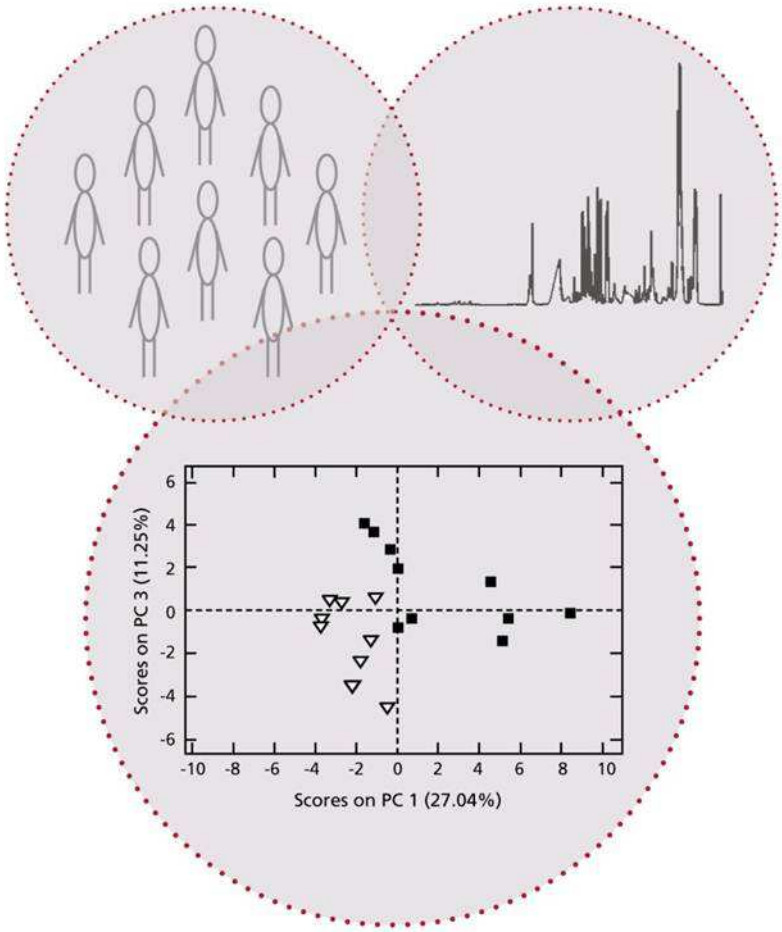
DEPARTMENT OF FOOD SCIENCE
FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

Metabolomics

from a chemometric point of view



PhD thesis by Maja H. Kamstrup-Nielsen • 2013

DEPARTMENT OF FOOD SCIENCE
FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

Metabolomics

from a chemometric point of view

Maja H. Kamstrup-Nielsen

PhD Thesis

2013

University of Copenhagen · Faculty of Science ·
Department of Food Science (FOOD) · Quality & Technology
Rolighedsvej 30 · 1958 Frederiksberg C · Denmark

Title
Metabolomics from a chemometric point of view

Submission
26 August 2013

Defence
25 November 2013

Supervisor
Professor Rasmus Bro
Quality and Technology, Department of Food Science
Faculty of Science, University of Copenhagen, Denmark

Opponents
Assistant Professor Jeroen Jansen
Radboud Universiteit Nijmegen, Netherlands

Chemometrician/Data analyst, PhD Carsten Ridder
Lattec I/S, Denmark

Associate Professor Thomas Skov
Quality and Technology, Department of Food Science
Faculty of Science, University of Copenhagen, Denmark

Cover illustration
Britt Friis and Maja H. Kamstrup-Nielsen

"A man should look for what is, and not for what he thinks should be"

Albert Einstein

Preface

This thesis has been carried out at the Quality and Technology group, Department of Food Science (FOOD), Faculty of Science, University of Copenhagen under supervision by Professor Rasmus Bro as a requirement for obtaining the PhD degree.

I am beyond grateful to my supervisor Rasmus for his endless support and supervision during the work of my PhD. He has guided me through the world of chemometrics, listened to my frustrations and supported me throughout the completion of this thesis.

The publications in the thesis have involved many co-authors. I am grateful to all for their contributions and collaboration. Much of my thesis has been conducted in collaboration with Louise Hansen and Anja Olsen from the Danish Cancer Society and Professor Lars Dragsted from Department of Human Nutrition. I truly appreciate our collaboration and our many fruitful discussions and pleasant meetings.

I would also like to thank all my colleges at Q&T. Q&T is a wonderful place work both from an academic and a social point of view. Especially, thanks to Francesco Savorani and Flemming H. Larsen for assistance and guidance with NMR spectroscopy. Also great thanks to Lea G. Johnsen for fantastic collaboration, to Abelrhani Mourhib for indispensable help with preparing plasma samples and to Bekzod Khakimov for unlimited assistance with the GC-MS. Special thanks to Anders Lawaetz and Morten Rasmussen for collaboration on papers, but especially for endless discussions on chemometrics, help in Matlab and comments on parts of this thesis. José Amigo is also thanked for commenting parts of the thesis.

I have had several office mates to whom I am all grateful: Minah Mosele, Hanne Winning, Peter I. Hansen and especially Parvaneh Ebrahemi, for being a good friend and support through the finalization of this thesis. A special thanks to Lotte B. Lyndgaard – you have been one of my best friends throughout my entire life at university from the day we started studying food

science. We have shared our problems and successes whenever needed, especially during our time as PhD students.

Finally, I would like to thank my family and friends – particularly Henrik and August – for always believing in me, supporting me and for reminding me what life is really about.

Maja H. Kamstrup-Nielsen

Frederiksberg, August 2013

Summary

Metabolomics is the analysis of the whole metabolome and the focus in metabolomics studies is to measure as many metabolites as possible. The use of chemometrics in metabolomics studies is widespread, but there is a clear lack of validation in the developed models. The focus in this thesis has been how to properly handle complex metabolomics data, in order to achieve reliable and valid multivariate models. This has been illustrated by three case studies with examples of forecasting breast cancer and early detection of colorectal cancer based on data from nuclear magnetic resonance (NMR) spectroscopy (**Paper II**), fluorescence spectroscopy (**Paper III**) and gas chromatography coupled to mass spectrometry (GC-MS).

The principles of the three data acquisition techniques have been briefly described and the methods have been compared. The techniques complement each other, which makes room for data fusion where data from different platforms can be combined.

Complex data are obtained when samples are analysed using NMR, fluorescence and GC-MS. Chemometrics methods which can be used to extract the relevant information from the obtained data are presented. Focus has been on principal component analysis (PCA), parallel factor analysis (PARAFAC), PARAFAC2 and partial least squares discriminant analysis (PLS-DA) all being described in depth. It can be a challenge to determine the appropriate number of components in PARAFAC2, since no specific tools have been developed for this purpose. **Paper I** is a presentation of a core consistency diagnostic aiding in determining the number of components in a PARAFAC2 model. It is of great importance to validate especially PLS-DA models and if not done properly, the developed models might reveal spurious groupings. Furthermore, data from metabolomics studies contain many redundant variables. These have been suggested to be eliminated using an approach termed reduction of redundant variables (RRV), which is time consuming but efficient, since the curse of dimensionality is reduced and the risk of over-fit is decreased.

The use of appropriate multivariate models in metabolomics studies has been presented in the three case studies. In the first case study, plasma samples from healthy individuals have been analysed by NMR. Some have developed breast cancer later in life and these have been separated from healthy individuals by means of a properly validated PLS-DA model based on NMR data with RRV and known risk markers. The sensitivity and specificity values are 0.80 and 0.79, respectively, for a test set validated model.

The second case study is based on plasma samples with verified colorectal cancer and three types of control samples analysed by fluorescence spectroscopy. The acquired data have been analysed by PARAFAC models and the components from the PARAFAC models have been used as variables in seven PLS-DA models in order to separate the cancer samples from the control groups. Sensitivity and specificity values of approximately 0.75 make fluorescence spectroscopy a potential tool in early detection of colorectal cancer.

Finally, plasma samples have been analysed using GC-MS. The method requires extensive sample preparation and therefore the study can only be considered a feasibility study with room for optimization. However, 14 plasma samples were analysed and the results indicate that GC-MS-based metabolomics in combination with PARAFAC2 modelling is applicable for extracting relevant biological information from the plasma samples.

Overall, the work in this thesis shows that suitable and properly validated chemometrics models used in metabolomics are very useful in forecasting and early detection of cancer. The use of chemometrics in metabolomics can e.g. increase the understanding of the underlying etiology of cancer and could be extended to cover other diseases as well.

Resumé

Metabolomics er undersøgelse af hele metabolomet, og fokus i metabolomics-studier er at måle så mange metabolitter som muligt. Brugen af kemometri i metabolomics-studier er udbredt, men der er en klar mangel på validering af de udviklede modeller. Fokus i denne afhandling har været, hvordan metabolomics-data bliver håndteret korrekt, således at pålidelige og valide multivariate modeller opnås. Dette er blevet illustreret ved tre casestudier med eksempler på forudsigelse af brystkræft og tidlig detektering af kolorektalkræft baseret på data fra nuklearmagnetisk resonans (NMR) spektroskopi (**Artikel II**), fluorescens-spektroskopi (**Artikel III**) og gaskromatografi koblet til massespektrometri (GC-MS).

Principperne bag de tre måleteknikker er blevet kort beskrevet og sammenlignet. Teknikkerne komplementerer hinanden, hvilket giver anledning til datafusion, hvor data fra forskellige platforme kombineres.

Når prøver analyseres ved NMR, fluorescens og GC-MS, opnås komplekse data. Kemometriske metoder, der kan anvendes til at udtrække den relevante information fra de opnåede data, er præsenteret. Fokus har været på principal component analysis (PCA), parallel factor analysis (PARAFAC), PARAFAC2 og partial least squares discriminant analysis (PLS-DA). Det kan være en udfordring at bestemme antallet af komponenter i PARAFAC2, idet ingen specifikke værktøjer er blevet udviklet til dette formål. **Artikel I** er en præsentation af en core consistency diagnostik, der hjælper til at bestemme antallet af komponenter i en PARAFAC2-model. Det er utrolig vigtigt at validere især PLS-DA-modeller, og hvis det ikke er gjort ordentligt, vil de udviklede modeller vise ukorrekte grupperinger. Ydermere indeholder data fra metabolomics-studier mange redundante variable. Det er foreslået, at disse elimineres ved at anvende en metode kaldet reduktion af redundante variable (RRV), som er tidskrævende, men effektiv, da "the curse of dimensionality" reduceres, og risikoen for overfit derved mindskes.

Brugen af egnede multivariate modeller i metabolomics er blevet præsenteret i de tre casestudier. I det første studie er plasmaprøver fra raske individer

blevet analyseret med NMR. Nogle har udviklet brystkræft senere i livet, og disse er blevet adskilt fra de raske individer ved hjælp af en PLS-DA-model baseret på data fra NMR med RRV og kendte risikomarkører. Sensitiviteten og specificiteten er henholdsvis 0,80 og 0,79 for en testsætvalideret model.

Det andet studie er baseret på plasmaprøver med verificeret kolorektalkræft og tre typer kontrolprøver analyseret med fluorescens-spektroskopi. De opnåede data er blevet analyseret med PARAFAC-modeller, og komponenterne fra PARAFAC-modellerne er blevet brugt som variable i syv PLS-DA-modeller for at separere kræftprøverne fra kontrolgrupperne. Sensitiviteten og specificiteten på cirka 0,75 viser, at fluorescens-spektroskopi er et potentielt redskab til tidlig detektering af kolorektalkræft.

Til sidst er plasmaprøver blevet analyseret med GC-MS. Teknikken kræver omfattende prøveforberedelse, og derfor kan studiet kun betragtes som et forstudie med plads til optimering. Dog blev 14 plasmaprøver analyseret, og resultaterne indikerer, at GC-MS-baseret metabolomics kombineret med PARAFAC2-modellering kan anvendes til at udtrække relevant biologisk information fra plasmaprøverne.

Alt i alt viser arbejdet i denne afhandling, at egnede og ordentligt validerede kemometriske modeller anvendt i metabolomics er meget anvendelige til at forudsige og tidligt detektere kræft. Brugen af kemometri i metabolomics kan eksempelvis øge forståelsen af den underliggende kræftetiologi og kunne meget vel udvides til andre sygdomme.

