the regularization by the number of factors has to be considered (Figure 32).

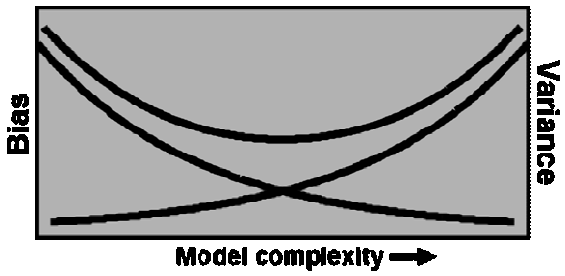


Figure 32. Bias-variance trade off.

The issue block weighting haunts the multi-block approaches. This subject is considered equally important as the model complexity since the weighting will influence the final outcome of the model. A (potentially automated) procedure of the factor and weight selection was pursued by employing the jackknife estimates of the parameter uncertainty[7] in combination with the prediction error of cross validation (RMSECV). This approach is illustrated by using the data from Nielsen and coworkers[151] explained in example I in the matrix correlation section. The matrix correlation example concluded that selecting the two blocks - SNV pre-treated NIR spectra and the raw NIR spectra - will cover the range of spectroscopic information in relation to laser size distribution and chemical measurements.

A weighting scheme in the range of 0 to 1 is selected where the weights for the two blocks sum up to 1. For the blocks, X_{NIR} the weights are Sw_1 from 1.0 to 0.0 in 15 equidistant steps and for block, X_{NIRSNV} the weights are Sw_2 0.0 to 1.0, resulting in a total of 15 regression models. The regression model is performed as an ordinary PLS model. The overall result of a PLS model equals the result of a consensus MBPLS regression[115,127] and in order to save time regular PLS regression was chosen. In Figure 33 the principles of how to pick the optimal factor and weighting combination are illustrated.

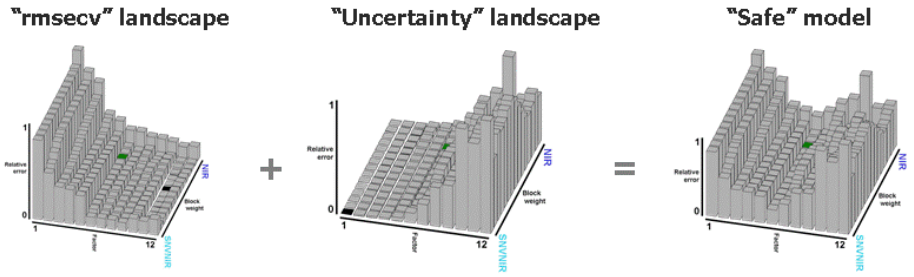


Figure 33. Prediction of the laser size distribution. RMSECV landscapes and uncertainty landscapes of 12 factors × 15 weighing schemes PLS models based on different factor and weighting combinations for the two blocks, X_{NIR} and X_{NIRSNV}. The black square indicates the minimum and the green square is the actual factor/weight combination for the "safe model".

The "RMSECV" landscape consists of factors from 1 to 12 on one axis and the block weight combinations (Sw_1+Sw_2). The RMSECV's are range scaled between 0 and 1. The front row expresses the model where the SNV pre-treated NIR spectra with the weight 1, and the last row represents the model with raw NIR spectra and the weight 1. It can be observed that the RMSECV decreases with an increasing number of factors as expected.

The uncertainty estimate is presented by the "uncertainty" landscape computed as a 2-norm of regression vector/matrix errors. The errors are range scaled from 0 to 1. The "uncertainty" landscape shows the opposite tendency as the RMSECV curves and this indicates that the variance in the prediction errors is higher with an increasing number of factors. For both the "RMSECV" and the "uncertainty" landscape it is not clear which weighting is the most appropriate and what model complexity is the best.

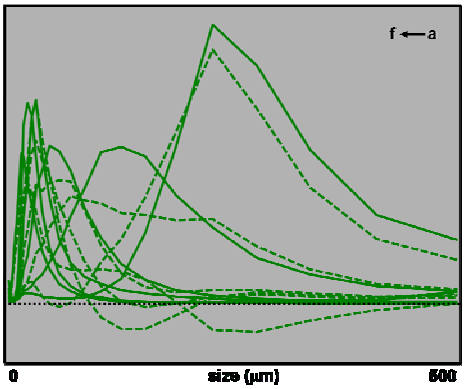


Figure 34. Actual (full line) and predicted (dotted line) and laser size distribution using 6 factor PLS regression with the block weights 0.21/0.79 (NIR/SNV-NIR).

Following the most applied rule of thumb would be to pick the model with the minimum prediction error (black square in Figure 33). Applying this rule may lead to serious overfitting and instead an approach using a combination of the two estimates is suggested. A combination of the two landscapes can provide a "safe model" given by the absolute minimum (green square in Figure 33). This model will have an acceptable predictive performance and it will take the different weighting possibilities into account. The predictive performance with 6 factors and the weighting scheme of 0.21/0.79 for NIR/SNVNIR, respectively give a prediction error of 1.61 and a correlation R of 0.94. An estimate of the predicted laser size distribution is given in Figure 34.The same approach is used to predict the dry matter content in Figure 35. The "safe model" suggests using 3 PLS factors and the weighting of the blocks is 0/1 for NIR/SNVNIR. This shows that only SNVNIR is relevant when predicting the dry matter content, which makes good sense from a spectroscopic perspective.

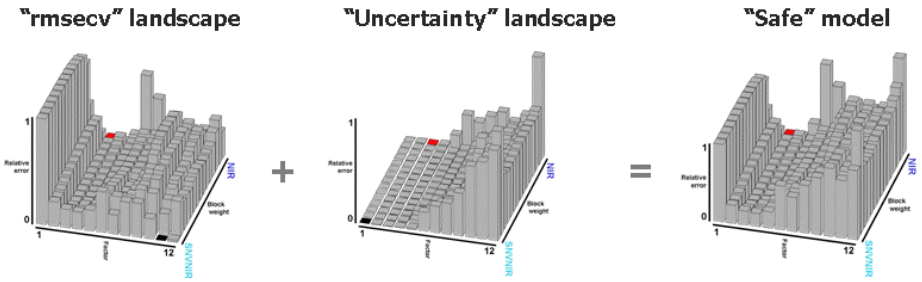


Figure 35. Prediction of the dry matter contents. RMSECV landscapes and uncertainty landscapes 12×15 PLS models based on different factor and weighting combinations for the two blocks, X_{NIR} and X_{NIRSNV}. The black square indicates the minimum and the red square is the actual factor/weight combination for the "safe model".

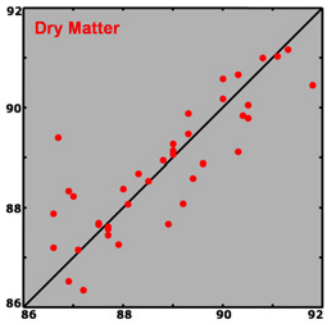


Figure 36. Predicted versus measured plot of the dry matter content.

The prediction using the suggested factor and weight combination gives a RMSECV of 0.78 and a correlation R of 0.84. A predicted versus measured plot is shown in Figure 36.

A closer look at all the predicted compounds (Table 3) shows that NIR spectroscopy can predict the chemical compounds satisfactorily and it is predominantly the SNV pre-treated NIR which explains the chemical composition. The laser size distribution on the other hand included a weight of 0.21 from the raw NIR spectra a combination which could have been overlooked when performing the ordinary one block PLS regression or performing equal weighting of the blocks.

Table 3. Overview of the parameter settings and results from prediction using NIR and SNV-NIR.

	Range (min/mean/max)	Factors	Weights (NIR/SNVNIR)	RMSECV	R
Laser	0.0/3.1/27.3	6	0.21/0.79	1.61	0.94
Dry matter	86.6/88.8/91.8	3	0.00/1.00	0.78	0.84
Ash	0.4/0.6/1.1	8	0.07/0.93	0.05	0.95
Protein	3.6/7.0/11.8	5	0.00/1.00	0.2	0.99
Starch	76.0/83.2/89.7	6	0.14/0.86	1.25	0.93
Damaged Starch	1.3/6.8/17.5	5	0.00/1.00	1.42	0.93

In this section I have presented some new ideas on how to treat multiple data blocks problems. Unfortunately we did not achieve the perfect automated weighting scheme as the presented "safe model" approach has some pitfalls. There is still a high risk of overfitting using this approach and many solutions can be classified as equal in Figure 33 and Figure 35. This highlights another aspect presented: different ways of visualizing the multi-block results. Emphasizing the graphical interpretation is something we have been very aware of when working with multi-block problems. The general aim for this multi-block work has been to make multi-block analysis more accessible[3]. Multi-block analysis is a more advanced chemometric method, but in the view of PAT it fits well as providing a way of obtaining better process understanding when several process inputs are present.

[3] During this work a Matlab multi-block toolbox was developed by Frans van den Berg. The toolbox is free and can be downloaded on the department website www.models.life.ku.dk.

8. Concluding Remarks and Perspectives

The increased amount of data collected in the industry and the increased complexity of the data are often more than what can be handled by standard multivariate methods such like PCA, PCR and PLS regression. There are still gaps between the actual number of multi-block and multi-way problems and how many are treated as such. This dissertation work has tried to solve some of these gaps by applying some of the more advanced chemometric methods included in the PAT framework in order to obtain focus on handling data structures in their natural state and make it easier to grasp for the ordinary user.

Two-way Chemometrics

NIR spectroscopy was proven to be a possible fast method for monitoring the ripening process in barrel salted herring by determining the protein content in the brine. The novelty of this work was to perform NIR analysis on the surrounding brine instead of the herring and it is the first demonstrated application of NIR spectroscopy for assessing the ripening of salted herring. There is still a great need for performing two-way analysis even though the food industry has overcome its reluctance towards applying two-way chemometric modeling such as PCA and PLS regression. NIR spectroscopy is well suited for on-line applications and is also one of the most commonly used on-line spectrophotometric methods used in PAT applications. Identifying processes where NIR spectroscopy can contribute by providing a faster feedback time during monitoring can help secure the food quality. This research also demonstrates the importance of instrument standardization on a daily basis as instrumental disturbances caused by unforeseen technical problems can occur. Due to a sudden break-down of a NIR instrument during an experiment the remaining samples had to be measured on another identical instrument from the same instrument vendor. The two instruments gave different spectral profiles and in order to cope with this problem a spectral pre-processing method subtracting the BSA standard spectra measured on the same day and additionally correction using extended inverse scatter correction (EISC) was developed. Standardization of instruments is still a common problem and ensuring that the instruments deliver an optimal measurement requires that procedures and standards are available. Measuring standard samples throughout an experiment provides information about the condition of the instrument. And if needed the standard spectra may be used to correct for variations due to drift or other day to day changes.

Pre-processing is performed in order to compensate for these deviations from linear relations in order to improve the linear relationship between the spectral signal and the analyte concentration

Multi-way Chemometrics

An important step in PAT is to find methods which quickly can provide information about food quality during processing for storage, e.g., as part of the process or in the retail step. Being able to handle more complex data structures such as multi-way data makes it possible to obtain levels of information in a much more direct manner. Handling three-way arrays in their three dimensional structures may be considered far too advanced for the industry today, and since two-way chemometrics just recently have been accepted in some line of works, it will take a little persuading and some good examples to make people appreciate the advantages of multi-way analysis.

This work presents an application of the method of SLICING on low field NMR relaxation profiles of potatoes. By constructing a three-way array from the two-way exponential relaxation profiles using the SLICING approach, PARAFAC analysis was able to separate potatoes according to the variety, This was not possible with the existing two-way data analysis approaches. This work illustrates that creating an "artificial" multi-way array can provide a higher level of information, and it has been recognized to some degree since the latest reported optimization of the SLICING has been published in 2007[70].

Fluorescence excitation-emission spectroscopy is another three-way array which is ideal for PARAFAC analysis due to its tri-linear structure and the directly translatable estimated spectral loadings. An approach to monitor stress in processed cheese is presented. Certain types of stress in cheese can lead to oxidation and a combined fluorescence analysis and PARAFAC modeling has proven to provide detailed information about compounds which proves to be affected by the storage temperature and light exposure. Changes in temperature and light exposure are two stress factors which can be expected in retail. This research shows the potential for EEM fluorescence as a fast method for monitoring stress during storage. On-line fluorescence instruments already exist and pursuing this application all the way by developing a portable EEM fluorescence could provide a method which can be used in the production, at the distributor and in retail to get a fast indication of the quality of cheese, and it would be revealed if the product has been exposed to elevated levels of stress with a resulting degrade in quality.

Another monitoring application is given where EEM fluorescence spectroscopy is used to analyze brine from barrel salted herring. Applying fluorescence and PARAFAC gives more detailed information about the chemical composition of brine. It is revealed that a number of compounds are extracted to the brine during the ripening of the herring but for some of the constituents the concentration level in the brine is not increasing during the storage. This indicates that these compounds do not affect the sensory changes which happen during the entire storage period. Only the initial extraction which occurs within the first month of storage may influence the sensory development. Details about the protein composition are obtained quickly by fluorescence spectroscopy measurements, whereas under normal circumstances such protein information requires a chemical analysis and typically these are much more time consuming.

Both fluorescence applications were related to monitoring changes in the chemical composition during storage. Both studies support that fluorescence spectroscopy has a future as a relatively fast monitoring method for assessing the chemical composition in food. The next step is to apply fluorescence as an on-line method and measure fluorescence directly on the food product. An additional step will be to bring the fluorescence spectrophotometer into the true environment such as the production site for barrel salted herring. The area of fluorescence spectroscopy can also be expanded by testing other food products and production in order to study where fluorescence spectroscopy can be applied as a fast method to characterize chemical changes in the product. The future perspectives within this line of work is to obtain better understanding of how fluorescence spectroscopy provide information about food and food production and find ways to incorporate this understanding in the design phase and the process monitoring stage.

Multi-block Chemometrics:

Expanding the direction in the second dimension and thereby adding more variables in form of several data blocks with measurements is not an unusual situation. Multi-block analysis is designed to cope with this data structure. The work presents a multi-block approach on a ripening study of semi-hard cheese. The works illustrated how the block structure of multiple blocks of data can be decomposed in different ways and how each model composition can provide relevant information. Parameters important for the prediction of the sensory quality are identified and data reduction is performed. In my opinion multi-block analysis is the ideal PAT tool. Multi-block analysis is designed to provide information about block contribution as well as the overall information about all the blocks and their interactions. In the work of applying multi-block analysis it was emphasized to find a better way of determining how to scale and weight blocks and to make multi-block analysis more applicable.

Alternatives and Ideas as to Handle Multiple Blocks of Data

A block correlation method for analyzing several block contributions was developed. The RV coefficient was used to quantify relations between the blocks. The correlation between blocks was illustrated as a 3D bar plot or a 2D matrix plot. The method is not meant as a replacement for the multi-block analysis itself but it can provide prior knowledge before performing the multi-block analysis. This knowledge can be used to reduce the data material, provide means for performing the correct block scaling and weighting.

A genetic algorithm approach for improved data interpretation and block selection was developed. The aim was to obtain better prediction of blocked data by picking out blocks with special influence in the regression. The strongest/most important blocks and combinations of blocks were selected based on how much they influence the model in a given factor. With focus on graphical interpretation a plot was presented that illustrated the relations between the response variables, and including information about the importance of each block contribution for every calculated factor.

In order to overcome the multi-block methods weighting and scaling issue, a way to choose factors and weights providing a "safe model" was developed. A semi-automated approach on how to choose block weights and factors has been developed and tested. The selection of the "safe model" is based on the criteria of the RMSECV from models testing factor and weighting combinations and the uncertainty computing the induced 2-norm of regression vector/matrix error. Together the RMSECV and the uncertainty can give an estimate of the best factor and weight combination resulting in the lowest RMSECV and with the lowest uncertainty.

LS-parPLS$_c$ - A Solution to the Weighting and Scaling Issue

The weighting and scaling problems in many multi-block methods was essentially solved by Måge and coworkers who presented a least squares method (LSparPLSc). In this work I present the first practical application of LSparPLSc, where the method is made applicable for PAT applications by modifying the original method to exclude the design module and a program in Matlab has been made. The work documents that LSparPLSc is well suited for PAT applications. It provides the possibility to compare several block contributions and help to understand the differences and similarities in block structured data. A graphical overview of the optimal factor combination between blocks is presented.

Bringing PAT to the Next Level

PAT is a tool for gaining better understanding of a given process/product. Spectroscopy and chemometrics are some of the tools available in the "PAT toolbox". Spectroscopy can provide detailed information about structures and composition of food products and some spectroscopies are well suited for on-line monitoring. The multivariate nature of the spectra makes chemometrics an obvious companion. The food industry has overcome its reluctance towards applying multivariate analysis, and two-way chemometric modeling such as PCA and PLS regression is now frequently used. But to be able to approach the complex structures of food it requires more from the chemometric analysis than "just" performing PCA and PLS analyses. Since the food industry is less restricted than the pharmaceutical industry, they can afford to experiment and explore a variety of spectroscopy and chemometrics applications.

This dissertation work has applied advanced chemometrics methods for handling data in their true structure. Multi-block analyses, PARAFAC and SLICING analyses are advanced methods which can decompose complex data structures into intuitive interpretable solutions. The methods are strong tools in the understanding of food products and processes by keeping the natural structure of the analyzed data. In order to bring PAT to the next level, methods like the ones used in the current work are needed. Advanced chemometric methods are a way to obtain better process understanding of the more complex processes. Therefore it is crucial to maintain, test, and continuously developing the methods strength and field of applications and focus in order to expand PAT. Ideas on how to optimize and improve the multi-block methods in order to make them easier to perform and how to deal with instrumental disturbances occurring during the measurements have been included in the present work.

Time will pass before the more advanced chemometrics methods such as multi-way and multi-block analyses will be standard PAT tools like the PCA and PLS regression analyses are today. However, I am convinced that the advanced methods will prove their worth and will be used as standard methods in the future. So by keeping things simple when addressing the world of multivariate data, advanced chemometrics is inevitable.

9. References

(1) Bruun, S. W.; Søndergaard, I.; Jacobsen, S. Analysis of Protein Structures and Interactions in Complex Food by Near-Infrared Spectroscopy. 1. Gluten Powder. *Journal of Agricultural and Food Chemistry*, **2007,** *55,* 7234-7243.

(2) Karoui, R.; De Baerdemaeker, J. A Review of the Analytical Methods Coupled With Chemometric Tools for the Determination of the Quality and Identity of Dairy Products. *Food Chemistry*, **2007,** *102,* 621-640.

(3) Kulmyrzaev, A. A.; Karoui, R.; De Baerdemaeker, J.; Dufour, E. Infrared and Fluorescence Spectroscopic Techniques for the Determination of Nutritional Constituents in Foods. *International Journal of Food Properties*, **2007,** *10,* 299-320.

(4) Nicolai, B. M.; Beullens, K.; Bobelyn, E.; Peirs, A.; Saeys, W.; Theron, K. I.; Lammertyn, J. Nondestructive Measurement of Fruit and Vegetable Quality by Means of NIR Spectroscopy: A Review. *Postharvest Biology and Technology*, **2007,** *46,* 99-118.

(5) Munck, L. *The Revolutionary Aspect of Chemometric Technology. The Universe and Biological Cells As Computers*; Narayana Press, Denmark: Gylling, **2005**; pp 1-352.

(6) Martens, H.; Næs, T. *Multivariate Calibration*; Wiley: New York, **1989**; pp -442.

(7) Martens, H.; Martens, M. *Multivariate Analysis of Quality. An Introduction*; John Wiley and Sons Ltd.: West Suzzex, England, **2001**; pp 1-445.

(8) Li-Chan, E. C. Y.; Ismail, A. A.; Sedman, J.; van de Voors, F. R. Vibrational Spectroscopy of Food and Food Products. In *Handbook of Vibrational Spectroscopy*; Chalmers, J. M., Griffiths, P. R., Eds.; John Wiley & Sons Ltd: Chichester, UK, **2002**.3662

(9) Norris, K. H.; Butler, W. L. Techniques for Obtaining Absorption Spectra on Intact Biological Samples. *IRE transactions on bio-medical electronics*, **1961,** *8,* 153-157.

(10) Huang, H. B.; Yu, H. Y.; Xu, H. R.; Ying, Y. B. Near Infrared Spectroscopy for on/in-Line Monitoring of Quality in Foods and Beverages: A Review. *Journal of Food Engineering*, **2008,** *87,* 303-313.

(11) Gjerde, B.; Martens, H. Predicting Carcass Composition of Rainbow Trout by Near-Infrared Reflectance Spectroscopy. *Journal of Anima Breeding and Genetics*, **1987,** *104,* 137-148.

(12) Mathias, J. A.; Williams, P. C.; Sobering, D. C. The Determination of Lipid and Protein in Fresh-Water Fish Using Near-Infrared Reflectance Spectroscopy. *Aquaculture*, **1987,** *61,* 303-311.

(13) Rasco, B. A.; Miller, C. E.; King, T. L. Utilization of Nir Spectroscopy to Estimate the Proximate Composition of Trout Muscle With Minimal Sample Pretreatment. *Journal of Agricultural and Food Chemistry*, **1991,** *39,* 67-72.

(14) Lee, M. H.; Cavinato, A. G.; Mayes, D. M.; Rasco, B. A. Noninvasive Short-Wavelength Near-Infrared Spectroscopic Method to Estimate the Crude Lipid-Content in the Muscle of Intact Rainbow-Trout. *Journal of Agricultural and Food Chemistry*, **1992,** *40,* 2176-2181.

(15) Jørgensen, B. M.; Jensen, H. S. Can Near-Infrared Spectroscopy Be Used to Measure Qualtiy Attributes in Frozen Cod. In *Seafood From Producer to Consumer, Integrated Approach to Quality*; J.B.Luten, T.Børresen, J.Oehlenschläger, Eds.; Elsevier Science B.V.: **1997**.496

(16) Uddin, M.; Ishizaki, S.; Okazaki, E.; Tanaka, M. Near-Infrared Reflectance Spectroscopy for Determining End-Point Temperature of Heated Fish and Shellfish Meats. *Journal of the Science of Food and Agriculture*, **2002,** *82,* 286-292.

(17) Solberg, C.; Saugen, E.; Swenson, L. P.; Bruun, L.; Isaksson, T. Determination of Fat in Live Farmed Atlantic Salmon Using Non-Invasive NIR Techniques. *Journal of the Science of Food and Agriculture*, **2003,** *83,* 692-696.

(18) Sollid, H.; Solberg, C. Salmon Fat-Content Estimation by Near-Infrared Transmission Spectroscopy. *Journal of Food Science*, **1992,** *57,* 792-793.

(19) Wold, J. P.; Jakobsen, T.; Krane, L. Atlantic Salmon Average Fat Content Estimated by Near-Infrared Transmittance Spectroscopy. *Journal of Food Science*, **1996,** *61,* 74-77.

(20) Wold, J. P.; Isaksson, T. Non-Destructive Determination of Fat and Moisture in Whole Atlantic Salmon by Near-Infrared Diffuse Spectroscopy. *Journal of Food Science*, **1997,** *62,* 734-736.

(21) Huang, Y.; Cavinato, A. G.; Mayes, D. M.; Bledsoe, G. E.; Rasco, B. A. Nondestructive Prediction of Moisture and Sodium Chloride in Cold Smoked Atlantic Salmon (Salmo Salar). *Journal of Food Science*, **2002,** *67,* 2543-2547.

(22) Huang, Y.; Cavinato, A. G.; Mayes, D. M.; Kangas, L. J.; Bledsoe, G. E.; Rasco, B. A. Nondestructive Determination of Moisture and Sodium Chloride in Cured Atlantic Salmon (Salmo Salar) (Teijin) Using Short-Wavelength Near-Infrared Spectroscopy (SW-NIR). *Journal of Food Science*, **2003,** *68,* 482-486.

(23) Lin, M. S.; Cavinato, A. G.; Huang, Y. Q.; Rasco, B. A. Predicting Sodium Chloride Content in Commercial King (Oncorhynchus Tshawytscha) and Chum (O-Keta) Hot Smoked Salmon Fillet Portions by Short-Wavelength Near-Infrared (SW-NIR) Spectroscopy. *Food Research International*, **2003,** *36,* 761-766.

(24) Nortvedt, R.; Torrissen, O. J.; Tuene, S. Application of Near-Infrared Transmittance Spectroscopy in the Determination of Fat, Protein and Dry Matter in Atlantic Halibut Fillet. *Chemometrics and Intelligent Laboratory Systems*, **1998,** *42,* 199-207.

(25) Shimamoto, J.; Hiratsuka, S.; Hasegawa, K.; Sato, M.; Kawano, S. Rapid Non-Destructive Determination of Fat Content in Frozen Skipjack Using a Portable Near Infrared Spectrophotometer. *Fisheries Science*, **2003,** *69,* 856-860.

(26) Shimamoto, J.; Hasegawa, K.; Hattori, S.; Hattori, Y.; Mizuno, T. Non-Destructive Determination of the Fat Content in Glazed Bigeye Tuna by Portable Near Infrared Spectrophotometer. *Fisheries Science*, **2003,** *69,* 1247-1256.

(27) Shimamoto, J.; Hasegawa, K.; Sato, M.; Kawano, S. Non-Destructive Determination of Fat Content in Frozen and Thawed Mackerel by Near Infrared Spectroscopy. *Fisheries Science*, **2004,** *70,* 345-347.

(28) Uddin, M.; Okazaki, E.; Ahmad, M. U.; Fukuda, Y.; Tanaka, M. NIR Spectroscopy: A Non-Destructive Fast Technique to Verify Heat Treatment of Fish-Meat Gel. *Food Control*, **2006,** *17,* 660-664.

(29) Uddin, M.; Okazaki, E.; Fukushima, H.; Turza, S.; Yumiko, Y.; Fukuda, Y. Nondestructive Determination of Water and Protein in Surimi by Near-Infrared Spectroscopy. *Food Chemistry*, **2006,** *96,* 491-495.

(30) Vogt, A.; Gormley, T. R.; Downey, G.; Somers, J. A Comparison of Selected Rapid Methods for Fat Measurement in Fresh Herring (Clupea Harengus). *Journal of Food Composition and Analysis*, **2002,** *15,* 205-215.

(31) Nielsen, D.; Hyldig, G.; Nielsen, J.; Nielsen, H. H. Lipid Content in Herring (Clupea Harengus L.) - Influence of Biological Factors and Comparison of Different Methods of Analyses: Solvent Extraction, Fatmeter, NIR and NMR. *Lwt-Food Science and Technology*, **2005,** *38,* 537-548.

(32) Isaksson, T.; Swensen, L. P.; Taylor, R. G.; Fjaera, S. O.; Skjervold, P. O. Non-Destructive Texture Analysis of Farmed Atlantic Salmon Using Visual/Near-Infrared Reflectance Spectroscopy. *Journal of the Science of Food and Agriculture*, **2002,** *82,* 53-60.

(33) Nilsen, H.; Esaiassen, M.; Heia, K.; Sigernes, F. Visible/Near-Infrared Spectroscopy: A New Tool for the Evaluation of Fish Freshness? *Journal of Food Science*, **2002,** *67,* 1821-1826.

(34) Ritthiruangdej, P.; Suwonsichon, T. Relationships Between NIR Spectra and Sensory Attributes of Thai Commercial Fish Sauces. *Analytical Sciences*, **2007,** *23,* 809-814.

(35) Nilsen, H.; Esaiassen, M. Predicting Sensory Score of Cod (Gadus Morhua) From Visible Spectroscopy. *Lwt-Food Science and Technology*, **2005,** *38,* 95-99.

(36) Warm, K.; Martens, H.; Nielsen, J.; Martens, M. Sensory Quality Criteria for Five Fish Species Predicted From Near-Infrared (NIR) Reflectance Measurement. *Journal of Food Quality*, **2001,** *24,* 389-403.

(37) Andersen, C. M.; Rinnan, A. Distribution of Water in Fresh Cod. *Lebensmittel-Wissenschaft Und-Technologie-Food Science and Technology*, **2002,** *35,* 687-696.

(38) Andersen, E.; Andersen, M. L.; Baron, C. P. Characterization of Oxidative Changes in Salted Herring (Clupea Harengus) During Ripening. *Journal of Agricultural and Food Chemistry*, **2007,** *55,* 9545-9553.

(39) Isaksson, T.; Togersen, G.; Iversen, A.; Hildrum, K. I. Nondestructive Determination of Fat, Moisture and Protein in Salmon Fillets by Use of Near-Infrared Diffuse Spectroscopy. *Journal of the Science of Food and Agriculture*, **1995,** *69,* 95-100.

(40) Cozzolino, D.; Chree, A.; Murray, I.; Scaife, J. R. The Assessment of the Chemical Composition of Fishmeal by Near Infrared Reflectance Spectroscopy. *Aquaculture Nutrition*, **2002,** *8,* 149-155.

(41) Khodabux, K.; Sophia, M.; L'Omelette, S.; Jhaumeer-Laulloo, S.; Ramasami, P.; Rondeau, P. Chemical and Near-Infrared Determination of Moisture, Fat and Protein in Tuna Fishes. *Food Chemistry*, **2007,** *102,* 669-675.

(42) Xiccato, G.; Trocino, A.; Tulli, F.; Tibaldi, E. Prediction of Chemical Composition and Origin Identification of European Sea Bass (Dicentrarchus Labrax L.) by Near Infrared Reflectance Spectroscopy (NIRS). *Food Chemistry*, **2004,** *86,* 275-281.

(43) Rinnan, A.; Nørgaard, L.; van den Berg, F.; Thygesen, J.; Bro, R.; Engelsen, S. B. *Infrared Spectroscopy for Food Quality Analysis and Control*; **Elsevier** : **2008**.

(44) Martens, H.; Jensen, S. A.; Geladi, P. Multivariate linearity transformations for near infrared reflectance spectroscopy. O.H.J.Christie. Proceedings of the Nordic Symposium on Applied Statistics , 205-234. 1983. Stavanger, Norway, Stokkland Forlag.

(45) Martens, H.; Stark, E. Extended Multiplicative Signal Correction and Spectral Interference Subtraction: New Preprocessing Methods for Near Infrared Spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, **1991,** *9,* 625-635.

(46) Pedersen, D. K.; Martens, H.; Nielsen, J. P.; Engelsen, S. B. Near-Infrared Absorption and Scattering Separated by Extended Inverted Signal Correction (EISC): Analysis of Near-Infrared Transmittance Spectra of Single Wheat Seeds. *Applied Spectroscopy*, **2002,** *56,* 1206-1214.

(47) Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy*, **1989,** *43,* 772-777.

(48) Savitzky, A.; Golay, M. J. E. Smoothing + Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, **1964,** *36,* 1627-&.

(49) Lakowicz, J. R. *Principles of Fluorescence Spectroscopy*; Kluwer Academic/Plenum Publishers: New York, **1999**; pp 1-698.

(50) Schulman, S. G. Luminescence Spectroscopy: An Overview. In *Molecular Luminescence Spectroscopy*; Sculman, S. G., Ed.; John Wiley & Sons: **1985**.29

(51) Wolfbeis, O. S. Fluorescence of Organic Natural Products. In *Molecular Luminescence Spectroscopy*; Schulman, S. G., Ed.; John Wiley & Sons: **1985**.370

(52) Skoog, D. A.; Leary, J. J. Molecular Fluorescence, Phosphorescence and Chemiluminescence Spectroscopy. In *Principles of Instrumental Analysis*; Skoog, D. A., Leary, J. J., Eds.; Harcourt Brace College Publishers: **1992**; Chapter 9.195

(53) Jiji, R. D.; Booksh, K. S. Mitigation of Rayleigh and Raman Spectral Interferences in Multiway Calibration of Excitation-Emission Matrix Fluorescence Spectra. *Analytical Chemistry*, **2000,** *72,* 718-725.

(54) Thygesen, L. G.; Rinnan, Å.; Barsberg, S.; Møller, J. K. S. Stabilizing the PARAFAC Decomposition of Fluorescence Spectra by Insertion of Zeros Outside the Data Area. *Chemometrics and Intelligent Laboratory Systems*, **2004,** *71,* 97-106.

(55) Purcell, E. M.; Torrey, H.; Pound, R. V. Resonance Absorption by Nuclear Magnetic Moments in a Solid. *Physical Review*, **1946,** *69,* 37-38.

(56) Bloch, F.; Hansen, W.; Packard, M. E. Nuclear Induction. *Physical Review*, **1946,** *70,* 474.

(57) Freeman, R. A Short History of NMR. *Chemistry of Heterocyclic Compounds*, **1995,** *31,* 1004-1005.

(58) Alberti, E.; Belton, P. S.; Gil, A. M. Applications of NMR to Food Science. *Annual Reports on NMR Spectroscopy*, **2002,** *47,* 109-148.

(59) Harris, R. K. *Nuclear Magnetic Resonance Spectroscopy - A Physiochemical View*; Longman Scientific & Technical: Essex, England, **1989**; pp -260.

(60) Eads, T. M.; Davis, E. A. Nuclear Magnetic Resoonance and Electron Spin Resonance. In *Introduction to the Chimical Analysis of Foods*; S.S.Nielsen, Ed.; Chapman & Hall: New York, **1994**; Chapter 27.399

(61) Abragam, A. *Principles of Nuclear Magnetism*; Oxford University Press Inc.: New York, **1983**; pp 1-599.

(62) Carr, H. Y.; Purcell, E. M. Effects of Diffusion on Free Precession in Nuclear Magnetic Resonance Experiments. *Physical Review*, **1954,** *94,* 630-638.

(63) Meiboom, S.; Gill, D. Modified Spin-Echo Method for Measuring Nuclear Relaxation Times. *The review of scientific instruments*, **1958,** *29,* 688-691.

(64) Belton, P. S.; Ratcliffe, R. G. NMR and Compartmentation in Biological Tissue. *Progress in NMR Spectroscopy*, **1985,** *17,* 241-279.

(65) Pedersen, H. T.; Bro, R.; Engelsen, S. B. Towards Rapid and Unique Curve Resolution of Low-Field NMR Relaxation Data: Trilinear SLICING Versus Two-Dimensional Curve Fitting. *Journal of Magnetic Resonance*, **2002,** *157,* 141-155.

(66) Windig, W.; Antalek, B. Direct Exponential Curve Resolution Algorithm (DECRA): A Novel Application of the Generalized Rank Annihilation Method for a Single Spectral Mixture Data Set With Exponentially Decaying Contribution Profiles. *Chemometrics and Intelligent Laboratory Systems*, **1997,** *37,* 241-254.