List of Publications

Paper I

Maja H. Kamstrup-Nielsen, Lea G. Johnsen, Rasmus Bro (2013): "Core consistency diagnostic in PARAFAC2", *Journal of Chemometrics*, 27(5), 99-105

Paper II

Rasmus Bro, **Maja H. Kamstrup-Nielsen**, Søren Balling Engelsen, Francesco Savorani, Flemming H. Larsen, Morten A. Rasmussen, Louise Hansen, Anja Olsen, Anne Tjønneland, Lars O. Dragsted (2013): "Forecasting breast cancer development using metabolomics and bio-contours", *submitted*

Paper III

Anders Juul Lawaetz, Rasmus Bro, **Maja Kamstrup-Nielsen**, Ib Jarle Christensen, Lars N. Jørgensen, Hans J. Nielsen (2012): "Fluorescence spectroscopy as a potential metabonomic tool for early detection of colorectal cancer", *Metabolomics*, 8, 111-121

List of Abbreviations

1D	One Dimensional
2D	Two Dimensional
AUC	Area Under the Curve
BMI	Body Mass Index
CC	Core Consistency
CRC	ColoRectal Cancer
CV	Cross-Validation
CPMG	Carr-Purcell-Meiboom-Gill
EEM	Excitation Emission Matrix
GC	Gas Chromatography
*i*PLS	Interval Partial Least Squares
LC	Liquid Chromatography
LOO	Leave-One-Out
LV	Latent Variable
MCR	Multivariate Curve Resolution
MS	Mass Spectrometry
NMR	Nuclear Magnetic Resonance
NOESY	Nuclear Overhauser Effect SpectroscopY
OPLS-DA	Orthogonal Projections to Latent Structures-Discriminant Analysis
PARAFAC	PARAllel FACtor analysis
PARAFAC2	PARAllel FACtor analysis 2
PC	Principal Component
PCA	Principal Component Analysis
PLS	Partial Least Squares
PLS-DA	Partial Least Squares-Discriminant Analysis
ROC	Receiver Operating Characteristic
RRV	Reduction of Redundant Variables
TIC	Total Ion Chromatogram
XC-MS	Gas or liquid Chromatography-Mass Spectrometry

Contents

Chapter 1

Introduction

1.1. Background

Over the past decades, studies of the human metabolism in relation to various diseases have evolved. Metabolomics is now a well-established term in life sciences and the number of publications are greatly increasing [1;2]. One focus in metabolomics is the research regarding detection of cancer and the common analytical techniques in metabolomics research are nuclear magnetic resonance (NMR) and gas or liquid chromatography coupled to mass spectrometry (XC-MS). A method used as a supplement to the common techniques is fluorescence spectroscopy, which is rarely applied in metabolomics, but has a potential. All the methods provide complex data, and several multivariate mathematical methods are applicable in order to extract the information of interest from these complex data.

However, there is especially one issue in (some) metabolomics studies that need to be handled: Doing proper data analysis. In many research areas, classic univariate statistical methods are applied. However, univariate models are not adequate when tens of thousands of metabolites are to be compared simultaneously to reveal differences between for example healthy and diseased subjects. Picking out selected metabolites on the sole assumption that this or these particular metabolites are the only important ones is unfortunate, and potential relevant information might be lost. Fortunately, most metabolomics studies have embraced the usefulness of multivariate models and chemometrics, where all measured metabolites are investigated simultaneously, but sometimes single metabolites are picked out (e.g. Sreekumar *et al.* from 2009 [3]). However, despite the increasing role of multivariate data analysis in metabolomics, much of the developed models lack reliability. A thorough review of the applications of multivariate data analysis and chemometrics in metabolomics studies has been published by Madsen *et al.* in 2010 [1]. In the review, the authors pinpoint that missing validation is generally a major issue in the data analysis in many studies. It

is of utmost importance that prediction models are capable of predicting for example the cancer status of *new* samples and not just the status of the samples used to build the model. Otherwise, nothing general can be revealed from the research.

In this thesis, development of proper multivariate models will be exemplified using data from plasma samples of breast cancer and colorectal cancer. Breast cancer is the most common type of cancer among women in the Western part of the world. In Denmark, approximately 4,700 women are diagnosed with breast cancer each year [4]. Despite a good prognosis in many cases (the five-year survival rate is higher than 80%) the greatest chance of survival is early detection of the tumour. Today, mammography screenings are offered to middle aged women in many Western countries. One important (and controversial) drawback of mammography is the risk of too many false positives. Some detected (and treated) tumours would never progress to a stage that would affect the well-being of the patient. This is called over-diagnosis, and has been heavily discussed in recent years [5]. Colorectal cancer is one of the most frequent malignant diseases in the Western part of the world among men and women. In Denmark, 1,575 men and women on average were diagnosed with colorectal cancer each year from 2006 – 2010 [6]. For colorectal cancer, the five-year survival rate is 56% for men and 60% for women. The survival rate is very likely to increase if an increasing fraction of colorectal cancer is detected in earlier stages. Today, screening for colorectal cancer is carried out by home sampling of faeces to detect the presence of occult blood. In case of a positive sample, the test is followed up by colonoscopy. This procedure is uncomfortable for many people resulting in low compliance of the screening programme. In addition, only six out of ten individuals with colorectal cancer are detected using this method [7]. With the drawbacks of today's primary screening methods, there is a considerable need for the development of methods which are able to non-invasively detect cancer at an *early* stage of the disease or maybe even anticipate cancer before the disease can be detected clinically.

1.2. Aim of thesis

Multivariate data analysis in metabolomics offer many challenges, and the work in this thesis focuses on how to properly handle metabolomics data and on the development of valid models. This will be illustrated using different examples: (i) forecasting breast cancer status from healthy females by NMR spectroscopy with data reduction, (ii) early detection of colorectal cancer by

fluorescence spectroscopy, and (iii) early detection of colorectal cancer by GC-MS.

1.3. Thesis outline

This thesis is based on three quite diverse papers all published in or submitted to peer-reviewed journals, and a feasibility study (not published). The main focus in the thesis is on the multivariate methods applied on complex metabolomics datasets, using examples from NMR on healthy plasma samples with later development of breast cancer, fluorescence spectroscopy on plasma samples with colorectal cancer and GC-MS on standard plasma samples.

Chapter 2 describes the term metabolomics and discusses why it is an interesting and relevant research area.

In Chapter 3, the three data acquisition techniques NMR spectroscopy, fluorescence spectroscopy and GC-MS are briefly described.

Chapter 4 is devoted to description and discussion of the relevant models in multivariate data analysis, and is one of the main cores of this thesis. The chapter is divided into several major topics: an explorative part concerning PCA, MCR, PARAFAC and PARAFAC2 (Paper I), a classification part concerning PLS-DA and OPLS-DA, and the challenges when using these models. Also variable reduction and selection will be discussed and finally the term data fusion will be discussed.

Chapter 5 is the second core of this thesis, where two specific applications from Paper II (NMR) and Paper III (fluorescence) are illustrated in detail, and unpublished work on acquiring GC-MS data from plasma samples is presented. Issues concerning sample preparation prior to GC-MS analysis are also discussed.

Finally, in Chapters 6 and 7, the conclusions from the work of this thesis, and perspectives and ideas for future research, are presented.

Chapter 2

What is Metabolomics?

The terms metabonomics and metabolomics were first introduced in the late 1990s and the early 2000s, respectively [8;9]. The two terms are very much alike and are used interchangeably by scientists in the field. However, there is a slight distinction between the two terms. According to the paper by Nicholson *et al.* 1999 [9], metabonomics is defined as "*the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification*" – a response or reaction when state changes occur in the organism are measured quantitatively from changes in the metabolites. Fiehn (2001) [8] is addressing the term metabolomics to be "*a comprehensive and quantitative analysis of all metabolites…*" focussing on the study of the metabolome. Throughout this thesis the term metabolomics will be used for consistency, but either term could be equally appropriate.

The metabolome is the complete set of low molecular weight compounds of a specific physiological state found in cells [10]. There are different approaches to the analyses of the metabolic response, and these can roughly be divided in four groups [8;11;12]: (i) metabolite targeted profiling, (ii) metabolite profiling, (iii) metabolomics and (iv) metabolic fingerprinting. Metabolite targeted profiling is the analysis of specific changes of the metabolome directly related to for example enzyme activity. Metabolite profiling (sometimes denoted metabolic profiling) is the analysis of a group of metabolites associated with a specific pathway. Metabolomics is the analysis of the *whole* metabolome, and finally metabolic fingerprinting is the classification of samples based on their biological relevance, where it is not necessary to determine the levels of all metabolites. In this thesis, the aim is to detect the widest possible range of metabolites, and therefore the metabolomics approach is called for.

The metabolomics approach was first applied in plant science [13] and toxicology [14], but the field quickly emerged to other areas, for example disease diagnosis in humans. Most metabolomics studies in disease diagnosis analyse body fluids or tissue samples to detect changes in metabolites related

to the specific disease. The analyses are carried out using different analytical techniques. The most commonly applied methods are Nuclear Magnetic Resonance (NMR) spectroscopy, Liquid Chromatography-Mass Spectrometry (LC-MS) and Gas Chromatography-Mass Spectrometry (GC-MS). The technique fluorescence spectroscopy is not commonly applied in metabolomics studies, due to the limited range of metabolites which are detectable by fluorescence spectroscopy. However, the method has previously been applied to classify breast cancer samples [15].

In the present thesis, metabolomics studies of forecasting and early detection of cancer are used as examples, and metabolomics in cancer diagnosis and detection has been investigated in numerous publications – see for example [16-22]. Some studies on early detection of cancer have been published [16;23], but no studies – to the knowledge the author – have been published on forecasting cancer before the disease is clinically detectable. In a paper by van der Greef *et al.* from 2004 [24], the relevance of metabolomics as a potential tool for detection of early metabolite pertubations in relation to diseases before symptoms are appearing is presented.

One general drawback of the majority of cancer related metabolomics studies so far, is the low number of samples. In the above mentioned publications, the minimum number of samples included in the analysis is 14 cancer patients and ten control samples. The remaining analyses included from 40 to 160 samples. One of the case studies presented in this thesis (Paper II) is based on samples from 838 samples. This number greatly increases the robustness of the developed models as will be discussed in Chapters 4 and 5. Additionally, the publications above present data from patients with diagnosed cancer. In Paper II, the subjects were all healthy when participating in the study and half of them developed breast cancer later in life.

Chapter 3

Data Acquisition Techniques

In this chapter, the principles of proton NMR spectroscopy, fluorescence spectroscopy and GC-MS will be described briefly. Additionally, the advantages and disadvantages for each technique will be summarized.

3.1. Nuclear Magnetic Resonance spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy was first introduced in the 1930s by Breit and Rabi [25]. NMR is a non-invasive, quantitative and highly reproducible analytical method, which provides detailed information about the metabolic profile of the analysed samples. Especially the possibility of measuring the contributions of small organic molecules has made proton NMR (^{1}H NMR) an attractive tool in metabolomics research, where changes in the biochemical composition can be monitored during progression of a disease [26]. With current ^{1}H NMR spectroscopy technology, it is common to measure down to concentration levels about one ppm, but this is highly dependent on the strength of the magnetic field, the number of scans of each sample (affecting the total acquisition time), and the number of identical protons originating from one molecule (many identical protons will give a higher signal).

There are several different experiments that can be used for acquiring spectra in NMR. For 1D analyses, the two most common techniques for bio-fluids are Carr-Purcell-Meiboom-Gill (CPMG) and 1D Nuclear Overhauser Effect Spectroscopy (NOESY) [27]. In this section, the principle of ^{1}H NMR will be described briefly and the differences between CPMG and NOESY will be illustrated. A more detailed description of ^{1}H NMR can be found in for example [28].

3.1.1. Principle

In ^{1}H NMR, the interaction between the spin states of protons and radio waves in an external magnetic field is studied. Protons are positively charged and have spin or rotation and can therefore be considered magnetic dipoles. The orientation of these dipoles will be along the rotational axis. When exposed to an external magnetic field, these dipoles will only be directed in two different states which are parallel or anti-parallel to the vertical axis. In order to obtain an observable signal, the protons are radiated with a radio frequency pulse equal to the frequency of the proton called the Larmor frequency. This absorbed energy induces the orientation of the protons spin states to be flipped to a horizontal position, and the return to an equilibrium status will result in the emission of an observable radio frequency signal, namely *Free Induction Decay* (FID) that is acquired. The FID is then converted from the time domain to the frequency domain by means of Fourier transformation giving rise to a spectrum which contains signals (peaks) characteristic of the different populations of protons. The position of these signals is determined by the *chemical shift*, which is expressed in parts per million (ppm). The chemical shift is the frequency difference between the studied proton and a selected reference molecule. The shift occurs due to *shielding*. Due to interactions between the protons and other nuclei within a molecule, different degrees of shielding are possible resulting in different resonating frequencies and therefore different chemical shifts. In the reference molecule – usually TSP (3-(trimethylsilyl)-propionic acid-d4) for aqueous samples – the proton shielding is almost complete. Hence, TSP proton signal will be located at 0 ppm and the studied protons will have positive ppm values, because of a lower degree of shielding.

The number of scans, which increases the signal intensity summing a new FID on top of the previous ones, is important for the *signal-to-noise ratio*. More scans will result in a higher signal-to-noise ratio; however the time of conducting the experiment will increase and it is therefore a trade-off between increasing the signal-to-noise ratio and fast data acquisition.

Most biological samples contain large amounts of water. Since ^{1}H NMR measures protons, the water in such samples will contribute to a major signal dominating the spectra and affecting its dynamic range in such a way that it will be impossible to detect small molecules. It is therefore necessary to induce water suppression during acquisition. This can be achieved using pre-saturation, where the water signal is irradiated and suppressed by a "long" radio frequency pulse.