(67) Engelsen, S. B.; Bro, R. PowerSlicing. *Journal of Magnetic Resonance*, **2003,** *163,* 192-197.

(68) Sanchez, E.; Kowalski, B. R. Generalized Rank Annihilation Factor-Analysis. *Analytical Chemistry*, **1986,** *58,* 496-499.

(69) Sanchez, E.; Kowalski, B. R. Tensorial Resolution: A Direct Trilinear Decomposition. *Journal of Chemometrics*, **1990,** *4,* 29-45.

(70) Andrade, L.; Micklander, E.; Farhat, I.; Bro, R.; Engelsen, S. B. DOUBLESLICING: A Non-Iterative Single Profile Multi-Exponential Curve Resolution Procedure - Application to Time-Domain NMR Transverse Relaxation Data. *Journal of Magnetic Resonance*, **2007,** *189,* 286-292.

(71) Harshman, R. A. Foundation of the PARAFAC Procedure: Model and Conditions for an "Explanatory" Multi-Mode Factor Analysis. *UCLA Working Papers in phonetics*, **1970,** *16,* 1.

(72) Carroll, F. D.; Chang, J.-J. Analysis of Individual Differences in Multidimensional Scaling Via N-WAY Generalization of "Eckart Young" Decomposition. *Psychometrika*, **1970,** *35,* 283-319.

(73) Bro, R. PARAFAC. Tutorial and Applications. *Chemometrics and Intelligent Laboratory Systems*, **1997,** *38,* 149-171.

(74) Smilde, A. K.; Bro, R.; Geladi, P. *Multi-Way Analysis With Applications in the Chemical Sciences*; John Wiley & Sons Ltd.: West Sussex, England, **2004**; pp 1-381.

(75) Bro, R. Multiway Calibration. Multilinear PLS. *Journal of Chemometrics*, **1996,** *10,* 47-61.

(76) Andersen, C. M.; Bro, R. Practical Aspects of PARAFAC Modeling of Fluorescence Excitation-Emission Data. *Journal of Chemometrics*, **2003,** *17,* 200-215.

(77) Bro, R.; Sidiropoulos, N. D. Least Squares Algorithms Under Unimodality and Non-Negativity Constraints. *Journal of Chemometrics*, **1998,** *12,* 223-247.

(78) Bro, R.; DeJong, S. A Fast Non-Negativity-Constrained Least Squares Algorithm. *Journal of Chemometrics*, **1997,** *11,* 393-401.

(79) Tomasi, G.; Bro, R. PARAFAC and Missing Values. *Chemometrics and Intelligent Laboratory Systems*, **2005,** *75,* 163-180.

(80) Bro, R. Review on Multiway Analysis in Chemistry - 2000-2005. *Critical Reviews in Analytical Chemistry*, **2006,** *36,* 279-293.

(81) Warner, I. M.; Callis, J. B.; Davidson, E. R.; Gouterman, M.; Christian, G. D. Fluorescence Analysis - New Approach. *Analytical Letters*, **1975,** *8,* 665-681.

(82) Warner, I. M.; Callis, J. B.; Davidson, E. R.; Christian, G. D. New Data Reduction Strategies for Multicomponent Fluorescence Analyses. *Abstracts of Papers of the American Chemical Society*, **1976,** *172,* 26.

(83) Warner, I. M.; Davidson, E. R.; Christian, G. D. Quantitative-Analyses of Multicomponent Fluorescence Data by Methods of Least-Squares and Nonnegative Least Sum of Errors. *Analytical Chemistry*, **1977,** *49,* 2155-2159.

(84) Malinowski, E. R.; Mccue, M. Qualitative and Quantitative-Determination of Suspected Components in Mixtures by Target Transformation Factor-Analysis of Their Mass-Spectra. *Analytical Chemistry*, **1977,** *49,* 284-287.

(85) Ho, C. N.; Christian, G. D.; Davidson, E. R. Application of Method of Rank Annihilation to Quantitative-Analyses of Multicomponent Fluorescence Data From Video Fluorometer. *Analytical Chemistry*, **1978,** *50,* 1108-1113.

(86) Appellof, C. J.; Davidson, E. R. Strategies for Analyzing Data From Video Fluorometric Monitoring of Liquid-Chromatographic Effluents. *Analytical Chemistry*, **1981,** *53,* 2053-2056.

(87) Russell, M. D.; Gouterman, M. Excitation-Emission-Lifetime Analysis of Multicomponent Systems .1. Principal Component Factor-Analysis. *Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy*, **1988,** *44,* 857-861.

(88) Ross, R. T.; Arcelay, A. R.; Collins, J. M.; Davis, C. M.; Desai, T. S.; Marchiarullo, M. A.; Holmer, B. K. Resolution of the Spectra of Photosynthetic Pigments by Means of Factor-Analysis. *Biophysical Journal*, **1985,** *47,* A421.

(89) Christensen, J. H.; Hansen, A. B.; Mortensen, J.; Andersen, O. Characterization and Matching of Oil Samples Using Fluorescence Spectroscopy and Parallel Factor Analysis. *Analytical Chemistry*, **2005,** *77,* 2210-2217.

(90) Jiji, R. D.; Andersson, G. G.; Booksh, K. S. Application of PARAFAC for Calibration With Excitation-Emission Matrix Fluorescence Spectra of Three Classes of Environmental Pollutants. *Journal of Chemometrics*, **2000,** *14,* 171-185.

(91) Banaitis, M. R.; Waldrip-Dail, H.; Diehl, M. S.; Holmes, B. C.; Hunt, J. F.; Lynch, R. P.; Ohno, T. Investigating Sorption-Driven Dissolved Organic Matter Fractionation by Multidimensional Fluorescence Spectroscopy and PARAFAC. *Journal of Colloid and Interface Science*, **2006,** *304,* 271-276.

(92) Martins, J. A.; Sena, M. M.; Poppi, R. J.; Pessine, F. B. T. Fluorescence Piroxicam Study in the Presence of Cyclodextrins by Using the PARAFAC Method. *Applied Spectroscopy*, **1999,** *53,* 510-522.

(93) Xie, H. P.; Jiang, J. H.; Chu, X.; Cui, H.; Wu, H. L.; Shen, G. L.; Yu, R. Q. Competitive Interaction of the Antitumor Drug Daunorubicin and the Fluorescence Probe Ethidium Bromide With DNA As Studied by Resolving Trilinear Fluorescence Data: the Use of PARAFAC and Its Modification. *Analytical and Bioanalytical Chemistry*, **2002,** *373,* 159-162.

(94) Bro, R.; Heimdal, H. Enzymatic Browning of Vegetables. Calibration and Analysis of Variance by Multiway Methods. *Chemometrics and Intelligent Laboratory Systems*, **1996,** *34,* 85-102.

(95) Nørgaard, L. Spectral Resolution and Prediction of Slit Widths in Fluorescence Spectroscopy by Two- and Three-Way Methods. *Journal of Chemometrics*, **1996,** *10,* 615-630.

(96) Munck, L.; Nørgaard, L.; Engelsen, S. B.; Bro, R.; Andersson, C. A. Chemometrics in Food Science - a Demonstration of the Feasibility of a Highly Exploratory, Inductive Evaluation Strategy of Fundamental Scientific Significance. *Chemometrics and Intelligent Laboratory Systems*, **1998,** *44,* 31-60.

(97) Bro, R. Exploratory Study of Sugar Production Using Fluorescence Spectroscopy and Multi-Way Analysis. *Chemometrics and Intelligent Laboratory Systems*, **1999,** *46,* 133-147.

(98) Baunsgaard, D.; Andersson, C. A.; Arndal, A.; Munck, L. Multi-Way Chemometrics for Mathematical Separation of Fluorescent Colorants and Colour Precursors From Spectrofluorimetry of Beet Sugar and Beet Sugar Thick Juice As Validated by HPLC Analysis. *Food Chemistry*, **2000,** *70,* 113-121.

(99) Baunsgaard, D.; Nørgaard, L.; Godshall, M. A. Fluorescence of Raw Cane Sugars Evaluated by Chemometrics. *Journal of Agricultural and Food Chemistry*, **2000,** *48,* 4955-4962.

(100) Baunsgaard, D.; Munck, L.; Nørgaard, L. Analysis of the Effect of Crystal Size and Color Distribution on Fluorescence Measurements of Solid Sugar Using Chemometrics. *Applied Spectroscopy*, **2000,** *54,* 1684-1689.

(101) Møller, J. K. S.; Parolari, G.; Gabba, L.; Christensen, J.; Skibsted, L. H. Monitoring Chemical Changes of Dry-Cured Parma Ham During Processing by Surface Autofluorescence Spectroscopy. *Journal of Agricultural and Food Chemistry*, **2003,** *51,* 1224-1230.

(102) Guimet, F.; Ferre, J.; Boque, R.; Rius, F. X. Application of Unfold Principal Component Analysis and Parallel Factor Analysis to the Exploratory Analysis of Olive Oils by Means of Excitation-Emission Matrix Fluorescence Spectroscopy. *Analytica Chimica Acta*, **2004,** *515,* 75-85.

(103) Guimet, F.; Ferre, J.; Boque, R.; Vidal, M.; Garcia, J. Excitation-Emission Fluorescence Spectroscopy Combined With Three-Way Methods of Analysis As a Complementary Technique for Olive Oil Characterization. *Journal of Agricultural and Food Chemistry*, **2005,** *53,* 9319-9328.

(104) Guimet, F.; Ferre, J.; Boque, R. Rapid Detection of Olive-Pomace Oil Adulteration in Extra Virgin Olive Oils From the Protected Denomination of

Origin "Siurana" Using Excitation-Emission Fluorescence Spectroscopy and Three-Way Methods of Analysis. *Analytica Chimica Acta*, **2005,** *544,* 143-152.

(105) Lozano, V. A.; Ibanez, G. A.; Olivieri, A. C. Three-Way Partial Least-Squares/Residual Bilinearization Study of Second-Order Lanthanide-Sensitized Luminescence Excitation-Time Decay Data - Analysis of Benzoic Acid in Beverage Samples. *Analytica Chimica Acta*, **2008,** *610,* 186-195.

(106) Christensen, J.; Nørgaard, L.; Bro, R.; Engelsen, S. B. Multivariate Autofluorescence of Intact Food Systems. *Chemical Reviews*, **2006,** *106,* 1979-1994.

(107) Pedersen, D. K.; Munck, L.; Engelsen, S. B. Screening for Dioxin Contamination in Fish Oil by PARAFAC and N-PLSR Analysis of Fluorescence Landscapes. *Journal of Chemometrics*, **2002,** *16,* 451-460.

(108) Boubellouta, T.; Dufour, E. Effects of Mild Heating and Acidification on the Molecular Structure of Milk Components As Investigated by Synchronous Front-Face Fluorescence Spectroscopy Coupled With Parallel Factor Analysis. *Applied Spectroscopy*, **2008,** *62,* 490-496.

(109) Christensen, J.; Becker, E. M.; Frederiksen, C. S. Fluorescence Spectroscopy and PARAFAC in the Analysis of Yogurt. *Chemometrics and Intelligent Laboratory Systems*, **2005,** *75,* 201-208.

(110) Wold, J. P.; Bro, R.; Veberg, A.; Lundby, F.; Nilsen, A. N.; Moan, J. Active Photosensitizers in Butter Detected by Fluorescence Spectroscopy and Multivariate Curve Resolution. *Journal of Agricultural and Food Chemistry*, **2006,** *54,* 10197-10204.

(111) Jensen, K. N.; Jorgensen, B. M.; Nielsen, H. H.; Nielsen, J. Water Distribution and Mobility in Herring Muscle in Relation to Lipid Content, Season, Fishing Ground and Biological Parameters. *Journal of the Science of Food and Agriculture*, **2005,** *85,* 1259-1267.

(112) Manetti, C.; Castro, C.; Zbilut, J. P. Application of Trilinear SLICING to Analyse a Single Relaxation Curve. *Journal of Magnetic Resonance*, **2004,** *168,* 273-277.

(113) Manetti, C.; Casciani, L.; Castro, C. LF-NMR and Multivariate Data Analysis: Compression of Data to Classify Hydrogel Contact Lenses. *Journal of Biomaterials Science-Polymer Edition*, **2005,** *16,* 421-434.

(114) Westerhuis, J. A.; Kourti, T.; MacGregor, J. F. Analysis of Multiblock and Hierarchical PCA and PLS Models. *Journal of Chemometrics*, **1998,** *12,* 301-321.

(115) Qin, S. J.; Valle, S.; Piovoso, M. J. On Unifying Multiblock Analysis With Application to Decentralized Process Monitoring. *Journal of Chemometrics*, **2001,** *15,* 715-742.

(116) Vigneau, E.; Sahmer, K.; Qannari, E. M.; Bertrand, D. Clustering of Variables to Analyze Spectral Data. *Journal of Chemometrics*, **2005,** *19,* 122-128.

(117) Gerlach, R. W.; Kowalski, B. R.; Wold, H. O. A. Partial Least-Squares Path Modeling With Latent-Variables. *Analytica Chimica Acta-Computer Techniques and Optimization*, **1979,** *3,* 417-421.

(118) Frank, I. E.; Kowalski, B. R. Prediction of Wine Quality and Geographic Origin From Chemical Measurements by Partial Least-Squares Regression Modeling. *Analytica Chimica Acta*, **1984,** *162,* 241-251.

(119) Frank, I. E.; Feikema, J.; Constantine, N.; Kowalski, B. R. Prediction of Product Quality From Spectral Data Using the Partial Least-Squares Method. *Journal of Chemical Information and Computer Sciences*, **1984,** *24,* 20-24.

(120) Wold, S.; Hellberg, S.; Lundstedt, T.; Wold, H. PLS modeling with latent variables in two or more dimensions. Symposium, PLS Model Building: Theory and Application, Frankfurt Am Main, Sept. 23-25. 1987.

(121) Wold, S.; Kettaneh, N.; Tjessem, K. Hierarchical Multiblock PLS and PC Models for Easier Model Interpretation and As an Alternative to Variable Selection. *Journal of Chemometrics*, **1996,** *10,* 463-482.

(122) Vivien, M.; Sabatier, R. Generalized Orthogonal Multiple Co-Inertia Analysis(-PLS): New Multiblock Component and Regression Methods. *Journal of Chemometrics*, **2003,** *17,* 287-301.

(123) Rannar, S.; MacGregor, J. F.; Wold, S. Adaptive Batch Monitoring Using Hierarchical PCA. *Chemometrics and Intelligent Laboratory Systems*, **1998,** *41,* 73-81.

(124) AlGhazzawi, A.; Lennox, B. Monitoring a Complex Refining Process Using Multivariate Statistics. *Control Engineering Practice*, **2008,** *16,* 294-307.

(125) Wangen, L. E.; Kowalski, B. R. A Multiblock Partial Least Squares Algorithm for Investigating Complex Chemical Systems. *Journal of Chemometrics*, **1988,** *3,* 3-20.

(126) Westerhuis, J. A.; Coenegracht, P. M. J. Multivariate Modelling of the Pharmaceutical Two-Step Process of Wet Granulation and Tableting With Multiblock Partial Least Squares. *Journal of Chemometrics*, **1997,** *11,* 379-392.

(127) Westerhuis, J. A.; Smilde, A. K. Deflation in Multiblock PLS. *Journal of Chemometrics*, **2001,** *15,* 485-493.

(128) Berglund, A.; Wold, S. A Serial Extension of Multiblock PLS. *Journal of Chemometrics*, **1999,** *13,* 461-471.

(129) Muteki, K.; MacGregor, J. F.; Ueda, T. Estimation of Missing Data Using Latent Variable Methods With Auxiliary Information. *Chemometrics and Intelligent Laboratory Systems*, **2005,** *78,* 41-50.

(130) Muteki, K.; MacGregor, J. F. Multi-Block PLS Modeling for L-Shape Data Structures With Applications to Mixture Modeling. *Chemometrics and Intelligent Laboratory Systems*, **2007,** *85,* 186-194.

(131) Jørgensen, K.; Segtnan, V.; Thyholt, K.; Naes, T. A Comparison of Methods for Analysing Regression Models With Both Spectral and Designed Variables. *Journal of Chemometrics*, **2004,** *18,* 451-464.

(132) Måge, I. Modelling and optimisation of industrial processes with raw material variation. 2006. Norwegian University of Life Science, Department of Chemistry, Biology and Food Science.

(133) Eriksson, L.; Toft, M.; Johansson, E.; Wold, S.; Trygg, J. Separating Y-Predictive and Y-Orthogonal Variation in Multi-Block Spectral Data. *Journal of Chemometrics*, **2006,** *20,* 352-361.

(134) Trygg, J.; Wold, S. O2-PLS, a Two-Block (X-Y) Latent Variable Regression (LVR) Method With an Integral OSC Filter. *Journal of Chemometrics*, **2003,** *17,* 53-64.

(135) Skov, T.; Ballabio, D.; Bro, R. Multiblock Variance Partitioning: A New Approach for Comparing Variation in Multiple Data Blocks. *Analytica Chimica Acta*, **2008,** *615,* 18-29.

(136) Jaworski, A.; Wikiel, K.; Wikiel, H. Application of Multiblock and Hierarchical PCA and PLS Models for Analysis of AC Voltammetric Data. *Electroanalysis*, **2005,** *17,* 1477-1485.

(137) Gabrielsson, J.; Jonsson, H.; Airiau, C.; Schmidt, B.; Escott, R.; Trygg, J. The OPLS Methodology for Analysis of Multi-Block Batch Process Data. *Journal of Chemometrics*, **2006,** *20,* 362-369.

(138) MacGregor, J. F.; Jaeckle, C.; Kiparissides, C.; Koutoudi, M. Process Monitoring and Diagnosis by Multiblock Pls Methods. *Aiche Journal*, **1994,** *40,* 826-838.

(139) Kourti, T.; Nomikos, P.; MacGregor, J. F. Analysis, Monitoring and Fault-Diagnosis of Batch Processes Using Multiblock and Multiway Pls. *Journal of Process Control*, **1995,** *5,* 277-284.

(140) Kourti, T. Multivariate Dynamic Data Modeling for Analysis and Statistical Process Control of Batch Processes, Start-Ups and Grade Transitions. *Journal of Chemometrics*, **2003,** *17,* 93-109.

(141) Kohonen, J.; Reinikainen, S. P.; Aaljoki, K.; Perkio, A.; Vaananen, T.; Hoskuldsson, A. Multi-Block Methods in Multivariate Process Control. *Journal of Chemometrics*, **2008,** *22,* 281-287.

(142) Felicio, C. C.; Bras, L. P.; Lopes, J. A.; Cabrita, L.; Menezes, J. C. Comparison of PLS Algorithms in Gasoline and Monitoring With MIR and NIR. *Chemometrics and Intelligent Laboratory Systems*, **2005,** *78,* 74-80.

(143) Lopes, J. A.; Menezes, J. C.; Westerhuis, J. A.; Smilde, A. K. Multiblock PLS Analysis of an Industrial Pharmaceutical Process. *Biotechnology and Bioengineering*, **2002,** *80,* 419-427.

(144) Hwang, D. H.; Stephanopoulos, G.; Chan, C. Inverse Modeling Using Multi-Block PLS to Determine the Environmental Conditions That Provide Optimal Cellular Function. *Bioinformatics*, **2004,** *20,* 487-499.

(145) Vong, R. J.; Frank, I. E.; Charlson, R. J.; Kowalski, B. R. Exploratory Data-Analysis of Rainwater Composition. *Acs Symposium Series*, **1985,** *292,* 34-52.

(146) Vivien, M.; Verron, T.; Sabatier, R. Comparing and Predicting Sensory Profiles From NIRS Data: Use of the GOMCIA and GOMCIA-PLS Multiblock Methods. *Journal of Chemometrics*, **2005,** *19,* 162-170.

(147) Tenenhaus, M.; Vinzi, V. E. PLS Regression, PLS Path Modeling and Generalized Procrustean Analysis: a Combined Approach for Multiblock Analysis. *Journal of Chemometrics*, **2005,** *19,* 145-153.

(148) Tenenhaus, M.; Pages, J.; Ambroisine, L.; Guinot, C. PLS Methodology to Study Relationships Between Hedonic Judgements and Product Characteristics. *Food Quality and Preference*, **2005,** *16,* 315-325.

(149) Bras, L. P.; Bernardino, S. A.; Lopes, J. A.; Menezes, J. C. Multiblock PLS As an Approach to Compare and Combine NIR and MIR Spectra in Calibrations of Soybean Flour. *Chemometrics and Intelligent Laboratory Systems*, **2005,** *75,* 91-99.

(150) Ramsay, J. O.; ten Berge, J.; Styan, G. P. H. Matrix Correlation. *Phychometrika*, **1984,** *49,* 403-423.

(151) Nielsen, J. P.; Bertrand, D.; Micklander, E.; Courcoux, P.; Munck, L. Study of NIR Spectra, Particle Size Distributions and Chemical Parameters of Wheat Flours: a Multi-Way Approach. *Journal of Near Infrared Spectroscopy*, **2001,** *9,* 275-285.

PAPER I

Vibeke Tølbøl Povlsen and Connie Benfeldt

Application of Multiblock PLSR in the Dairy Industry.
PLS and Related Methods, Proceedings of the PLS'01 International Symposium, V. Esposito Vinzi, C. Lauro A Morineau, M. Tenenhaus (Eds.), 371-383, 2001

Application of Multiblock PLSR in the Dairy Industry

Vibeke T. Povlsen [1] and Connie Benfeldt [2]

[1]*Food Technology, Department of Dairy and Food Science, The Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958, Denmark, vip@kvl.dk.*

[2]*Arla Foods amba, Innovation Centre Brabrand, Rørdrumvej 2, DK-8820 Brabrand, Denmark, connie.benfeldt@arlafoods.com.*

Abstract

The handling of the large amount of data collected during a 35 week ripening period of semihard smear-ripened cheese plus the selection of important variables in relation to sensory quality from this quantity of data can be difficult. In the present work a method for handling such a problem is presented. A standard Partial Least Squares Regression (PLSR) model was decomposed into block contributions using multiblock PLSR (MBPLSR). The data was reduced by selecting the most descriptive variable blocks based on the block weights given in the MB algorithm. Using MBPLSR1 models for variable selection, an improvement in the predictive performance was observed, whereas MBPLSR2, in this case, was not well-suited for variable selection when predicting several sensory variables simultaneously.

Keywords: Multiblock PLSR, block decomposition, block selection, sensory evaluation, cheese quality.

Introduction

In the food industry large amounts of data are collected in order to improve and control the food production processes. The data is often difficult to handle, visualize and explore due to the size of the data matrices and often much of the data is neglected due to lack of time for proper data analysis. Food is generally a complex system, and it can be difficult to investigate the influence and importance of different measured variables. Multiblock methods (MB) as described in the literature [1] [2] can provide a useful tool in such situations. These are data mining tools, which provide a quick graphic overview of data that consists of several blocks.

Multiblock methods have the feature of modelling several blocks simultaneously while still providing information about the individual blocks. The restriction of MB analysis is that one mode in all blocks has to be common, e.g. the same set of samples for each data table. In standard Partial Least Squares Regression (PLSR) the descriptor variables are collected in one large block and it can be difficult to identify the role of individual blocks or variables in the regression model. In MBPLSR an additional level is introduced called the "super" level. The "super" level contains the augmented block, which give information about the block contribution, whereas the "lower" level contains the individual blocks showing contributions from individual variables. The principle of MB modelling is illustrated in Figure 1. In the MBPLSR the relationship between the response block **Y** and the descriptor block **T** at the "super" level is found. The **T** block is a function of the original descriptor **X**-blocks at the "lower" level. In the MBPLSR the same tools are present as in the standard PLSR: The scores and loadings, the percentage of variation explained in **X** and **Y** of each block, etc. But when performing MB regression additional information about each of the individual blocks is provided in the loading weight $\mathbf{W_T}$ at the "super" level. This information can be used e.g. to select the most explanatory variables among the descriptor blocks.

Before performing MB analysis. it is important to consider the scaling and weighting of blocks, as this will influence the outcome model.

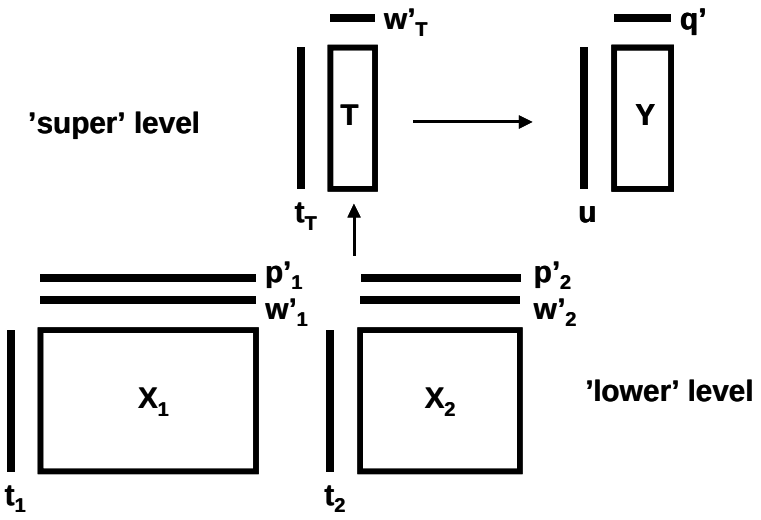


Figure 1. Overview of multiblock PLSR.

In this work the MBPLSR method with deflation on the super scores is used [4]. Previous work has shown that using super score deflation of X is identical to performing a standard PLSR with all the predictors in one block. The results from the standard PLSR can then be decomposed into block contributions, provided that the blocks are equally scaled and weighed [2,5].

The objective of this study was to explore the application of MB methods for prediction of the quality of semihard smear-ripened cheese. The cheese ripening process is a multi-step reaction involving the formation of rather large and well-defined peptides, their subsequent digestion into smaller peptides and free amino acids before final transformation into various aroma compounds [3]. Thus, the development of body and flavour in the cheese results from the action of various enzymes involved in proteolysis and amino acid degradation, their interactions, and the physical conditions e.g. pH in the cheese and the storage temperature. Data from various physical and chemical analyses were collected and used for MBPLSR with the objective to identify and perform a block selection of the important chemical and physical analyses in relation to prediction of the sensory attributes.

Experimental

The ripening characteristics of four batches each comprising six smear-ripened Danbo-type cheeses were monitored after the first 4, 10 and 16 weeks of a 35-week ripening period. The 24 cheeses were subjected to chemical analyses (pH and the amount of fat, dry matter salt, protein, intact casein, soluble peptides and various aroma compounds and physical measurements (compression, stretch, texture profile analysis and oscillation). Cheese quality was monitored by sensory evaluation after 35 weeks of ripening using 1 odour (overall cheese intensity odour), 5 flavour (overall cheese intensity flavour, aftertaste, acid, sharp and unclean flavour) and 6 textural attributes (soluble, sticky, elastic, breakable, cutable and hard texture).

The analytical results were examined using standard PLSR and MBPLSR and were validated by leave one out cross-validation [6]. All data analysis and modelling was performed using the software Unscambler 7.5 (Camo) for Principle Component Analysis (PCA) and Matlab 5.3 (Mathworks) for Windows and algorithms from the MBtoolbox (www.models.kvl.dk).

Results and Discussion

All data blocks were individually analysed using PCA to view the different physical and chemical analyses obtained at each sampling time. Outliers were detected and the behaviour of the batches was noted. A clear batch variation in the four batches was observed, which is believed to result primarily from variations in the microbial flora in the cheese and from small deviations from the cheese production scheme.

Multiblock PLSR models

The dependent regression block, **Y**, consists of 12 sensory attributes. At first, a selection of sensory variables was made based on the correlation of the 12 sensory attributes. A score plot of factor 1 and factor 2 from a PCA performed on the 12 sensory attributes is presented in Figure 2.

Four clusters appear in the score plot (Figure 2) suggesting that attributes within each of these four groups are somehow correlated. Correlation between variables can be advantageous for predicting, as the information in these variables will be more or less the same, yielding better overall predictions. The group in the upper left corner consisting of the four sensory attributes (overall cheese intensity odour (odour), overall cheese intensity flavour (intensity), sharp flavour and aftertaste) was selected as example predictor variables in this paper.

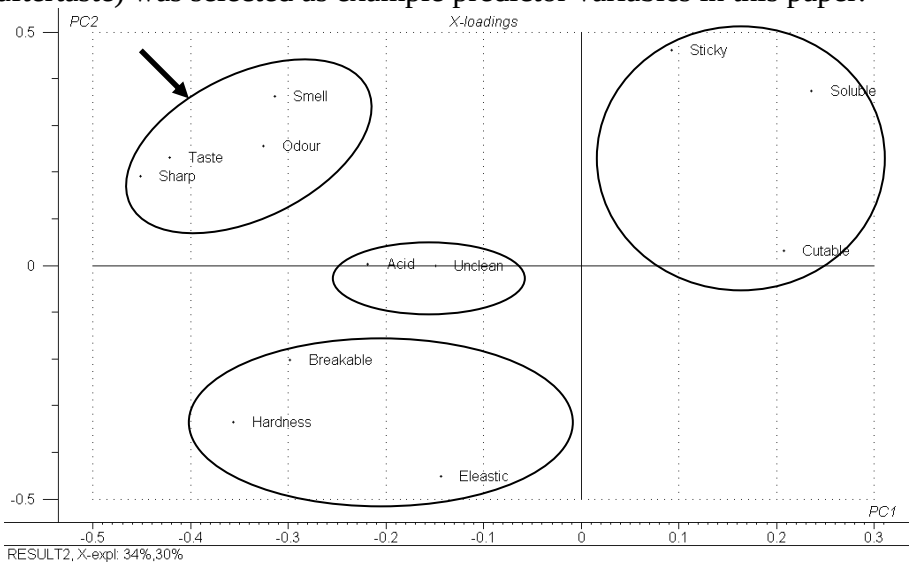


Figure 2. Score plot of factor 1 (35 % explained variance) vs. factor 2 (30% explained variance) of a PCA performed on the 12 sensory attributes. The arrow mark the selected four sensory attributes: Odour (overall cheese flavour odour), intensityste (overall cheese intensity flavour), sharp taste, aftertaste.

The **X**-block was composed of two measurement types: chemical and physical, and includes 153 and 150 variables, respectively. The data was autoscales, but no weighting was performed, as the blocks were approximately the same size. Thus, expected to carry almost equal weight. A standard PLSR1 model on the entire **X**-matrix, displaying an acceptable predicting ability, was used as a starting point for the MB analysis, which was continued using the decomposition pattern illustrated in Figure 3. The block decomposition was based on the information given by the structure of the experiment. In step 1, the **X**-block consisted of all the measured variables combined, whereas step

2 illustrates the decomposition of the **X**-block into two block contributions: chemical and physical variables. This split is interesting, as information about the contribution of the two types of analysis is retrieved. Step 3 displayed the information hidden in data obtained by analyses performed at the three sampling times, whereas step 4 provided information of each type of analytical measurement. In step 5 a selection of blocks was made due to the information given in the four previous steps.

The criteria for block selection are of course a function of the specific goal of the experiment. The block selection in this work was based on the block weights, the explained variance of X and Y, and the correlation coefficients between predicted and reference values.

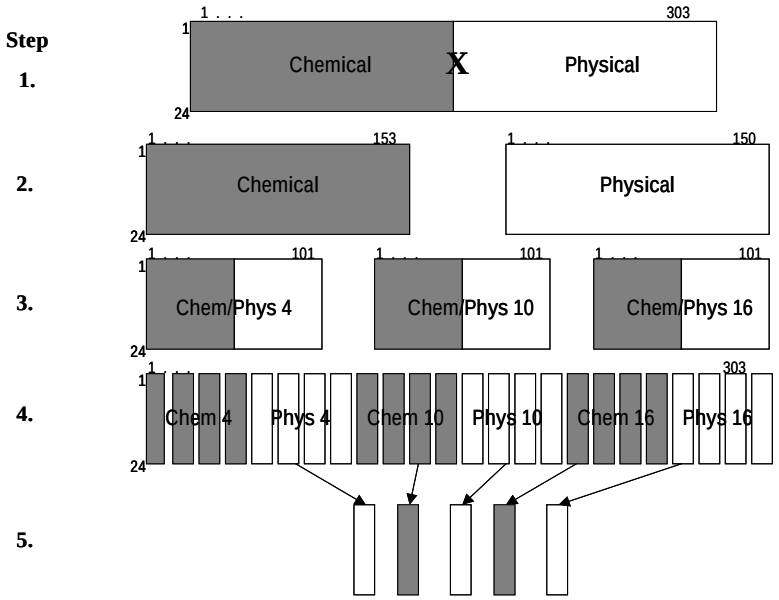


Figure 3. Decomposition of the X-matrix [24 x 303] into conceptual meaningful blocks. Step 1: The entire X-matrix. Step 2: Decomposition into chemical and physical measurements. Step 3: Decomposition due to sampling time, week 4, week 10 and week 16. Step 4: Decomposition into individual chemical and physical measurements. Step 5: Selection of blocks based on contribution and importance in the models from step 1 to 4.

In the following, we will focus on the sensory attribute, overall cheese intensity odour. A PLSR1 using the entire **X**-matrix as predictor block indicated the predicting ability that could be expected from the data. The data was autoscaled but no weighting of the individual measurement types were made. A four-factor model describing 49% of **X**-variance and 96% of the **Y**-variation was selected based on cross validation. The correlation coefficient was r = 0.71 when predicting the overall cheese intensity odour.

This model was then decomposed and the loading weights $\mathbf{W_T}$ of the "super" level of the three decomposition steps are presented in Figure 4. Figure 4A shows $\mathbf{W_T}$ of the chemical and physical blocks. The physical block dominates the two first factors whereas the chemical block has the highest influence in factors three and four. This indicates that both types of measurement are used to describe the variation in the sensory attribute. In the next step, the influence of storage time was examined. This is illustrated in Figure 4B, where the $\mathbf{W_T}$ of week 4, week 10 and week 16 are shown. Week 4 only primarily seems to influence the model in factor 3, whereas week 10 is very dominant in the first two factors. Week 16 clearly dominates the model in the fourth factor. Comparing the profiles of $\mathbf{W_T}$ in Figure 4A and Figure 4B, the influence of the chemical block seems to be caused by the data obtained analyzing the 10 week old cheeses. The influence of the physical block in factor 3 can be assigned to analyses performed on 4-week-old cheeses.