The two types of techniques applied in the work of this thesis are CPMG and NOESY. The two methods complement each other, but many molecules are

observed in both methods. In CPMG, the bulky and wide signals due to the larger molecules such as proteins are suppressed, which results in spectra with a very flat baseline. The advantage of CPMG is that far more sharp signals ascribable to small compounds will be observable in the spectra. Figure 1 (top) shows an example of an average CPMG profile of human plasma. In NOESY, most compounds are observable, including the large proteins. Generally, NOESY gives a good overview of all molecules [27], but smaller molecules may be lost due to the fluctuating baseline. In Figure 1 (bottom) the average NOESY profile of human plasma is shown.

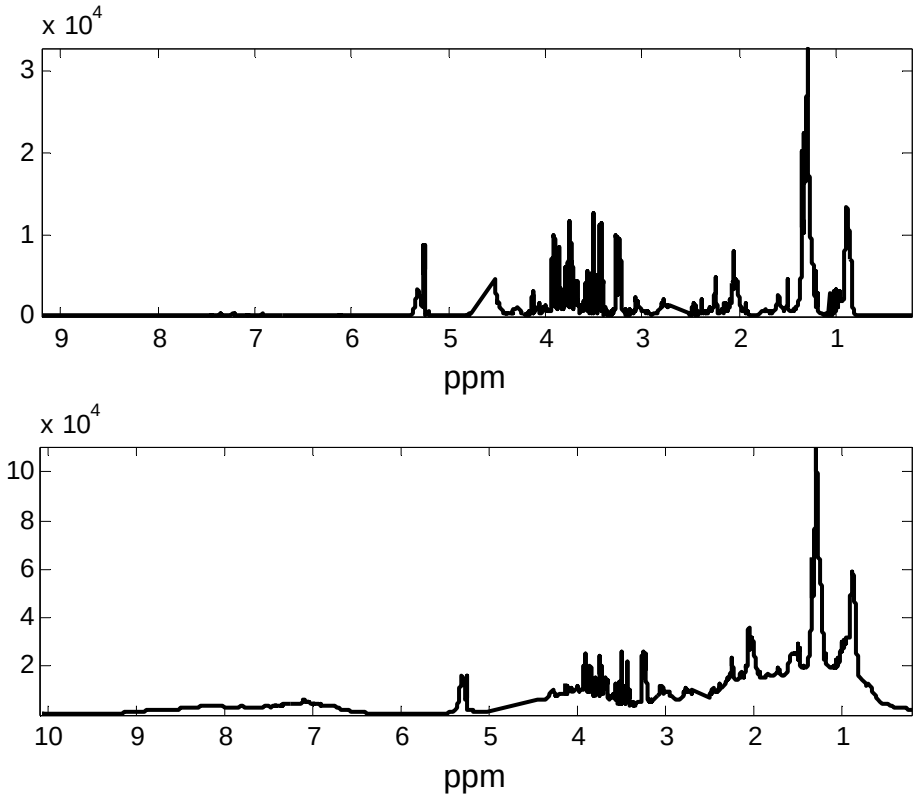


Figure 1. Top: Average CPMG profile of human plasma with suppression of large molecules (i.e. protein). Bottom: Average NOESY profile of human plasma

3.1.2. Pre-processing

After acquiring the spectra, a couple of aspects need to be considered prior to data analysis. Due to unavoidable experimental variations and small pH changes, there might be small shifts in the position of equal protons (unwanted changes in the chemical shift), baseline, etc. During the work of this thesis (Paper II), two pre-processing tools were applied prior to data

analysis in addition to automatically implemented phase correction: Alignment and normalization.

3.1.2.1. Alignment

In metabolomics studies, it is essential that shifts in the NMR signals are corrected. Otherwise, it will be difficult to locate and identify signals of interest. There are several ways to align spectral shifts, e.g. by binning, but the sole method applied here is interval-correlation-shifting (*i*coshift) [29]. *i*coshift is based on alignment of user-defined intervals in the NMR spectra. A starting point is to align according to the reference compound (TSP). However, alignment according to α-glucose has been performed prior to the data analysis in Paper II, since TSP binds the proteins present in blood and is therefore not reliable.

3.1.2.2. Normalization

^{1}H NMR is a quantitative method, and ideally the integrals of the signals are directly proportional to the concentration of the corresponding molecule. Due to possible experimentally induced variations and biological variations in body fluids, the concentration and signal of specific compounds can vary considerably. This can most easily be exemplified when considering the variation in urine concentration with water intake. Urine will be diluted with increased water intake, and the signal intensity will therefore be affected and unwanted variations will occur. These unwanted variations make it difficult to compare spectra between samples, and some adjustments are needed in order to reduce the variations. Normalization can aid in reducing the intensity or concentration variations by normalizing with a factor expressing the variations. Commonly, each spectrum is set to have a unit total intensity, where each data point is expressed as a fraction to the total spectral integral [30]. However, Craig *et al.* [30] state that concentration related problems are limited in plasma samples. The work in Paper II is based on normalized data (2-norm) on plasma samples, which is a common normalization technique in metabolomics [31]. However, the differences between the results in the data analysis performed on normalized and on raw spectra were minimal (not shown).

3.2. Fluorescence spectroscopy

The measurements in Paper III are obtained using fluorescence spectroscopy. The principle of fluorescence was discovered by Herschel in 1845 [32], but especially during the last two decades fluorescence spectroscopy has emerged in biological sciences [33]. Fluorescence spectroscopy is an analytical technique, where the intensity of light from different molecules is measured. A very brief description of the principles will be presented in the following.

3.2.1. Principle

When electrons of a molecule are excited from the ground state, there is a gap in the energy levels between the two states. If a molecule is exposed to light with a wavelength equal to this gap, the electrons of the molecule will go from the ground state to an excited state, which is also known as the molecular absorbance of light. After being excited, the molecule will decay from the excited state back to the ground state. This relaxation will for some molecules result in emission of light, also known as fluorescence. The energy states can be visualised using a Jablonski diagram, see for example [33]. The energy from emission is lower than the energy from excitation hence the wavelength at which emission occurs is longer than the wavelength at excitation. The difference between the excitation and emission wavelengths is called the Stoke's shift. When a wide range of wavelengths are recorded for excitation and emission their properties can be expressed by an excitation-emission matrix (EEM) represented by a fluorescence landscape. An example of a fluorescence landscape of human plasma can be seen in Figure 2.

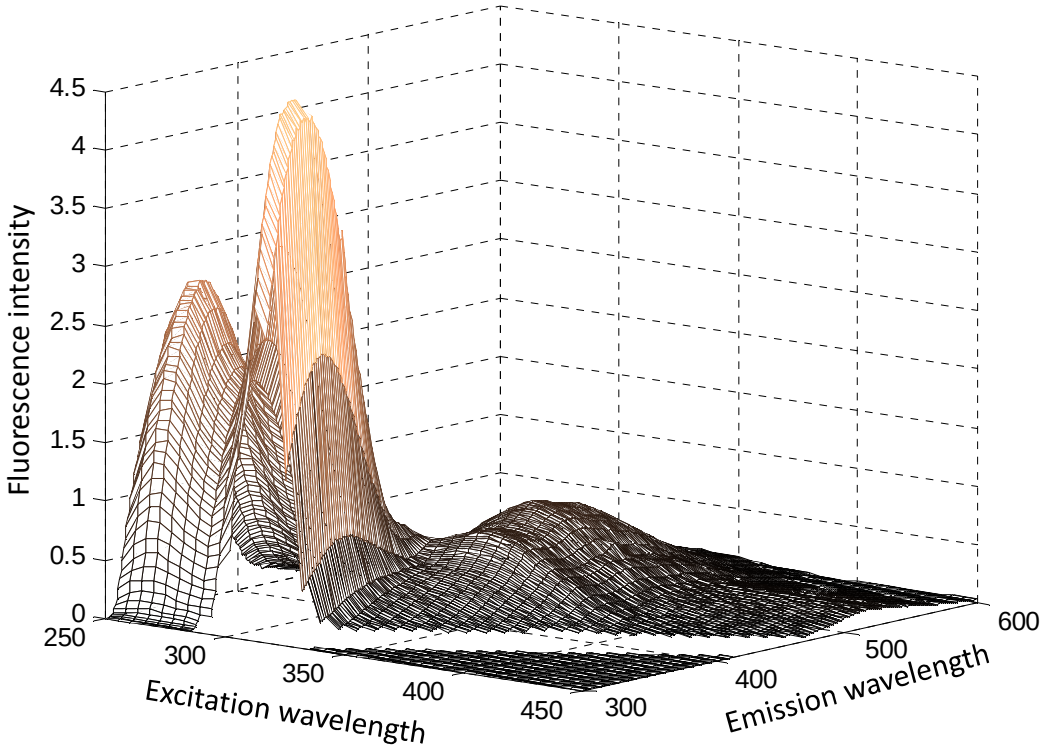


Figure 2. Example of a fluorescence landscape acquired from human plasma

Only few molecules exhibit fluorescent behaviour. These are called fluorophores. Typically, fluorescence occurs from aromatic molecules or conjugated double bonds. In human plasma, the most important naturally occurring fluorophores are NAD(P)H, flavins, tryptophan and tyrosine [34].

3.2.2. Pre-processing

As in NMR, acquired data from fluorescence need to be pre-processed prior to data analysis. First, the non-chemical phenomenon Rayleigh scatter [35] needs to be eliminated from the EEM. This has been done in Paper III by replacing the Rayleigh scatter with missing numbers, which is also visible in the fluorescence landscape in Figure 2.

Additionally, the samples should be intensity corrected. Intensity correction is performed to obtain comparable results between different instruments. In Paper III, the intensities of the samples have been calibrated according to the integral of the water Raman signal [36].

3.3. Gas Chromatography-Mass Spectrometry

Gas Chromatography (GC) was introduced in 1952 [37] and the coupling to mass spectrometry (MS) has been applied since the 1970s for analysing metabolites in plasma and urine [38]. The technique is two to four times more sensitive than ^{1}H NMR, but the samples need to be heavily pre-treated prior to analysis. The sensitivity is greatly dependant on injection volume, separation and detection techniques, but it will usually detect concentrations down to approximately 10 ppb. Using GC-MS, it is possible to get detailed information of especially hydrophobic molecules and, depending on the specific pre-treatment, carbohydrates and amino acids can also be measured. The principles of GC-MS are described briefly in this section, and details can be found elsewhere [39;40].

3.3.1. Principle

In GC-MS, vaporized compounds in a mixture are separated followed by detection in a mass detector. The compounds are separated based on their volatility or ability to evaporate. This is achieved by adhesion of the compounds to the surface of a slightly heated GC column (stationary phase). The column is placed in an oven, where heating and cooling of the column can take place. After adhesion, the column is heated and the compounds are released from the column according to volatility and carried to the mass detector by a carrier gas (mobile phase). Hence, the most volatile compounds are released first. When the compounds reach the mass detector, ionization and fragmentation of the compounds take place. This can be achieved using different ionization techniques, but the ionization applied in the feasibility study presented in Chapter 5 is the electron-impact (EI). EI has a good sensitivity and unique fragmentation; however, this heavy fragmentation can result in reduced possibilities of identification of unknown compounds [41]. The fragments (which are ions) are detected by their mass to charge ratio (m/z). In the feasibility study, the applied mass analyzer is a quadrupole. Unique fingerprints of the compounds are then obtained by mass spectra and elution profiles. An example of a GC-MS landscape, a mass spectrum and an elution profile from a human plasma sample is seen in Figure 3.

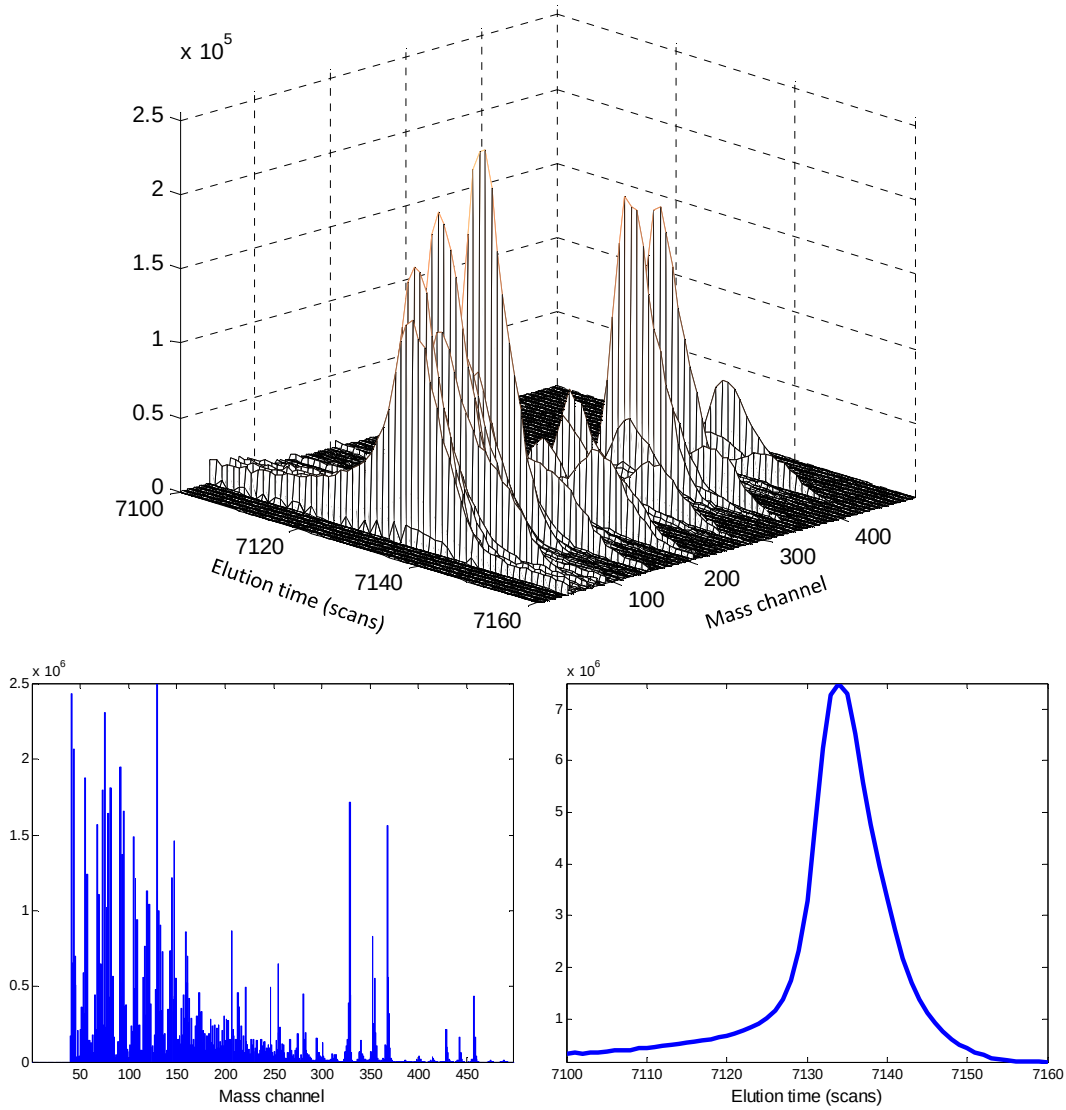


Figure 3. Raw GC-MS data for one human plasma sample. Top: Landscape of a selected area over 61 elution times and 499 mass channels. Bottom left: Profile of the mass spectra summed over the 61 elution times. Bottom right: Elution time profile summed over the 499 scans

Before the compounds can reach the column, they need to be brought to a volatile state. Polar metabolites are not volatile by nature, and therefore the samples need to be prepared prior to injection to the GC. Additionally, the compounds of interest in blood need to be extracted from the blood matrix. These aspects are the subjects of the GC-MS feasibility study in Chapter 5.

3.3.2. Pre-processing

The pre-processing steps applied in this thesis include peak selection, since it is difficult to model GC-MS data by PARAFAC2 (see Chapter 4) on the entire chromatogram. This can be done by manual inspection of the chromatogram, where intervals with visible peak areas are selected. In Paper I, chromatographic datasets of wine and apple samples were manually inspected resulting in 50 intervals selected in the wine dataset and 26 intervals selected in the apple dataset.

In GC-MS-based metabolomics studies, deconvolution is often applied for pre-processing of the chromatograms [42]. Deconvolution is a method to extract the pure signals of the metabolites. The method is not applied in the work of this thesis, and will therefore not be touched upon further.

3.4. Advantages and disadvantages

The above mentioned three methods are quite different from each other and all three methods present advantages and disadvantages. For comparison, the advantages and disadvantages are listed in Table 1 [33;43].

Table 1. Overview of advantages and disadvantages of the three data acquisition techniques

Technique	Advantages	Disadvantages
^{1}H NMR	Non-destructive No need for sample pre-treatment Fast acquisition Measures many compounds Many metabolites are already identified	Low sensitivity Requirement of relatively large sample size Cannot detect compounds without protons
Fluorescence	Non-destructive Fast acquisition Very sensitive	Only few compounds can be measured Complete metabolomics profile impossible
GC-MS	Sensitive Low sample size needed (down to 50-100 μL) Measures many compounds Software for metabolite identification	Extraction of metabolites necessary Derivatization of samples necessary Time consuming sample preparation Slow acquisition

The advantages and disadvantages listed prove the applicability of all three methods and that they complement rather than surpass each other. In the presented case studies in Chapter 5, the methods are applied for different purposes, but the relevance of applying all three methods on the same sample sets is discussed as a future perspective. The acquired data from the methods can be combined, and the strengths of the methods can be extracted by data fusion (see Chapter 4). Hence, the advantages of the methods might be enhanced, whereas the disadvantages might be reduced.

Chapter 4

Multivariate Data Analysis

4.1. Exploring the data

The very first step in analysing any set of data is *always* to explore the raw data. Plotting the raw data in a suitable way (line plots, landscapes, histograms, etc.) can e.g. give a rough overview of structures and noisy parts in the data, expected biomarkers can be located and outlying or deviating samples can be detected. For example, if an extreme sample is included in the analysis e.g. if one blood sample differs remarkably from the others – it could be a sample containing a large amount of ethanol – this particular sample will most likely have an extreme intensity in the metabolite relating to ethanol compared to the remaining samples. This will give an indication of a possible outlying sample, and if so, an explanation of the outlier is found directly in the raw data.

However, small deviations in the samples are most often not visible directly from the raw data, especially if many samples are measured. Furthermore, looking at variables one or a few at a time will mostly not reveal important relations *between* variables. Therefore, it will be useful to apply multivariate models to investigate these more subtle variations in the data – models that can reveal underlying patterns in the data structure.

The first part of this chapter explains the concepts and uses of the explorative multivariate models PCA, MCR, PARAFAC and PARAFAC2, which are all applied to investigate and extract the features "hidden" in the raw data. In these methods, the major variations in data are investigated.

In the second part, the classification model PLS-DA is presented. Here, the model development is based on the relation between the samples and the target of the investigation. In addition, the importance of proper validation of classification models will be discussed followed by an explanation of how to quantify the quality of PLS-DA classification models. Specifically, *sensitivity*, *specificity* and *ROC curves* will be explained in some detail, as these are often used for evaluation of such models.

Considering the huge amount of data obtained in metabolomics studies, it is often of great interest to reduce the number of redundant variables and to include all relevant information. In the last three parts of this chapter, preliminary variable reduction, variable selection and data fusion are discussed.

4.2. Explorative models

4.2.1. Principal Component Analysis

Principal Component Analysis (PCA) [44-46] is a classical method for exploring two-way data arrays (a matrix). PCA can be used on any dataset with a two-way structure originating from for example 1D NMR data, integrated NMR or integrated XC-MS data, unfolded XC-MS data and data extracted from questionnaires.

In PCA, a data matrix $\mathbf{X}$ of size $I \times J$ (I being the number of samples and J being the number of variables) is decomposed into a systematic part ($\mathbf{TP}^{\mathrm{T}}$) and a residual part ($\mathbf{E}$). This can be expressed by the equation

$$\mathbf{X} = \mathbf{TP}^{\mathrm{T}} + \mathbf{E} \qquad (1)$$

Here $\mathbf{T}$ ($I \times G$) denotes the score matrix (G being the number of principal components – see below), $\mathbf{P}^{\mathrm{T}}$ ($G \times J$) denotes the transposed loading matrix and $\mathbf{E}$ ($I \times J$) denotes the residuals, which is the part of the data that is not explained by the model. The loadings hold information of e.g. the metabolites in blood, while the scores contain information of the amount, or importance, of the loadings for each sample. The systematic part holds the main variation in $\mathbf{X}$ expressed by fewer latent variables termed *principal components* (PC). Principal components consist of the score vectors ($\mathbf{t}$) and the loading vectors ($\mathbf{p}$). The score vectors are the columns in $\mathbf{T}$ and the loading vectors are the rows in $\mathbf{P}^{\mathrm{T}}$ as illustrated in Figure 4. The score and loading vectors are orthogonal; that is

$$\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_{\mathrm{j}} = 0, \mathbf{p}_i^{\mathrm{T}}\mathbf{p}_j = 0, \forall_{i \neq j} \qquad (2)$$

which makes the PCA solution unique. In Figure 4, the PCA model is illustrated using two components. The first principal component captures the largest variation in the data, $\mathbf{X}$, (it could be the difference between men and women based on measurements on blood samples); the second principal component captures the second largest variance. The number of components cannot exceed the rank of the data matrix, but generally much fewer components are needed to extract the relevant features in data.

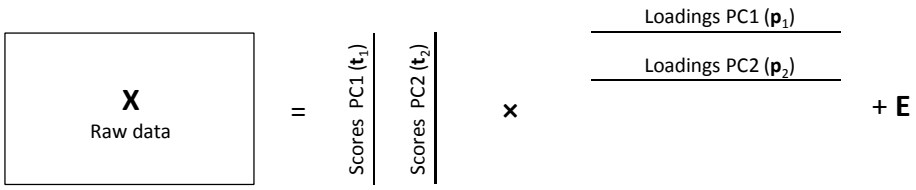


Figure 4. Schematic illustration of a two-component PCA model, where the raw data matrix (**X**) is decomposed into score vectors (**t**) and loading vectors (**p**). **E** represents the residuals

Prior to any multivariate data analysis, it is important to pre-process the data in a meaningful way. Commonly, centering of data (mean centering) is applied [47]. In this case, the average value of a variable is subtracted from the variable for the individual samples. The loadings form a new coordinate system, and when data are mean centred, the scores will have their centre in origo. The concept of mean centering is illustrated in Figure 5. Another pre-processing method is scaling. When the variables are scaled, each variable is normally divided by the standard deviation within the variable. Hence, scaling scales the variable to unity standard deviation. The combination of mean centering and scaling is commonly applied as a pre-processing technique and is denoted autoscaling [47]. Using autoscaling, all variables are weighted equally in a least squares sense. This is important when the dataset is composed of discrete variables, possibly measured in different units (e.g. height in cm, weight in kg, carbohydrate intake in grams per day, etc.). Autoscaling is also illustrated in Figure 5.

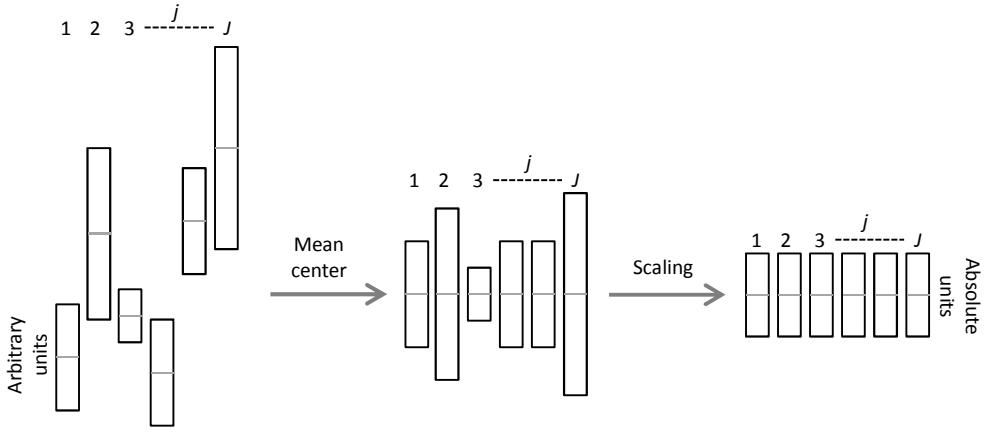


Figure 5. Illustration of mean centring and autoscaling, *J* being the number of variables

4.2.1.1. Visualization of the PCA model

A dataset with 100 variables can be considered as a 100-dimensional sub-space where each dimension represents a variable. A four-component PCA model based on these 100 variables can be represented in four dimensions

within this 100-dimensional space; hence the data representation is reduced from 100 to four dimensions. Commonly, the model is depicted by the first two components, since these describe the majority of the variance, in score and loading plots. However, plotting all combinations of the four components is beneficial, as not all relevant variance is captured in the first components. The samples are easily visualised in score plots based on the score vectors presented in a coordinate system – a "map" of the samples. From these plots, groupings and trends in the data can be revealed and extremely deviating samples (outliers) can be detected. Samples situated close to each other have similar loading profiles. Likewise, the variables can be visualized in a loading plot based on the loading vectors. The score and loading plots are connected in the sense that the position or location of the samples in the score plot can be explained in terms of the variables in the loading plot. For example, variables with high positive loading values in the first component explain samples with high positive scores in the first component. However, variables close to origo have no influence on the model and do not explain anything concerning the samples.

In order to illustrate the PCA model, a part of the questionnaire data applied in Paper II is used. The present PCA model is based on a total of 43 physical measurements and eating habits for ten men and nine women. Hence, **X** is individuals × discrete numbers (19 × 43), each representing a specific feature (e.g. height, alcohol intake, etc.). A separation between the males (squares) and females (triangles) is evident from the score plot of a PCA model on autoscaled data as seen in Figure 6.