Figure 4C shows the $\mathbf{W_T}$ of the chemical and physical blocks for week 4, 10 and 16. The physical measurements of week 10 dominate factor 1 and factor 2. In the third factor, the chemical block of week 10 is important and in the fourth factor the physical measurements of week 16 show a high influence. This decomposition shows that the physical domination in the first two factors is due to the physical block of week 10, and the chemical influence in factor three is caused by the chemical block of week 10, while factor 4 is dominated by physical analysis of week 16.

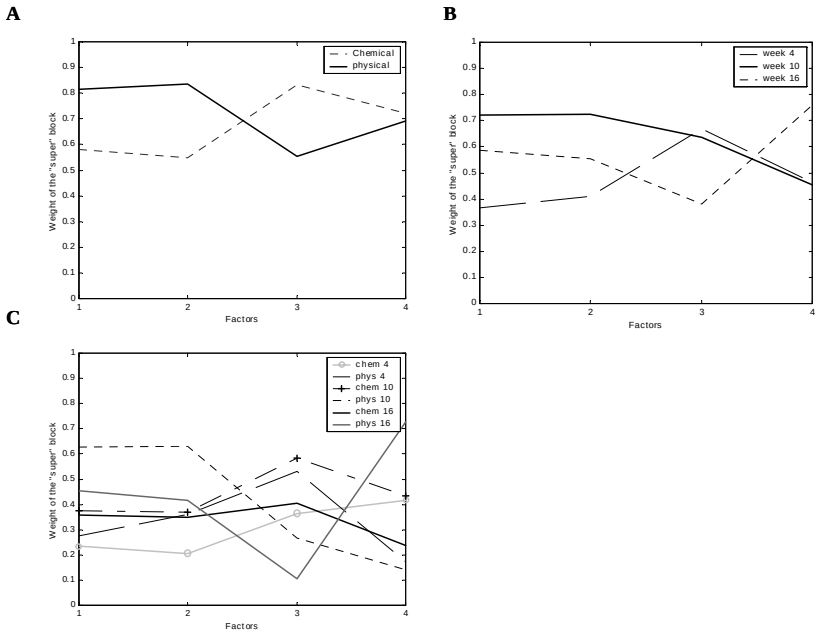


Figure 4. Loading weights (W_T) of the X-block when decomposed into block contribution. A. Chemical and physical block. B. Week 4, 10 and 16. C. Chemical and Physical blocks at week, 10 and 16.

In the next step (step 5, Figure 3), further decomposition into individual measurements types was performed. There were 8 different types of data blocks and they were measured at three times yielding 24 blocks. This decomposition is crucial in the final selection step, since the importance of each analytical analysis is provided. The loading weights of the 24 blocks are shown in Figure 5. From Figure 5 it appears that some of the blocks have a low influence on the model, whereas other blocks show a high predictive performance. The blocks with a low influence are the physical measurements of compression and stretch at all three sampling weeks and the texture profile analysis at weeks 4 and 16. This is consistent with previous models, which revealed a high contribution of the physical block obtained after 10 weeks of ripening. The block containing the five chemical variables: pH, fat, drymatter content, salt, and protein displays a very low predictive performance. The chemical blocks with high influence in the model are the peptide analysis of all three sampling weeks and the

aroma analysis measured after 10 and 16 weeks of ripening. The physical measurement of oscillation for all three weeks shows a high contribution to the model as does the textile profile analysis of week 10.

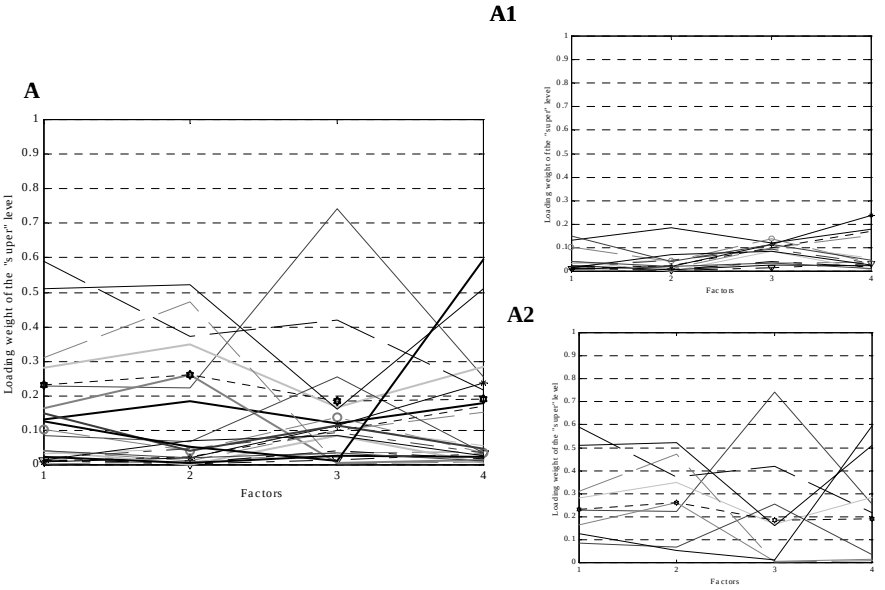


Figure 5. Loading weights (W_T) of the 4 blocks corresponding to the analytical measurements performed at week 4, 10, and 16. A) All 24 WT. A1) W_T of the 15 blocks with the lowest influence. A2) W_T of the 9 blocks with the highest influence.

A stepwise selection of blocks can be carried out by first dismissing the blocks with the lowest influence. Further selection of blocks can be performed by raising the limit of block contribution to the model. In Table 1 the correlation coefficients of cross validation of the PLSR1 performed on the entire **X**-matrix and PLSR1 on selected blocks selected are listed.

Table 1.View of the block selection for the sensory attribute, overall cheese flavour odour.

No.	Blocks																							
	Week 4								Week 10								Week 16							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X			X	X	X	X	X		X			X	X	X	X		X			X	X	X	X
3							X	X		X				X	X	X		X					X	X
4							X			X					X	X		X					X	X
5										X					X	X		X					X	X

Table 2. Correlation coefficients of the predictions using standard PLSR1 of 5 models on selected blocks (Table 1)

No.	Factors/correlation
1	4/0.714
2	4/0.730
3	3/0.807
4	3/0.808
5	3/0.811

Model no.1 in Table 1 includes all 24 blocks and in model 2 the second (compression) and the third (stretch) blocks in all three weeks are excluded. A minor improvement from 0.714 to 0.730 can be observed. Model 3 corresponds to the blocks in Figure 5A2 where only the blocks showing high influence are included. In this model the predictability increases to 0.807. In the two models 4 and 5, blocks are excluded according to the block weights in Figure 4B and C. Here it was shown that the influence of the analytical analysis performed after 10 and 16 weeks of ripening was higher than a ripening period of 4 weeks. The predictions of the blocks were almost equal resulting in a correlation of 0.811 for the model using six descriptor blocks: aroma, peptides, and oscillation of weeks 10 and 16. The peptide distribution and the composition of the aroma compounds in the cheeses usually display some correlation due to the fact that the formation of aroma compounds is highly dependent on the proteolytic process in the cheese during the ripening period. The amount of intact casein in cheeses after 10 weeks of ripening is also included in model 3 as the loading weight indicated a contribution, but is then left out in the following models. The amount of intact casein left in the product is usually high in the beginning of the ripening process after which these proteins are sequentially degraded to peptides and free amino acids

during the ripening process. Since the caseins are related to the distribution and content of peptide, this may explain why the analysis can be left out without decreasing the prediction. The oscillation measurements play a crucial role in the prediction of the overall cheese intensity odour, which could result from proteolysis, since the proteins degradation is responsible of the aroma formation, structure and consistency.

The procedure described above was repeated for the entire **Y**-block, i.e. overall cheese intensity flavour, sharp flavour and aftertaste. The correlation of the PLSR1 models including all the blocks and the reduced model using selected blocks are presented in Table 3. All four sensory attributes display an improved predictive performance observed when reducing the number of blocks in the model. The number of blocks included in the four models varies from 4 to 9 blocks, which is a reduction in the number of variables of approximately 75 to 30 %. This clearly shows that a feasible reduction in the performed measurements can be made.

Table 3. Correlation coefficient of the PLSR1 when predicting: overall cheese flavour odour (Y1), overall cheese flavour taste (Y2), sharp taste (Y3), and aftertaste (Y4).

Method	Model	Y1		Y2		Y3		Y4	
		Fac	r	Fac	r	Fac	R	Fac	r
PLSR1	All blocks	4	0.714	4	0.802	4	0.831	4	0.840
	Selected blocks	3	0.811	4	0.831	4	0.876	3	0.858

The predictor variables were selected due to their mutual correlation, as this should be an advantage when performing PLSR2. In Table 4 the correlations of PLSR2 based on all 24 blocks and on the selected blocks (5 blocks) are presented.

Table 4. Correlation coefficient of the PLSR2 when predicting: overall cheese flavour odour (Y1), overall cheese flavour taste (Y2), sharp taste (Y2), and aftertaste (Y4) simultaneously.

Method	Model	fac	Y1	Y2	Y3	Y4
PLS2	All blocks	7	0,7119	0,7949	0,8379	0,8552
	Selected blocks	5	0,8429	0,7906	0,7961	0,8403

In this model, only the sensory attribute, overall cheese flavour odour improves. The correlation coefficient of the three other sensory variables stays the same or decreases. A comparison of the PLSR1 and PLSR2 models of all 24 blocks reveals a slight improvement in the prediction when predicting the entire **Y**-block simultaneously. The PLSR2 can be used to get an overview of the prediction ability of the descriptor block, when predicting correlated response variables. But a slightly better prediction can be gained when predicting a single regression variable at a time by using MBPLSR to select the important blocks

Conclusion

A standard PLSR model can be decomposed into block contributions without changing the predicting performance in order to get a better view of the data. Multiblock PLSR provides a good tool to interpret and select important descriptor blocks. A reduction in the number of predictor variables of 70 to 30 % could be gained by using the multiblock decomposition procedure in this case. Differences in the predictive performance were seen when predicting the sensory response attributes individually or blockwise. The predictive performance using the entire descriptor block is slightly better when predicting a block of correlated sensory variables. But the optimization by blocks selection of high contribution blocks is better when predicting one response variable at a time.

References

[1] Westerhuis, J. A., Kourti, T., and Macgregor, J. F. (1998): *Analysis of multiblock and hierarchical PCA and PLS models*. Journal of Chemometrics, 5, 12, 301-321

[2] Westerhuis, J. A. and Coenegracht, P. M. J. (1997): *Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares*. Journal of Chemometrics, 5, 11, 379-392

[3] Fox, P. F., Singh, T. K., and McSweeney, P. L. H. (1994): Proteolysis in chesse during ripening. In: Andrews, A. T. and Varley, J. (eds.), Biochemistry of Milk Products, p. 1-31, Royal Society of Chemistry, London

[4] Rännar, S., Macgregor, J. F., and Wold, S. (1998): *Adaptive batch monitoring using hierarchical PCA*. Chemometrics and Intelligent Laboratory Systems, 41, 73-81

[5] Westerhuis, J. A. and Smilde, A. K. (2001): *Short Communication. Deflation in multiblock PLS*. Journal of Chemometrics, 15, 485-493

[6] Eastman, H. T. and Krzanowski, W. J. (1982): *Cross-validatory Choice of the Number of Components From a Principal Component Analysis*. Technometrics, 1, 24, 73-77

PAPER II

Vibeke Tølbøl Povlsen, Åsmund Rinnan, Frans van den Berg, Henrik J. Andersen, and Anette K. Thybo

Direct decomposition of NMR relaxation profiles and prediction of sensory attributes of potato samples.
LEBENSMITTEL-WISSENSCHAFT UND-TECHNOLOGIE-FOOD SCIENCE AND TECHNOLOGY, 36 (4), 423-432, 2003

Lebensm.-Wiss. U.-Technol. 36 (2003) 423–432

lwt

www.elsevier.com/locate/lwt

Direct decomposition of NMR relaxation profiles and prediction of sensory attributes of potato samples

V.T. Povlsen[a], Å. Rinnan[a,*], F. van den Berg[a], H.J. Andersen[b], A.K. Thybo[c]

[a] *Department of Dairy and Food Science, The Royal Veterinary and Agricultural University (KVL), Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark*

[b] *Department of Animal Quality, Danish Institute of Agricultural Sciences (DJF), DK-8830 Tjele, Denmark*

[c] *Department of Horticulture, Danish Institute of Agricultural Sciences (DJF), DK-5792 Aarslev, Denmark*

Received 13 March 2002; accepted 12 January 2003

Abstract

In this paper the decomposition of low-field Carr–Purcel–Meiboom–Gill (CPMG) NMR relaxation measurements on 23 raw potato categories was investigated. The potato categories were formed from five different cultivars, each binned in 2 or 3 dry matter intervals, sampled at two storage times. A novel data analytical tool—called SLICING—revealed that different amounts of four distinct proton relaxation profiles could describe the main variation in the data set. Magnitudes (scores) of the third and fourth profile separated the potato cultivars, storage times, and dry matter content indicating that properties related to fast relaxation times explain the differences between cultivars and storage times for the potatoes. The concept of direct decomposition using SLICING on low-resolution NMR data is a new approach in potato analysis and a promising tool for obtaining more information about the structure and water distribution in food products.

Furthermore, the texture-related sensory attributes, hardness, cohesiveness, adhesiveness, mealiness, graininess, and moistness of cooked potatoes were predicted by partial least-squares regression (PLSR). Four different types of predictor variables derived from the NMR relaxation curves were compared in the regression models: (i) the raw CPMG curves, (ii) the parameters from the traditional bi-exponential fitting, (iii) the results from a distribution analysis, and (iv) the scores from the SLICING model. The predictions based on the distribution analysis performed worse than the first three procedures, which all showed similar prediction ability. The advantage of the SLICING approach is in the possibility to interpret physical properties, e.g. water distribution of the potato samples.

Keywords: Potato; Low-field NMR; NMR relaxation; PARAFAC; PLSR

1. Introduction

The texture of cooked potatoes is an important quality attribute when assessing potato quality. In the potato industry great interest lies in both improving and developing rapid methods to determine this quality. Special interest lies in assessing the raw potato samples and relating them to the sensory quality of cooked potatoes. The potential perspective could be an early sorting of the raw material according to quality prior to packaging or processing. The texture of cooked potatoes is related to the size and amount of starch, rigidity and chemistry of the cell walls, enzyme activities, minerals, heating, water content as well as the subsequent heating process (Gould, 1999). Evaluation of potato texture and quality can be performed by mechanical, analytical and/or sensory methods (VanMarle, DeVries, Wilkinson, & Yuksel, 1997; Thybo & Martens, 1999; Ulrich, Hoberg, Neugebauer, Tiemann, & Darsow, 2000). Using sensory evaluation, information about the human perception of potato quality is obtained, as the senses of sight, smell, taste, touch and hearing are studied. In sensory analysis, the texture is evaluated in terms of moistness, adhesiveness, mealiness, etc. In addition, mechanical measurements, for example uniaxial compression and nuclear magnetic resonance (NMR) relaxation, have been applied in the texture analysis of vegetables (Tang, Belton, Ng, Waldron, & Ryden, 1999; Thybo & Martens, 1999; Tang, Godward,

*Corresponding author.
E-mail address: aar@kvl.dk (Å. Rinnan).

0023-6438/03/$30.00
doi:10.1016/S0023-6438(03)00023-9

& Hills, 2000). NMR has been shown to provide useful information about molecular structure within a sample and has become a powerful nondestructive analytical tool in chemistry (Hemminga, 1992; Ruan & Chen, 1998). In food science, NMR techniques have been used to study the texture and the state of water in food samples (Hills & Le Floc'h, 1994; Seow & Teo, 1996; Hills, Goncalves, Harrison, & Godward, 1997; Ruan et al., 1997; Tang et al., 1999; Tang et al., 2000) and for the analysis of fats and oils (Pedersen, Munck, & Engelsen, 2000). Previous work by Thybo and Martens (1999) showed a higher correlation between sensory quality of cooked potatoes and ^{1}H NMR on raw potatoes compared to using ^{1}H NMR on cooked potatoes. This work forms the basis for the present study, where the objective was to compare the SLICING method (Pedersen, Bro, & Engelsen, 2001) to existing methods for analysing low-field proton NMR signals (^{1}H-NMR) from Carr–Purcel–Meiboom–Gill (CPMG) pulse relaxation curves of raw potatoes. The comparison was based on the interpretability in data analysis and the predictive performance of sensory quality on cooked potatoes using multivariate regression. The SLICING procedure has previously shown good results for data analysis purposes when estimating the underlying relaxation curves of fish (Andersen & Rinnan, 2002). When handling low-field NMR data, these underlying relaxation curves ideally correspond to the different chemical states of water in the measured samples. Thus, SLICING makes it possible to interpret the data directly on a physical basis because the model separates the measured signal, a mixture of exponential curves, into physically meaningful uni-exponential contributions. It is noted that the SLICING method assumes that a fairly low number of such curves are sufficient for describing the actual measurement signal, in contrary to, e.g. distribution analysis, where it is assumed that the data consist of a sufficiently large number of distinguishable exponentials, such that a distribution of these can be computed. In the bi-exponential fitting method two contributing exponentials are assumed sufficient to describe the measured signal. However, the two last methods assume no relationship between samples—treating each sample individually—and thus differ from the factor-based SLICING method. The discussion as to which of these alternative decomposition methods is most appropriate will not be the main issue of this paper. Rather, it will be shown that the SLICING approach as such provides a solution, which is scientifically sound and useful for interpretation and further modelling.

The relation between the NMR relaxation curves on raw potatoes and sensory attributes evaluated on cooked potatoes was studied by regression modelling. The prediction performance based on partial least-squares regression (PLSR, Martens & Næs, 1989) using the SLICING scores as predictors were compared to modelling on the raw low-field ^{1}H-NMR curves (CPMG PLSR). PLSR has previously been used on raw low-field ^{1}H-NMR curves for prediction of fish and potatoes sensory attributes, showing good performance (Thybo, Bechmann, Martens, & Engelsen, 2000; Thygesen, Thybo, & Engelsen, 2001). However, the CPMG PLSR results are less interpretable because the loadings do not have a direct physical meaning. The regression performance based on the model parameters retrieved from bi-exponential fitting and distribution analysis was also compared to the regression methods based on the SLICING scores. Bi-exponential fitting was applied because it constitutes one of the main alternatives to the SLICING approach, while distribution analysis was applied because it has been used with success in previous potato studies (Hills & Le Floc'h, 1994; Hills, Goncalves, Harrison, & Godward, 1997), as well as other areas of research (Tang et al., 2000).

2. Materials and methods

2.1. Potatoes

The material used in the experiments included five potato cultivars grown at an experimental field at the Danish Institute of Agricultural Sciences. Within the five cultivars the potatoes were graded in salt solutions according to 1% dry matter bins (Burton, 1989) in the range of 18.0–22.9%, as described by Thybo and Martens (1999). Potato samples harvested in September 1999 were analysed in November 1999, and in May 2000 after being stored at 4°C at 95% relative humidity. This selection procedure gave a total of 23 different potato samples (see Table 1).

2.2. Sensory analysis

The potatoes were peeled and boiled in water for 20–25 min until they were cooked through. The sensory analysis was performed on the cooked potatoes by a trained panel of ten assessors and evaluated on a scale from 0 to 15. The measurements were performed as described by Thybo and Martens (1999) using the average of the ten assessors times four sensory replicates. The sensory variables hardness, cohesiveness, adhesiveness, mealiness, graininess, and moistness were evaluated.

2.3. NMR measurements

The relaxation measurements of the water protons were performed on a Maran Bench top Pulsed ^{1}H-NMR Analyser (Resonance Instruments Ltd., Witney, UK) with a magnetic field strength of 0.47 T, corresponding

Table 1
Tuber samples used in the experiments

Cultivar	Dry matter bins (%) Storage time	
	November 1999	May 2000
Ditta	20.0–20.9	21.0–21.9
	21.0–21.9	22.0–22.9
Sava	18.0–18.9	18.0–18.9
	19.0–19.9	19.0–20.9
	20.0–20.9	21.0–21.9
Bintje, low dry matter	19.0–19.9	20.0–20.9
	20.0–20.9	21.0–21.9
	21.0–21.9	
Bintje, high dry matter	21.0–21.9	—
	22.0–22.9	
Berber	18.0–18.9	18.0–18.9
	19.0–20.9	19.0–20.9
	21.0–21.9	21.0–21.9

to a resonance frequency of 23.2 MHz. The instrument was equipped with an 18-mm temperature variable probe. The samples were sized in cylinders of $h \times d = 40 \times 14\,\text{mm}^2$. They were stamped longitudinally from the stem end of the potato, and placed in a cylindrical glass tube (14 mm in diameter and 50 mm in height). This tube fitted into the NMR temperature variable probe 18 mm in diameter. Before the measurement was performed, the sample was temperature controlled to 25°C in a water-bath for 15–20 min.

Transverse relaxation (T_2) was measured using the CPMG sequence (Carr & Purcell, 1954; Meiboom & Giil, 1958). The transversal relaxation measurements were performed with a τ value (time between 90° and 180° pulse) of 1000 μs. The data were acquired as four scan repetitions. The repetition delay between two succeeding scans was 4 s. The signal amplitude was measured every echo and the relaxation measurements were performed at 25°C.

2.4. Data treatment

Each potato sample (bin) was measured by NMR in a number of replicates (tubers) ranging from 12 to 15. If outliers were detected in any of the replicate series, they were removed before the computations. Outliers were defined as replicates that were significantly different from the other replicates in any of the following attributes: low initial value, slower relaxing curve or faster relaxing curve. The initial data consisted of a total of 324 measurements, which was reduced to 295 after removing the outliers. Each sample was now represented by 11–14 NMR measurement replicates. The sensory analysis was performed on only four replicates with no direct link to the tubers used in the NMR measurements. To compensate for differences between tubers from one category, the average of the sensory analysis was used together with the average of the NMR curves for each bin. In this study the difference between cultivars, and not between tubers, was of interest, hence using the average reduces the natural variety within the bins.

3. Data analysis and modelling

3.1. Description of the NMR curves

NMR relaxation signals can be expressed mathematically as a sum of exponential decays (see Eq. (1)):

$$I(t) = \sum_{n=1}^{N} M_{0,n} \exp\left(-\frac{t}{T_{2,n}}\right). \quad (1)$$

In this equation the profile $I(t)$ is parameterized such that N is the (expected) number of uni-exponentials, M_0 holds the N magnitude values, t is time, and T_2 is the time constants associated with each uni-exponential decay. For a set of curves, it is assumed that the quantitative information, amount of a specific proton signal, is carried by the M_0 values and the qualitative information, the type of proton signal, by the T_2 values. There are several methods to find these parameters. Three methods are evaluated in this article: bi-exponential fitting, distribution analysis and SLICING. In bi-exponential fitting, the assumption is that N in Eq. (1) is two for any sample and that the T_2 value can vary from sample to sample. In the SLICING N is not known beforehand but determined as part of the modelling step. It is assumed that all samples can be described by the same set of T_2 values. In distribution analysis, it is assumed that a distribution of T_2 values generates each profile. Hence, N is assumed to be very large indicating that each proton has its own distinct value. This assumption appears reasonable at first glance, but in practice distribution analysis can be hampered by numerical instabilities caused by the high amount of parameters to be determined from a limited data set with finite signal-to-noise ratio. The discrete methods, bi-exponential fitting and SLICING, on the other hand, assume an approximation, which may be valid in practice due to this limited signal-to-noise ratio and the similarity of the individual proton relaxations over samples. Hence, it is not possible on theoretical grounds to reject any of the proposed methods. One purpose of this investigation is to show empirically to what extent, these methods can provide reliable information on the current data. In the following the different modelling approaches for NMR data and regression are described.

3.2. Regression by PLS on the raw CPMG curves

One of the advantages of multivariate methods such as PLS regression (Martens & Næs, 1989) is that they handle correlated variables well. This feature makes them suitable for handling data such as NMR relaxation curves, where neighbouring time points are highly correlated. Using PLSR on raw data, focus is on the prediction ability of the model, but the interpretation of the models might not be as straightforward as the other methods described in this paper.

3.3. Bi-exponential fitting

A common approach to model NMR curves is bi-exponential fitting, yielding for each sample individual values for parameters $M_{0,1}$, $M_{0,2}$, $T_{2,1}$, and $T_{2,2}$ in Eq. (1). This approach is based on the assumption that any sample can be described as a weighted sum of two exponentials and the T_2 values are specific for this sample. The M_0 and T_2 values may be used for the prediction of the sensory attributes by the use of PLSR.

3.4. Distribution analysis

Another method for describing the NMR curves is by the use of distribution analysis. Distributed exponential fitting analysis was performed on T_2 relaxation data using the Win-DXP program for Matlab (Butler, Reeds, & Dawson, 1981). A continuous distribution of exponentials for a CPMG experiment can be defined by Eq. (1), setting N to a large number. To use this distribution information for regression analysis the results need to be transformed into a suitable set of variables. In this paper, the position and the amplitude of the peaks in the distribution were used for regression analysis.

4. SLICING

SLICING is a novel method for exploring NMR relaxation curves (Pedersen et al., 2001). The method decomposes the relaxation curves from NMR measurements into a few individual archetype proton contributions. It is based on increasing the dimensionality of the data from a two-way to a three-way array by a proper rearrangement. The rearranged data cube (three dimensional) will ideally follow the so-called tri-linear model. Performing a tri-linear decomposition of the rearranged data will directly yield a set of normalized exponential decays (i.e. T_2 values) as well as the corresponding amounts/magnitudes of these decays for each sample (M_0 values).

In SLICING the assumption is that all samples can be represented by a weighted sum of a number of exponentials, conforming Eq. (1). Thus, there is no predefined number of exponentials as in the bi-exponential fitting. On the other hand, it is assumed that all samples are sums of the same exponentials, which is not the case for bi-exponential fitting.

The SLICING algorithm uses the principles of direct exponential curve resolution algorithm (DECRA, Windig & Antalek, 1997). The idea is to split the CMPG relaxation curves (see Fig. 1a) into two (or more) overlapping parts (slabs), where the size of the overlap is determined by the lag term, generating a three-dimensional array. Most of the original relaxation curve is present in both slabs. This operation is illustrated in Fig. 1a. Next, PARAllel FACtor analysis (PARAFAC) is performed on the three-dimensional array (Bro, 1997). The PARAFAC model is described by the following equation:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{if} c_{kf} + e_{ijk}$$
$$(i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, K). \qquad (2)$$

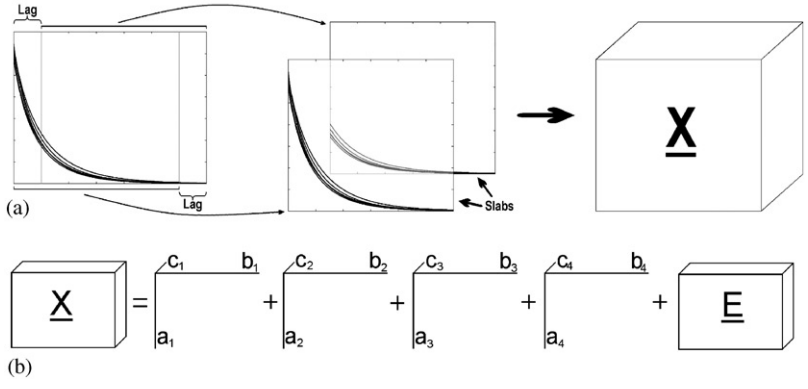


Fig. 1. Going from NMR signal to the data cube for PARAFAC modelling: (a) illustrating the principles of creating a three-way array from NMR relaxation curves; (b) the data cube $\boldsymbol{X}$ [$23 \times 1991 \times 3$] is decomposed into four triads with sample scores **a** (23 exponential loadings), **b** (1991) and slab loadings **c** (3), plus residual cube $\boldsymbol{E}$ ('noise').

The element x_{ijk} is the original value in the position (i,j,k) of the data cube $\boldsymbol{X}$. The parameter $\mathbf{a}_{\bullet\mathbf{f}}$ is the object score (magnitude) for factor f (first mode), $\mathbf{b}_{\bullet\mathbf{f}}$ is the exponential decay curve for the pure component f (second mode), and loading $\mathbf{c}_{\bullet\mathbf{f}}$ gives the ratio between the different slabs (third mode). The term e_{ijk} contains residual variation not captured by the model. The data cube $\boldsymbol{X}$ is decomposed into F different components (triads) and a residual cube $\boldsymbol{E}$ (Fig. 1b). In the PARAFAC algorithm used here, the factors (triads) are found simultaneously via an alternating least-squares algorithm (Bro, 1997). If the model is correctly specified, the residual of the exponential loadings indicates how much structural information remains unmodelled. If the residuals show random behaviour and no systematic trend, only noise is left unexplained and hence the N estimated profiles explain the variation in the data up to the noise. Furthermore, if the model is adequate each loading is described by a single exponential. If too many components are extracted, the estimated curves will reflect this (one or more being nonexponential). The residuals were used together with the appearance of the relaxation loadings to estimate the correct number of components. The object scores from the SLICING were then used for prediction of the sensory attributes.

In this study the data matrix $\boldsymbol{X}$ held the CPMG relaxation curves of the 23 samples. The SLICING was performed by splitting the relaxation curves into three slabs; with a lag of 0, 1, and 4 data points, respectively. This choice of lags was based on a subjective selection from initial investigations. The dimension of the rearranged data cube was 23 objects $\times$ 1991 relaxation variables $\times$ 3 slabs.

4.1. Validation

The validation of the regression models for the CPMG PLSR, the SLICING, the bi-exponential, and the distribution analysis predictions were all performed by the leave one subset out cross validation (Eastman & Kranowski, 1982; Martens & Næs, 1989). In this method the data are split into equally sized, randomly selected subsets. One subset is left out and a model is built from the remaining data. The properties of the left-out objects are then predicted using this model, and the residuals are calculated for models of increasing model complexity (number of factors). In the next step a new subset is removed and the procedure is repeated until every subset has been left out once. The root mean square error of cross validation (RMSECV, see Eq. (3)) indicates the difference between the predicted and the measured values. In the following equation, y is the measured values, $\hat{y}$ is the predicted value, while n represents the number of samples:

$$\mathrm{RMSECV} = \sqrt{\frac{\sum(y-\hat{y})^2}{n}}. \tag{3}$$

In this study the data sets were divided into four subsets. RMSECV and the correlation coefficients (r, upon plotting measured versus predicted) were used as indicators of the model's predictive ability.

All data analysis and modelling were performed using Matlab 5.3 software (Mathworks) for Windows with algorithms taken from the PLS-Toolbox (www.eigenvector.com) and the N-way Toolbox (Andersson & Bro, 2000). A dedicated SLICING toolbox is available at www.models.kvl.dk, but was not yet available at the start of this investigation.

5. Results and discussion

To get an impression on the way the sensory attributes discriminate potato cultivars a principal component analysis (PCA) is performed (Martens & Næs, 1989). Fig. 2 shows the bi-plot of sample scores and attribute loadings. In this figure clear grouping of cultivars and storage times are observed, as well as for the dry matter bins. This proves that the data set contains information which can distinguish these design variables. It was of interest to investigate the possibility to extract the same information from the NMR measurements via multivariate data analysis, without the requirement of sensory panel input.

In the present work the region from 12 to 4000 ms of the NMR measurement signal was used in the analysis. The first five data points were considered unreliable due to noise and the last 2000 points had a signal close to zero, not contributing any significant information. The average CPMG relaxation curves of the raw potatoes were investigated prior to any analysis. Upon studying the raw data, a variation in the decays for the five potato cultivars and dry matter bins was observed (not shown). The potato samples of the cultivar Berber were distinct from the rest of the cultivars showing a slower exponential decay. The difference was observed throughout the entire signal, and indicates a deviation in the composition and distribution of water compared to the other four cultivars. Within the five cultivars, the two storage times, November 1999 and May 2000, appeared different, where the storage time May 2000 showed a faster decay. This implies changes in the water distribution due to storage time.

5.1. Data analysis using SLICING

A SLICING model of the CPMG curves was computed. The NMR profile loadings for the optimal SLICING model, consisting of four factors, are shown

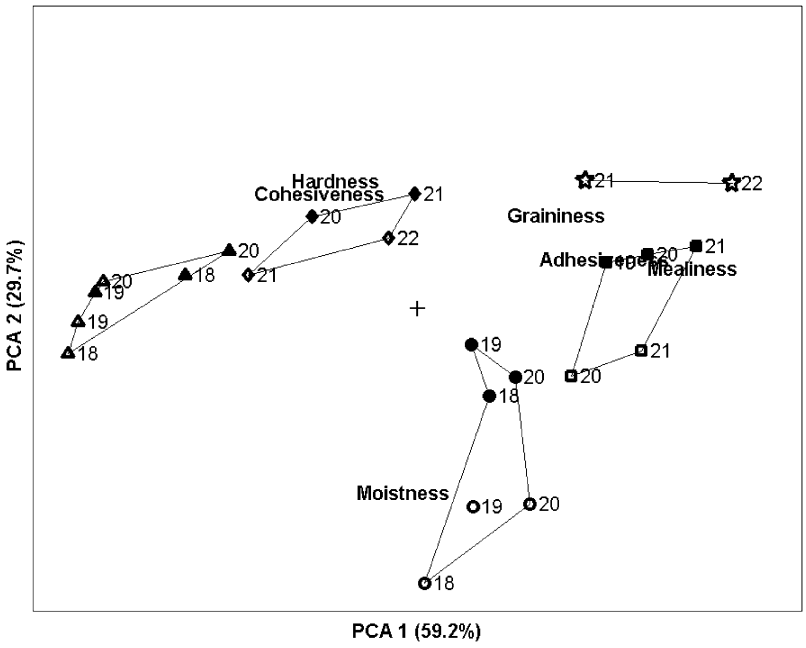


Fig. 2. Bi-plot from PCA on the sensory data. Ditta (◇), Sava (Δ), Bintje—low dry matter (□), Bintje—high dry matter (★), and Berber (○) for storage 1999 (open) and 2000 (filled). The numbers from 18 to 22 represent the % dry matter bins (see Table 1) where the range of the % dry matter bin is 18: 18.0–18.9, 19: 19.0–19.9, 20: 20.0–20.9, 21: 21.0–21.9 and 22: 22.0–22.9.

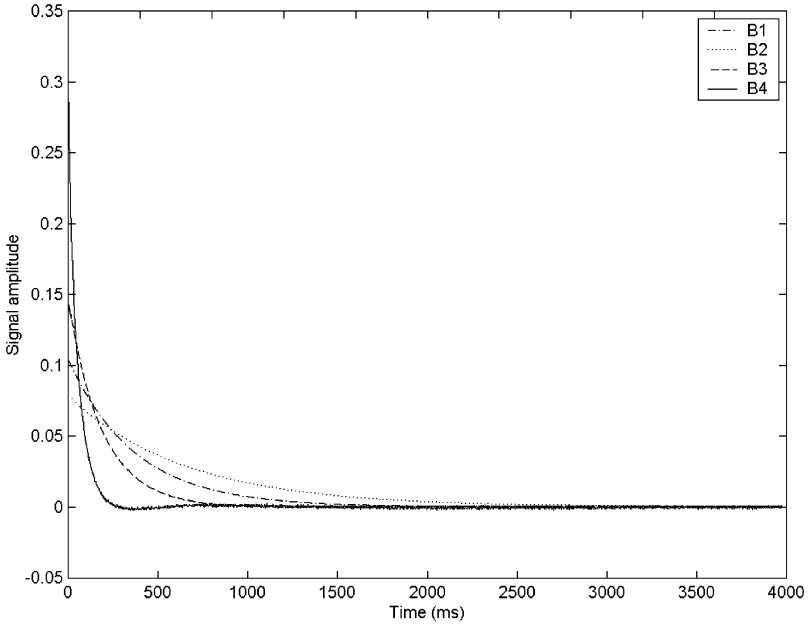


Fig. 3. Exponential loadings for components one to four from the PARAFAC model.

in Fig. 3. The four loadings are all exponentials as expected. This was further verified using a Monte Carlo approach, where 97% of 1000 randomly selected split-half tests resulted in the same four exponential loadings (Harshman & De Sarbo, 1994). In each split-half run, the data set was split into two parts, each part containing 12 and 11 samples, respectively. Both of these data sets were then modelled individually. Obtaining similar results from two such completely independent sets of data implies that the results are reproducible in a scientific sound way. I.e. the components are not merely an arbitrary result from a specific set of samples, but rather a fundamental property of all similar samples. This indicates that a valid estimate of the CPMG relaxation curves was derived from SLICING and hence the loadings could be associated with the water distribution in the potatoes. Previous studies have made use of bi-exponential fitting of the raw CPMG

relaxation curves to explain the different states of water in potatoes (Thygesen et al., 2001). The $T_{2,1}$ and $T_{2,2}$ relaxation times from bi-exponential of the raw CPMG relaxation curves using Eq. (1) ($N = 2$) are listed in Table 2, together with the uni-exponential fitting of the four loadings from the SLICING model.

The transversal relaxation times $T_{2,1}$ and $T_{2,2}$ from bi-exponential fitting for the 23 objects range from 130 to 180 and 430 to 540 ms, respectively. The relaxation times for the four exponential loadings in Fig. 3 show that the fourth loading has the fastest decay with a $T_{2,B4}$ of 52 ms followed by the third loading $T_{2,B3}$ of 192 ms ("B" indicating that these T_2 values are calculated from the B-loadings of PARAFAC). The $T_{2,B}$ for the first two exponential loadings are 378 and 646 ms, respectively. The results from the distribution analysis gave two peaks, where the first peak ranged from 56 to 82 ms and the second peak from 404 to 540 ms (not shown). A comparison of the four relaxation time constants showed that the relaxation times $T_{2,B}$ for the SLICING model span wider than the $T_{2,1}$ and $T_{2,2}$ from the bi-exponential fitting. Both the bi-exponential fitting and the distribution analysis have a peak around 480 ms, while the SLICING estimated decays at 378 and 646 ms. The first peak from the distribution analysis corresponds approximately with the fastest relaxing component from the SLICING. In the bi-exponential fitting, the fastest component lies in between the two fastest components from the SLICING.

By looking at the average residual over time for each of the three decomposition methods, it became clear that the residual from bi-exponential fitting was roughly four times larger than the residual from distribution analysis, which was twice as large as the average residual from the SLICING model. The reason may be caused by a smoothing constrain in the distribution algorithm. These observations imply that loadings derived from the SLICING model provide more information about the data than a simple bi-exponential fitting or a distribution analysis. In a four-factor SLICING model, the best description of the potato cultivars was given by the sample scores for factors 3 and 4 (the two fastest decays), where a clear distinction of the five cultivars was seen. This is shown in the score plot in Fig. 4, where the samples are marked due to cultivar, storage, and dry

Table 2
Overview of the transversal relaxation time (T_2)

Curves	Fitted against	T_2 (ms)
Exponential loadings (T_{2B}) from SLICING	T_{2B1}	378
	T_{2B2}	646
Fitted by uni-exponentials	T_{2B3}	192
	T_{2B4}	52
Raw curves fitted by bi-exponentials	$T_{2,1}$, fast decay	130–180
The range of the 23 potato samples	$T_{2,2}$, slow decay	430–540
Raw curves fitted by distribution analysis	$T_{2,1}$, fast decay	56–82
The range of the 23 potato samples	$T_{2,2}$, slow decay	404–540

$T_{2,B}$ represent the relaxation times of the uni-exponential fitting of the four relaxation slicing loadings and the $T_{2,1}$ and $T_{2,2}$ represent the relaxation times of the bi-exponential fitting of the raw CPMG relaxation curves. The raw curves show the range of the 23 potato samples.

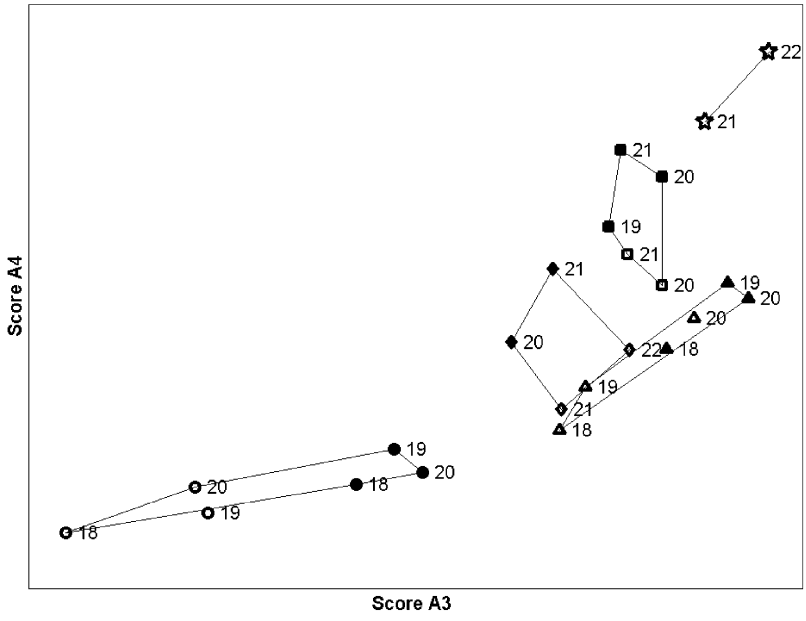


Fig. 4. Score plot of sample score 3 and 4 for the PARAFAC . Ditta (◇), Sava (Δ), Bintje—low dry matter (□), Bintje—high dry matter (★), and Berber (○) for storage 1999 (open) and 2000 (filled). The numbers from 18 to 22 represent the % dry matter bins (see Table 1) where the range of the % dry matter bin is 18: 18.0–18.9, 19: 19.0–19.9, 20: 20.0–20.9, 21: 21.0–21.9 and 22: 22.0–22.9.

matter bins. Within the cultivars the storage times for each bin are explained in the direction from the upper right corner to the lower left corner. The opposite direction from the left lower corner to the upper right corner describes the increase in dry matter bins within the same cultivar. The separation can be related to the diversity in the structure of the five cultivars and the varying water content and distribution of water (Table 1). There is no such clear distinction between the cultivars using the results from either the bi-exponential fitting or the distribution analysis (not shown).

Relaxation times can be related to the distribution of water within the samples. A high mobility of water makes it more available, and it will take a long time before it reaches the equilibrium state, giving rise to a high T_2. Thus, the highest T_2 value ($T_{2,B2}$) may reflect the water used for gelatinization and can be expected to be of major importance for texture differences. However, the variation in the potatoes is not captured by the two slower decaying loadings, indicating that this type of water is not important for the description of the differences in the five cultivars. The description of the cultivars in the SLICING scores 3 versus 4 indicates that the clear difference in the cultivars is caused by the distribution of the water with low mobility in the potato tubers. These low-mobile water components are assumed to describe the less mobile diffusion-hindered water hypothesized to be located in, e.g. the cell walls, entrapped in pectin, in sites with high ionic strengths, and in the vascular tissue (water transport tissue).

Several states and locations of water are possible within potatoes. Water compartments may be found in the cytoplasm in the cells and in the pectin network in the cell walls. Furthermore, very different tissue segments within a potato tuber exist. This makes the investigation of the distribution of water in potatoes very complex. Hills and Le Floc'h (1994) made a thorough study of the water in potatoes as they froze them down. Their study give an explanation to three of the four components found using SLICING. The first one is similar to a peak they find at about 50 ms, coming from water in cell walls, while the next two resembles peaks they find at around 200 and 400 ms, which they state is from water in the cytoplasm. However, they do not find any component higher than ca. 400 ms. Tang et al. (2000), on the other hand, found a peak at around 50 ms upon studying water saturated starch granules, so the exact cause of the fastest decaying component cannot be given. By the application of the SLICING a more direct method is introduced. This makes it possible to get a quick estimate of the parameters related to the quality, instead of high-cost laboratory analyses.

5.2. Regression models

PLSR has previously been used for the prediction of sensory attributes and potato quality from CPMG relaxation curves (Thybo et al., 2000; Thygesen et al., 2001). In this work six texture-related sensory attributes—hardness, cohesiveness, adhesiveness, mealiness, graininess, and moistness—of cooked potatoes were predicted using four different types of predictor variables. First, the CPMG PLSR was performed on the raw data set. The right number of components—four—was selected using the RMSECV values, the exponential loadings, and the exponential residuals as diagnostics. The second approach was the bi-exponential fitting predictions, which was based on the M_0 and T_2 values as independent variables in a PLSR model, and the third was the predictions using the results from distribution analysis. Two peaks were found from the distribution analysis, and the predictions were based upon the position and the amplitude of these two peaks. The last type of predictors was the four scores from the SLICING model referred to as SLICING prediction. The model complexity for prediction based on bi-exponential fitting and distribution analysis ranges from one to four components, depending on the sensory attribute being regressed. In Table 3, the RMSECV and correlation coefficients (predicted versus reference

Table 3
RMSECV and correlation coefficients (r) for CPMG PLSR prediction (PLSR), the bi-exponential fitting prediction (Bi-exp. fitting) models, prediction using the parameters from distribution analysis and SLICING on the six sensory attributes hardness, cohesiveness, adhesiveness, mealiness, graininess, and moistness

Sensory variables		PLSR[a]		Bi-exp. fit[b]		Distrib. anal.[b]		SLICING[a]	
Attributes	Range[c]	RMSECV	r	RMSECV	r	RMSECV	r	RMSECV	R
Hardness	4.9	1.19	0.69	1.22	0.67	1.53	0.33	1.15	0.69
Cohesiveness	5.7	1.10	0.78	1.01	0.83	1.74	0.29	1.31	0.71
Adhesiveness	4.7	1.27	0.58	1.01	0.73	1.14	0.64	1.27	0.57
Mealiness	7.4	1.49	0.74	1.26	0.83	2.00	0.48	1.33	0.79
Graininess	5.2	1.25	0.54	1.07	0.63	1.10	0.64	1.13	0.58
Moistness	7.0	1.11	0.76	0.86	0.87	0.71	0.91	1.05	0.79

[a] Four-factor models.
[b] Both M_0 and T_2 values used. Optimal regression results shown.
[c] Effective range on a scale from 0 to 15.

values) for the four regression models predicting the six sensory attributes are shown. The correlation coefficients are in the range of 0.29–0.91 and the RMSECV is between 0.71 and 2.00. The CPMG PLSR and the SLICING prediction show almost equal predicting performance, whereas the bi-exponential prediction in some cases gave a slightly better result. Distribution analysis gave the most varying results, ranging from the worst to the best predictions. In general, the six sensory attributes are not well predicted by any of the four methods except for the moistness attribute where the bi-exponential model gives a correlation coefficient of $r = 0.84$ and an acceptable RMSECV is observed. This is sensible as the relaxation curves express the water content and distribution within the potato starch cells, whereby the predictions indicate that this attribute was expressed in the CPMG relaxation curves. The correlation coefficient of the attributes cohesiveness and mealiness are also acceptable for all four methods, but taking into consideration the RMSECV and the range of the scale used by the assessors, the overall prediction is not impressive.

6. Conclusion

For the investigation of the differences in potatoes and potato texture by low-field NMR, this study compared new and established modelling methods to analyse NMR data: CPMG PLSR, bi-exponential fitting, distribution analysis, and SLICING. The work consists of two parts: a qualitative data analysis of the potato samples where the interpretation of the loadings was of special interest. Secondly regression analysis was performed using six sensory attributes as predictor variables.

In the data analysis part, the results show that the SLICING method is superior to CPMG PLSR, bi-exponential fitting, and distribution analysis. The SLICING method decomposed the CPMG relaxation curve into four uni-exponential components describing all the variation in the data set up to the noise. It is possible to interpret the exponential decaying loadings, and directly relate them to the design variables: cultivar, dry matter and storage time. The distinction between the five potato cultivars is caused by properties related to the fast decaying loadings, as the properties of water related to a long transversal relaxation time do not seem to have the same influence on the separation of the groups. To understand the role of the water component more research is required.

In the regression analysis the predictions from CPMG PLSR and the SLICING scores were very similar. There is no gain using PLSR on the raw curves if data analysis is of interest. The predictions using bi-exponential fitting gave slightly better results (RMSECV ranging from 0.86 to 1.26 and correlation coefficients ranging from 0.63 to 0.87) than the predictions using the CPMG PLSR or the SLICING (RMSECV ranging from 1.05 to 1.49 and correlation coefficients ranging from 0.54 to 0.79). The predictions using the results from the distribution analysis gave varying results, and in general these results are inferior to the results from bi-exponential fitting.

Acknowledgements

Povlsen wish to thank the STVF (Danish Research Council) for financial support of the project Applied Quality Monitoring in the Food Production Chain (AQM), while Rinnan wish to thank the STVF for financial support through Project 1179. Thanks to professor Rasmus Bro (KVL) for valuable discussions, to Ole Dahl Pedersen (DJF) for performing the NMR analysis, and to the sensory panel (DJF).

References

Andersen, C. M., & Rinnan, Å. (2002). Distribution of water in fresh cod. *Lebensmittel-Wissenschaft und-Technologie*, *35*, 687–696.

Andersson, C. A., & Bro, R. (2000). The N-way Toolbox for MATLAB. *Chemometrics and Intelligent Laboratory Systems*, *52*, 1–4.

Bro, R. (1997). PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, *38*, 149–171.

Burton, W. G. (1989). *The potato* (600pp.). New York, USA: Longman Scientific and Technical.

Butler, J. P., Reeds, J. A., & Dawson, S. V. (1981). Estimating solutions of first kind integral equations with nonnegative constraints and optimal smoothing. *SIAM Journal on Numerical Analysis*, *18*, 381–397.

Carr, H. Y., & Purcell, E. M. (1954). Effects of diffusion on free precession in nuclear magnetic resonance experiments. *Physical review*, *94*, 630–638.

Eastman, H. T., & Kranowski, W. J. (1982). Cross-validatory choice of the number of components from a principal component analysis. *Technometrics*, *24*, 73–77.

Gould, W. A. (1999). *Potato, Production, Processing, and Technology* (259pp.). Maryland, USA: CTI Publications Inc.

Harshman, R. A., & De Sarbo, W. S. (1994). An application of PARAFAC to a small sample problem, demonstrating preprocessing, orthogonality constraints and split-half diagnostic techniques. In H. G. Law, C. W. Snyder, J. A. Hattie, & R. P. McDonald (Eds.), *Research methods for multimode data analysis* (pp. 1–2). New York: Praeger.

Hemminga, M. A. (1992). Introduction to NMR. *Trends in Food Science and Technology*, *3*, 179–186.

Hills, B. P., & Le Floc'h, G. (1994). NMR studies of non-freezing water in cellular plant tissue. *Food Chemistry*, *51*, 331–336.

Hills, B. P., Goncalves, O., Harrison, M., & Godward, J. (1997). Real time investigation of the freezing of raw potato by NMR microimaging. *Magnetic Resonance in Chemistry*, *35*, 29–36.

Martens, H., & Næs, T. (1989) *Multivariate calibration* (419pp.). New York, USA: Wiley.

Meiboom, S., & Gill, D. (1958). Modified spin-echo method for measuring nuclear relaxation times. *Review of Scientific Instruments*, *29*, 688–691.

Pedersen, H. T., Bro, R., & Engelsen, S. B. (2001). SLICING—A novel approach for unique deconvolution of NMR relaxation decays. In G. A. Webb, P. S. Belton, A. M. Gill, & I. Delgadil (Eds.), *Magnetic resonance in food science: A view to the future* (pp. 202–209). Cambridge, MA: Royal Society of Chemistry.

Pedersen, H. T., Munck, L., & Engelsen, S. B. (2000). Low-field H-1 nuclear magnetic resonance and chemometrics combined for simultaneous determination of water, oil, and protein contents in oilseeds. *Journal of the American Oil Chemists Society*, *77*, 1069–1076.

Ruan, R. R., & Chen, P. L. (1998). *Water in foods and biological materials. A nuclear magnetic resonance approach* (298pp.). Lancaster, USA: Technomic Publishing Co. Inc.

Ruan, R. R., Zou, C., Wadhawab, C., Martinez, B., Chen, P. L., & Addis, P. (1997). Studies of hardness and water mobility of cooked wild rice using nuclear magnetic resonance. *Journal of Food Processing and Preservation*, *21*, 91–104.

Seow, C. C., & Teo, C. H. (1996). A comparative study by firmness and pulsed NMR measurements. *Starch/Stärke*, *3*, 90–93.

Tang, H. R., Belton, P. S., Ng, A., Waldron, K. W., & Ryden, P. (1999). Solid state H-1 NMR studies of cell wall materials of potatoes. *Spectrochimica Acta, Part A—Molecular and Biomolecular Spectroscopy*, *55*, 883–894.

Tang, H. R., Godward, J., & Hills, B. (2000). The distribution of water in native starch granules—a multinuclear NMR study. *Carbohydrate Polymers*, *43*, 375–387.

Thybo, A. K., Bechmann, I. E., Martens, M., & Engelsen, S. B. (2000). Prediction of sensory texture of cooked potatoes using uniaxial compression, near infrared spectroscopy and low field H-1 NMR spectroscopy. *Lebensmittel-Wissenschaft und-Technologie*, *33*, 103–111.

Thybo, A. K., & Martens, M. (1999). Instrumental and sensory characterization of cooked potato texture. *Journal of Texture Studies*, *30*, 259–278.

Thygesen, L. G., Thybo, A. K., & Engelsen, S. B. (2001). Prediction of sensory texture quality of boiled potatoes from low-field 1H NMR of raw potatoes. The role of chemical constituents. *Lebensmittel-Wissenschaft und-Technologie*, *34*, 469–477.

Ulrich, D., Hoberg, E., Neugebauer, W., Tiemann, H., & Darsow, U. (2000). Investigation of the boiled potato flavor by human sensory and instrumental methods. *American Journal of Potato Research*, *77*, 111–117.

VanMarle, J. T., DeVries, R. V. D. V., Wilkinson, E. C., & Yuksel, D. (1997). Sensory evaluation of the texture of steam-cooked table potatoes. *Potato Research*, *40*, 79–90.

Windig, W., & Antalek, B. (1997). Direct exponential curve resolution algorithm (DECRA): A novel application of the generalized rank annihilation method for a single spectral mixture data set with exponentially decaying contribution profiles. *Chemometrics and Intelligent Laboratory Systems*, *37*, 241–254.

PAPER III

Jakob Christensen, Vibeke Tølbøl Povlsen and John Sørensen

Application of Fluorescence Spectroscopy and Chemometrics in the Evaluation of Processed Cheese During Storage.
JOURNAL OF DAIRY SCIENCE, 86 (4), 1101-1107, 2003

J. Dairy Sci. 86:1101–1107

Application of Fluorescence Spectroscopy and Chemometrics in the Evaluation of Processed Cheese During Storage

J. Christensen*, V. T. Povlsen*, and J. Sørensen†
*Food Technology, Department of Dairy and Food Science,
The Royal Veterinary and Agricultural University,
Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark
†Arla Foods Innovation, Søderupvej 26, DK-6920 Videbæk, Denmark

ABSTRACT

Front face fluorescence spectroscopy is applied for an evaluation of the stability of processed cheese during storage. Fluorescence landscapes with excitation from 240 to 360 nm and emission in the range of 275 to 475 nm were obtained from cheese samples stored in darkness and light in up to 259 d, at 5, 20 and 37°C, respectively. Parallel factor (PARAFAC) analysis of the fluorescence landscapes exhibits four fluorophores present in the cheese, all related to the storage conditions. The chemometric analysis resolves the fluorescence signal into excitation and emission profiles of the pure fluorescent compounds, which are suggested to be tryptophan, vitamin A and a compound derived from oxidation. Thus, it is concluded that fluorescence spectroscopy in combination with chemometrics has a potential as a fast method for monitoring the stability of processed cheese.
(**Key words:** cheese, chemometrics, fluorescence spectroscopy, PARAFAC)

Abbreviation key: GC-MS = gas chromatography-mass spectrometry, **PARAFAC** = parallel factor analysis.

INTRODUCTION

The development of undesirable flavor caused by lipid oxidation and nonenzymatic browning are critical quality factors during storage of processed cheese. The deterioration of the cheese product is dependent on the handling in the post manufacturing processes. Since cheese mainly consists of protein, fat, minerals and water, oxidation is reflected in the composition of these constituents. Monitoring the changes in structure and composition of the cheese constituents, especially protein and fat, will help understand the effect of stress factors during storage. Common stress factors in the distribution retails and production are light exposure and varying temperature, which can result in reduced shelf life partly due to increased formation of free radicals. Therefore, processed cheese samples stored under different light and heat conditions are investigated in the present study.

Many methods have been developed to shed light on the degree of oxidation of dairy products, a process that consists of several stages. The early stage of lipid oxidation can form hydroperoxides, which normally are measured by HPLC or by evaluation of the peroxide value (Emmons et al., 1986). Secondary oxidation products can be analyzed by static or dynamic headspace GC-MS (Sunesen et al., 2002) or methods using thiobarbituric acid (Kristensen et al., 2001). Methods based on electron spin resonance spectrometry were recently suggested for monitoring the formation of radicals during the oxidation of processed cheese (Kristensen and Skibsted, 1999). All these methods for evaluation of the oxidative levels of dairy products have in common, that they are destructive and time consuming. In this study, the potential of front face fluorescence, measured directly on the cheese surface were investigated, as an alternative, fast and nondestructive method. Theoretically the potential of fluorescence seems sound, since the cheese product contains well known fluorescent compounds in form of aromatic amino acids, vitamin A and riboflavin (Duggan et al., 1957), which all have been reported to be affected during structural changes in cheese (Dufour et al., 2001) or during light and heat exposure (Kristensen et al., 2001; Whited et al., 2002; Wold et al., 2002).

Fluorescence spectroscopy is a sensitive, rapid and noninvasive analytical technique that can provide information on the presence of fluorescent molecules and their environment in all sorts of biological samples. The development and improvement of chemometric methods (Bro, 1996; Bro, 1997; Andersson and Bro, 2000) combined with the technical and optical development of spectrofluorometers have in recent years increased the possibilities for the use of fluorescence spectroscopy.

Received September 2, 2002.
Accepted October 25, 2003.
Corresponding author: Jakob Christensen; e-mail: jach@kvl.dk.

Thus, online monitoring sensors that enable measurements of complete excitation emission spectra (fluorescence landscapes) are now commercially available.

In the last years, a few studies have focused on the potential of using front face fluorescence of dairy products without any pretreatment of the samples. Previously heat treatment and structural changes during coagulation have successfully been investigated in milk using fluorescence spectroscopy (Dufour and Riaublanc, 1997; Birlouez-Aragon et al., 1998; Herbert et al., 1999). Changes in fat and protein composition and structure have been characterized by the means of measuring the tryptophan and vitamin A fluorescence of cheeses during ripening (Dufour et al., 2000; Mazerolles et al., 2001) and for identification of different cheeses at a molecular level (Dufour et al., 2000; Herbert et al., 2000). Wold et al. (2002) demonstrated the potential of fluorescence spectroscopy for measuring the light-induced oxidation, ascribed to the photodegradation of riboflavin.

Common to all these studies is that basic chemometric tools like Principal Component Analysis and Partial Least Squares Regression are applied for the evaluation of single excitation or emission fluorescence spectra. The multivariate approach increases the extracted information and is very useful when handling the fluorescence signal of complex food products. Even more information can be obtained, if the fluorescence measurements are not limited to single emission or excitation spectra. The possibilities when measuring whole fluorescence landscapes (excitation emission matrices) will be investigated here. New chemometric methods (Andersson and Bro, 2000) make it possible to handle fluorescence landscapes keeping the 2-dimensional data structure of each measurement. The techniques are known as N-way or multiway chemometrics, and in the case of fluorescence signals, a 3-way (samples × excitation × emission) data analysis is an obvious choice. The advantage of the multiway analysis is that one can utilize the original and true structure in data, which can stabilize the decomposition of the data, and potentially increase the interpretability (Bro, 1996; Bro, 1997).

In the present study Parallel Factor analysis (**PARAFAC**) (Bro, 1997) is applied on the fluorescence landscapes of processed cheese exposed to light and varying temperature during storage. PARAFAC analysis of fluorescence data is previously used with success on model system of mixtures of fluorophores and in other food applications like sugar and fish (Bro, 1999; Baunsgaard et al., 2000a; Baunsgaard et al., 2000b; Pedersen et al., 2002) to investigate the present fluorescent compounds in complex matrices. PARAFAC is based on the decomposition of the fluorescence data represented in a three-way array, into a few spectral loadings expressing the common structure of the data. The feature of PARAFAC is that the retrieved loading spectra can be directly related to the original fluorescence characteristics of the present fluorophores, which means that the emission and excitation maximum of the loadings can be used in the interpretation and identification of the fluorophores (Bro, 1997).

Thus, the overall objective of the present investigation is to use multivariate analysis on fluorescence spectra keeping the 3-dimensional structure and extract information about the product at hand regarding age and storage conditions. This is pursued by using a non-destructive and rapid high-sensitive fluorescence method, which is simple to perform, and does not involve sample preparation.

MATERIALS AND METHODS

Processed Cheese: Product and Storage Conditions

The product and storage conditions are identical to the experimental plan used by Kristensen et al., 2001. A batch of processed cheese spread samples (density approximately 1.1 g/mL) with 65% fat in dry matter was obtained from Arla Foods amba, Denmark. The processed cheese was produced according to standard production of processed cheese and was constituted of bovine milk, starter culture, salt and emulsifier. After production the product was filled without any headspace (140 g) in transparent glass containers and sealed with a metal lid. The samples were stored for 10 months at three temperatures 5, 20, and 37°C and were exposed by placing the samples at a distance of approx. 55 cm from a fluorescent lamp or protected from light by wrapping the glass container in tin foil. The light source was fluorescent tubes (Phillips TLD 18/83 W) with a light intensity of 2000 lx as measured by a Topcon IM-1 illumination meter (Tokyo Kogaku Kikai K.K.). Samples were taken out at the beginning of the experiment and then after 14, 28, 56, 84, 112 and 256 d. Only the 1 cm outer layer which had been in contact with the wall of the containers were used and each of the samples were taken from the glass jars by breaking the original seal prior to freezing at –80°C. The samples were frozen for a year before being thawed. Two cheese samples from each treatment were withdrawn for each analysis time.

Fluorescence Spectroscopy and Sampling

All samples were measured on a Perkin-Elmer LS 50B spectrometer equipped with a Front Surface Accessory and controlled with FLDM software. The stored cheese samples were mixed thoroughly before spread-

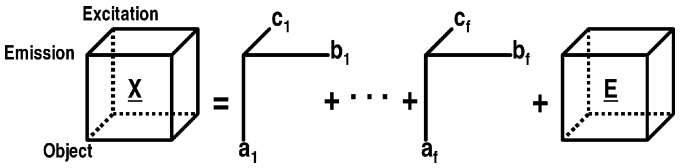


Figure 1. Illustration of the decomposition scheme into *f* number of components of the PARAFAC model for the data array **X**. The cube **E** represents the residual.

ing directly onto the quartz window of a powder cell, which was then assembled and placed in the light path in an angle of around 60°. The spectral range of the experiment was selected upon an exploratory basis. A preliminary investigation measuring excitation wavelengths from 200 to 600 nm, and emission wavelengths from 220 to 800 nm on different cheeses were performed, and resulted in focusing on excitation wavelengths in the UV region. Strong fluorescence signals were obtained from the cheese samples in this area, leaving no signal from higher excitation and emission wavelengths when using this technique and set-up. The selected spectral range of the excitation wavelength was 240 to 360 nm with 20 nm intervals. Emission was obtained for every nm from 275 to 475 nm. The slit width was 6 nm for excitation and 5 nm for the emission and a 1% attenuation filter was used.

It should be noted that the selected spectral range does not cover riboflavin fluorescence, which exhibit emission around 520 nm (Duggan et al., 1957), despite it would be an obvious compound to monitor throughout storage. However, the preliminary studies on cheese samples showed that no detectable signal was obtained in this spectral area when using the described measuring set-up.

Data Analysis—PARAFAC

PARAFAC decomposes the fluorescence spectra, into tri-linear components according to the number of fluorophores present the cheese samples (objects). The number of fluorophores present in the samples is equal to the minimal number of factors ($f = 1, \ldots, F$) needed to describe the fluorescence matrix **X**.

A graphical illustration of the decomposition of the data array **X** is given in Figure 1. The object mode is expressed by the A-scores ($a_1, \ldots, a_f$) and the two spectral loadings excitation and emission are expressed as B loadings ($b_1, \ldots, b_f$) and C loadings ($c_1, \ldots, c_f$), respectively. The loadings in a spectral bilinear decomposition reflect the pure spectra of the fluorophores and the true underlying spectra can be recovered in the single components.

The principle behind the PARAFAC decomposition is to minimize the sum of squares of the residual e_{ijk}, see Equation 1.

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

$$(i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K; f = 1, \ldots, F)$$

The element x_{ijk} represents the raw fluorescence excitation/emission spectra (**X**) of the stored cheese, where i is the number of measured samples, j is the number of excitation wavelengths, k is the number of emission wavelengths and f is the number of factors. $a_{.f}$ is the object score (magnitude of the fluorophore) for factor f (first mode), $b_{.f}$ is the excitation loading for factor f (second mode), and loading $c_{.f}$ express the emission spectra (third mode). e_{ijk} is the residual (**E**) and contains the variation not captured by the PARAFAC model (Bro, 1997). Split half analysis is suggested for validation of PARAFAC models by Bro (1997). The idea of this strategy is to divide the data set into two halves and make a PARAFAC model on both halves. Due to the uniqueness of the PARAFAC model one will obtain the same result—same loadings in the nonsplitted mode e.g., excitation and emission mode—on both datasets, if the correct number of components is chosen.

Calculating the PARAFAC Model

The following sampling was performed: 45 samples × 2 replicates × 2 repetitions = 190 samples. Seven samples were removed, as they were considered to be spectral outliers based on a preliminary data inspection and resulted in a total of 183 samples. The preliminary PARAFAC modelling indicated that nonnegativity constraints on all three modes (samples, excitation, and emission) were necessary. Validation of the PARAFAC modelling was performed with split half test, based on replicated samples, i.e. not splitting of the repetitions.

In addition to the split-half experiment, the residuals were inspected, and the results were judged, interpreted and compared with external knowledge.

All calculations were performed in Matlab version 6.1 (MathWorks, Inc.) with the N-way Toolbox (Andersson and Bro, 2000) and the PLS Toolbox (www.Eigenvector.com).

RESULTS AND DISCUSSIONS

The fluorescence landscapes of two cheese samples are shown in Figure 2. The two samples represent the extremes in the experimental plan, i.e., a fresh cheese sample (a) and a cheese sample stored under the most severe conditions (b). The highest fluorescence peak for

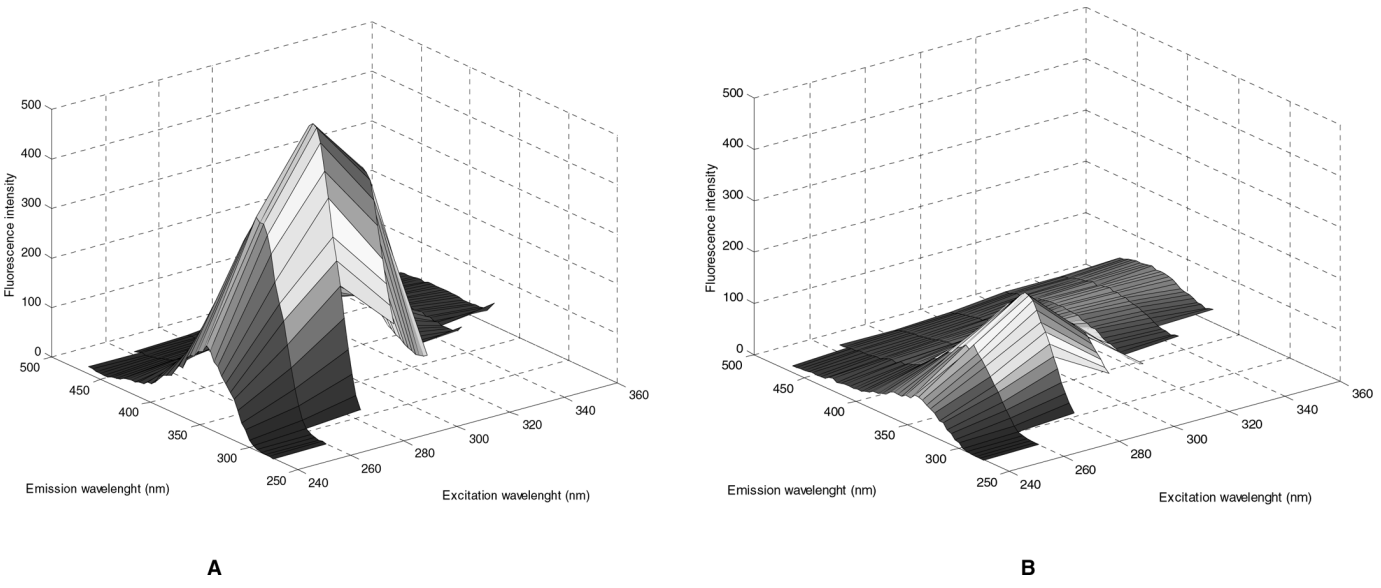


Figure 2. Three dimensional plot of fluorescence landscapes of processed cheese samples. a) fresh cheese sample, and b) cheese sample stored in 259 d at 37°C exposed to light.

both samples is seen with excitation around 280 nm and emission around 350 nm, with a significant higher and apparently broader signal from the fresh cheese. The excitation and emission characteristics indicate that the fluorescence peak corresponds to tryptophan fluorescence, which is reported to have excitation/emission wavelength maximum at 285/365 nm in pure solutions (Duggan et al., 1957), and previously measured in cheese products with excitation 290 nm and emission from 305–400 nm (Herbert et al., 2000; Dufour et al., 2001; Mazerolles et al., 2001). Apart from this major peak, a vague peak is observed in the higher wavelength region with excitation around 320 to 360 nm and emission round 400 to 460 nm, especially for the cheese sample stored for 259 d.