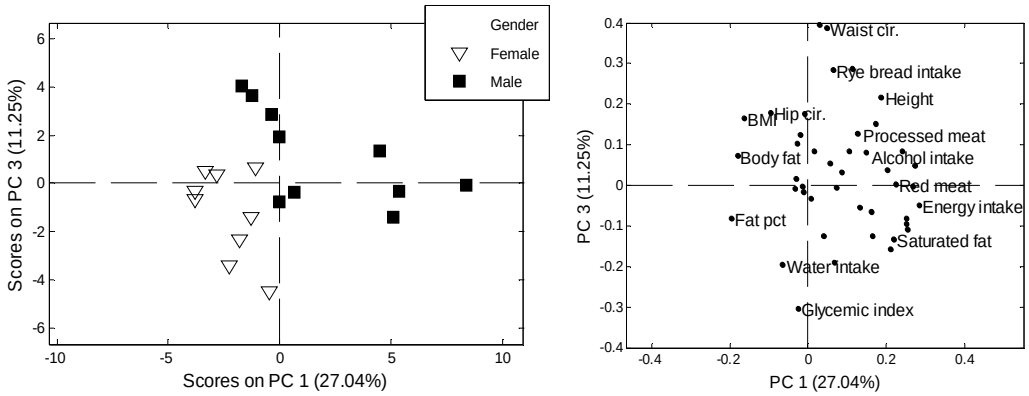


Figure 6. PCA model on data from questionnaires. The samples (men (squares) and women (triangles)) are represented in the score plot and the variables (data from questionnaires) are represented in the loading plot

The separation between men and women is observed when plotting the scores from the first and third components against each other. The explanation of the separation can be found in the loading plot (the number of variable labels has been reduced for simplicity), and the trend is that these women i.e. have a higher fat percent, more body fat and drink more water. The men, on the other hand, consume more alcohol, have a higher energy intake and are taller.

As mentioned, outlying samples can be observed in the score plot. However, samples with semi-extreme profiles are not always visible in the score plot. Outliers can be detected based on Hotelling T^2 [45] values and on the residuals. The Hotelling T^2 value is calculated from the square of the score values, the standard deviation of the score values and the number of components. The residuals, $\mathbf{E}$, are measures of the samples' distance to the model and in a perfectly fitted model, $\mathbf{E}$ is all zero. If the values are too extreme in one sample compared to the remaining samples, the influence of this sample should be carefully investigated and understood, as it can strongly affect the model. If the sample turns out to be a true outlier (it strongly deviates from the other samples), it should be eliminated from the analysis, or more samples exhibiting the same type of variation should be included. However, there are no final answers when to eliminate a sample with outlying behaviour from the analysis. The most feasible approach is to apply a priori knowledge of the samples and common sense.

4.2.1.2. PCA in metabolomics

The PCA model is frequently applied in metabolomics studies, and the model often gives a reasonable overview of the data structure. Many metabolomics studies with focus on cancer diagnosis are based on few samples and the representation of samples in a score plot is clear [22;48]. When many samples and thousands of variables are included in the model, there are a couple of drawbacks when using the PCA model. In the complex data from metabolomics studies, the interesting information is buried deep within the data. For example, measurement techniques like NMR and GC-MS detect many metabolites in blood, but the specific metabolite(s) related to cancer is difficult to locate. Many PCA components might be needed to capture the variation in the data related to cancer. Using many components (say above 30), the number of samples also needs to be correspondingly large. A rough rule of thumb is to include four to five samples for each component as a minimum, if there is no a priori knowledge about the data. A PCA model with 30 components will therefore require a dataset containing at least 120 samples, and if too many components are included, with respect to sample

size, there is a risk of over-interpretation. In Figure 7, a PCA model of NMR data from Paper II are presented. Data consist of 838 samples (controls and future breast cancer patients) and 129 discrete NMR variables; the data are described in detail in Chapter 5. The control samples seem to be slightly more spread out in the score plot compared to the cancer samples. However, the separation is not clear and the high number of samples does not facilitate the interpretation of the plot, and this PCA model should therefore be interpreted with caution.

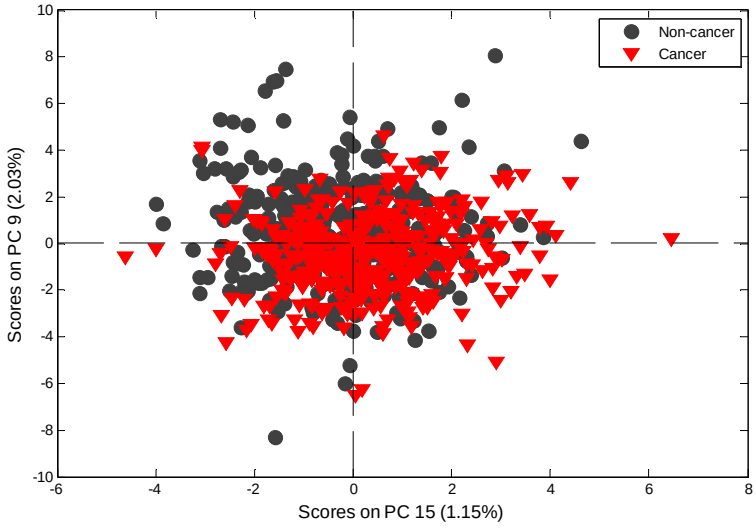


Figure 7. PCA model of NMR data from Paper II. A minor distinction can be observed, but due to the large amount of samples, the visual separation becomes poor

4.2.2. Multivariate Curve Resolution

Multivariate Curve Resolution (MCR) [49] is a curve resolution technique which can be used as an alternative to PCA. An MCR model decomposes a data matrix into two matrices by fewer MCR components. The structure of MCR also resembles that of PCA and with I samples and J variables, the equation for a G-component MCR model can be written as

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \quad (3)$$

Here, $\mathbf{X}$ ($I \times J$) denotes the raw data, $\mathbf{C}$ ($I \times G$) is a matrix holding the concentration loadings and $\mathbf{S}^{\mathrm{T}}$ ($G \times J$) is a matrix holding the spectral loadings. $\mathbf{E}$ ($I \times J$) is the part of the data not explained by the model. MCR is oftentimes applied when spectral signals are overlapping. The main purpose of MCR is, then, to estimate the true concentration profiles from the samples and the true spectral profiles from the variables. This, indeed, is the main difference from PCA. While PCA gives an account of the variance (rank),

MCR finds the true underlying physicochemical information often by using the previously determined rank.

However, one of the main drawbacks of MCR models compared to PCA is that the solution as written above is not unique. Several solutions, which fit the data equally well, can be found due to rotational freedom and intensity issues. It is therefore often necessary to impose constraints such as non-negativity (assuming that the spectral profiles and the concentrations are non-negative). This constraint will aid in finding the real solution describing the underlying chemistry in the data.

Figure 8 shows an example of an MCR model. Plasma samples from 3419 individuals have been measured by NMR, and the presented dataset consists of 3419 samples and a small region of the NMR spectrum (61 ppm values). In the upper part of the figure, every 20^{th} sample is plotted and it seems that there is an overlap between two compounds in this region. In the lower part of the plot, a two-component MCR model has been calculated. To the left, the concentration profiles are shown ($\mathbf{C}$) for and to the right, the two loading profiles ($\mathbf{S}$) are shown. In Paper II, MCR has been applied as an integration technique on selected NMR regions. This is further described in Chapter 5.

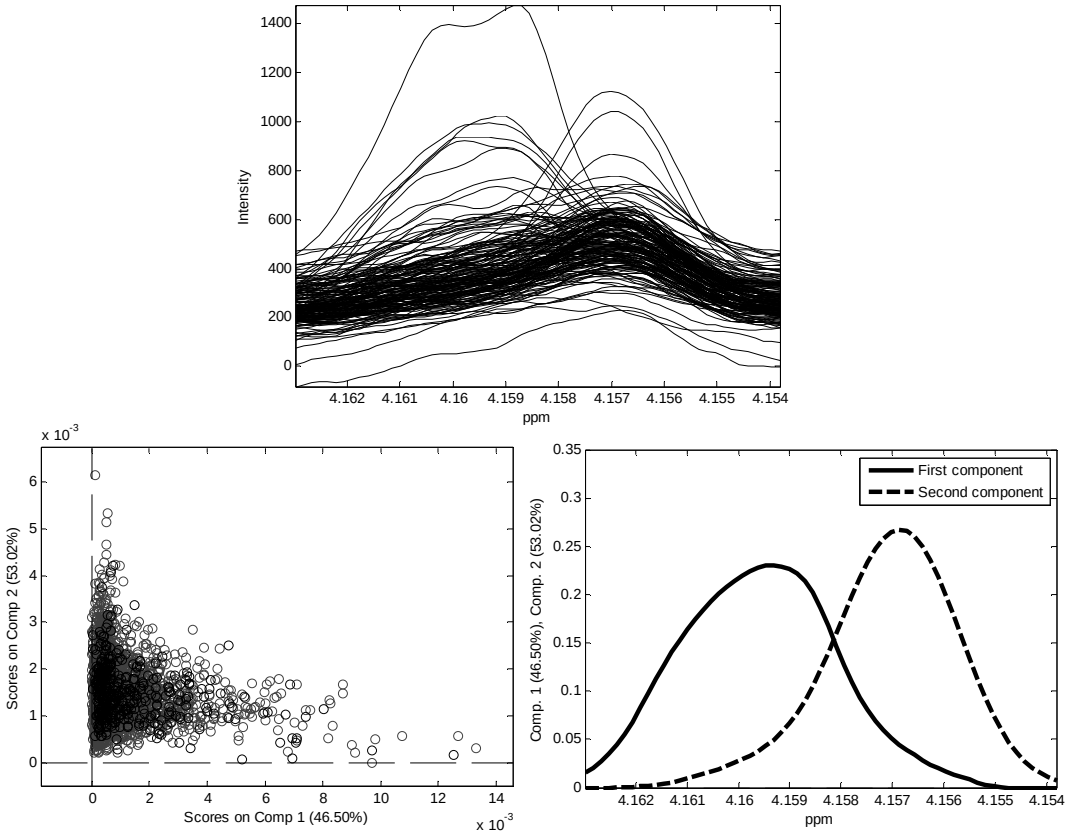


Figure 8. Example of a two-component MCR model. Raw data (top), concentration profiles (bottom left) and spectral profiles (bottom right)

4.2.3. Parallel Factor Analysis

For higher order data (> 2D), PCA is inapplicable, if the multi-way structure of the data is to be maintained. PARAllel FACtor analysis (PARAFAC) [50;51] is an extension of PCA handling higher order data. The theory is described in some detail in Paper I and will be briefly discussed and illustrated in this section.

An $I \times J$ data matrix $\mathbf{X}_k$, with $k = 1,\ldots, K$ as the kth slab of an $I \times J \times K$ three-way array $\underline{\mathbf{X}}$ with a low-rank trilinear structure is suitable for decomposition by PARAFAC. Data are decomposed into three loading matrices, $\mathbf{A}$ ($I \times G$), $\mathbf{B}$ ($J \times G$), and $\mathbf{C}$ ($K \times G$); however, the matrix holding the loadings of the samples can be termed the score matrix to make the terminology similar to that of PCA. The extension from PCA lies in the third loading matrix $\mathbf{C}$ which is the additional dimension. In for example fluorescence excitation-emission spectroscopy, the dimensions are the samples, the emission wavelengths and the excitation wavelengths. The differences between PCA and PARAFAC are illustrated in Figure 9, where residuals are excluded for simplicity. The

matrix $\mathbf{D}_k$ ($G \times G$) is a diagonal matrix containing parameters of the kth row of $\mathbf{C}$ on its diagonal.

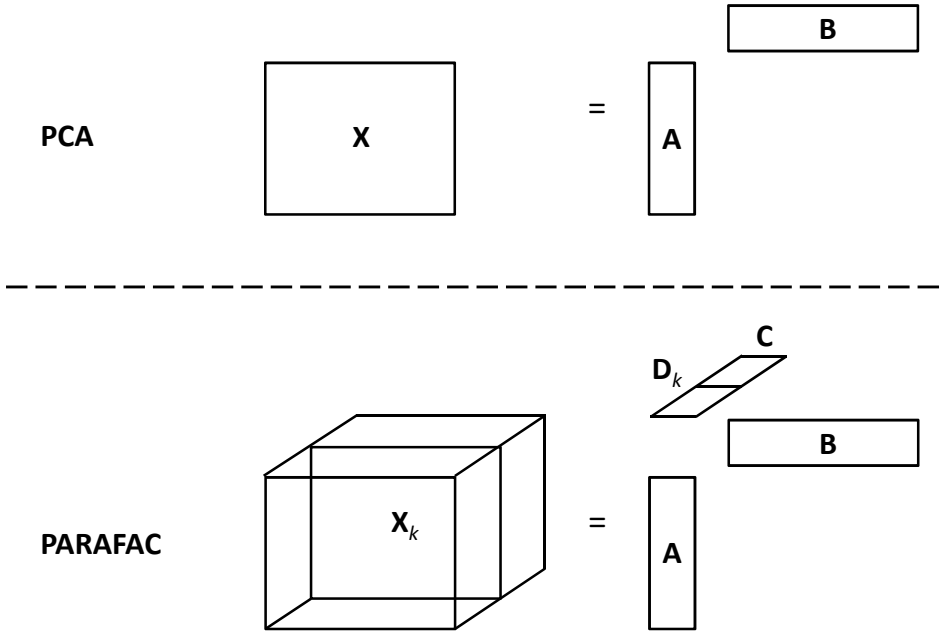


Figure 9. Illustration of the differences between PCA and PARAFAC. The extension of PARAFAC lies in the matrix $\mathbf{C}$ with $\mathbf{D}_k$ on the diagonal

The PARAFAC model decomposes data into fewer PARAFAC components in a manner very similar to that of PCA. In PARAFAC, the loading matrices provide unique estimates of the underlying main variations in the data. However, in order to obtain a solution estimating the pure chemical profiles, it is very important to use the optimal number of components. The optimal number of components can be found using the core consistency diagnostic [52], originally presented by Bro (1998) [53]. As already mentioned, data must be low-rank trilinear and core consistency is a measure of how well only low-rank trilinearity is captured by the model. This means that if there are three types of low-rank variations in data, the data must be modelled with three components. If this is the case, the core consistency will be high (close to 100). If the model is over-fitted – if too many components are included – the core consistency will be close to zero or negative. However, it is important to note that the core consistency diagnostic can only be used as an *indicator* of the number of components. The PARAFAC model obtained should be carefully inspected regarding loadings and residuals and any a priori knowledge concerning the chemical rank should be included in the decision as well. If the correct number of components is not determined, the estimated loadings from the PARAFAC model will not only reflect the true nature of the data, despite

the unique properties of the PARAFAC model [50]. An example from Paper I is the frequently applied dataset of three different amino acids in five samples analysed by fluorescence spectroscopy. Figure 10 shows a PARAFAC model with five components calculated on the amino acids dataset. For a five-component PARAFAC model, the core consistency becomes negative, indicating that this model is over-fitted. This is supported by inspection of the loadings in the first two modes, where neither the emission profiles nor the excitation profiles are correctly estimated.