The aforementioned patterns in the fluorescence landscapes were investigated further by the use of PARAFAC analysis with the objective to resolve the fluorescence signal into the contributions of each of the fluorescent compounds present in the set of samples, i.e. estimate the excitation and emission profiles of fluorophores directly from the three-dimensional fluorescence landscapes. PARAFAC models of the fluorescence data were estimated with one to five components, but the four-component model was chosen based on split half analysis (Bro, 1997). A high explained variation of 99.76% is captured by the PARAFAC model, and the resulting PARAFAC components are shown in Figure 3. The model indicates that four different fluorophores are present in the cheese samples with the excitation and emission profiles shown in the figure. The excitation/emission maximum for the two compounds are 300/347 nm and 280/339 nm, respectively, as listed in Table 1. The loading profiles of the second PARAFAC component corresponds quite well with the characteristics of tryptophane, whereas the excitation maximum of the first component seems a little too high for tryptophan. Having the rather low resolution of 20 nm in the excitation mode in mind, and knowing that the fluorescence properties of protein-bound amino-acids are known to be affected by the structure of protein (Lakowicz, 1999), we dare to suggest that the first PARAFAC components is also due to tryptophan fluorescence, but simply shifted due to inclusion to different protein structures.

The score values in the first column of Figure 3 represent the concentration mode for each of the fluorophores, and since the excitation and emission loadings are normalized when calculating the PARAFAC model, the contribution for each of the components can be compared to the overall variation based on the level of the scores. The score values are arranged so the development of the fluorophores easily can be caught throughout the storage time. Looking at the two proposed tryptophan components, a significant decrease is observed throughout the storage period for the samples stored at 37°C. This shows that alterations in the protein structure, monitored by the decrease in tryptophan fluorescence, somehow can reflect the conditions of the cheese samples during storage. The samples exposed to light during storage show a systematically higher tendency to be degraded throughout the storage than the samples stored in the dark. Compared with the

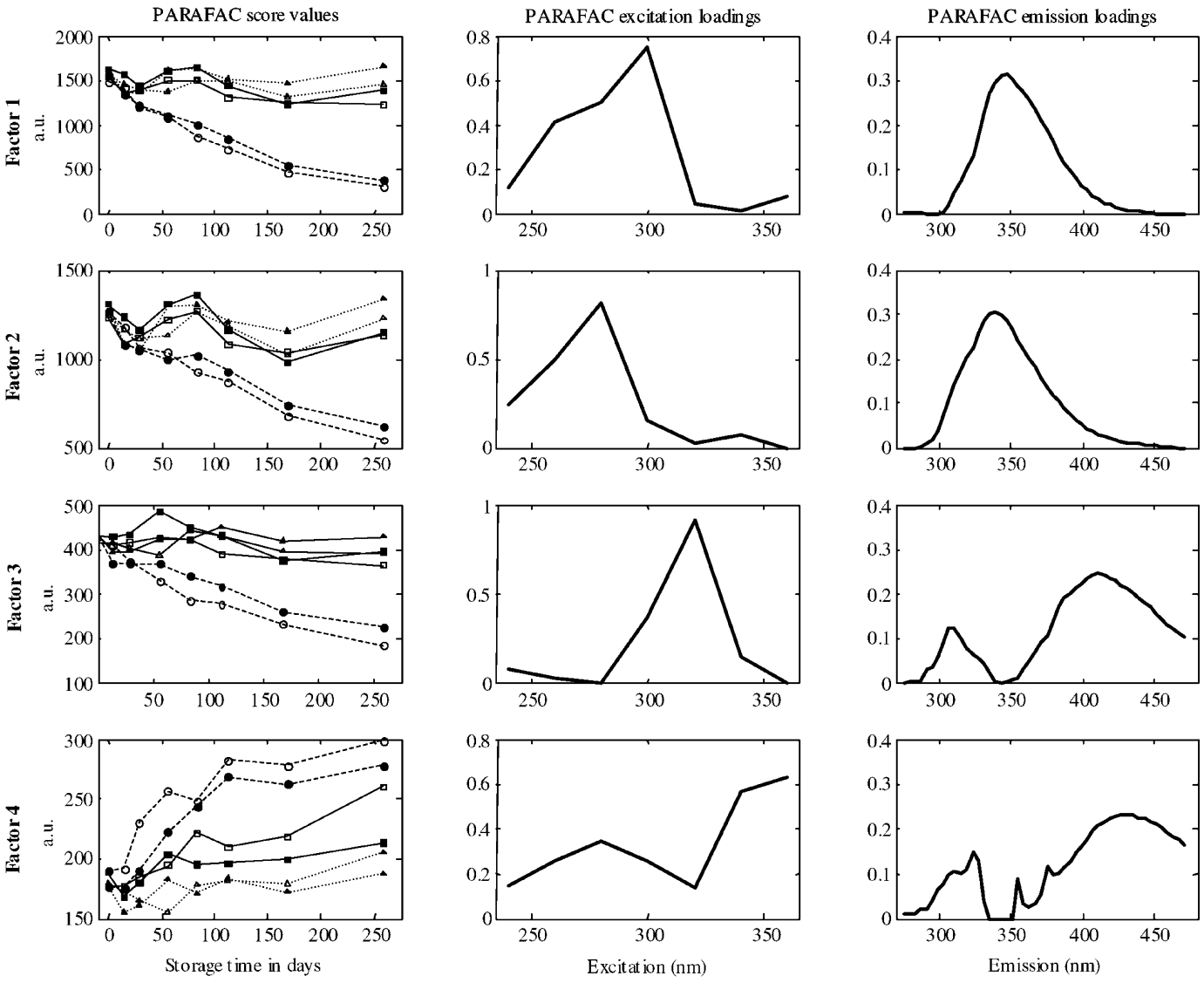


Figure 3. A- (scores), B- (excitation), and C- (emission) loadings of a four component PARAFAC model, based on the fluorescence landscapes of 183 processed cheese samples. Samples stored at 5°C are indicated with triangles, and connected with dotted line (·····). Samples stored at 20 and 37°C are shown with squares and dashed line (– – –) and circles connected with a full line (——), respectively. Open signs represent samples stored in light, and filled sign illustrates storage in darkness.

effect of different temperatures, the light exposure seems negligible for the two tryptophan components, though. The same storage experiment showed a similar tendency of light exposure having little, if any influence on the browning of cheese (Kristensen et al., 2001), and thereby indicate that the observed differences in the protein structure are somehow related to the browning reaction i.e. the formation of Maillard products from the protein and lipid oxidation products in cheese, even though tryptophan itself may not be part of the browning reaction scheme. As indicated by the first visual inspection of the fluorescence landscapes, the level of the score values for the two first components are much higher than the third and fourth component, simply showing that the development in the tryptophan signal represents the major variation in the fluorescence data.

The development of the third estimated fluorophore (score values of the third PARAFAC component) shows a similar pattern as the decrease in the tryptophan signal. Thus, the cheese samples stored at 37°C contain less of this component throughout the storage, especially the samples exposed to light during storage. Comparing the fluorescence profiles seen in Figure 3 and the excitation/emission wavelength maximum of 320/411 nm (Table 1) with observed maximum of 325/470 nm in pure solution reported (Duggan et al., 1957) and

Table 1. Excitation and emission maximum of four components in the PARAFAC model of 45 different cheese samples.

Component	λ_{max} (nm) Excitation	Emission
1	300	347
2	280	339
3	320	411
4	360	431

322/412 nm for dairy products (Dufour et al., 2001; Herbert et al., 2000), vitamin A is an obvious suggestion for the third component. This is underlined by the fact, that the observed decrease in vitamin A fluorescence signal throughout the storage period corresponds well to reported vitamin A degradation during light exposure in dairy products (Whited et al., 2000).

The fourth PARAFAC component reveals an opposite and very interesting trend in the score values, as seen in Figure 3. The level of this fluorophore increases throughout the storage period, especially for the cheese samples stored in light. The excitation and emission loadings look somewhat noisier with several small peaks, probably caused by the fact that the fluorescence signal is very low, as can be seen from the levels of the score values. Scattering effects might be the reason for the extra emission peak observed around 320 nm for both the third and the fourth component. The identification of the fourth fluorescent compound, showing an increase signal during the oxidation of the cheese samples, give rise to more doubt. Taking the increasing concentration of the fourth component throughout storage in consideration, it is obvious to suggest that the fourth component can be attributed to some kind of oxidation product. So-called "Advanced Maillard Products" in milk samples have been reported (Birlouez-Aragon et al., 1998) to excite around 350 nm with emission at 440 nm, which is almost identical to the peak observed in the fourth component. Another suggestion could be that the fourth component is a secondary oxidation product developed when carbonyl compounds produced by lipid oxidation interacts, as reported by Dillard and Tappel (1971) with a fluorescent compound from lipid peroxidation with excitation maximum at 360 nm, and emission maximum at 430 nm, which is even closer to the fluorescent characteristics of the fourth component. Finally, Stapelfeldt and Skibsted (1994), demonstrated that the reaction between secondary lipid oxidation products from milk products and β-lactoglobulin in a model system yielded a fluorescent condensation products with excitation/emission maximum at 350/410 nm, which could also form an educated guess for identification of the fourth PARAFAC component.

CONCLUSIONS

This exploratory study of processed cheese demonstrates the potential of fluorescence spectroscopy and chemometrics applied to the analysis of dairy products. The rapid fluorometric analysis reveals information at a molecular level about the stability of the cheese when exposed to manufacture handling stress like light and temperature changes. PARAFAC analysis provides a unique mathematical decomposition of four fluorescent compounds present in the cheese samples all showing a change in the fluorescence signal corresponding the storage time and the grade of oxidation.

The fluorescent signal from the processed cheese samples is suggested to derive from tryptophan, vitamin A and an oxidation product. Thus, the suggested analytical method provides a fast and simultaneous determination of the fluorescence level of all these compounds. The observed results still remain to be validated with chemical reference analyses in order to proof the identification of the fluorophores, but this investigation certainly underlines the potential of fluorescence spectroscopy in combination with chemometrics, as a fast, nondestructive innovative method, that can be applied to dairy products for monitoring oxidation, screening studies and perhaps in development of new fast quantitative analyses of vitamin A.

ACKNOWLEDGMENTS

This study is partly sponsored by an EU innovation programme, OPUS (IN30905 I), with the aim to adapt and apply advanced fluorescence measurement techniques for online analysis to traditional food industries. Povlsen wish to thank FØTEK 3 (93S-2444-Å01-00100) for financial support. Furthermore the authors wish to thank Joanna Zeppelin and Vibeke Birk from Arla Foods for making this publication possible, and for carrying out all the experimental work. Finally we are grateful to Professor Rasmus Bro for valuable discussion.

REFERENCES

Andersson, C. A., and R. Bro. 2000. The N-way Toolbox for MATLAB. Chemometrics Intell. Lab. Syst. 52:1–4.

Baunsgaard, D., C. A. Andersson, A. Arndal, and L. Munck. 2000a. Multi-way chemometrics for mathematical separation of fluorescent colorants and colour precursors from spectrofluorimetry of beet sugar and beet sugar thick juice as validated by HPLC analysis. Food Chem. 70:113–121.

Baunsgaard, D., L. Munck, and L. Norgaard. 2000b. Analysis of the effect of crystal size and color distribution on fluorescence measurements of solid sugar using chemometrics. Appl. Spec. 54:1684–1689.

Birlouez-Aragon, I., M. Nicolas, A. Metais, N. Marchond, J. Grenier, and D. Calvo. 1998. A rapid Fluorimetric Method to estimate the heat treatment of liquid Milk. Int. Dairy J. 8:771–777.

Bro, R. 1996. Multi-way Calibration Multi-Linear PLS. J. Chemometrics. 10:47–62.
Bro, R. 1997. PARAFAC. Tutorial and applications. Chemometrics Intell. Lab. Syst. 38:149–171.
Bro, R. 1999. Explorative study of sugar production using fluorescence spectroscopy and PARAFAC analysis. Chemometric Intell. Lab. Syst. 46:133–147.
Dillard, C. J., and A. L. Tappel. 1971. Fluorescent Products of Lipid Peroxidation of Mitochondria and Microsomes. Lipids. 6:715–721.
Dufour, E., M. F. Devaux, and S. Herbert. 2001. Delineation of the structure of soft cheeses at the molecular level by fluorescence spectroscopy—relationship with texture. Int. Dairy J. 11:465–473.
Dufour, E., G. Mazerolles, M. F. Devaux, G. Duboz, M. H. Duployer, and N. M. Riou. 2000. Phase Transition of triglycerides during semi-hard cheese ripening. Int. Dairy J. 10:81–93.
Dufour, E., and A. Riaublanc. 1997. Potentiality of spectroscopic methods for the characterisation of dairy products. I. Front-face fluorescence study of raw, heated and homogenised milks. Lait 77:657–670.
Duggan, D. E., R. L. Bowman, B. B. Brodie, and S. Udenfriend. 1957. A Spectrophotofluorometric Study of Compounds of Biological Interest. Arch. Biochem. and Biophys. 16:1–14.
Emmons, D. B., G. J. Paquettte, D. A. Froehlich, D. C. Beckett, H. W. Modler, G. Butler, Brackenbridge, and G. Daniels. 1986. Oxidation of butter by low intensities of fluorescent light in relation to retail stores. J. Dairy Sci. 69:2437–2450.
Herbert, S., A. Riaublanc, B. Bouchet, D. J. Galllant, and E. Dufour. 1999. Fluorescence Spectroscopy Investigation of Acid- or Rennet-Induced Coagulation of Milk. J. Dairy Sci. 82:2056–2062.
Herbert, S., N. M. Riou, M. F. Devaux, A. Riaublanc, B. Bouchet, D. J. Galllant, and E. Dufour. 2000. Monitoring the identity and the structure of soft cheeses by fluorescence spectroscopy. Lait. 80:621–634.
Kristensen, D., E. Hansen, A. Arndal, R. A. Trinderup, and L. H. Skibsted. 2001. Influence of light and temperature on the colour and oxidative stability of processed cheese. Int. Dairy J. 11:837–843.
Kristensen, D., and L. H. Skibsted. 1999. Comparison of Three Methods Based on Electron Spin Resonance Spectrometry for Evaluation of Oxidative Stability of Processed Cheese. J. Agric. Food Chem. 47:3099–3104.
Lakowicz, J. R. Protein fluorescence. Pages 446–485 in Principles of fluorescence spectroscopy. Kluwer Academic/plenum Publishers, New York.
Mazerolles, G., M. F. Devaux, G. Duboz, M. H. Duployer, N. M. Riou, and E. Dufour. 2001. Infrared and fluorescence spectroscopy for monitoring protein structure and interaction changes during cheese ripening. Lait. 81:509–527.
Pedersen, D. K., L. Munck, and S. B. Engelsen. 2002. Screening for dioxin contamination in fish oil by PARAFAC and N-PLSR analysis of fluorescence landscapes. J. Chemometrics. In press.
Stapelfeldt, H., and L. H. Skibsted. 1994. Modification of β-lactoglobulin by aliphatic aldehydes in aqueous solution. J. Dairy Res. 61:209–219.
Sunesen, L. O., P. Lund, J. Sørensen, and G. Hølmer. 2002. Development of volatile compounds in processed cheese during storage. Lebensmittel-Wissenschaft Und-Technologie. 35:128–134.
Whited, L. J., B. H. Hammond, K. W. Chapman, and K. J. Boer. 2002. Vitamin A Degradation and Light-Oxidized Flavor Defects in Milk. J. Dairy Sci. 85:351–354.
Wold, J. P., K. Jørgensen, and F. Lundby. 2002. Nondestructive Measurement of Light-induced Oxidation in Dairy Products by Fluorescence Spectroscopy and Imaging. J. Dairy Sci. 85:1693–1704.

PAPER IV

Vibeke Tølbøl Svensson, Henrik Hauch Nielsen and Rasmus Bro

Determination of the protein content in brine from salted herring using near-infrared spectroscopy.

LEBENSMITTEL-WISSENSCHAFT UND-TECHNOLOGIE-FOOD SCIENCE AND TECHNOLOGY, 37 (7), 803-809, 2004

ELSEVIER

Lebensm.-Wiss. u.-Technol. 37 (2004) 803–809

lwt

www.elsevier.com/locate/lwt

Research note

Determination of the protein content in brine from salted herring using near-infrared spectroscopy

Vibeke T. Svensson[a,*], Henrik Hauch Nielsen[b], Rasmus Bro[a]

[a] *Department of Food Science, Food Technology, The Royal Veterinary and Agricultural University, Rolighedsvej 30, Frederiksberg C DK-1958, Denmark*

[b] *Department of Seafood Research, Danish Institute for Fisheries Research, Ministry of Food, Agricultural and Fisheries, Blgd 121, Technical University of Denmark, Lyngby DK-2800, Denmark*

Received 19 September 2003; received in revised form 8 March 2004; accepted 11 March 2004

Abstract

Near-infrared reflectance (NIR) spectroscopy in the spectral range of 1000–2500 nm, was measured directly on brine from barrel salted herring, to investigate the potential of NIR as a fast method to determine the protein content. A principal component analysis performed on the NIR spectra shows two groups, separating the first 100 days of storage from the storage time exceeding 100 days. A partial least-squares regression model between selected regions of the NIR spectra and the protein content yields a correlation coefficient of 0.93 and a prediction error (RMSECV) of 0.25 g/100 g. The results clearly indicate that NIR spectroscopy has a potential as a fast and noninvasive method for assessing the protein content in brine from barrel salted herring, which again may be used as an indicator for the ripening quality of barrel salted herring.

Keywords: Barrel salted herring; Ripening; Protein; Multivariate calibration; NIR spectroscopy

1. Introduction

Barrel salted herring is an important product in the Nordic fishery industry and the manufacturing process is bound by tradition, based on human knowledge and experience (Voskrensensky, 1965). Scientific knowledge about the ripening process is still limited despite the progress that has been achieved by a large research effort in the Nordic and European countries. A great interest lies in obtaining a better understanding of the unique taste and texture development that occurs during the several months of ripening period. During storage, degradation of protein takes place and is believed to be one of the main factors reflecting the ripening of salted herring. It is known that both digestive and muscle proteases participate in the proteolytic degradation of the muscle proteins (Nielsen, 1995; Stefansson et al., 2000). The proteins are divided into peptides and free amino acids, small peptides and myofibrillar proteins, which will be extracted into the surrounding brine (Nielsen 1995; Nielsen & Børresen, 1997). The addition of salt affects the proteins and contribute to the sensory quality (Gudmundsdottir & Stefansson, 1997). A previous study by Nielsen, Bro, Stefansson, and Skåra (1999) aimed to gather knowledge from three Nordic institutes in order to investigate if further information about the salting and ripening of herring could be derived. Principal component analysis (PCA) models showed correlation between a number of the basic chemical analyses and important sensory parameters. A striking result was the correlation between the protein concentration in the brine and the sensory attribute, softness. The softness of salted herring is an important quality parameter of the final product related to the ripening quality of the fish and as this correlates well with the protein concentration in brine, it suggests that the protein concentration in brine may be used as an indicator variable for the ripening process in barrel salted herring. It is known that the protein content in brine correlates with trichloroacetic acid-soluble nitrogen in muscles, which expresses the degree of protein degradation during ripening of salted herring (Bro,

*Corresponding author. Tel.: +45-35-28-35-01; fax: +45-35-28-32-45.

E-mail address: vip@kvl.dk (V.T. Svensson).

0023-6438/$30.00
doi:10.1016/j.lwt.2004.03.004

Nielsen, Stefansson, & Skåra, 2002; Nielsen et al., 1999). This particular study is the main motivation for the present study. Sampling of brine is more accessible and representative than sampling of a whole fish, where the inhomogeneity of the fish has to be taken into account (Andersen & Rinnan, 2002), so the change in protein concentration in brine instead of the protein degradation in the fish may be used as an indicator variable for the ripening of barrel salted herring.

Near-infrared (NIR) spectroscopy is well suited for determining the major components of foods such as water, fat, and protein (Osborne, Fearn, & Hindle, 1993). NIR spectroscopy is based on vibrational modes of molecules. These vibrations can be observed in the NIR spectra as overtones and combinations. The reason why NIR spectroscopy is well suited when assessing the presence of water and protein is due to the specificity of O–H and N–H bindings. In the overtone region from 1000 to 1900 nm water can be observed around 1400–1550 nm, and this overlap to some extent with the N–H regions from 1490 to 1600 nm. In the combinations region water absorption can be expected in the region of 1900–2000 nm and protein can be detected from 2000 to 2100 and 2150 to 2200 nm (Williams, 1987; Osborne et al., 1993). It is known that high concentration of salt may cause the absorptions band to change shape and the spectra may even shift (Lin and Brown, 1992). However, this is not considered to cause problems in the present study, as the variation in the salt concentration will be small. NIR spectroscopy is nondestructive, fast and easy to implement. NIR spectroscopy has previously been used to assess fish and its quality (Wold, Esbensen, & Geladi, 1987; Wold, Jakobsen, & Krane, 1996; Jørgensen & Jensen, 1997; Solberg & Fredriksen, 2001; Bøkness, Jensen, Andersen, & Martens, 2002).

The objective of the present study is to test if NIR spectroscopy can be used to determine the protein content in brine from traditionally barrel-salted herring. Previous studies have used NIR spectroscopy to assess the protein content of whole fish with satisfactory results (Isaksson, Tøgersen, Iversen, & Hildrum, 1995; Solberg, 1997; Bechmann & Jørgensen, 1998; Pink, Naczk, & Pink, 1999). On that basis and due to the short sampling time, NIR spectroscopy may be considered as a possible fast method for assessing protein in brine. Multivariate-data analysis and prediction modelling between NIR spectroscopy and the protein concentration is used as evaluation tools to study the relation between the spectroscopic measurements and the protein content.

2. Materials and methods

Two ripening experiments were carried out. In experiment 1 (ex. 1), herring caught by local fishermen in The Sound between Sweden and Denmark in August and September 2001 were used. In experiment 2 (ex. 2), the herring caught by local fishermen in The Sound in February 2002 was used. The herring was salted by a herring manufacturing company. One hundred kilograms of whole-headed herring was mixed with 10 kg of salt. After 1 day the barrel was filled with saturated brine and stored at 0–5°C. In ex. 1 eight barrels were used and in ex. 2 four barrels were used. In Table 1 the experimental set-up is listed and the days of storage are specified. At each sampling time 20–25 ml brine was taken for analysis from each barrel. Upon sampling, the brine was centrifuged at 10,000 g for 20 min at 5°C to remove tissue parts and insoluble matter and kept at −80°C until analyses were carried out.

The two experiments are combined in one overall dataset (38 samples), in the attempt to make a model including the seasonal variation.

2.1. Protein content

The protein content of the brine was determined by the Kjeldahl method (Total N × 6.25) (AOAC, 1996).

2.2. NIR measurements

The NIR spectra were measured with an InfraProver, II Fourier transform spectrometer (Bran and Luebbe, Germany) using a cuvette with a lightpath of 2 mm (Hellma fluorescence cell with four windows). Spectral range of the NIR spectroscopy from 1000 to 2500 nm (10,000 to 4000 cm^{-1}) was used.

2.3. Multivariate-data analysis

Initial multivariate-data analysis was performed with PCA (Martens & Næs, 1989). This method can be used

Table 1
Experimental set-up of barrels with salted herring and sampling of brine

Ripening ex.	Barrel	Salting date	Sampling date/days from salting							
1	A, B	23-08-01	04-10-01	42					16-02-02	178
	C, D	28-08-01	04-10-01	37			30-10-01	68		
	E, F	05-09-01	04-10-01	29			30-10-01	55		
	G, H	20-09-01			12-10-01	22	30-10-01	40	16-02-02	149
2	A, B	08-01-02	03-03-02	54	02-04-02	82	25-04-02	112	22-05-02	133
	C, D	14-01-02	03-03-02	48	02-04-02	76	25-04-02	106	22-05-02	127

for exploring the NIR data, to see if information about the ripening of the salted herring can be extracted from the data. Multivariate calibration on the NIR spectra was performed using partial least-squares (PLS) regression (Martens et al., 1989) and protein concentration as the dependent variable. The PLS model was validated using leave-one-out cross-validation (Wold, 1978; Eastman & Krzanowski, 1982). The principle of leave-one-out cross-validation is to leave one sample out and make a PLS model using the rest of the samples. This procedure is repeated until all the samples have been left out once. The predictive performance is tested using the root mean-squared error of cross-validation (RMSECV) and the correlation (r) of the predicted versus measured y-values. The RMSECV is given by comparing the predicted value and the reference value as shown in Eq. (1).

$$\mathrm{RMSECV} = \sqrt{\frac{\sum (y - \hat{y})^2}{n}}. \quad (1)$$

The y and $\hat{y}$ represent the measured reference value and the predicted value, respectively, and n is the number of samples. In order to test if the two experiments cover the same variation range, test-set validation was used. The principle of test-set validation is to predict the samples of the test set with a model based on the calibration data. If the test set is well described using the same complexity and loadings given by the calibration, the test set spans the same space as the calibration set. If the test set is not well described by the calibration set model, the calibration set does not explain the same variation as the test set and this is reflected in poor predictions (Martens & Næs, 1989). The validation is expressed by the Root Mean-Squared Error of Prediction (RMSEP). The data analysis was performed using The Unscrambler® Ver. 7.8 (Camo, Norway) and MatLAB 6.5, Matworks Inc.

3. Results and discussion

Total nitrogen content in brine expressed as protein increases up to approx. 5–6 g/100 g (Table 2) during the storage period. This is in agreement with previous work by Nielsen (1995), who found similar increase in protein contents in brine from salted herring. In the raw NIR spectra, scatter effects are present and spectral pre-treatment by a 2-window first derivative Savitzky–Golay filter (Martens & Næs, 1989) was performed. In Fig. 1, raw NIR spectra and pre-treated spectra in the region from 1000 to 2500 nm are shown. As the protein will be degraded during the storage of barrel salted herring, the brine will consist of a mixture of peptides and amino

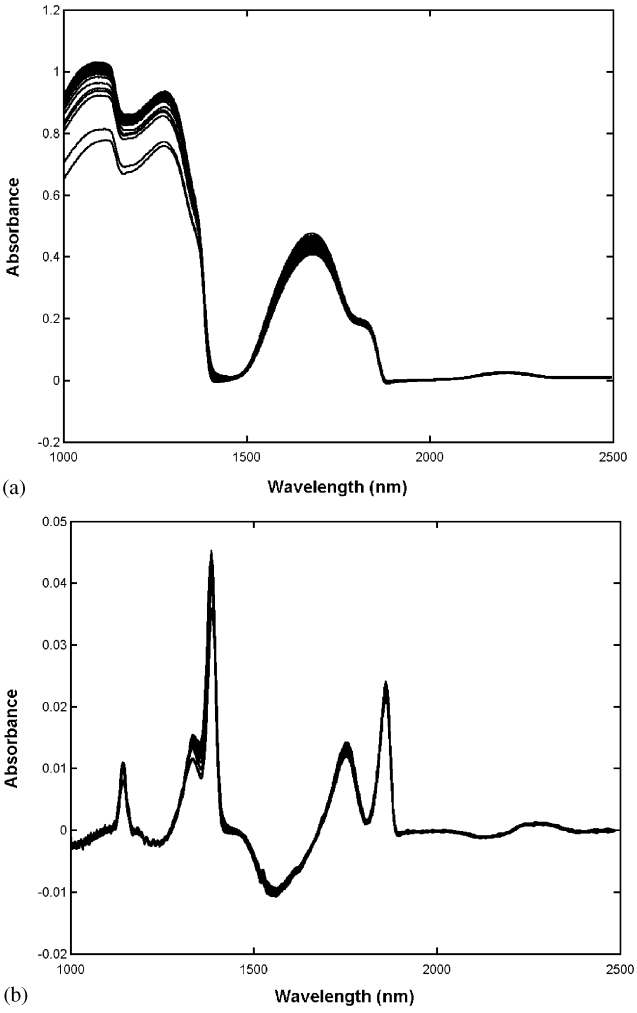


Fig. 1. NIR spectra, (a) raw NIR spectra in the range from 1000 to 2500 nm and (b) first derivative pre-treated spectra in the range from 1000 to 2500 nm.

Table 2
Protein content (g/100 g) in barrels determined by Kjeldahl method (N × 6.25)

Barrels ex. 1									Barrels ex. 2				
Sampling date	A	B	C	D	E	F	G	H	Sampling date	A	B	C	D
04-10-01	3.55	3.66	3.04	3.42	3.45	3.58			03-03-02	3.08	3.35	3.46	3.00
12-10-01							2.59	3.33	02-04-02	3.69	3.31	3.78	3.62
30-10-01			3.30	3.40	3.77	3.85	2.89	3.62	25-04-02	4.56	3.93	6.40	4.17
16-02-02	4.48	4.76					4.81	4.79	22-05-02	4.50	4.09	4.57	5.37

acids. The protein absorption will be due to absorption of these molecules. In general amides absorb in the region of 1460–1490 nm, and 1570 nm which corresponds well with the absorptions band observed in the region from 1470 to 1587 nm. Amide I absorbs around 2150–2180 nm, amide II around the region of 1960–2050 nm and amide III around 2050, 2110 and 2150–2180 nm. This also makes sense as absorption can be seen in the region from 2110 to 2300 nm (Williams, 1987; Osborne et al., 1993).

A preliminary PCA was performed to study the data. Two samples were removed as outliers and optimization of the model by selecting a region in the NIR spectra from 1294 to 1902 nm based on loadings, was performed. The resulting PCA model used three principal components (PC) and explained 96.8% of the variance. In Fig. 2, two score plots from the three-component PCA model are shown. The score plots can provide information about changes that happen during storage of herring and the capability of NIR spectroscopy to describe it. The two parameters of special interest are the storage time and the seasonal variation (batch variation). The seasonal variation can be observed in Fig. 2a, where PC1 and PC2 tends to explain the experiments performed in autumn (ex. 1) and winter (ex. 2). Fig. 2b describes information related to the storage time, where storage exceeding 100 days can be distinguished from the shorter storage times from 0 to 100 days by PC1 and PC3. Within the two groups it is not possible to distinguish the storage time further. From the PCA model it is obvious that the NIR measurements provide information related to the changes taking place in herring during the storage period.

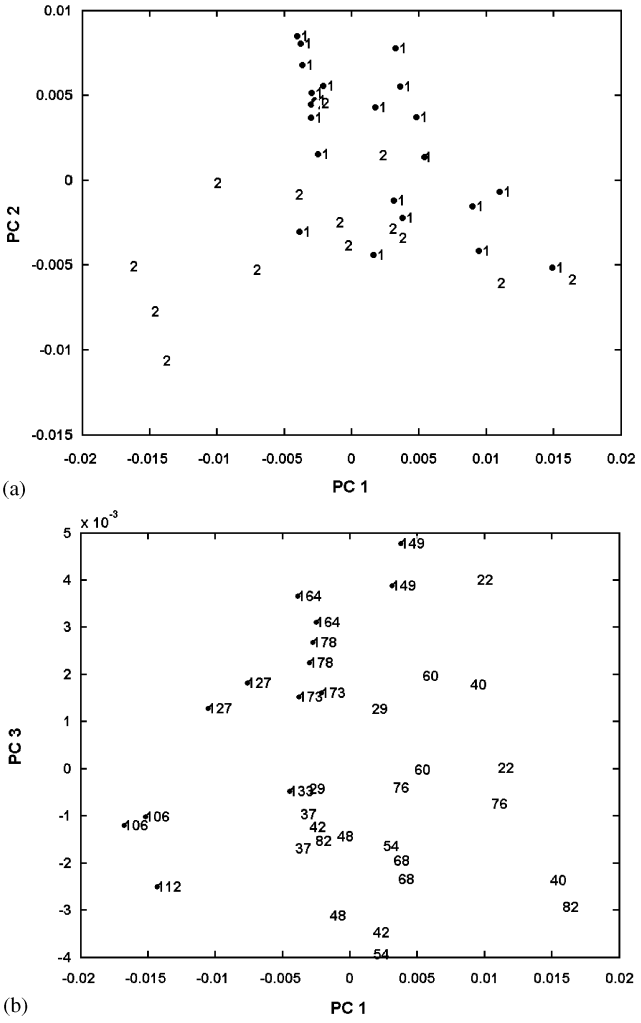


Fig. 2. Score plot from a three factor PCA model, (a) Score plot of PC1 versus PC2 explaining the batch/seasonal variation and (b) Score plot of PC1 versus PC3, separating the samples stored less than 100 days and more than 100 days.