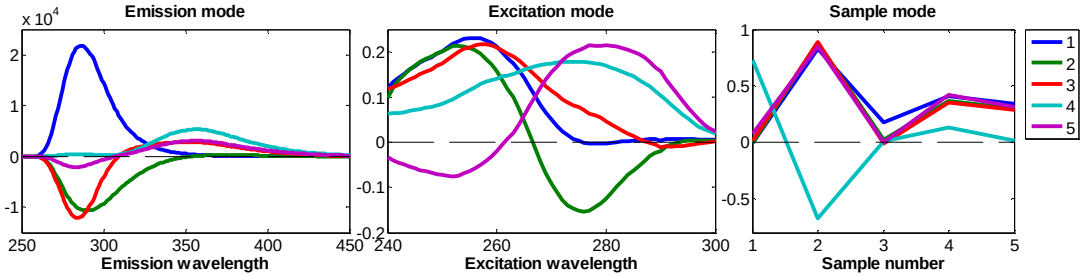


Figure 10. Example of a five-component PARAFAC model. The plot to the left shows the emission loadings, the middle plot shows the excitation loadings and the far right plot shows the sample mode. The core consistency is negative, which indicates over-fit

One limitation of PARAFAC is that the second mode loadings, **B**, are assumed to be equal for each slab of $\mathbf{X}_k$. This is often not the case in e.g. GC-MS, where identical compounds may elute at slightly different retention times from run to run. If a chromatogram of a single analyte with shifting retention times is modelled by PARAFAC, the solution is most likely to include additional components in order to describe the shift. This is shown in Paper I, where an example based on data from a study by Amigo *et al.* [54] is used to illustrate the problem in determining the number of components in such cases. The problem can be handled by extending PARAFAC to PARAFAC2 where the second mode loadings are no longer assumed to be equal across different samples.

4.2.4. Parallel Factor Analysis 2

If the abovementioned example with shifts in the retention time is modelled by PARAFAC2, the solution will most likely handle the shift as seen in Paper I. PARAFAC2 [55;56] is closely related to PARAFAC, which is able to e.g. mathematically separate overlapping peaks in a chromatogram and to handle batch data with differing temporal duration. Multi-way data acquired from e.g. GC-MS have many overlapping peaks as well as retention time shifts and

PARAFAC2 is therefore suitable for estimating the underlying structures of such data. Throughout this section PARAFAC will be denoted PARAFAC1 for a clearer distinction from PARAFAC2 [55].

The extension in PARAFAC2 from PARAFAC1 lies in the second mode loadings, **B**, and the PARAFAC2 model of $\underline{\mathbf{X}}$ $(I \times J \times K)$ can be expressed as

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}_k^{\mathrm{T}}, k = 1, \ldots, K \quad (4)$$

As for PARAFAC1, $\mathbf{A}$ $(I \times G)$ holds the loadings in the sample mode, $\mathbf{B}_k$ $(G \times J)$ holds the second mode loadings, and $\mathbf{D}_k$ $(G \times G)$ holds the information from the last mode loadings. The PARAFAC2 model is schematically illustrated in Figure 11, the only difference being the extension of the B-loading matrices.

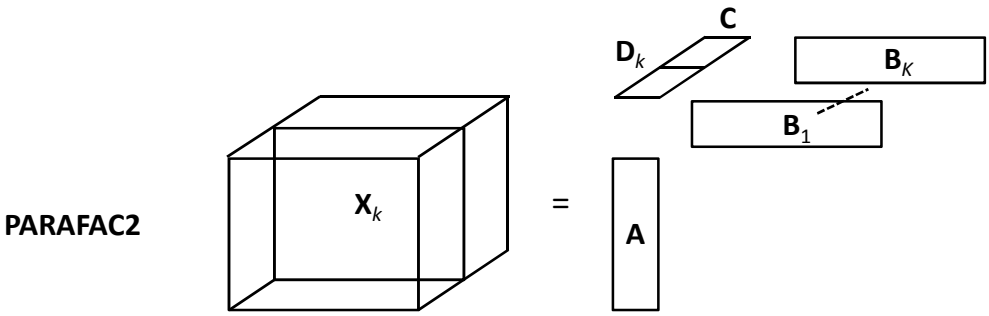


Figure 11. Illustration of PARAFAC2. The extension from PARAFAC1 lies in the B-loading matrices

With the introduction of the sample specific B-loading matrices, PARAFAC2 is then capable of handling the differences (shifts) in the second mode of the data [57] and fewer components are needed to describe the data. The proper number of components to be included in PARAFAC2 can be determined more or less as in PARAFAC1 evaluating for example residuals and loadings. The essence of Paper I is a proposal of a core consistency diagnostic to aid in the determination of the number of components in a similar fashion as the core consistency diagnostic in PARAFAC1 [58]. The paper also describes the difference between PARAFAC1 and PARAFAC2 and shows how, theoretically, a PARAFAC1 model is "fitted" inside PARAFAC2. This theory makes it possible to calculate a core consistency value for PARAFAC2.

In Paper I it has been shown that core consistency can actually be used for determining the appropriate number of components in a PARAFAC2 model, but it is suggested that loadings and residuals are also inspected as for the original development of core consistency for PARAFAC1. The use of core consistency in PARAFAC2 sometimes leads to including more components

than first anticipated, but with the addition of detecting all chemical information from the data. The theory is tested on three different real datasets as well as a simulated dataset. One of the datasets consists of 24 samples of red wine analysed by GC-MS. The dataset contain several peak regions and it is therefore necessary to divide the data into smaller parts. This is done manually, resulting in 50 different peak regions. In interval 31, retention time shifts are present as well as a low signal-to-noise ratio. These features make it extremely difficult to allocate the correct number of components. Figure 12 shows a figure from the paper depicting the raw data from interval 31 (Figure 12A) along with a two-component (Figure 12B) and a five-component (Figure 12C) PARAFAC2 model. In addition, the difference between the estimates of the main peak using two and five components, respectively, is shown (Figure 12D).

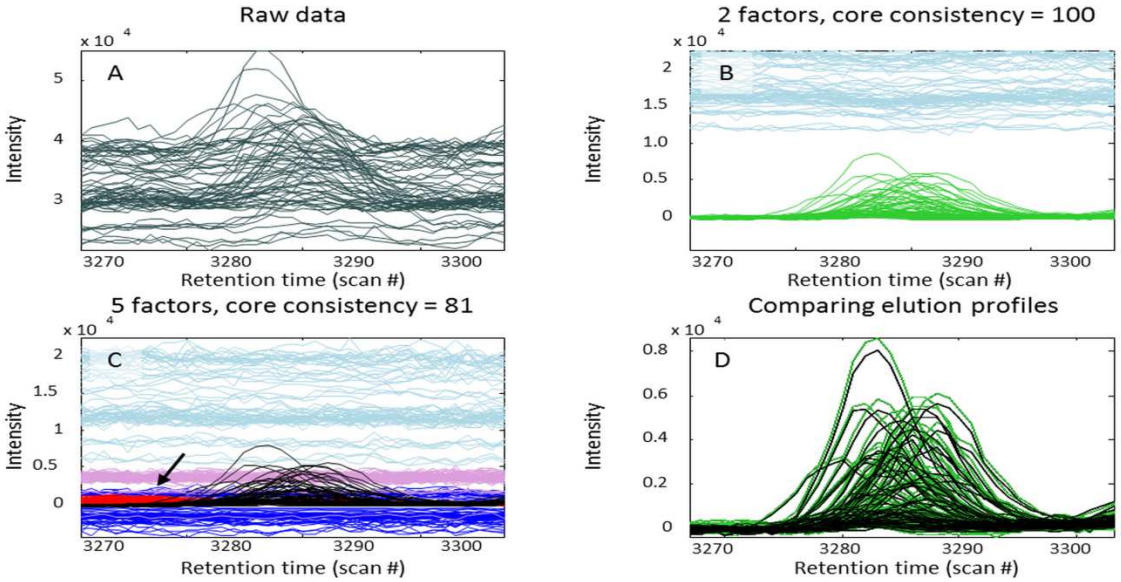


Figure 12. A: Raw data from wine interval 31. B: Estimated retention time profiles from a PARAFAC2 model with two components. C: Estimated retention time profiles from a PARAFAC2 model with five components. D: Estimated profiles of the main peak with two components (green) and five components (black). The figure is from Paper I

The first visual inspection of the region will most probably lead to the inclusion of two components. The model seems apparently over-fitted when more components are included. However, the core consistencies for both models are also presented and both values are rather high – 100 for the model with two components and 81 for the model with five components. This indicates that five components might be appropriate compared to only including two components. When the model with five components is inspected even further, it is noticed that a small peak is actually estimated by the fifth

component (Figure 12C, arrow). This is one of the evidences on the usefulness of core consistency in PARAFAC2 models. Inclusion of five components will give estimates of *all* the chemical information in the data and the estimated structure of the main peak is maintained as seen in Figure 12D.

Here, the usefulness of core consistency in PARAFAC2 is illustrated using an example from a dataset consisting on samples of red wines. As this thesis is mainly focusing on data from metabolomics, it should be mentioned that PARAFAC2 modelling can be used on e.g. GC-MS data from blood samples as well. In order to estimate the profiles of metabolites in the blood, the issue of determining the appropriate number of components will most likely also be a challenge. A small feasibility study on plasma samples has been conducted and is described in Chapter 5.

4.3. Classification models

In metabolomics studies the aim is oftentimes to investigate changes in the metabolism related to a response to a certain treatment or a certain disease. In classification models, the relation between the measurements of the metabolism and a categorical response is investigated. There are many different classification models and in metabolomics commonly used ones include PLS-DA and OPLS-DA.

4.3.1. Partial Least Squares-Discriminant Analysis

Partial Least Squares-Discriminant Analysis (PLS-DA) [59;60] is a classification method used for two-way data, based on the principle of the regression model, Partial Least Squares regression (PLS) [61-64]. To set the stage for PLS-DA, a brief introduction to PLS is presented.

In PLS, two sets of variables are investigated: The (predictor) variables in $\mathbf{X}$ ($I \times J$) and the dependent variables in $\mathbf{Y}$ ($I \times K$), and the purpose is to predict $\mathbf{Y}$ from $\mathbf{X}$. The two matrices are each decomposed into score matrices and loading matrices. The pairs of scores and loadings in PLS are latent variables termed PLS components ranging from $g = 1,\ldots G$, G being the number of components. Formally, the decompositions can be written as

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \qquad (5)$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \qquad (6)$$

In the first equation, $\mathbf{X}$ is decomposed into scores, $\mathbf{T}$ ($I \times G$), loadings, $\mathbf{P}$ ($J \times G$) and residuals, $\mathbf{E}$ ($I \times J$). In the second equation, $\mathbf{Y}$ is decomposed into the score and loading matrices in $\mathbf{Y}$, $\mathbf{U}$ ($I \times G$) and $\mathbf{Q}$ ($K \times G$), respectively, and the residual matrix, $\mathbf{F}$ ($I \times K$). These two equations resemble two PCA models, but that is not quite the case. In PLS, the aim is to find a direction, $\mathbf{t}$ (= $\mathbf{Xw}$), that describes $\mathbf{X}$ and has maximum covariance with $\mathbf{Y}$. The X-scores, $\mathbf{t}$, are calculated from the residuals of $\mathbf{X}$, where the contributions from the previous components have been subtracted. The Y-scores, $\mathbf{u}$, are calculated in the same manner based on the residuals of $\mathbf{Y}$, respectively:

$$\mathbf{t}_g = \mathbf{E}_{g-1}\mathbf{w}_g \quad \mathbf{u}_g = \mathbf{F}_{g-1}\mathbf{q}_g \,, g = 1 \dots G \quad (7)$$

$\mathbf{w}_g$ are the columns in $\mathbf{W}$ ($J \times G$). The inner relation between $\mathbf{X}$ and $\mathbf{Y}$ in PLS can be expressed as

$$\mathbf{U} = \mathbf{T}\mathbf{R}_{\text{diag}} \quad (8)$$

where $\mathbf{R}_{\text{diag}}$ is a diagonal matrix holding the inner correlation coefficients between the X- and Y-scores for each PLS component on the diagonal. Estimates of the Y-values can be calculated from the mixed relation

$$\mathbf{Y} = \mathbf{T}\mathbf{R}_{\text{diag}}\mathbf{Q}^{\text{T}} + \mathbf{F} \quad (9)$$

New samples can be predicted by using equation (9), which is based on the X-scores or directly from $\mathbf{X}$ by compiling weights and loadings into the regression coefficients, $\mathbf{B}$ ($J \times G$):

$$\mathbf{B} = \mathbf{W}(\mathbf{P}^{\text{T}}\mathbf{W})^{-1}\mathbf{Q} \quad (10)$$

$\mathbf{B}$ is the matrix, where the columns hold the regression vector for each Y-variable and the prediction of $\mathbf{Y}$ from new samples can be expressed as

$$\mathbf{Y}_{\text{new}} = \mathbf{X}_{\text{new}}\mathbf{B} \quad (11)$$

$\mathbf{Y}_{\text{new}}$ is the predicted values from the new set of samples, $\mathbf{X}_{\text{new}}$. $\mathbf{B}$ is the regression coefficient matrix from the calibration model at the selected number of PLS components. A schematic overview of the PLS model is seen in Figure 13, where the different steps are highlighted.