3.1. Multivariate regression

Multivariate regression was performed in order to correlate the actual protein content to first derivative NIR spectra. Data inspection based on the *y*-values revealed one sample as an outlier, which was caused by high protein content for the specific sample. The deviation was considered to be a sampling or a laboratory error and the sample was removed from the data set. Together with the two samples removed during the PCA, a total of three samples were removed before performing the PLS analysis. A PLS regression, on the entire spectral range is described by five factors, explaining 90.5% of the *X*-variance and 78.4% of the *Y*-variance, *r* is 0.87 and the RMSECV is 0.34 g/100 g. An uncertainty test (Martens & Martens, 2002) was performed in order to refine the model and to find the spectral regions contributing the most to the predictive performance. The use of this procedure can reduce the number of spectral variables and thereby often the model complexity. Sixteen regions consisting of variable regions or individual variables were selected by the uncertainty test approach. This is illustrated in Fig. 3, where the grey areas are the selected informative regions. The PLS regression using the uncertainty test performed equally well, as the first model, which was based on the entire NIR spectra. This confirms that the removed wavelengths did not contribute in explaining the protein content. The relevant part of the spectra was explained by four factors, fitting 95.1% of the *X*-variance and 87.7% of the *Y*-variance. In Fig. 4 the prediction error, expressed by RMSECV, is given as a function of the number of PLS components for the PLS model based on the entire spectra and the PLS model using the uncertainty test. The first two PLS factors explain the storage time, in a similar way as the first and third factor of the PCA model. In Fig. 5 the predictive performance of the PLS regression model is illustrated with predicted protein (cross-validated) content versus the actual protein content. By the uncertainty test

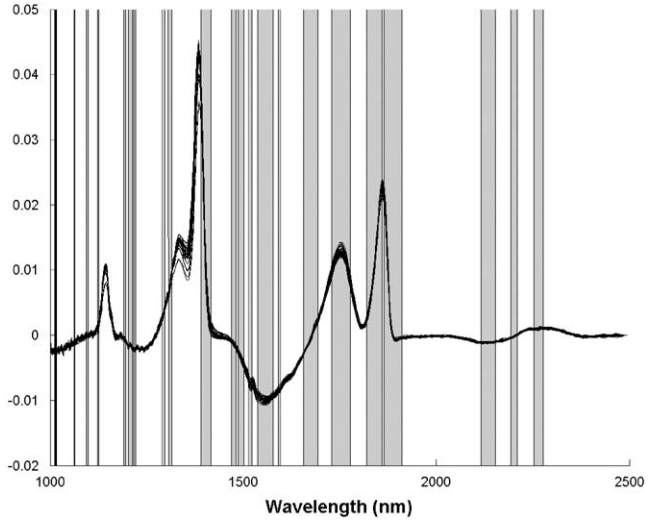


Fig. 3. First derivative NIR spectra marked with the selected regions based on Martens and Martens (2002) uncertainty test.

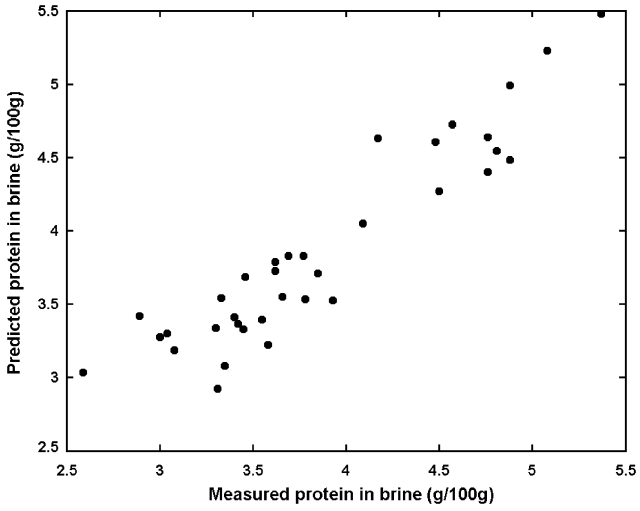


Fig. 5. Predicted versus measured plot of protein concentration in brine of barrel salted herring for a four factor PLS model based on selected NIR spectra regions.

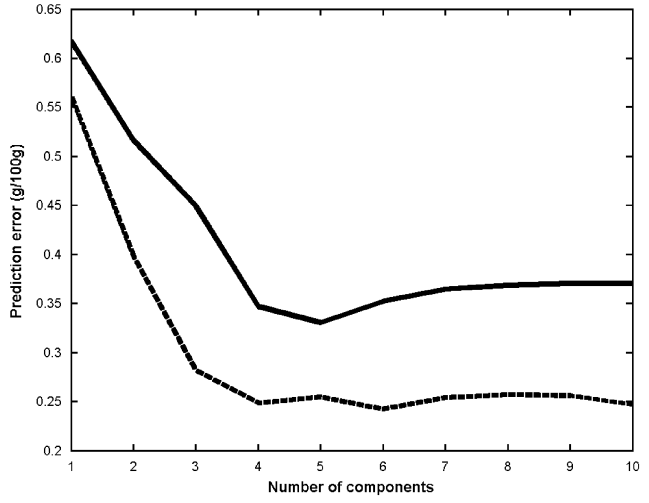


Fig. 4. Prediction errors (RMSECV) versus the model complexity of the model based on the entire spectra (—) and the model using the uncertainty test approach (- - -).

approach the resulting model is a four factor model with r of 0.93 and RMSECV of 0.25 g/100 g. Interpretation of the informative spectral regions in relation to chemical assignments corresponds well with the spectral regions related to specific N–H vibrations. This indicates that the NIR spectroscopy is highly correlated to the protein content in brine and that the protein content can be determined despite the high water and salt content present in brine of salted herring.

Another issue is how well each of the two experiments can describe the variation in the other experiment, which was tested by test-set validation using the entire NIR spectra. Two approaches were tested. In the first approach; ex. 1, was selected as the calibration set and tested using ex. 2, and in the second approach using ex. 2 as calibration set and ex. 1 as test set. Ex. 1 performed poor, when predicting ex. 2, whereas ex. 2 is capable of explaining ex. 1 in a five-component model, explaining 85% of X and 69% of Y, r is 0.87 and RMSEP of 0.4 g/100 g. Comparing this result with the regression model based on both experiments, it is obvious that ex. 2, is dominant in the overall model and that good sampling is needed to provide proper validation.

It is interesting to investigate if storage time can be predicted from the NIR spectra. During storage the protein matrix within the fish and the brine changes, due to diffusion between the fish and the brine and due to presence of salt. Furthermore, an enzymatic degradation of the larger proteins into smaller amino acids occurs. The changes in protein concentration and the degradation of protein are dependent on the storage time. In Fig. 6, a plot of the storage time versus the protein concentration is shown. The correlation between the storage time and the protein concentration is 0.87. The figure indicates a relation between protein concentration and the time of storage, where the protein concentration increases by time. A PLS regression performed on the entire NIR spectra and storage time, gives a 3 factor model describing 79.7% of the X-variance and 59.8% of the Y-variance, r is 0.76 and RMSECV is 33 days. It is known that the concentration of soluble proteins in herring during the ripening is not a linear function of the storage time (Olsen & Skåra, 1997). This may reflect the low correlation and the relative high RMSECV given by the PLS model. On the other hand, the scatter plot in Fig. 6 does indicate linearity in the selected storage period, and to some extent it seems that the relation between storage time and the protein concentration can be approximated by a linear function.

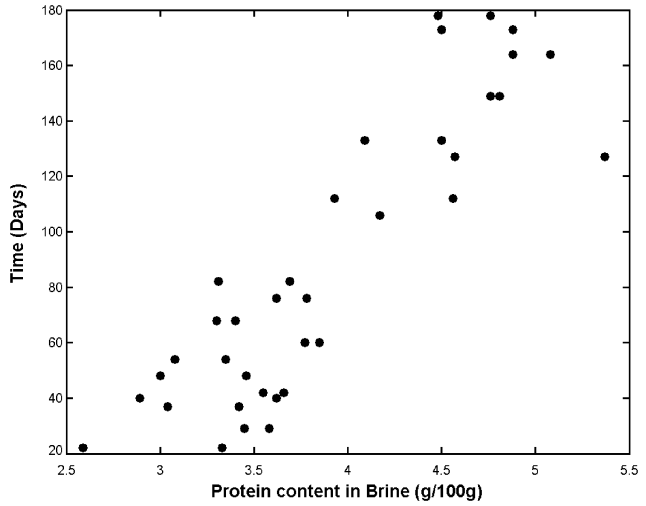


Fig. 6. Scatter plot of protein concentration in brine from barrel salted herring versus storage time, correlation is 0.87.

4. Conclusion

Sampling of and measurements on brine are easy to perform compared to sampling of whole fish. A PCA suggests that NIR spectroscopy of brine in the range of 1000–2500 nm, carry information related to changes in the nitrogen fraction of herring and thus can be a potential indicator for the ripening characteristics of salted herring, as brine from the early stage of ripening e.g. before 100 days, can be separated from brine the late stage of ripening e.g. after 100 days. Furthermore multivariate-regression modelling shows that the protein content in brine can be predicted by NIR spectroscopy using the region from 1000 to 2500 nm. Further optimization by variable selection results in a PLS regression model with the correlation coefficient of 0.93 and a prediction error of 0.25 g/100 g. The study shows that NIR spectroscopy is an obvious alternative to the time consuming chemical analysis to determine the protein concentration. Further studies using the brine of salted herring instead of sampling a whole fish has to be performed to investigate whether or not more detailed information about the ripening of salted herring can be extracted from the brine using spectroscopy.

Acknowledgements

The authors want to thank Lykkeberg A/S for providing salted herring and to Jørgen Andersen for valuable discussion. The authors also want to thank Karin Reimers for excellent technical assistance and Bo Jørgensen for valuable help and advice. Svensson wishes to thank FØTEK 3 (93S-2444-Å01-00100) for financial support.

References

Andersen, C., & Rinnan, Å. (2002). Distribution of water in fresh cod. *Lebensmittel-Wissenschaft Und-Technologie*, *35*, 687–696.

AOAC. (1996). Method no. 39.1.19 981.10. Crude protein in meat. Block digestion method. In: P. Cunnitt (Ed.) *Official methods of analysis*, *16th ed*. Gaithesberg, MD: AOAC International.

Bechmann, I. E., & Jørgensen, B. M. (1998). Rapid assessment of quality parameters for frozen cod using near infrared spectroscopy. *Lebensmittel-Wissenschaft Und-Technologie*, *31*, 648–652.

Bøkness, N., Jensen, K. N., Andersen, C. M., & Martens, H. (2002). Freshness assessment of thawed and chilled cod fillets packed in modified atmosphere using near-infrared spectroscopy. *Lebensmittel-Wissenschaft Und-Technologie*, *35*, 628–634.

Bro, R., Nielsen, H. H., Stefansson, G., & Skåra, T. (2002). A phenomenological study of ripening of salted herring. Assessing homogeneity of data from different countries and laboratories. *Journal of Chemometrics*, *16*, 81–88.

Eastman, H., & Krzanowski, W. J. (1982). Cross-validatory choice of the number of components from a principal component analysis. *Technometrics*, *24*, 73–77.

Gudmundsdottir, G., & Stefansson, S. (1997). Sensory and chemical changes in spice-salted herring as affected by handling. *Journal of Food Science*, *62*, 894–897.

Isaksson, T., Tøgersen, G., Iversen, A., & Hildrum, K. I. (1995). Non-destructive determination of fat and moisture and protein in salmon fillets by use of near-infrared diffuse spectroscopy. *Journal of the Science of Food and Agriculture*, *69*, 95–100.

Lin, J., & Brown, C. W. (1992). Near-IR spectroscopic determination of NaCl in aqueous solution. *Applied Spectroscopy*, *46*, 1809–1815.

Jørgensen, B. M., & Jensen, H. S. (1997). Can near-infrared spectrometry be used to measure quality attributes in frozen cod? In J. B. Luten, T. Børresen, & J. Oehlenschlager (Eds.), *Seafood from producers to consumer, integrated approach to quality* (pp. 491–496). Amsterdam: Elsevier Science BV.

Martens, H., & Martens, M. (2002). Multivariate analysis of quality. An Introduction (445p). New York: Wiley.

Martens, H., & Næs, T. (1989). Multivariate calibration (417p). New York: Wiley.

Nielsen, H. H. (1995). *Proteolytic enzyme activities in salted herring during cold storage*. Ph.D. thesis, Danish Institute for Fisheries Research, Technical University of Denmark, Lyngby.

Nielsen, H. H., & Børresen, T. (1997). The influence of intestinal proteinases on ripening of salted herring. In J. B. Luten, T. Børresen, & J. Oehlenschlager (Eds.), *Seafood from producer to consumer, integrated approach to quality* (pp. 193–304). Amsterdam: Elsevier Science.

Nielsen, H. H., Bro, R., Stefansson, G., & Skåra, T. (1999). *Salting and ripening of herring—collection and analysis of research results and industrial experience within the Nordic Countries*. TemaNord, Nordic Council of Ministers, Copenhagen.

Olsen, S. O., & Skåra, T. (1997). Chemical changes during ripening of north sea herring. In J. B. Luten, T. Børresen, & J. Oehlenschlager (Eds.), *Seafood from producer to consumer, integrated approach to quality* (pp. 305–318). Amsterdam: Elsevier Science.

Osborne, B. G., Fearn, T., & Hindle, P. H. (1993). Practical NIR spectroscopy with application in food and beverage analysis (200 p). Essex, UK: Longman, Scientific & Technical.

Pink, J., Naczk, M., & Pink, D. (1999). Evaluation of the quality of frozen Minched Red Hake: Use of fourier transform near-infrared spectroscopy. *Journal of Agricultural Food Science*, *47*, 4280–4284.

Solberg, C. (1997). NIR—A rapid method for quality control. In J. B. Luten, T. Børresen, & J. Oehlenschläger (Eds.), *Seafood from producer to consumer, integrated approach to quality* (pp. 529–534). Amsterdam: Elsevier Science BV.

Solberg, C., & Fredriksen, G. (2001). Analysis of fat and drymatter in capelin by near infrared transmission spectroscopy. *Journal of Near Infrared Spectroscopy*, *9*, 221–228.

Stefansson, G., Nielsen, H. H., Skåra, T., Schubring, R., Oehlenschalger, J., Luten, J., Derrick, S., & Gudmundsdottir, G. (2000). Frozen herring as raw material for spice-salting. *Journal of the Science of Food and Agriculture*, *80*, 1319–1324.

Voskrensensky, N. A. (1965). Salting of herring. In G. Borgström (Ed.), *Fish as food*, Vol. III (pp. 107–131). New York: Academic Press.

Williams, P. C. (1987). Implementation of near-infrared technology. In P. Williams, & K. Norris (Eds.), *Near-infrared technology. In the agricultural and food industries* (2nd ed.) (pp. 145–169). American Association of Cereal Chemists, USA: St. Paul, MN.

Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal component models. *Technometrics*, *20*, 397–405.

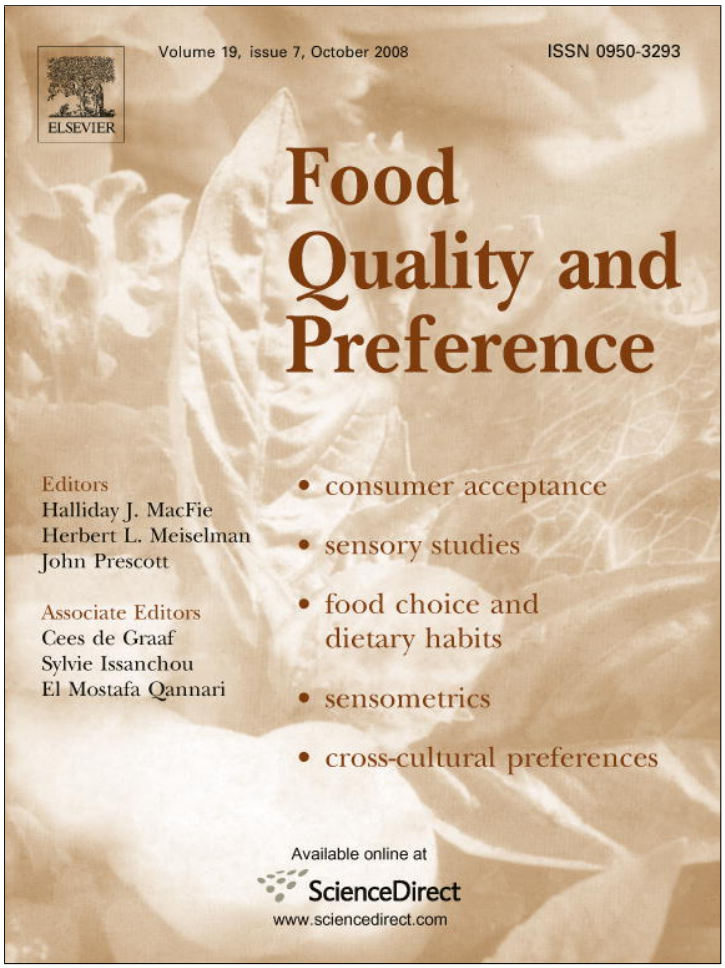
Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*, 37–52.

Wold, J. P., Jakobsen, T., & Krane, L. (1996). Atlantic salmon average fat content estimated by near-infrared transmittance spectroscopy. *Journal of Food Science*, *61*, 74–77.

PAPER V

Stine Kreutzmann, Vibeke Tølbøl Svensson, Anette K. Thybo, Rasmus Bro and Mikael A. Petersen

Prediction of sensory quality in raw carrots (Daucus Carota L.) using multi-block LS-ParPLS, Food Quality and Preference 19, 609-617, 2008



Food Quality and Preference 19 (2008) 609–617

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Prediction of sensory quality in raw carrots (*Daucus carota* L.) using multi-block LS-ParPLS

Stine Kreutzmann [a,*], Vibeke T. Svensson [b], Anette K. Thybo [a], Rasmus Bro [b], Mikael A. Petersen [b]

[a] Faculty of Agricultural Sciences, University of Aarhus, Research Centre Aarslev, Kirstinebjergvej 10, P.O. Box 102, DK-5792 Aarslev, Denmark
[b] Faculty of Life Sciences, University of Copenhagen, DK-1958 Frederiksberg C, Denmark

ARTICLE INFO

Article history:
Received 25 April 2007
Received in revised form 4 March 2008
Accepted 31 March 2008
Available online 18 April 2008

Keywords:
Daucus carota
Sensory profiling
Multi-block analysis
Principal component analysis (PCA)
LS-ParPLSc

ABSTRACT

The relations between the sensory quality of 24 different carrot genotypes and content of dry matter, non-volatile and volatile compounds were studied using a multi-block approach called LS-parPLS. The prediction of five sensory attributes; bitterness, sweetness, terpene flavour, green flavour and carrot flavour, gave prediction errors (RMSECV) between 0.98 and 1.36 and correlation coefficients (r) between 0.60 and 0.81. The explained Y-variances were between 15.1% and 66.0%. The highest prediction error was observed for the attribute carrot flavour whereas green flavour gave the best prediction. The attributes green flavour, bitterness and terpene flavour showed fairly good predictions (r/RMSECV/% exp-Y = 0.81/0.98/66.0, 0.79/1.23/62.3 and 0.71/1.04/50.2) whereas sweetness gave an unexpected poor prediction (r/RMSECV/% exp-Y = 0.67/1.36/44.6). Non-volatile compounds found to be important predictors were chlorogenic acid (5-CQA), sucrose, 6-methoxymellein (6-MM), falcarindiol (FaDOH), and falcarinol (FaOH). The volatile compounds found to be important predictors are considered as key flavour compounds of raw carrots: terpinolene, β-pinene, sabinene, γ-terpinene, α-pinene, β-bisabolene, caryophyllene and cuparene. In general, the overall results show that the sensory quality variation in the material regarding bitterness, green flavour and terpene flavour are explained by relatively few parameters. Despite that the results revealed some reliable relationships between the sensory attributes, aroma and chemical analysis, a large variance (about 40%) in the sensory block of variables remained unexplained and still needs further investigation for an in-depth understanding of sensory quality. LS-ParPLSc is shown to be feasible for handling several types of data blocks in one regression model.

1. Introduction

Horticulturists have been working on genetic approaches for improving nutrient content and visual appeal of vegetables in hopes of increasing consumer consumption of beneficial phytochemicals. Consumers have shown an increased interest in healthy food and an increased demand for diversity in vegetables (Jongen, 2000) and in order to be able to act on changing consumer demands, it is important to have quantitative means for assessing quality and quality changes in vegetables. Biological materials such as carrots are complex substances that usually have uncontrollable variations in quality. These variations can arise from genetics, environmental conditions or be due to pre-processing (Baardseth et al., 1996; Hogstad, Risvik, & Steinsholt, 1997; Rosenfeld, Aaby, & Lea, 2002; Seljåsen, Bengtsson, Hoftun, & Vogt, 2001a; Simon, Peterson, & Lindsay, 1980). Carrot genotypes with different colours (orange, red, yellow, purple and white) are now available on the market and they show a large diversity in quality (Alasalvar, Grigor, Zhang, Quantick, & Shahidi, 2001; Kreutzmann, Christensen, & Edelenbos, 2008; Kreutzmann, Thybo, & Bredie, 2006).

Carrots have a number of chemical and physical quality characteristics and the eating quality can be probed directly by sensory methods or indirectly by chemical, mechanical, or optical measurements. Important properties are contents of sugars, dry matter, non-volatile bitter compounds and volatile compounds. However, it is a challenge to understand exactly which chemical compounds and combination of compounds that affect the sensory-perceived quality.

Multivariate data analysis using ordinary principal component analysis (PCA) and partial least squares (PLS) regression are well known when trying to understand the sensory quality in relation to physical and chemical properties of carrots (Kreutzmann et al., 2008; Kreutzmann, Thybo, Christensen, & Edelenbos, in press; Rosenfeld et al., 2002; Seljåsen et al., 2001a). Trying to identify which chemical analysis and individual components that are having high impact on the sensory attributes can be overwhelming even with general PCA and PLS regression. Multi-block (MB) analysis is one approach capable of handling data where several types of analysis or blocks structures are present. Several approaches for

* Corresponding author. Tel.: +45 89 99 34 13; fax: +45 89 99 34 95.
E-mail address: Stine.Kreutzmann@agrsci.dk (S. Kreutzmann).

0950-3293/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.foodqual.2008.03.007

performing MB analysis have been suggested and compared through out the years. Wangen and Kowalski (1988) suggested a MBPLS algorithm based on the MBPLS algorithm from Wold, Martens, and Wold (1984). Wold, Kettaneh, and Tjessam (1996) published an approach referred to as hierarchical MBPLS. The hierarchical MBPLS and the MBPLS are similar in the way they combine the common information in a super level described by super scores and super loadings. Westerhuis, Kourti, and MacGregor (1998) showed that MBPLS provides the same predictions as an ordinary PLS under the condition that the weighting and scaling are the same. Furthermore they suggested another approach where deflation is to be performed on the super scores instead of the block scores of the X blocks as Wangen and Kowalski originally suggested. Westerhuis and Smilde (2001) altered the deflation step of the method suggested by Westerhuis et al. (1998) using only the super scores to deflate the Y response and not change the X blocks. An alternative MB approach is the serial-PLS (S-PLS) suggested by Berglund and Wold (1999). This method shows resemblance to the iterative least squares PLS (LS–PLS) purposed by Jørgensen, Segtnan, Thyholt, and Næs (2004), Jørgensen, Mekvik, and Næs (2007). Måge (2006) followed up on the LS–PLS and suggested the LS-parallel-PLS (LS-parPLS) and LS-parPLS with common components (LS-parPLSc). These methods will be described in detail later in the data analysis section.

One of the main problems with most MB methods have been how to determine the proper block scaling. Depending on how the data blocks are scaled, very different results, hence interpretations can be obtained. The LS-parPLS and LS-parPLSc methods both avoid the block scaling effect as they are invariant to scaling and are capable of handling different complexities of the individual block contributions. Måge (2006) compared the most frequently used MB methods based on two case studies with design variables and parallel blocks of spectroscopic measurements. The predictive performance and interpretability were compared. The study concludes that the models performed equally well in their predictive performance if properly used whereas differences were be found in the ability to provide meaningful interpretation of the models. When the information in the X blocks was overlapping the LS-parPLSc model reflected the true data structure in the best way whereas situations with no overlapping information was best modelled by the LS-parPLS. The present study will focus on the LS-parPLSc method. Originally, this method was suggested when modelling designed data where information in addition to a design can be split up into intuitively meaningful blocks (e.g. volatile and non-volatile compounds) (Mekvik, Jørgensen, Måge, & Næs, submitted for publication). The information in the dependent variables related to the design matrix is extracted from additional blocks before modelling the influence of these. If the experiment is not designed the design step can be eliminated and the remaining blocks are then analysed in such a way that any common structure between the blocks is analysed separately from unique information. LS-parPLSc provides an excellent visualizing and data interpretation tool, as scores and loadings for each data block contribution are given. This is where the LS-parPLSc distinguishes itself from the ordinary PLS regression as it calculates a regression model based on all data material but keep the structure of individual block contributions. An ordinary PLS regression has to be performed either on one single block at a time or if many blocks are present they have to be combined in on big block.

The aim of the present work was to investigate which chemical compounds or combinations of compounds are important for the sensory quality of raw carrots. In order to predict the sensory quality from the raw material measurements the non-volatile compounds, contents of sugars, dry matter (Block I) are combined with the volatile compounds (Block II) in one regression model using LS-parPLSc. The main goal is to find a good model that explains the variation in the end product well and to interpret the given model parameters in order to obtain information about which variables are important and in what way they seem to affect the sensory quality.

2. Materials and methods

2.1. Plant material and sample preparation

Twenty-four different carrot genotypes were selected to represent a large variation in odour and taste by sensory screening of 50 genotypes. The genotypes were grown in Denmark, Norway and Holland during 2004 and harvested at the end of October 2004. The carrots obtained from Denmark were cultivated at Research Centre Aarslev, those from Norway were cultivated at Plante Forsk, The Norwegian Crop Research Institute, Hedmark and those obtained from Holland were cultivated by Bejo Zaden B.V., Warmenhuizen. The roots were transported to Research Centre Aarslev and stored at 1 °C until February 2005 at >95% relative humidity (RH). All roots were stored in an ethylene free atmosphere except for Bolero. A sample of this cultivar was moved to apple storage facilities and exposed to ethylene generated by the apples one month before sensory evaluation (Kidmose et al., 2004; Seljåsen, Hoftun, & Bengtsson, 2001b). The root weight varied from approximately 50–150 g. However, the most representative size within each genotype was selected for analysis. Samples (8 kg) of carrots were taken from each genotype and divided into sub-samples of 1.5–2.0 kg carrots of first class quality, i.e. carrots with no visible damage representing each replicate. The carrots were then carefully washed, manually hand-peeled and trimmed. Approximately 0.65–1.00 mm of the periderm was removed by peeling and 2 cm of the tip and 2 cm of the top was also removed by trimming. The peeled carrots were cut into $2 \times 2 \times 20$ mm sticks using a food processor (Robot Coupe CL50, Vincennes Cedex, France), carefully mixed and samples of 1500 g were taken for immediately analysis of sensory quality, volatile compounds and phenolic acids. All analysis was carried out in three replicates. The rest of the raw carrots were frozen at −24 °C until analyzed for polyacetylenes, isocoumarin, sugars and dry matter content 2–4 months later.

2.2. Sensory analysis

Quantitative descriptive analysis was performed as previously described (Kreutzmann et al., 2008). A panel consisting of 10 trained assessors (5 females/5 males, aged from 26 to 54 years) evaluated the sensory quality in terms of 4 odour attributes, 7 flavour attributes, 2 taste attributes, and 1 aftertaste attribute. The fourteen attributes were: terpene aroma, carrot aroma, green aroma, faded aroma, terpene flavour, carrot flavour, green flavour, faded flavour, nutty flavour, soapiness, sickenly sweet flavour, bitterness, sweetness and burning aftertaste. The sensory laboratory and the computer screens were illuminated with red light during evaluation to mask visual differences between samples. The panellists evaluated the samples at individual speed by descriptive analysis on an unstructured 15 cm line scale with intensity ratings ranging from low (value 0) to high intensity (value 15). All data was registered on a direct computerised registration system (FIZZ, ver. 2.00 M, Couternon, F).

2.3. Chemical analysis

Extraction and quantification of the polyacetylenes falcarindiol (FaDOH), falcarindiol 3-acetate (FaDOAc), falcarinol (FaOH) and 6-methoxymellein (6-MM) were performed by solvent extraction and reversed phase-high performance liquid chromatographic

(RP–HPLC) according to the method of Kreutzmann et al. (2008). Polyacetylenes and 6-MM were identified by peak addition of authentic standards and quantified using calibration curves of authentic standards isolated from carrots.

Extraction and quantification of phenolic acids were performed by solvent extraction and RP–HPLC according to the method of Kreutzmann et al. (2008). Phenolic acids were identified on an Agilent HPLC–DAD–MS station also previously described by Kreutzmann et al. (2008). Phenolic acids derived from *p*-coumaric acid, ferulic acid and caffeic acid were identified based on retention time and their LC–MS and UV-data.

Extraction and quantification of sugars were performed with ultra pure water and analysed by analytical high performance anion exchange chromatography (HPAEC) according to the method of Kaack, Christensen, Hansen, and Grevsen (2004). Quantification was performed on a Dionex Series 300DX ion chromatograph previously described by Kreutzmann et al. (2008). Sugars (fructose, glucose and sucrose) were identified by authentic standards and quantified using calibration curves. In total 18 non-volatile compounds were identified and quantified.

2.4. Volatile analysis

Volatile compounds were collected from 50 g fresh-cut carrots by dynamic headspace sampling according to the method described by Kjeldsen, Christensen, and Edelenbos (2001). Volatile compounds were quantified by gas chromatography (GC) and identified by GC–mass spectrometry (MS). The individual volatiles were tentatively quantified from the FID peak areas relative to that of the internal standard ((*E*)-2-hexen-1-ol). The response factor was set to 1 for all compounds. Compounds suggested by the MS database (NIST, 1998) were verified by comparison of the relative retention indices (RI) and mass spectra of authentic reference compounds unless noted. Thirty volatile compounds were identified and quantified.

2.5. Dry matter

Dry matter was determined gravimetrically by weighing before and after lyophilisation in a ventilated oven at 80 °C for 20 h (Lytzen A/S, Herlev, Denmark).

2.6. Data analysis

Principal component analysis (PCA) (The Unscrambler 9.2, CAMO ASA, Trondheim, Norway) was performed to describe the correlation between sensory quality and instrumental measurements of volatile and non-volatile compounds. Average sensory response values over replicates were used in data analysis. Results are presented by score and loading plots (Martens & Næs, 1989). GC-data and HPLC-data was standardised (each variable divided with its standard deviation) prior to data analysis and leave one out cross validation was used as validation criterion (Eastman & Kranowski, 1982).

2.6.1. LS-ParPLS with common loadings (LS-ParPLSc)

The model LS-ParPLSc aims to model the predictive information in several blocks of data simultaneously. The principle is to split the information in each block into three parts:

1. Information in a possible design matrix which can be used to predict is firstly determined by a least squares regression (LS part).
2. Additional common information between the two blocks (c-part).

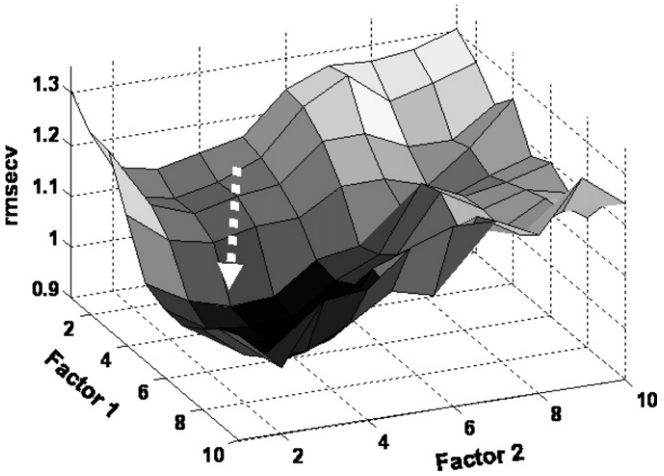


Fig. 1. RMSECV versus the number of components for the each of the two blocks contributions for the prediction of green flavour. The arrow shows the selected factor combination (block I: #3 and block II: #5).

Table 1
List of attributes used for sensory profiling raw carrots with descriptions

Characteristics		Description
Aroma[a]	Terpene aroma	Odour related to mixtures of turpentine-like flavour (α-pinene, caryophyllene, terpinolene)
	Carrot aroma	Relate to a carrot odour
	Silage aroma	Relate to faded hay odour
	Green aroma	Smell like green carrot leaves
	Hay aroma	Relate to hay odour
Flavour[b]	Terpene flavour	Flavour related to mixtures of turpentine-like flavour (α-pinene, caryophyllene, terpinolene)
	Carrot flavour	Like the flavour of carrot
	Green flavour	Like the flavour of green carrot leaves
	Soapiness	Like the flavour of soft soap
	Nutty flavour	Flavour of fresh hazelnut (green)
Taste[c]	Sweetness	Taste related to the taste of sucrose
	Bitterness	Sharp taste related to the taste of caffeine
Aftertaste[d]	Burning aftertaste	Related to sharp, burning taste in the mouth after 60 s

[a] The term "aroma" is used for retro-nasal odour perception.
[b] The term "flavour" is used for intra-nasal detection of responses.
[c] The term "taste" is used for basic taste responses on the tongue.
[d] The term "aftertaste" is used for responses received after evaporating the sample.

Table 2
Mean values, standard deviation, min and max values for each sensory attribute

Sensory attribute	Mean (std.)	Min–Max
Carrot aroma	8.55 (1.08)	6.30–10.39
Terpene aroma	8.06 (1.66)	4.80–10.89
Green aroma	3.90 (1.15)	1.66–6.37
Carrot flavour	9.14 (1.33)	5.97–11.27
Terpene flavour	7.59 (1.32)	4.82–10.13
Sickenly sweet	1.96 (0.88)	0.75–4.44
Green flavour	5.56 (1.56)	2.10–8.89
Nutty flavour	2.99 (0.69)	1.49–4.41
Sweetness	7.18 (1.64)	4.13–10.79
Bitterness	4.30 (1.89)	2.13–10.35
Burning aftertaste	3.77 (0.80)	2.82–5.64

The data are averaged over assessors and replicates.

3. Parallel extraction of additional individual predictive information in the two blocks (ParPLS part).

Step 1 is avoided if, as here, there is no design and step 2 is avoided if the blocks do not have common variation. The algorithm for LS-ParPLSc is given below (Måge, 2006).