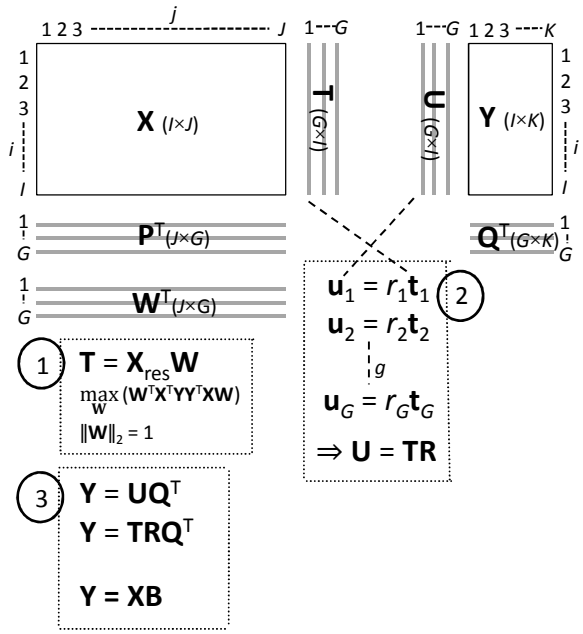


Figure 13. Schematic illustration of the relation between **X** and **Y** in a PLS model. **T**: Score matrix from **X** with columns $\mathbf{t}_g$. $\mathbf{P}^T$: Loading matrix from **X**. $\mathbf{W}^T$: X-weights. **U**: Score matrix from **Y** with columns $\mathbf{u}_g$. $\mathbf{Q}^T$: Loading matrix from **Y**. r: correlation coefficients between **t** and **u**. I: Number of samples. J: Number of X-variables. K: Number of Y-variables. G: Number of PLS components

The PLS-DA model can be expressed in the same manner by the combined relation between **X** and **Y**. However, for PLS-DA the dependent Y-variable is constructed as a dummy variable consisting of row vectors with values ones and zeros (or minus ones). One indicates membership of a class and zero indicates *no* membership of a class. The number of classes minus one determines the number of columns in **Y**. If only two classes are to be separated, the dependent variable can be constructed as a vector, **y**, of zeros and ones. The classification part of PLS-DA is connected to Linear Discriminant Analysis (LDA). LDA was developed in 1936 by Fisher [65] and is a discrimination method, where the between-group variance is maximized. Barker and Rayens [59] have shown that PLS-DA can be considered as an inverse least-squares approach to LDA. The results are similar, but PLS-DA has the benefit of noise reduction, the ability of handling correlated variables and the reduction of variables represented as latent variables [66]. Details concerning the theoretical connection between PLS-DA and LDA can be found in [59].

The purpose of the PLS-DA model is assignment of class membership for a sample. The class membership is based on the predicted y-value ($\hat{\mathbf{y}}$) and a specific threshold set for the PLS-DA model. A threshold value could be, but is not restricted to 0.5, when the dependent y-vector consists of zeros and

ones. If $\hat{\mathbf{y}}$ is larger than 0.5, the sample will be estimated as belonging to the class. If $\hat{\mathbf{y}}$ is smaller than or equal to 0.5, the sample will be estimated as *not* belonging to the class. The threshold can be adjusted depending on the problem, which will affect the performance values of the model. The performance of the PLS-DA model is described later.

The application of the PLS-DA model can be exemplified using the questionnaire example with ten men and nine women used in the PCA analysis. If we want to make a model that can classify if the measurements represent a man or a woman, then $\mathbf{y}$ is a vector of zeros and ones (men and women, respectively) and the data matrix $\mathbf{X}$ will be the same as the one analysed by PCA. The PLS-DA model maximizes the covariance between $\mathbf{X}$ and $\mathbf{y}$, and hence the direction where the discrimination between men and women is largest is located. The predicted class membership can be seen in Figure 14. One sample is misclassified in a PLS-DA model with three PLS components. The threshold is here set to 0.62 rather than 0.50, the former being the value where most samples are classified correctly.

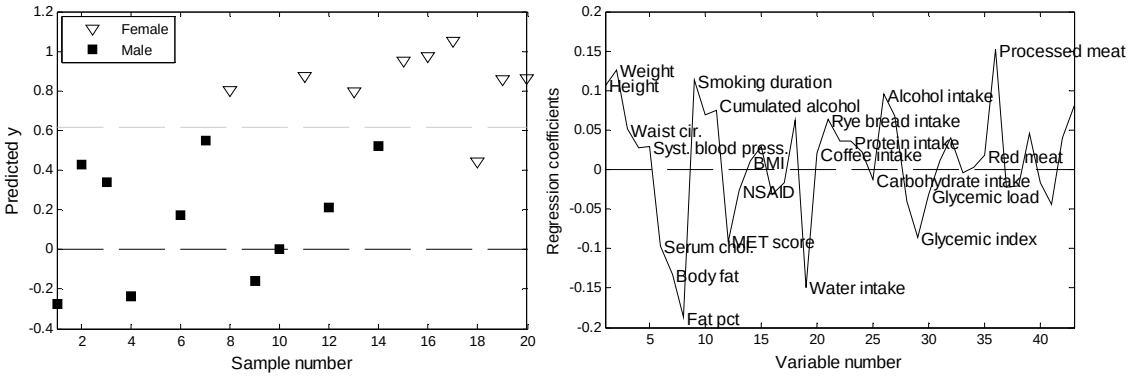


Figure 14. Left: Prediction of class membership of men (squares) and women (triangles) based on questionnaire data. The threshold (grey dashed line) is 0.62, where most samples are classified correctly. Right: Regression coefficients for the PLS-DA model with three PLS components

Interpretation of the model can be done by evaluation of the loadings, weights and regression coefficients, and Figure 14 presents the regression coefficients for the PLS-DA model. However, it is important to note that the variables important for the classification should be considered as a whole and should not be interpreted one by one when data are empirical. The importance of not selecting single variables as sole descriptors of class membership (i.e. male/female) is described in detail in Paper II and in Chapter 5.

An alternative to PLS-DA is Orthogonal PLS-DA (OPLS-DA) [67]. OPLS-DA is a variant of PLS-DA, where the variation in **X** orthogonal to **y** is removed. This means that all variation in **X** not linearly related to **y** is eliminated from the model. In practice, the components in OPLS-DA are divided into two groups: The number of *predictive* components and the number of *orthogonal* components. The predictive components are those related to the correlation between **X** and **y**, and the orthogonal components contain information orthogonal to **y**. According to a study by Tapp and Kemsley [68], the sum of the variance explained by the predictive and orthogonal components of OPLS-DA is equal to the variance explained by the components of PLS-DA. Additionally, the performance statistics (described later) related to the predictive performance are also often equal for the two models. Explicitly, this means that an OPLS-DA model with one predictive component and two orthogonal components will predict class membership equally well compared to a PLS-DA model with three PLS components. This means that with regard to prediction, OPLS-DA does not outperform PLS-DA and vice versa [68]. The question is then which method to apply. Given the above mentioned similarities between the models, it is difficult to favour one method over the other. It must be kept in mind though that OPLS-DA will not result in improved predictions compared to PLS-DA, which is unfortunately sometimes anticipated for or interpreted from OPLS-DA models [41;69].

However, there is one major issue that needs to be handled carefully in order to obtain consistent methods and that is proper validation. If not properly validated, neither PLS-DA nor OPLS-DA will give reliable results because there is a considerable risk of over-fit. The problem becomes even more pronounced, when analysing large datasets with many variables due to the curse of dimensionality [70], as is the case in metabolomics. Refraining from doing proper validation will result in spurious and misleading groupings; hence validation is an indispensable subject for discussion.

4.3.2. Validation

It is of utmost importance to achieve reliable and valid models, meaning that the model can be applied to what it was developed for. Making reasonable decisions is fundamental when building multivariate models and is part of the validation, for example choosing the optimal number of components and to find potential outliers. Choosing the optimal number of components is essential when developing classification models, in order to avoid both over-fit and under-fit (inclusion of too many PLS components and too few, respectively) and also for the development of robust models. A properly validated classification model should be capable of predicting class

membership of samples obtained from a new dataset; however, it is a prerequisite that there is an actual differentiation between the samples.

There are different ways of validating multivariate models. Cross-validation (CV) [71] is a common approach to validate a model, and typically either the leave-one-out (LOO) method or subsets are applied. LOO simply means that one sample is left out of the analysis, the model is calculated and the sample left out is predicted onto the model. This is performed until all samples have been excluded from the analysis once. However, LOO cross-validation should only be carried out on small datasets with fewer than, say, 20 samples. Otherwise the model is very prone to over-fit, and the validation then becomes unreliable [72;73]. In the case of using subsets, a defined number of samples are left out of the analysis and predicted onto the model until all samples have been left out. The subsets can for example be arranged in contiguous blocks, blocks chosen in a system resembling "venetians blinds" (e.g. every fifth sample) or randomly chosen blocks with predefined block size. Using the latter approach, the cross-validation can be iterated a certain number of times to smooth out noise.

Validation of classification models is very important [60]. For an illustrating purpose, a data matrix was generated with 20 × 100 random numbers. The first ten "samples" were assigned to a class of ones and the last ten "samples" were assigned to a class of zeros. A PLS-DA model will always find a direction, where the class membership of two random classes will be perfectly predicted. It is therefore important to be aware of what is plotted when looking at e.g. predicted class membership. Plotting the samples from the calibration model (Figure 15, left) can show a perfect prediction of the samples. However, if samples which were left out of the model building are predicted onto the calibration model, these will reflect the true predictive power and in this case the predicted class membership will be random as seen in Figure 15, right. The latter can only be obtained, if the model is validated.

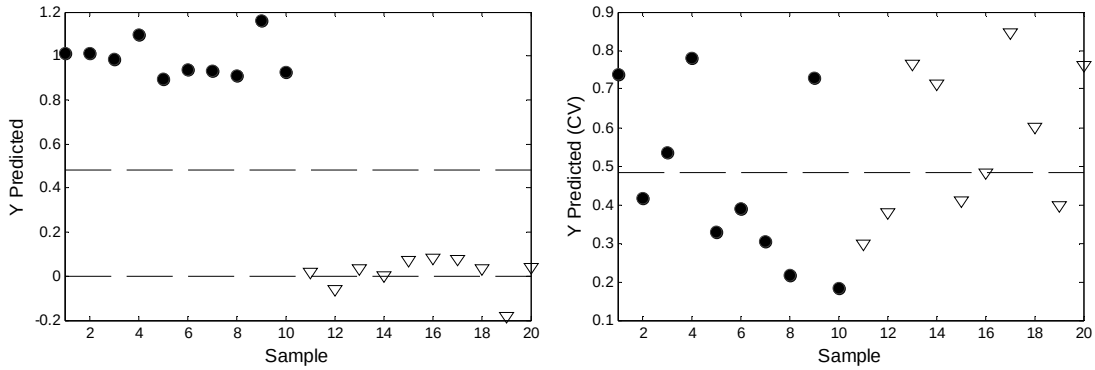


Figure 15. Classification of random numbers. Perfect prediction of random numbers from the calibration model (left), and random prediction when new samples are inspected (right)

Cross-validation is strictly an internal validation of the calibration set. In reality, it does not state anything about new samples measured at different times or using different instruments. Performing cross-validation on a calibration set can sometimes lead to perfect separation of groups due to spurious correlations. It is therefore necessary to test the performance of a calibration model on a test set. A test set can be obtained either by extracting a part of the acquired dataset – for this, a relatively large sample set is required – or by acquiring a new dataset with the same conditions. Obtaining a new dataset might not always be feasible due to limited access to sample material, operating expenses, etc., and therefore a test set from the original data is often applied. A key point in test set validation is to leave out the test set during preparation of the data, variable selection and when building the calibration model. Hence, the dataset should be divided in two parts as an initial step of the data analysis. One part is used to build the calibration model and only when the model is optimised, the remaining part can be applied to test the performance of the model. If the dataset is divided into a calibration part and a test part *after* building the model, the test set loses its function and becomes unreliable as a measure of model performance.