1. Remove design related information in **y**: Fit **y** to the design matrix **D** using LS regression, calculating the residual $\mathbf{f} = \mathbf{y} - (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{y}$.
2. Remove design information from Data set 1 ($\mathbf{X}_1$) and Data set 2 ($\mathbf{X}_2$) by orthogonalization against **D**. $\mathbf{X}_{1,\mathbf{orth}} = \mathbf{X}_1 - \mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{X}_1$ and $\mathbf{X}_{2,\mathbf{orth}} = \mathbf{X}_2 - \mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{X}_2$.
3. Predict the extracted residual (**f**) calculated in step 1 by PLS regression using the orthogonalised Data set 1 ($\mathbf{X}_{1,\mathbf{orth}}$) and Data set 2 ($\mathbf{X}_{2,\mathbf{orth}}$).
   a1. The residual **f** is fitted to ($\mathbf{X}_{1,\mathbf{orth}}$) – the A1 first scores **T1**, loadings **P1** and weights **W1** are calculated.
   a2. The residual **f** is fitted to($\mathbf{X}_{2,\mathbf{orth}}$) – the A2 first scores **T2**, loadings **P2** and weights **W2** are calculated.
   b. The scores T1 and T2 are analyzed by CCA. Using the scores, the CCA only considers the relevant information describing Y. If common components are present (evalu-

Table 3
LS-parPLSc correlation coefficients, RMSECV, explained Y-variance and factor combination for block I and II of five selected sensory attributes

Sensory attribute	Correlation coefficients (r)	RMSECV	# Fac	Exp. Y-variance CV model (%)
Bitterness	0.79	1.23	Block I -3 Block II -6	62.30
Sweetness	0.67	1.36	Block I -3 Block II -2	44.64
Green flavour	0.81	0.98	Block I -3 Block II -5	66.03
Terpene flavour	0.71	1.04	Block I -1 Block II -5	50.21
Carrot flavour	0.60	1.37	Block I -1 Block II -5	15.08

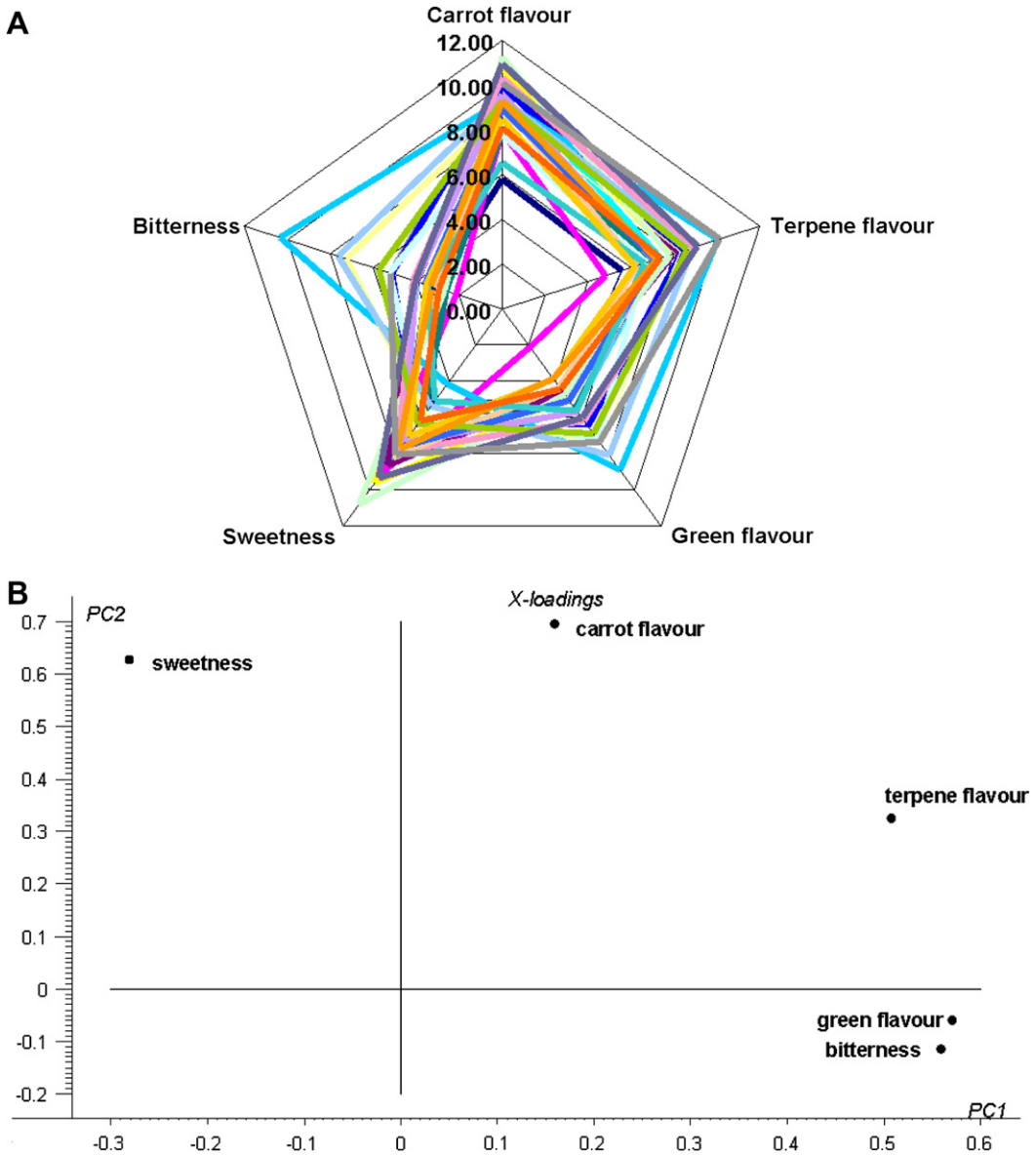


Fig. 2. (A) Sensory profile of 24 different carrot genotypes. The radar plot includes the sensory attributes analysed by LS-parPLSc. (B) PCA loadings plot for factor 1 versus factor 2.

ated by a threshold on the correlation) canonical weights **A** and **B** are obtained. The common scores will be $\mathbf{T_c} = \mathbf{T_1A/2} + \mathbf{T_2B/2}$.

c. Combine $\mathbf{T_c}$ and the design $[\mathbf{D}\ \mathbf{T_c}]$ and fit **y** on these using LS regression in order to calculate new residuals **f'**.

d1. The residual **f'** is fitted to $\mathbf{X}_{1,\mathbf{orth}}$ by PLS regression. $\mathbf{X}_{1,\mathbf{orth}}$ is orthogonalised against both **D** and $\mathbf{T_c}$ scores. The A_{1U} scores $\mathbf{T}_{2U}$, loadings $\mathbf{P}_{1U}$ and weight $\mathbf{W}_{2U}$ are calculated.

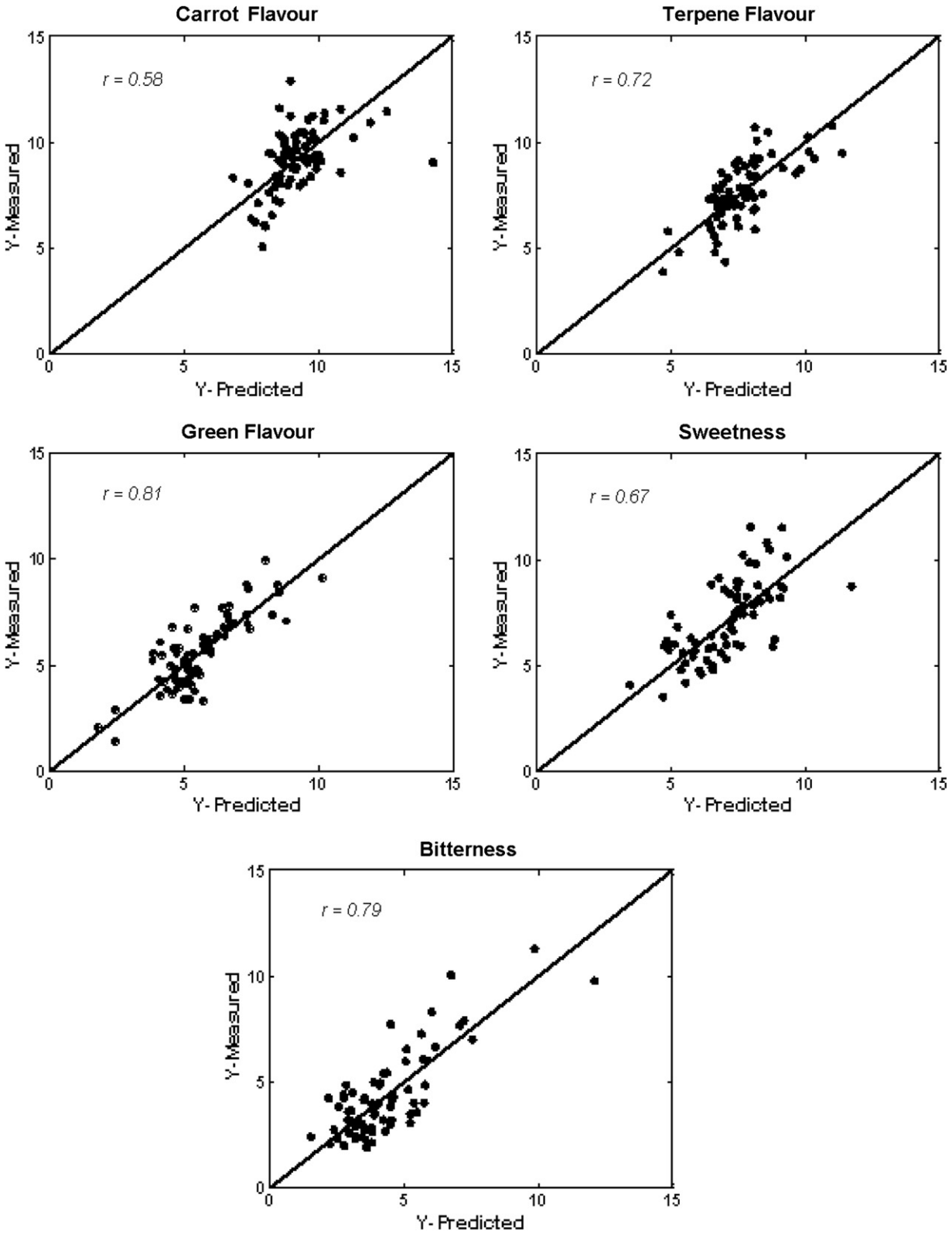


Fig. 3. Correlation plots of Y-predicted vs. Y-measured for the five sensory attributes.

d2. The residual **f'** is fitted to $\mathbf{X}_{2,\mathbf{orth}}$ by PLS regression. $\mathbf{X}_{2,\mathbf{orth}}$ is orthogonalised against both **D** and $\mathbf{T}_{1U}$. The $\mathbf{A}_{2U}$ scores $\mathbf{T}_{2U}$ loadings $\mathbf{P}_{2U}$ and weights $\mathbf{W}_{2U}$ are calculated.

e. Final LS regression step combining the design matrix (**D**) and the scores $\mathbf{T}_{1U}$, $\mathbf{T}_{2U}$ and $\mathbf{T}_c$ and fit **y** against [**D** $\mathbf{T}_c$ $\mathbf{T}_{1U}$ $\mathbf{T}_{2U}$].

The final LS-parPLSc model can be expressed:

$$\mathbf{y} = \mathbf{D}\beta_{\mathbf{D}} + \mathbf{T_C}\beta_{\mathbf{C}} + \mathbf{T_{1U}}\beta_{\mathbf{1U}} + \mathbf{T_{2U}}\beta_{\mathbf{2U}} + \mathbf{e}$$

Note that the canonical components in step 3b are determined from the score matrices rather than the raw material. Hence spurious correlations are avoided as the noise is essentially filtered off beforehand. Also, CCA is only used if high correlations exists within the score matrices. If no common components are present the method is equal to the LS-parPLS (Måge, 2006).

The calculations were performed in MATLAB R2006a (The Mathworks Inc). Before performing LS-ParPLSc, the data was divided into two block contributions. Block I consisted of 18 variables – dry matter and HPLC analyses. Block II was a larger block with 30 aroma variables analysed by GC–MS. The data set is not based on a designed experiment, thus the design part of the algorithm (step 1 and 2) is left out. A test using the LS-ParPLSc algorithm to find the optimal number of components for each data block was performed testing 1–10 factors on Block I and 1–10 factors on Block II. The test was evaluated by leave one out cross validation and the prediction error $\left(\mathrm{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}\right)$. Fig. 1 illustrates the RMSECV versus the number of components for the two blocks. The factor combination is picked based on a plateau in the RMSECV curve, absolute or local minimum. The data was studied in order to detect possible outliers and one outlier was removed due to high leverage in the aroma analysis. After removing the outlier the test was rerun to find the right number of factors (see Table 1).

3. Results and discussion

3.1. Sensory profile

The carrot material was selected to span a relevant variation in sensory quality of carrots and consists of 24 different carrot genotypes. Mean values, standard deviation and the range of variation (min–max values) of each sensory attribute are given in Table 2. The attributes nutty flavour and burning aftertaste reveal the smallest differences, while the attributes more characteristic for carrot flavour, like terpene aroma, green flavour, sweetness and bitterness show the largest differences between genotypes. In the model used for LS-parPLSc the attributes interesting for carrot quality and for the variation in the material will be analysed, namely carrot flavour, terpene flavour, green flavour, sweetness and bitterness. The interpretation of the sensory analysis is visualized by a radar plot in Fig. 2a. The data is averaged over assessors and replicates. Fig. 2a shows the differences in range for the sensory attributes which are included in the analysis. The figure indicates that genotype influences the variation in sensory quality. Previous studies have also revealed that genotype largely influences the sensory quality (Kreutzmann et al., 2006, 2008, in press; Seljåsen et al., 2001a; Simon, Peterson, & Lindsay, 1982). From Fig. 2a it is observed that the large differences are not caused by one extreme genotype but by the fact that the genotypes used have a large variation span. Furthermore PCA confirms the large variation between the 24 genotypes and the sensory attributes (Fig. 2b). The PCA explaining 47% in component one and 22% explained variance in component two shows bitterness and green flavour are highly correlated but they are also slighty correlated to terpene flavour. Component one mainly describes the harsh flavour variation e.g. bitterness and green flavour whereas the sweeter flavour components are described by component e.g. sweetness and carrot flavour.

3.2. Prediction of sensory quality from instrumental measurements

The selection of factor combinations for prediction of the five sensory attributes, bitterness, sweetness, terpene flavour, green flavour and carrot flavour was based on RMSECV and are listed in Table 3. The same number of factors from block I are selected for bitterness, sweetness and green flavour, whereas the factor combinations from block II are six, two and five, respectively. Terpene flavour and carrot flavour both have a factor combination of one factor for block I and five factors for block II. Fig. 1 shows RMSECV versus the number of components for each of the two blocks contributions for the prediction of green flavour. The arrow shows the selected factor combination (block I: #3 and block II: #5).

The search for common factors did not result in any common scores between the sensory attributes and the two blocks. Thus, the following data interpretation is based on the individual block contributions from block I to block II.

The prediction of the five sensory attributes gave prediction errors (RMSECV) between 0.98 and 1.36 and correlation coefficients (r) between 0.60 and 0.81 (Fig. 3, Table 3). The explained Y-variances were between 15.1% and 66.0% (Table 3). The worst prediction was observed for the attribute carrot flavour and the best prediction was seen for green flavour. The attributes green flavour, bitterness and terpene flavour showed fairly good predictions (r/RMSECV/% exp-Y = 0.81/0.98/66.0, 0.79/1.23/62.3 and 0.71/1.04/50.2) whereas an unexpected poor prediction was seen for sweetness (r/RMSECV/% exp-Y = 0.67/1.36/44.6). This was unexpected since the carrot samples showed a large span in the total sugar content ranging from 5.53 to 9.32 mg/100 g FW and the sensory evaluation had a wide span in sensory score of sweetness (Table 2). During the sugar analysis, the individual sugars (sucrose, fructose and glucose) content was determined, though none of these variables contributed to the prediction of sweetness in accordance with results by Rosenfeld, Samuelsen, and Lea (1998). In contrast, Seljåsen et al. (2001a, 2000b) found a correlation between sweet taste and sucrose. Similarly to sweetness, carrot flavour showed poor prediction (r/RMSECV/% exp-Y = 0.60/1.37/15.1). Carrot flavour is a complex sensory attribute that is likely to be determined by terpene content but also sweetness could play a role for carrot flavour. The result indicates that carrot flavour might represent an attribute difficult for the judges to evaluate. In general a better overall prediction between the flavour attributes and the volatile compounds was expected since sensory-perceived flavours in raw carrots such as green, carrot and terpene have earlier been cor-

Table 4
Important predictors based on biplots and regression coefficients

Bitterness	Green flavour	Terpene flavour
Dry matter	Dry matter	5-CQA
5-CQA	5-CQA	Sucrose
Sucrose	6MM	FaDOH
FaDOH	Sucrose	FAOH
Terpinolene	FaDOH	6 MM
γ-Terpinene	FAOH	Terpinolene
β-Bisabolene	Terpinolene	γ-Terpinene
α-Pinene	γ-Terpinene	Caryophyllene
β-Pinene	Cuparene	β-Bisabolene
Caryophyllene	Limonene	α-Pinene
Sabinene	β-Myrcene	β-Pinene
Cuparene	β-Pinene	Cuparene
	Sabinene	

The list is not ordered according to importance.

related well with the volatile compounds (Rosenfeld et al., 2002; Seljåsen et al., 2001b; Simon et al., 1980).

The flavour attributes are expected to be described mainly by the volatile compounds (Rosenfeld et al., 2002; Seljåsen et al., 2001b; Simon et al., 1980). The present study does confirm previous findings to a certain extent but it also shows a contribution from dry matter and non-volatile compounds (block I). The presence of block I in the prediction model improves the prediction with the exception for carrot flavour where less than 1% is explained by block I. Green flavour is described by both blocks but block II is responsible for explaining 60% of the explained Y-variance. A smaller contribution from block I is observed for terpene flavour where the ratio between the two blocks is 1–9, respectively. The same ratio between blocks is observed for bitterness. Sweetness is the only attribute where block I constitutes to a higher explained Y-variance of approximately 10% in comparison to block II.

One of the forces of LS-parPLSc is the ability to provide information about individual block contributions and significant variables. In order to take a closer look at the relationships between the sensory quality, dry matter, non-volatile and the volatile measurements for the best performing predictions, green flavour, terpene

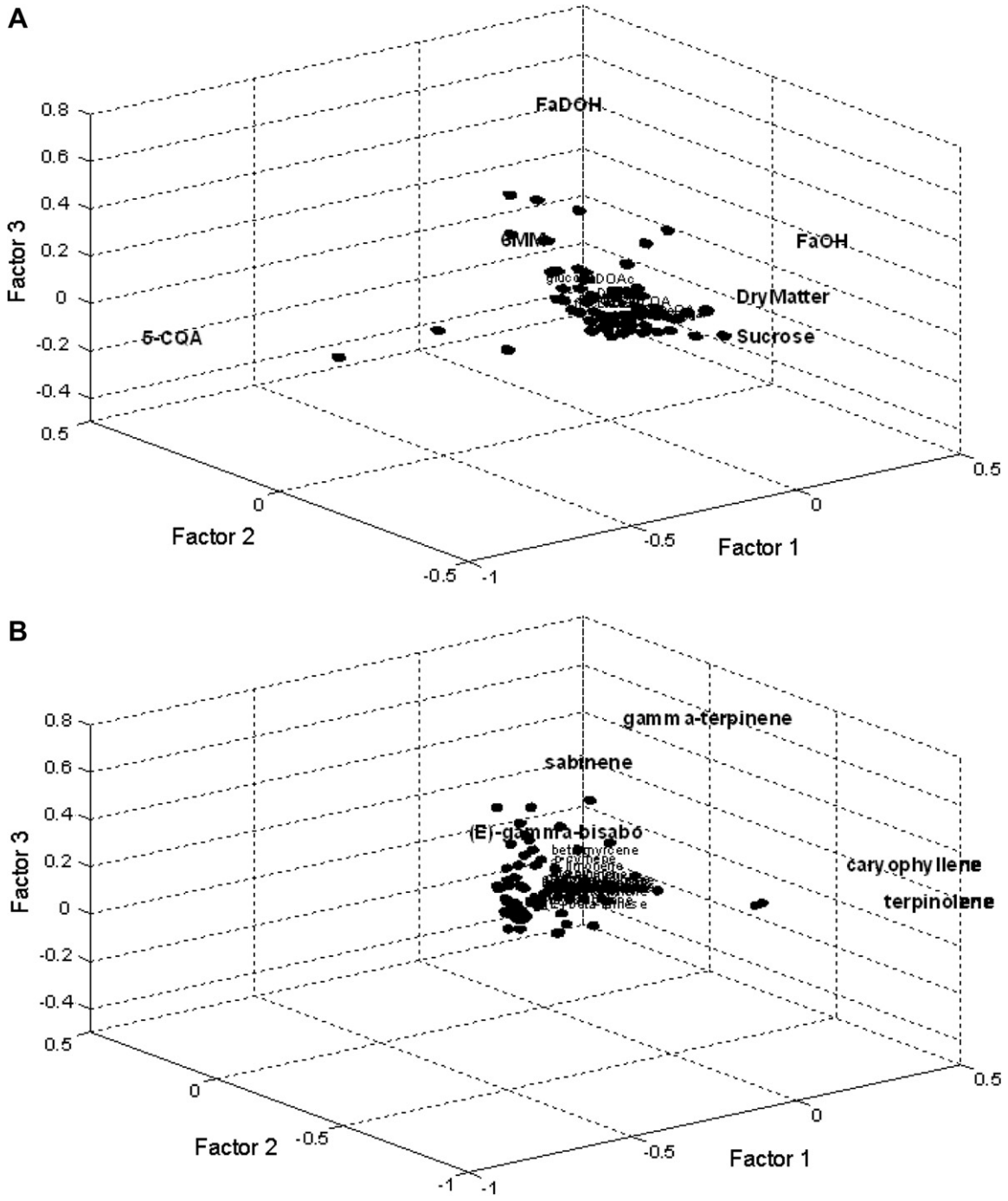


Fig. 4. Bi-plots of factor 1–3 for (A) block I and (B) block II. The descriptors in bold represent the variables spanning the space of the PLS regression model of green flavour in block I and block II.

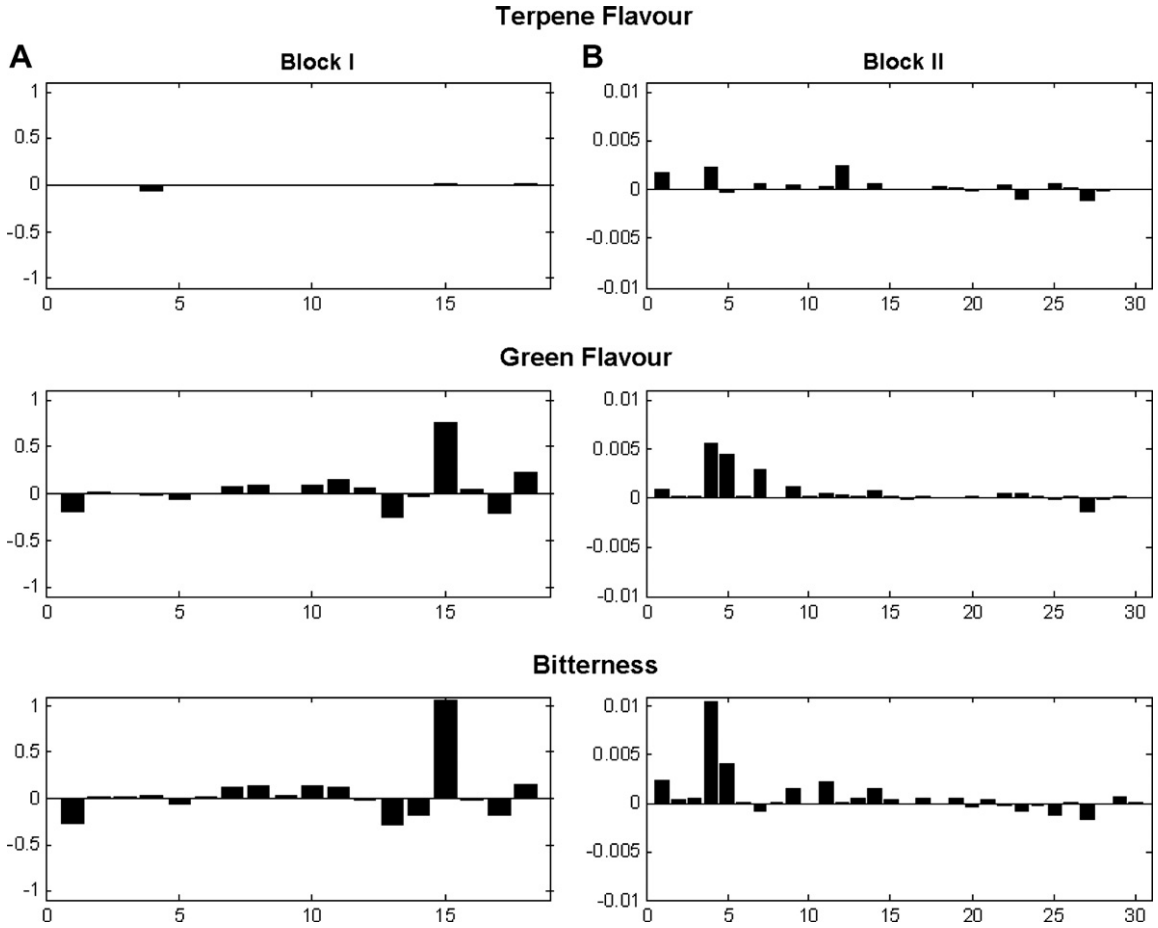


Fig. 5. Illustration of (A) regression coefficients for block I. (B) regression coefficients for block II.

flavour and bitterness are further investigated. Table 4 lists the important predictors for bitterness, green flavour, and terpene flavour.

For green flavour a model with dry matter and non-volatiles (block I) explains one third of the Y-variance. According to loadings and regression coefficients (Figs. 4A and 5A) the important predictors are dry matter, chlorogenic acid (5-CQA), 6-MM, FaDOH, and FaOH and sucrose. The main contributors in explaining the remaining Y-variance from block II are terpinolene, sabinene, β-pinene, γ-terpinene, β-myrcene, limonene and cuparene (Figs. 4B and 5B).

For terpene flavour, the major contributors are 5-CQA, 6MM, FaDOH, FaOH and sucrose based on loadings and regression coefficients (Table 4 and Fig. 5A). The second block main contributors are terpinolene, γ-terpinene, α-pinene, β-pinene, caroyphyllene, cuparene and β-bisabolene. The two most outstanding volatile compounds are terpinolene, which mainly seems to influence the first two factors, and γ-terpinene, which has higher impact on the higher factors. Both compounds show a fairly reasonable correlation to terpene flavour. A three factors model on block I show that dry matter, 5-CQA, FaDOH and sucrose are most important variables in describing bitterness. Block II gives the largest contribution when explaining bitterness using six factors, which count for more than 80% of the explained Y-variance. The important variables are terpinolene, β-pinene, sabinene, γ-terpinene, α-pinene, β-bisabolene, caryophyllene and cuparene.

Green flavour and bitterness seem to have common features when it comes to chemical and physical properties because the profiles of the regression coefficients of block I appear to be very similar (Fig. 5A). On the contrary, the aroma profiles seem to separate the two attributes as the regression coefficients differ for block II (Fig. 5B). The similarity in prediction of these two sensory attributes is supported by the initial PCA (Fig. 2B) which also indicated that the two attributes were highly correlated. Intercorrelated sensory attributes with nearly equal regression coefficients are assumed to represent some redundancy and likely explain the same basic information. The prediction of bitterness, green flavour and terpene flavour were explained by almost similar chemical components, due to the high correlation between these sensory attributes.

Dry matter is naturally related to texture attributes. However, Table 4 indicates that this variable is an important predictor for the sensory quality with respect to flavours and bitterness. Previously, high correlation between sucrose and dry matter was found and relationship between bitter tasting flavour compounds and dry matter were observed in carrots grown at high temperatures (Rosenfeld, Samuelsen, & Lea, 1998).

Previously, falcarindiol have been correlated to bitterness in raw carrots whereas falcarinol, the most abundant and bioactive of the polyacetylenes, was inversely correlated to bitterness of carrots (Kreutzmann et al., 2008). Kreutzmann et al. (2008) did not find 5-CQA to be correlated to bitterness or other flavours and the contribution of 5-CQA and other phenolic acids to taste in raw carrots is still unclear.

The volatile compounds found to be important predictors are all considered as key flavour compounds of raw carrots (Kjeldsen, Christensen, & Edelenbos, 2003; Rosenfeld et al., 2002; Seljåsen et al., 2001b). Harsh flavour attributes (terpene flavour, green flavour, bitterness and burning aftertaste) are found to increase in

intensity with increasing terpene content (Kreutzmann et al., in press). According to Burdock (2002) caryophyllene is characterised by terpene odour, like cloves or turpentine while γ-terpinene is characterised as having citrus-like odour and significant correlations between green flavour and terpene flavour and γ-terpinene and caryophyllene, respectively, have previously been found (Rosenfeld et al., 2002; Seljåsen et al., 2001b). The rest of the volatile compounds isolated in this study were highly intercorrelated and did not contribute to further explanation of the sensory quality (Fig. 4).

4. Conclusion

The overall results show that the sensory quality variation in the material regarding bitterness, green flavour and terpene flavour can be reasonably predicted by relatively few volatile and non-volatile compounds. Non-volatile compound predictors were cholorogenic acid (5-CQA), sucrose, 6-methoxymellein (6-MM), falcarindiol (FaDOH), and falcarinol (FaOH) whereas the volatile compounds with the highest prediction impact are the flavour compounds terpinolene, β-pinene, sabinene, γ-terpinene, α-pinene, β-bisabolene, caryophyllene and cuparene. Despite that the results revealed some reliable relationships, a large variance (about 40%) in the sensory block of variables remained unexplained and still needs further investigation for an in-depth understanding of the sensory quality.

LS-ParPLS/LS-parPLSc is shown to be a useful tool when handling several types of data blocks in one regression model. The method does not pose mathematically induced scaling problems and associated interpretation issues as in some of the existing multi-block regression methods where scaling can influence the regression model and especially interpretation significantly. The study confirmed that the predictive performance of the LS-parPLS is not improved compared to the ordinary PLS but information on the individual block level is easy to access.

Acknowledgements

We gratefully acknowledge the excellent technical assistance of Birgitte Foged, Kim G. Vitten and Leon Hansen. The investigation is a part of a Ph.D. study supported financially by The Danish Institute of Agricultural Sciences/University of Aarhus.

References

Alasalvar, C., Grigor, J. M., Zhang, D., Quantick, P. C., & Shahidi, F. (2001). Comparison of volatiles, phenolics, sugars, antioxidant vitamins, and sensory quality of different Colored carrot varieties. *Journal of Agricultural and Food Chemistry, 49*, 1410–1416.

Baardseth, P., Rosenfeld, H. J., Sundt, T. W., Skrede, G., Lea, P., & Slide, E. (1996). Evaluation of carrot varieties for production of deep fried carrot chips – II. Sensory aspects. *Food Research International, 5*, 215–224.

Berglund, A., & Wold, S. (1999). A serial extension of multiblock PLS. *Journal of Chemometrics, 13*, 461–471.

Burdock, G. A. (2002). *Fenaroli's handbook of flavor ingredients* (5th ed.). London, UK: CRC Press.

Eastman, H. T., & Kranowski, W. J. (1982). Cross-validatory choice of the number of components from a principal component analysis. *Technometrics, 24*, 73–77.

Hogstad, S., Risvik, E., & Steinsholt, K. (1997). Sensory quality and chemical composition in carrots: A multivariate study. *Acta Agriculturae Scandinavica, Section B: Soil and Plant Science, 47*, 253–264.

Jongen, W. M. F. (2000). Food supply chains: From productivity toward quality. In R. L. Shewfelt & B. Brückner (Eds.), *Fruit and vegetable quality. An integrated view* (pp. 3–20). CRC Press LLC.

Jørgensen, K., Segtnan, V., Thyholt, K., & Næs, T. (2004). A comparison of methods for analysing regression models with both spectral and designed variables. *Journal of Chemometrics, 28*, 451–464.

Jørgensen, K., Mekvik, B.-H., & Næs, T. (2007). Combining designed experiments with several blocks of spectroscopic data. *Chemometrics and Intelligent Laboratory Systems, 88*, 154–166.

Kaack, K., Christensen, L. P., Hansen, S. L., & Grevsen, K. (2004). Non-structural carbohydrates in processed soft fried onion (*Allium cepa* L.). *European Food Research and Technology, 218*, 372–379.

Kidmose, U., Hansen, S. L., Christensen, L. P., Edelenbos, M., Larsen, E., & Nørbæk, R. (2004). Effects of genotype, root size, storage, and processing on bioactive compounds in organically grown carrots (*Daucus carota* L.). *Journal of Food Science, 69*, S388–S394.

Kjeldsen, F., Christensen, L. P., & Edelenbos, M. (2001). Quantitative analysis of aroma compounds in carrot (*Daucus carota* L.) cultivars by capillary gas chromatography using large-volume injection technique. *Journal of Agricultural and Food Chemistry, 49*, 4342–4348.

Kjeldsen, F., Christensen, L. P., & Edelenbos, M. (2003). Changes in volatile compounds of carrots (*Daucus carota* L.) during refrigerated and frozen storage. *Journal of Agricultural and Food Chemistry, 51*, 5400–5407.

Kreutzmann, S., Christensen, L. P., & Edelenbos, M. (2008). Investigation of bitterness in carrots (*Daucus carota* L.) based on quantitative chemical and sensory. *LWT, 41*, 193–205.

Kreutzmann, S., Thybo, A. K., Christensen, L. P., & Edelenbos, M. (in press). The role of volatiles on aroma and flavour perception in coloured carrot genotypes. *International Journal of Food Science and Technology*.

Kreutzmann, S., Thybo, A. K., & Bredie, W. L. P. (2006). Training of a sensory panel and profiling of winter hardy and coloured carrot genotypes. *Food Quality and Preference, 18*, 482–489.

Mekvik, B.-H., Jørgensen, K., Måge, I., & Næs, T. (submitted for publication). LS–PLS regression: Combining categorical design variables with blocks of spectroscopic measurements.

Måge, I. (2006). Modelling and optimisation of industrial processes with raw material variation. Doctor Scientiarum Thesis, Norwegian University of Life Sciences.

Martens, H., & Næs, T. (1989). *Multivariate calibration*. Wiley. p. 419.

NIST. (1998). *The NIST/EPA/NIH mass spectral database*, version 6.0; Washington DC, USA: National Institute of Standards and Technology.

Rosenfeld, H. J., Aaby, K., & Lea, P. (2002). Influence of temperature and plant density on sensory quality and volatile terpenoids of carrot (*Daucus carota* L.) root. *Journal of the Science of Food and Agriculture, 82*, 1384–1390.

Rosenfeld, H. J., Samuelsen, R. T., & Lea, P. (1998). The effect of temperature on sensory quality, chemical composition and growth of carrots (*Daucus carota* L.). II. Constant diurnal temperatures under different seasonal light regimes. *Journal of Horticultural Science and Biotechnology, 73*, 578–588.

Seljåsen, R., Bengtsson, G. B., Hoftun, H., & Vogt, G. (2001a). Sensory and chemical changes in five varieties of carrot (*Daucus carota* L.) in response to mechanical stress at harvest and post-harvest. *Journal of the Science of Food and Agriculture, 81*, 436–447.

Seljåsen, R., Hoftun, H., & Bengtsson, G. B. (2001b). Sensory quality of ethylene-exposed carrots (*Daucus carota* L. cv 'Yukon') related to the contents of 6-methoxymellein, terpenes and sugars. *Journal of the Science of Food and Agriculture, 81*, 54–61.

Simon, P. W., Peterson, C. E., & Lindsay, R. C. (1980). Genetic and environmental influences on carrot flavor. *Journal of the American Society for Horticultural Sciences, 105*(3), 416–420.

Simon, P. W., Peterson, C. E., & Lindsay, R. C. (1982). Genotype, soil, and climate effects on sensory and objective components of carrot flavor. *Journal of the American Society for Horticultural Sciences, 107*(4), 644–648.

Wangen, L. E., & Kowalski, B. R. (1988). A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of Chemometrics, 3*, 3–20.

Westerhuis, J. A., Kourti, T., & MacGregor, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics, 12*, 302–321.

Westerhuis, J. A., & Smilde, A. K. (2001). Deflation in multiblock PLS. *Journal of Chemometrics, 15*, 485–493.

Wold, S., Martens, H., & Wold, H. (1984). In S. Wold (Ed.), Muldast Proceedings, Technical Report, Research Group for Chemometrics, Umeå Universitet.

Wold, S., Kettaneh, N., & Tjessam, K. (1996). Hierarchical multiblock PLS and PC models for easier interpretation and as an alternative to variable selection. *Journal of Chemometrics, 10*, 463–482.

PAPER VI

Vibeke Svensson and Charlotte Møller Andersen

Characterization of Brine from Salted Herring using Fluorescence Spectroscopy.
LEBENSMITTEL-WISSENSCHAFT UND-TECHNOLOGIE-FOOD SCIENCE AND TECHNOLOGY, (submitted), 2008

EEM fluorescence spectroscopy as a fast method to assess the brine composition of salted herring

Abstract

Ripening of barrel-salted herring (*Clupea harengus*) is evaluated by the use of fluorescence spectroscopy and protein determinations. During ripening, protein degradation takes place in the herring and protein is extracted into the brine. The present study aims at identifying parameters which are correlated to the ripening characteristics of barrel-salted herring and which can provide a better understanding of the ripening process. Front face fluorescence landscapes were obtained by measuring directly on the brine from barrel-salted herring. These data were analyzed by parallel factor analysis (PARAFAC), which revealed four fluorophores, tryptophan (two states), vitamin B6 and riboflavin. All four parameters showed an increase in concentration during the storage period corresponding to an increase in protein content that varied from 3g/100g at day 60 to 5g/100g after 277 days of storage. It was not possible to see a difference in the development of the four fluorophores during the ripening period. The protein content was predicted from the fluorescence landscapes by partial least squares (PLS). The use of unfolded fluorescence spectra gave an RMSECV of 0.26 g/100g and a correlation between the measured protein content and the predicted values of 0.86.

Introduction

Herring (*Clupea harengus*) has always been of special interest in the Scandinavian countries and the production of salted herring has been carried out for centuries (Cutting, 1955). In recent years, focus on promoting fish and fish products has increased, which has required improvement and optimization of the traditional production processes. In the herring industry, the understanding of how the herring changes during manufacturing is crucial in order to optimize the process.

Salting of herring is performed by an initial soaking of the fish in salt, which initiates the extraction of fluids from the fish (Vokresensky, 1965). After the soaking period, saturated brine consisting of water and salt is added and the herring is stored for a period of up to 12 months. During this time, the fish undergo changes which will result in the distinct taste and texture associated with salted herring. The proteins undergo a degradation process (Nielsen, 1995) in which the protein fractions will be extracted into the brine and the concentration of protein in the brine will increase as it decreases in the herring. This exchange in protein constituents was observed by Andersen and co-workers and Stefansson et al. (Stefansson, Nielsen and Gudmundsdottir, 1995; Andersen, Andersen and Baron, 2007). The degradation products released into the brine consist of soluble nitrogenous compounds like peptides, free amino acids, smaller peptides and myofibrillar proteins (Nielsen, 1995). As the changes in the protein concentration and the composition of nitrogenous compounds are believed to reflect the ripening process, monitoring the changes in protein in the brine could be a good indicator of the ripening stage of the herring.

Today's monitoring of barrel-salted herring typically consists of conventional chemical analysis or visual inspection. The result of the visual inspection depends solely on the experience of the process operator who performs the evaluation. Thus, monitoring tools, which secure a uniform evaluation of the ripening process of herring, are needed to secure high product quality. A variety of well-suited analyses

for chemical composition (Nielsen, 1995) and texture analysis of fish exists (Nielsen, Hyldig, Nielsen and Nielsen, 2005), but the majority of the analyses are time-consuming and require laboratory skilled staff. In order to ensure better monitoring, an alternative method is needed. Such an alternative analysis would be required to be time efficient and easy to perform, thus not requiring specially trained laboratory staff. Spectroscopy may be able to meet these requirements. Spectroscopic analysis has the advantage of being fast and non-destructive. Near infrared (NIR) spectroscopy has previously shown potential as a screening method for sorting herring based on fat content (Nielsen et al., 2005). Furthermore, in a previous study of brine from salted herring, NIR spectra correlated well with protein concentration (Svensson, Nielsen and Bro, 2004). Fluorescence spectroscopy can be an alternative method due to its high specificity and sensitivity and the possibility of measuring small changes in protein concentrations. It has shown promising possibilities for assessing the quality of various fish products such as oxidation of fish oils (Hasegawa et al., 1992), determination of dioxin content (Pedersen et al., 2002), storage of canned sardines (Aubourg and Medina, 1997; Aubourg et al., 1997) and differentiation of fresh and aged cod, mackerel, salmon and whiting (Dufour et al., 2003). Performing fluorescence measurements at several excitation wavelengths and measuring the corresponding emission spectra will result in a fluorescence landscape also referred to as an excitation-emission matrix (EEM). Fluorescence landscapes can be particularly well analyzed by multi-way analysis, giving a unique decomposition of the underlying structure, as shown by Pedersen et al. (Pedersen, Munck and Engelsen, 2002) and Andersen and Bro (Andersen and Bro, 2003) .

The objective of the present study is to investigate whether fluorescence spectroscopic measurements on brine can provide useful information about the ripening process of barrel-salted herring. The brine and fish constitute a closed

system in which any compound that leaves the herring is extracted into the brine. Thus, the brine can be used to describe the changes taking place in the herring. The main focus of the paper will be on protein and its degradation products to see if they can be related to the ripening of salted herring. Parallel factor analysis (PARAFAC) will be applied to decompose the fluorescence landscapes and multivariate or multi-way partial least squares regression (PLS and N-PLS) will be used to correlate the protein content in the brine with the fluorescence measurements.

Material and Methods

Production of Barrel-Salted Herring

Herring (*Claupea harengus*) caught in the Baltic Sea by commercial vessels was brought to the manufacturing company, Lykkeberg A/S (Hørve, Denmark). The herring was processed according to Lykkeberg A/S production protocol for barrel-salted herring. 100 kg of whole-headed herring was mixed with 10 kg of salt and placed in a 100 L plastic barrel. After 24 hours, the barrel was filled with saturated brine and stored at 0-5˚C for 277 days. A total of ten barrels (batches) were monitored and sampling was performed six times during the storage period, yielding 60 samples. Sampling times were at days 60, 96, 123, 172, 213 and 277. At each sampling time, 20-25 ml brine was taken for analysis from each barrel. Upon sampling, the brine was centrifuged at 10,000 g for 20 min at 5˚C to remove tissue parts and insoluble matter. The samples were kept in the freezer until analyses were carried out.

Protein Analysis – Kjeldahl

The protein content was determined by the Kjeldahl method (Total N x 6.25) (AOAC, 1996).

Fluorescence Spectroscopy

All samples were measured on a Perkin-Elmer LS50B spectrometer (Beaconsfield; Buckinghamshire; UK) equipped with FLDM software. The brine samples were kept in plastic vials and placed to thaw for 10 to 15 min. at room temperature. 3 mL of brine was transferred to a quartz cell (1×1 cm) and the sample was measured at room temperature. Fluorescence measurements were performed in the range of 250-550 nm for excitation with 10 nm intervals (30 excitation wavelengths) and emission was measured for every nm in the interval from 260-650 nm (391 emission wavelengths). The slit width was 5 nm for both excitation and emission and a 1% attenuation filter was used. The measurements started with the highest excitation wavelength and ended with the lowest to minimize photodecomposition of the sample.

Parallel Factor Analysis (PARAFAC)

Multi-way analysis in the form of PARAFAC (Harshman 1970) was used to decompose the fluorescence landscapes. The fluorescence data were held in a three-way array of the size 60 × 391 × 30 corresponding to 60 samples (objects), 391 emission wavelengths and 30 excitation wavelengths. PARAFAC decomposes the fluorescence landscapes into a number of factors (F) by minimizing the sum of squares of the residual (e_{ijk}) (Eq. 1). The number of factors is a reflection on how many fluorescent compounds are present in the fluorescence data.

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \qquad (i=1,\ldots,I;\ j=1,\ldots,J;\ k=1,\ldots,K;\ f=1,\ldots,F) \qquad (1)$$

Each PARAFAC factor consists of A-scores ($a_1, \ldots, a_F$) and two sets of loadings, the B-loadings ($b_1, \ldots, b_F$) and the C-loadings ($c_1, \ldots, c_F$). The A-scores represent the sample direction and give the relative concentrations of the estimated components. The B-loadings are estimates of the underlying emission spectra and the C-loadings

are estimates of the corresponding excitation spectra. Rayleigh and Raman scatter was removed and the parameters constrained to be non-negative, in order for the estimated model to make chemical sense (Bro and Sidiropoulos, 1998). The non-negativity constraints were applied on all three dimensions, sample, emission and excitation.

The factor selection was based on split-half tests. The data set was split up in two sets of 30 samples and a PARAFAC model on each of the two sets was calculated. A valid model has approximately similar loadings in the two split-half models. In addition to the split-half test, the residuals and core consistency were assessed using common sense and prior knowledge about the ripening process and PARAFAC modeling of fluorescence spectroscopic data (Andersen and Bro, 2003)

Multivariate regression

Partial least squares regression (PLS) was applied to predict the chemically measured protein concentration from the fluorescence measurements. Multivariate calibration was performed using the scores from the PARAFAC model as well as the raw unfolded fluorescence spectra (Martens and Næs, 1989). Furthermore, N-PLS was also applied (Bro, 1996). The regression models were evaluated by the correlation coefficients between predicted protein and measured protein content and by the root mean squared error of cross validation (RMSECV) (Eq. 2) using segmented cross validation (Eastment and Krzanowski, 1982). The segments were selected as batches, thus one batch was left out each time.

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (2)$$

All calculations were performed in Matlab Version 7.4.0 (R2007a) (MathWorks, Inc.), the PLS Toolbox (www.Eigenvector.com) and The Unscrambler ® v. 9.2 (Camo A/S).

Results and Discussion

Protein concentration in the brine

The Kjeldahl analysis shows that the protein content in the brine increases to approx. 3 g/100g after 60 days of storage (Figure 1). From day 60 to day 277, an increase in the protein content from 3 g/100g to 5 g/100g is observed. The protein content still seems to increase until the last day of sampling, which was performed on day 277. The changes in the individual batches show that the protein concentration does not follow exactly the same development in all the batches, but the overall trend is similar.

These initial observations on the development in the protein content in brine are in agreement with the findings by Andersen et al. (2007) who reported the same development in protein content as a function of time. That study also showed that the dry matter content in brine increased with 4% w/w, mainly due to protein and peptides with a molecular weight below 55 kDa (Andersen et al., 2007).

Fluorescence measurements

The raw fluorescence landscapes show that fluorescence compounds are present in the brine samples. In Figure 2, a fluorescence landscape of a representative brine sample is shown. The landscape is shown without the Rayleigh and Raman scatter. The first peak observed is the most dominant peak with λex_{max} around 280-290 nm and λem_{max} at 330-350 nm. This region is normally associated with tryptophan

fluorescence. The same peak has a shoulder with $\lambda ex_{max}/\lambda em_{max}$ at 340/400 nm. Due to the strong signal of the first peak, the peaks in th2e higher spectral region can not be visualized, but a closer look reveals that more peaks are present with $\lambda ex_{max}/\lambda em_{max}$ around 400/450 nm and 460/525 nm.

PARAFAC modeling

A four-component PARAFAC model seems to perform best. The four components explain 99.6% of the variation in data. One may argue that a fifth component is present at 390/460 nm, but according to the split-half analysis with five factors, this is not plausible (results not shown). In Table 1, the λex_{max} and λem_{max} for the four estimated PARAFAC components are listed.

The aromatic amino acids tryptophan, tyrosine and phenylalanine are known to exhibit fluorescence in the region of $\lambda em_{max}/\lambda ex_{max}$ 260-300/340-350 nm. They all exhibit natural fluorescence and are known to be present in herring. However, phenylalanine and tyrosine show much less fluorescence than tryptophan (Burnstein, Vedenkina, and Ivkova, 1973; Wolfbeis, 1985).

The first and the second component with $\lambda em_{max}/\lambda ex_{max}$ 290/341 nm and 300/350 nm, respectively, can be assigned to tryptophan fluorescence. Different peak characteristics have been reported for tryptophan, depending on the state of the molecule and the molecular environment (Duggan, Bowman, Brodie and Udenfriend, 1957; Pajot, 1976; Duggan et al., 1957). There may be several reasons why there are two tryptophan contributions. When assessing the estimated excitation maxima, it has to be taken into account that the maximum fluorescence intensity of the fluorescence spectrophotometer was reached for samples with high tryptophan content and that excitation was performed with a 10 nm step between each excitation. Thus, it can be expected that both components would have excitation maximum between 290 and

300 nm. In addition, previous studies have shown that emission spectra of tryptophan can shift in accordance to its surroundings. The physico-chemical conditions of the surroundings such as polarity, viscosity and availability of charged groups will affect the tryptophan residue and thereby the fluorescence properties of tryptophan. The polarity of the solvent may cause the fluorescence maximum to shift to shorter or longer wavelengths. If tryptophan residues are exposed to water, the maximum emission wavelengths are found between 340-350 nm, whereas the totally buried residues will emit around 330 nm. Thus, both components describe tryptophan exposed to water, but they experience somewhat different solvent polarity (Lakowicz, 1999).

The third component has $\lambda ex_{max}/\lambda em_{max}$ at 330/394 nm. Two compounds in herring, collagen and pyridoxine (vitamin B6), have fluorescence characteristics similar to this. However, collagen is not soluble in water and it will not be present in the brine. Instead, the component is believed to be vitamin B6. Vitamin B6 covers the three vitamin B6 active components: pyridoxal, pyridoxine and pyridoxamine. Vitamin B6 is soluble in water and is present in plant and animal tissue (Swatland, 1987; Duggan et al., 1957; Torres-Sequeiros, Garda-Falcon and Simal-Gandara, 2001). Each of the vitamin B6 compounds show different fluorescence characteristics, depending on the surroundings such as pH (Wolfbeis, 1985). At pH 7, the following $\lambda ex_{max}/\lambda em_{max}$ have been reported for pyridoxal; 330/385 nm, pyridoxamin; 335/400 nm and pyridoxine; 340/400 nm (Duggan et al., 1957). These values are quite similar to the properties of component 3.

Riboflavin is the flavin with the most intense fluorescence and is believed to be described by the fourth component with λex_{max} at 390 and 440 nm and λem_{max} at 521 nm. In neutral aqueous solution, emission has been reported to be around 515 nm (Wolfbeis, 1985). Duggan et al. (1957) reported riboflavin to have λem_{max} at 520 nm

with λex_{max} at 270, 370 and 445 nm. The excitation in the lower region is not observed, but may be drowned by the signal from tryptophan. Furthermore, the characteristics of the riboflavin component are similar to the riboflavin component identified in yoghurt which showed λex_{max} at 370 nm and 445 nm and λem_{max} at 530 nm (Christensen, Becker and Frederiksen, 2005).

In Figure 3, the four excitation and emission loadings are presented. The third and fourth factor emission spectra look a bit peculiar in that they are not smooth and both spectra indicate that more that one peak are present. There is no obvious reason for this appearance, but it could be due to the low intensity of the two fluorophores in question - close to the signal-to-noise limit. Another guess would be that one more fluorophore is present, whose concentration correlates with the scores of components 3 and 4. The excitation loading for the third component has a shoulder at 380/390 nm and both emission spectra have a shoulder at 460 nm.

For the region above excitation 320 nm, it was difficult to decide how many compounds were present, probably due to a lower signal in this region. Local PARAFAC models based on selected areas in this region were calculated in order to see if more compounds were present (results are not shown). The conclusion was that the system is most adequately described as a four-component system.

The development in scores compared with the protein concentration as a function of time show that the scores follow the protein concentration fairly well with a general increase throughout the storage period. In the first 60 days, an increase of 3 % is observed, whereas the development in protein slows down after day 60 with an increase in protein of 2 % during the remaining 210 days. In Figure 5, it can be observed that batch 3 looks different from the others and one sampling time in batch 4 at day 123 deviates. A closer look at the fluorescence log book for batch 3 reveals

that problems were experienced with the instrument when measuring batch 3. Due to this, batch 3 is left out in the remainder of this study. Furthermore, day 123 in batch 4 is also considered as an outlier. In the initial data inspection, these outliers were not recognized, because the samples do not deviate from the overall distribution of samples. However, when looking at the scores and the development in time, it is obvious that these samples are different. The estimated loadings do not change when excluding the samples in question in the PARAFAC modeling. This shows that the batches and outlier sampling times do not introduce a shift in the fluorescence and the PARAFAC model is not affected. The increase in tryptophan scores as a function of time can be explained by the degradation of protein into aromatic amino acids, increasing the content of water soluble amino acids in the brine. The vitamin B6 and riboflavin also increase in the brine throughout the ripening period. It is not possible to distinguish between the four fluorophores based on their development in scores. The correlation between the protein content and the PARAFAC scores is calculated to get an indication of how the four factors are related to protein. The correlations of the factors assigned to tryptophan are 0.77 and 0.76 for factors 1 and 2, respectively. Factor 3 has a correlation of 0.74 and factor 4 has a correlation of 0.78. The relatively high correlation between factor 4 (riboflavin) and the protein content is remarkable, but riboflavin in the brine may be connected to the protein content and thus it must be riboflavin bound to protein which is present in the brine.

A difference between the concentrations of the four components in the brine was expected, especially in the beginning of the storage period. If sampling had been performed more frequently in the first weeks of the storage period, more information on this development of fluorophores in the early phase of the ripening might have been obtained. Sampling of fish can be rather difficult to do uniformly (Andersen and Bro, 2004; Jepsen, Pedersen and Engelsen, 1999) due to the non-homogeneity of fish.

Measurements obtained from the brine can hence be a simple alternative to measuring directly on the fish and will also give a more uniform picture of the changes that a batch undergoes as a whole. However, it is difficult to obtain a precise sample representation of the entire barrel. The brine is surrounding the herring and it is not possible to access the brine in the bottom of the barrel or in the middle of the barrel. Therefore, sampling was performed on the top brine, which may vary according to the location of the herring in the barrel.

Multivariate regression

Protein content in the brine is predicted from the fluorescence data to verify if the identified fluorophores are related to the protein content in brine. The prediction is performed in three ways: 1) using PARAFAC scores in a PLS regression, 2) using the raw fluorescence landscapes in an N-PLS regression and 3) using the unfold landscapes in a PLS regression. Regression was performed excluding the outliers mentioned above and two samples for which the protein determinations were missing. A total of 51 samples were included. The three regression approaches do not differ much in their performance. Table 2 presents the results from the multivariate calibration. All models required two PLS components. The correlation between the chemically measured protein content and the predicted values (r) and the RMSECV indicate that the model made on the unfolded fluorescence landscapes performed somewhat better than the other models. The model made using the PARAFAC scores gave the highest explained variance of X and the lowest explained variance of Y, which seems reasonable, since PARAFAC decomposed the fluorescence landscapes with the focus of obtaining the best description of X, whereas the two other methods did not decompose the data before the prediction of protein content. However, the predictive ability obtained using PARAFAC scores indicates that the four extracted fluorescence compounds are connected to the changes in protein concentration. The

results are satisfactory in comparison to a previous study which predicted the protein content using NIR spectroscopy. The error estimates of this study showed an RMSECV of 0.25 g/100g (Svensson et al., 2004). However, the NIR spectra do not provide detailed information about the smaller fractions present in herring or brine, such as the degradation products of protein like peptides and amino acids. In order to study the smaller protein fragments in brine, fluorescence spectroscopy can be a better possibility.

Overall, it is shown that fluorescence spectroscopy is a fast alternative to the standard methods used to assess the chemical composition of brine. It is fairly simple to perform and it can provide a detailed picture of the changes that occur in the constituents tryptophan, riboflavin, and vitamin B6. Probably, similar results or even better predictions could have been obtained if measurements had been performed directly on the herring. However, the brine has the advantage of being a fluid, which reduces the sampling heterogeneity and makes it easier to handle the samples when performing the measurements. Furthermore, measurements on brine can be considered as a non-destructive measurement. This opens up for the use of fluorescence spectroscopy in process control. Typically, it takes around 15 minutes to obtain a fluorescence landscape, depending on the instrumental settings. For most industrial purposes, this is too long to be used as a fast, on-line or at-line measurement. However, in relation to the measurement of ripening of herring, which can last for many months, it may not be a problem. Furthermore, with the knowledge of the specific fluorophores that are correlated to the ripening, it is possible to design dedicated fluorescence probes that only measure the most important excitation and emission wavelengths.

Conclusion

Front face fluorescence spectroscopy of brine from barrel-salted herring is measured and the data analyzed with PARAFAC. The modeling reveals the presence of four fluorophores in the brine. They are believed to be tryptophan (two states), vitamin B6 and riboflavin. All four compounds show similar overall development in scores throughout the ripening period. The protein content increases similarly from approx. 3g/100g at day 60 to 5g/100g after 277 days of storage. PLS regression on unfolded fluorescence spectra gives the best prediction (RMSECV of 0.26 g/100g and r of 0.86) of protein concentration compared to PLS regression using scores from a PARAFAC model and N-PLS performed on the raw fluorescence landscapes. The present study shows that fluorescence spectroscopy can reflect the overall changes in protein concentration in brine with the same accuracy as NIR.

Acknowledgement

The authors would like to thank Lykkeberg A/S for providing the brine samples. We acknowledge Lisbeth T. Hansen for her excellent laboratory work in carrying out the fluorescence measurements and Karin Reimers for performing the protein analysis. Åsmund Rinnan is thanked for valuable discussions about fluorescence spectroscopy and chemometric issues. FØTEK 3 (93S.2444-Å01-00100) is acknowledged for financial support during the experiment.

References

Andersen,C.M. & Bro, R. (2003). Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. *Journal of Chemometrics.* 17 (4), 200-215.

Andersen, C.M. & Bro, R. (2004). Quantification and handling of sampling errors in instrumental measurements: a case study. *Chemometrics and Intelligent Laboratory Systems.* 72 (1), 43-50.

Andersen,E., Andersen,M.L., & Baron,C.P. (2007). Characterization of oxidative changes in salted herring (Clupea harengus) during ripening. *Journal of Agricultural and Food Chemistry.* 55 (23), 9545-9553.

AOAC. 1996. Method no. 39.1.19.981.10. Crude protein in meat Block digestion method. Block digestion method. In Official methods of analysis. Gaithersburg, MD, USA. AOAC International.

Bro,R. (1996). Multiway calibration. Multilinear PLS. *Journal of Chemometrics.* 10 (1), 47-61.

Bro,R.,& Sidiropoulos,N.D. (1998). Least squares algorithms under unimodality and non-negativity constraints. *Journal of Chemometrics.* 12 (4), 223-247.

Burnstein,E.A., Vedenkina,N.S., & Ivkova,M.N. (1973). Fluorescence and the location of tryptophan residues in protein molecules. *Photochemistry and Photobiology.* 18 262-279.

Christensen,J., Becker,E.M., & Frederiksen,C.S. (2005). Fluorescence spectroscopy and PARAFAC in the analysis of yogurt. *Chemometrics and Intelligent Laboratory Systems.* 75 (2), 201-208.

Cutting,C.L. 1955. Fish Saving. Leonard Hill Limited, London.

Duggan,D.E., Bowman,R.L., Brodie,B.B., & Udenfriend,S. (1957). A Spectrophotofluorometric Study of Compounds of Biological Interest. *Archives of Biochemistry and Biophysics.* 68 1-14.

Eastment,H.T.,& Krzanowski,W.J. (1982). Cross-Validatory Choice of the Number of Components from A Principal Component Analysis. *Technometrics.* 24 (1), 73-77.

Jepsen,S.M., Pedersen,H.T., & Engelsen,S.B. (1999). Application of chemometrics to low-field H-1 NMR relaxation data of intact fish flesh. *Journal of the Science of Food and Agriculture.* 79 (13), 1793-1802.

Lakowicz,J.R. 1999. Principles of Fluorescence Spectroscopy. Kluwer Academic/Plenum Publishers, New York. 1-698 pp.

Martens,H., and Næs,T. 1989. Multivariate calibration. Wiley, New York. -442 pp.

Nielsen,D., Hyldig,G., Nielsen,J., & Nielsen,H.H. (2005). Lipid content in herring (Clupea harengus L.) - influence of biological factors and comparison of different methods of analyses: solvent extraction, Fatmeter, NIR and NMR. *Lwt-Food Science and Technology.* 38 (5), 537-548.

Nielsen,H.H. 1995. Proteolytic enzyme activities in salted herring during cold storage. Department of Biotechnology, The Technical University of Denmark, 1-131.

Pajot,P. (1976). Fluorescence of Proteins in 6-M Guanidine-Hydrochloride - Method for Quantitative-Determination of Tryptophan. *European Journal of Biochemistry.* 63 (1), 263-269.

Pedersen,D.K., Munck,L., & Engelsen,S.B. (2002). Screening for dioxin contamination in fish oil by PARAFAC and N-PLSR analysis of fluorescence landscapes. *Journal of Chemometrics.* 16 (8-10), 451-460.

Stefansson G., Nielsen H. H., & Gudmundsdottir G. (1995). Ripening of Spice-Salted Herring. Report No. 613, (1-45). Danish Institute of Fisheries Research, Department of Seafood Research. *Nordic Council of Ministers*

Svensson,V.T., Nielsen,H.H., & Bro,R. (2004). Determination of the protein content in brine from salted herring using near-infrared spectroscopy. *Lebensmittel-Wissenschaft Und-Technologie-Food Science and Technology.* 37 (7), 803-809.

Swatland,H.J. (1987). Effect of Excitation Wavelength on the Separation of Types I and III Collagen by Fiber Optic Fluorimetry. *Journal of Food Science.* 52 (4), 865-868.

Torres-Sequeiros,R.A., Garda-Falcon,M.S., & Simal-Gandara,J. (2001). Analysis of Fluorescent Vitamins Riboflavin and Pyiridoxine in Beverages with Added Vitamins. *Chromatographia.* 53 236-239.

Vokresensky,N.A. 1965. Salted herring. In G.Borgstrom. Fish as food, processing: part 1. (pp. 107-131). New York. Academic Press Inc

Wolfbeis,O.S. 1985. Fluorescence of organic natural products. In S.G.Schulman. Molecular Luminescence Spectroscopy. (pp. 167-370). John Wiley & Sons.

Figure Legends

Figure 1. Protein concentration in brine as a function of storage time

Figure 1. PARAFAC emission and excitation loadings from split-half test with 4 factors. (-) split-half model I and (--) split-half model II

Figure 2. Fluorescence landscapes of brine samples from barrel-salted herring. A) Raw spectra $\lambda ex_{max}/\lambda em_{max}$ of 250-550/270-650 nm. B) Zoom in on the landscape. The arrows indicate possible peaks.

Figure 3. Estimated PARAFAC loadings of four excitation and emission loadings

Figure 4. PARAFAC scores (1-4) and protein concentration (- - - -) as a function of time for 10 batches. Protein concentration and PARAFAC scores have been normalized. The legends for the PARAFAC scores are not given, as they show approximately the same increasing pattern.

Figure 1

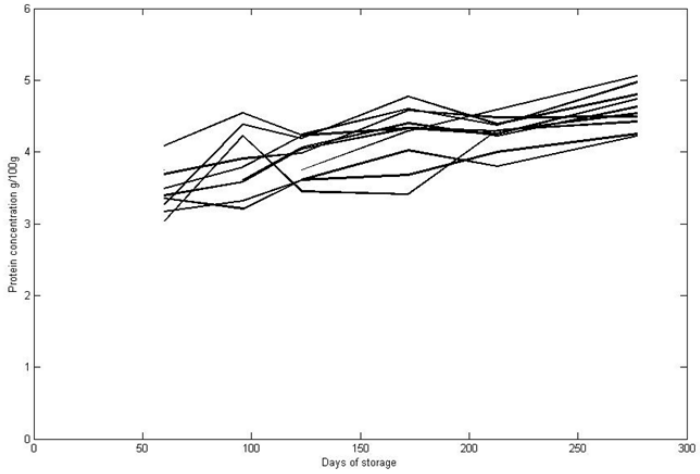


Figure 2

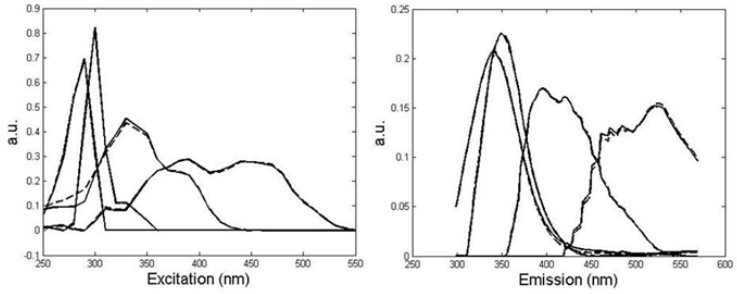


Figure 3

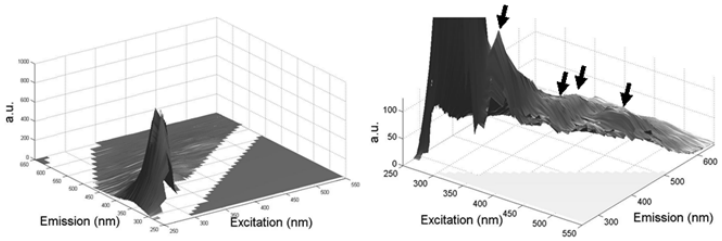


Figure 4

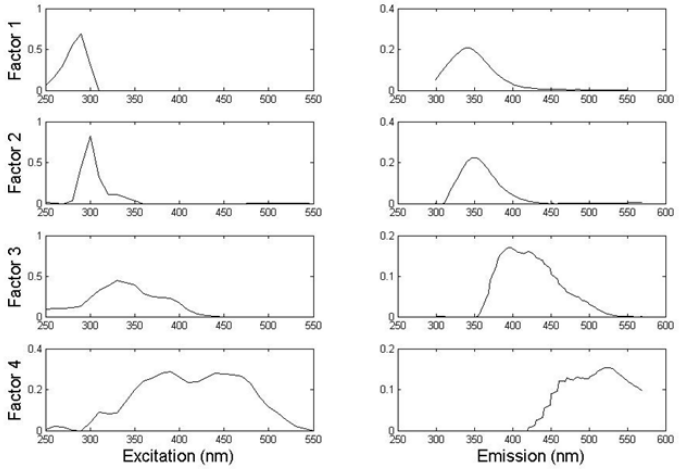


Figure 5

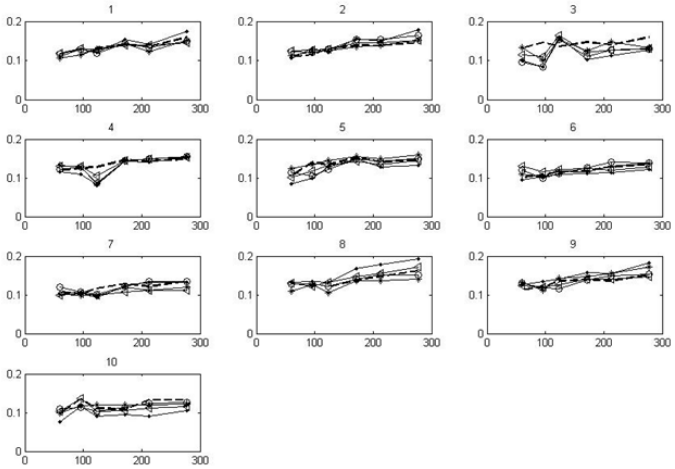


Table 1. λex_{max} and λem_{max} for a four-factor PARAFAC model

	λex_{max}	λem_{max}
Factor 1	290	341
Factor 2	300	350
Factor 3	330	396
Factor 4	390/440	521

Table 2. Results from PLS regression and N-PLS regression

	# Factors	r	RMSECV (g/100g)	Exp. X (%)	Exp. Y (%)
PLS on PARAFAC scores	2	0.818	0.287	95.9	69.2
PLS on unfold EEMs	2	0.855	0.257	82.9	75.0
N-PLS on EEM	2	0.828	0.282	76.3	77.58