An important issue in metabolomics is the presentation of PLS-DA and OPLS-DA score plots. It is very common to display score plots to demonstrate the performance of the classification [19;21;74]. This is perfectly alright *if* the model is properly validated and valid *and* the corresponding loadings or similar and performance statistics (described later) are presented. The problem with the score plots from PLS-DA and OPLS-DA is that neither of them reflects scores from new samples. Hence, the score plots look exactly the same whether or not the model has been cross-validated because validation does not change the model itself. However, considering the above mentioned example with the random numbers, a score plot of a test set validated model

with *test set validated scores* is reliable, as seen in Figure 16. To the left, the scores are from the calibration model and to the right, the scores are validated by predicting new score values from a random test set.

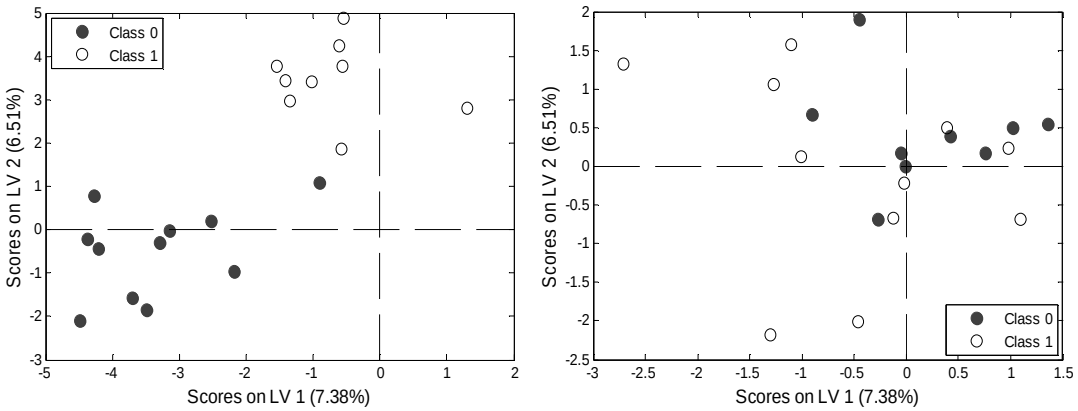


Figure 16. PLS-DA score plots of scores from a the calibration model (left) and score values from samples left out of the calibration model (right) based on random numbers

The essence of this is to avoid over-fit, which will occur if the calibrated scores are plotted, in cases where the calibrated and validated fits are widely different. If calibration scores are inspected and interpreted without support from the performance statistics, it will wrongly be concluded that the data can be perfectly separated.

All regression models must be presented using the optimal number of PLS components. If too many components are included, the model is over-fitted, and noisy parts of the data will be modelled. On the other hand, including too few components will result in a model that does not describe all the systematic variation in the data, leaving out possible important structures. In this case, the model is under-fitted. The appropriate number of PLS components to include in the analysis can be found based on model performance (described below) after validation. For classification models as PLS-DA, the classification error is commonly used as an indicator of how many PLS components to include in the model. The classification error rate is the percentage of misclassifications; hence the optimal number of latent variables should be selected for as few misclassifications as possible.

4.3.3. Performance statistics of the PLS-DA model

The aim of a PLS-DA model is correct prediction of class membership for a new sample. Several parameters are used in the evaluation of a PLS-DA model. One measure of the model performance is the receiver operating characteristic (ROC) curve which is based on the *sensitivity* and *specificity* values [75-77]. Sensitivity is the fraction of the true *cases* correctly identified by the model. Specificity is the fraction of the true *controls* correctly identified by the model. The cases are the samples with the condition of interest, whereas the controls are the samples without the condition of interest. In Paper II, the separation between women who have developed breast cancer from women who have not, has been investigated. The samples with breast cancer are the cases and the samples without cancer are the controls. If all samples are correctly identified, the sensitivity and specificity values are equal to one. The values are calculated from the confusion matrix (Table 2). The ROC curve is composed by plotting sensitivity against specificity, and the optimal set of values is chosen based on a user-defined (or software-defined) threshold value, depending on the purpose of the prediction model. If the model is used for e.g. screening of cancer, it is problematic with false positives [5], and therefore the threshold should be selected corresponding to a specificity of one. However, if the model is used to e.g. predict cancer five years ahead in time, both false positives and false negatives should be minimized. The threshold should then be set where both sensitivity and specificity have maximum values. The area under the ROC curve (AUROC) should be equal to one for perfect separation. If the samples are randomly classified, the ROC curve will be a diagonal line and AUROC will be 0.5. An example of an ROC curve can be seen in Figure 17.

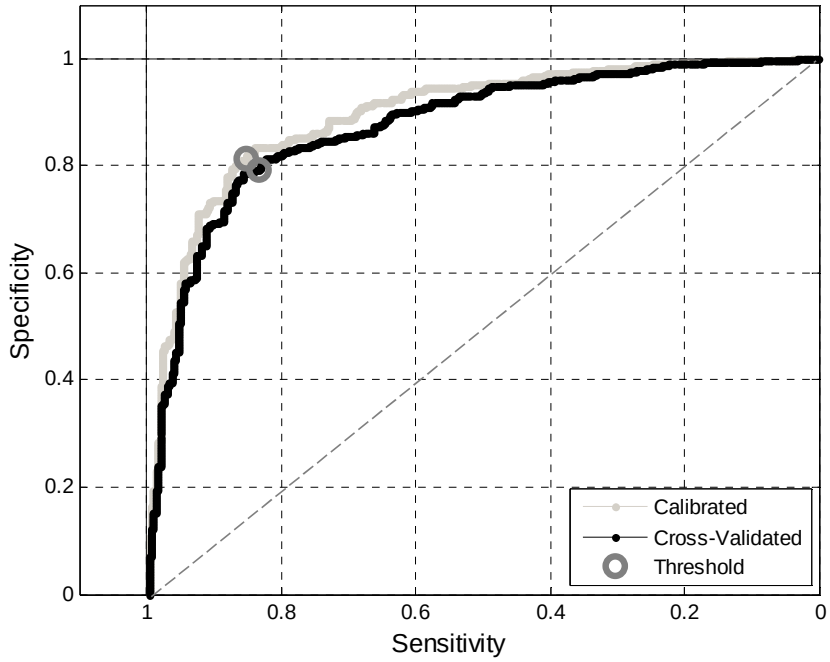


Figure 17. Example of the ROC curve. The light grey solid line is the calibrated ROC curve and the black line is the cross-validated. The dark grey circle marks the optimal model threshold. The dashed diagonal line represents a random classification

In the present example, the ROC curves for the calibrated and cross-validated models are very similar. This is an evidence of a robust model which is required for developing meaningful and reliable prediction models. However, the similarity of the curves is not evidence of a good model!

Sensitivity and specificity resembles the Type I and Type II errors (also known as α and ß errors) used in statistics [78]. These error measures are used when a hypothesis is being tested. Type I error is the risk of rejecting the null-hypothesis when it is actually true. This is also called a false positive, where the test states that e.g. a woman has breast cancer when she is healthy. Type II error is the risk of accepting the null-hypothesis when it is actually false. In this case, the test states that the woman is free of breast cancer, when she *is* sick. This is called a false negative. In addition, the power of the model is equal to the sensitivity.

The relation between sensitivity and specificity and the error types can be illustrated schematically by a confusion matrix as in Table 2.

Table 2. Confusion matrix illustrating sensitivity and specificity along with the error types

		Condition	
		Case	*Control*
Model/test	*Case*	True positive	False positive (Type I error)
	Control	False negative (Type II error)	True negative
		Sensitivity	Specificity

Besides sensitivity and specificity, other measures for the PLS-DA model performance should be mentioned. The number of misclassifications is simply the sum of false positives and false negatives. Q^2 is the goodness-of-fit or predictive ability of the model [79;80], and recently discriminant Q^2 has been suggested as a performance measure [81]. It has been discussed which parameters are most suitable in the evaluation of the PLS-DA model. A suggestion is that all of the above mentioned parameters are taken into account when testing model performance [82], but in general sensitivity and specificity are primarily chosen [16;19]. Evaluation of PLS-DA models throughout the work presented in this thesis is primarily based on sensitivity and specificity values and the ROC curve.

4.4. Reduction of redundant variables

The curse of dimensionality imposes a need for reducing the number of *redundant* variables in metabolomics studies in order to decrease the data dimensionality. Reduction of redundant variables (denoted RRV in the following) should be done prior to common variable selection (described below), where the class membership of the samples is known, in order to include *all* biologically meaningful variables – not only those related to the class of interest. Doing RRV in a blindfolded way, the temptation of searching for biomarkers directly related to the class membership is eliminated. When large datasets are being analysed – especially when the sample-to-variable ratio is low – there is a great need for RRV. Additionally, datasets generated in metabolomics very often contain noisy regions with no biological information, which is of no interest for multivariate data analytical purposes. It is very common to perform some degree of RRV in one way or another. In NMR studies, different binning techniques are applied [2] and in GC-MS studies, manual inspection and selection of peaks in the chromatogram can be performed [54]. However, in NMR studies, binning still includes all areas of the NMR spectra. The data analysis is then still based on the entire dataset (or integrals hereof) and the models are therefore greatly prone to over-fit and

to false predictions. In this section, an example on performing manual RRV on NMR data will be presented (Paper II).

Data from the Danish "Diet, Cancer and Health" cohort (presented in Chapter 5) have been analysed NMR resulting in more than 56,000 variables (chemical shifts). The first step in performing RRV, the NMR spectra from all samples should be thoroughly (and manually) assessed in order to locate *all* measured peaks. This will lead to a dataset only consisting of chemical information leaving out pure noise. In the NMR case study described in Chapter 5, 189 peaks of varying sizes and intensities were selected. An important point when performing the inspection is to leave out information of class membership of the samples i.e. cancer status to avoid inclusion and exclusion of assumed cancer and non-cancer related variables. When the selection is completed, all peaks should be properly integrated using for example the area of the peak, peak height or MCR. Area and MCR scores estimate the chemical concentration of a peak. Peak height is not a common integration technique, but provides a reasonable estimate of concentration (up to a scaling) when the line shapes of the peak in different samples are similar. Peak height can be used for integrating when e.g. baseline resolution is difficult to achieve, which is the case for many peaks in NMR. Integrating peaks will lead to one value instead of a line shape for each peak. In the case study, the integration resulted in 189 discrete values, each representing the concentration of a peak. The result of RRV is in this specific case a reduction of 56,000 continuous variables to 189 discrete variables, which minimizes the curse of dimensionality and eliminates redundant variables. Since the location of peaks and decisions on the integration approach for each peak are subjective, the result will most likely be different results each time the variable reduction is performed. Hence, the process needs to be repeated in order to obtain a consistent result. RRV is therefore very time consuming, but it handles the problems stated above, and the outcome is worth the effort.

4.5. Variable selection

After RRV, one of the most important steps in the development of reliable and robust prediction models is variable selection. The main difference between RRV and variable selection is that the class status is – or should be – unknown in the former and is used as a selection guideline in the latter. The main reason for performing variable selection after RRV is that much of the chemical information present in the measured metabolic profiles is irrelevant for the prediction of a specific outcome – many variables are redundant. A more parsimonious representation of data, where redundant variables have

been excluded will aid the variable selection. If there are too many variables compared to samples, the variable selection will be flawed and in the worst case be meaningless.

There are numerous ways to perform variable selection in chemometrics, for example variable influence on projection [83], regression coefficients [84], selectivity ratio [85;86] and Interval PLS (*i*PLS) [87]. *i*PLS is the sole applied variable selection technique applied in Paper II and is described in some detail below.

*i*PLS was developed for PLS models to extract relevant spectral regions for a given outcome. The idea is only to include intervals of variables relevant for the prediction of the outcome or to exclude intervals *not* relevant. There are two forms of *i*PLS; forward selection and reverse selection. In forward selection, one interval is included at a time meaning that a model is built on this specific interval and the error is calculated. Then *i*PLS includes an additional interval and a model is built on the two selected intervals. This procedure is continued until the model performance drops, hence only the intervals performing best are included in the model. In reverse selection, *all* variable intervals are included to begin with and then one interval is *excluded* at a time. Model performance is tested every time an interval has been excluded. In metabolomics, biomarkers are measured and after performing RRV and proper integration, each biomarker is represented by one variable. The intervals used for *i*PLS will therefore only include one variable contrary to spectral data, where spectral regions are binned in intervals.

It is important to note that during variable selection the models must be validated. Otherwise, there is a risk of achieving a dataset with unreliable variables leading to over-optimistic results. This leads back to the section suggesting the inherence of using test set validation and a model based on erratic variables will be found useless and proper action can be taken.

4.6. Data Fusion

In metabolomics, it is common to acquire information from many different sources. Already mentioned acquisition techniques such as NMR, GC-MS and fluorescence all complement each other, and additional data from questionnaires and anthropometrics provide even more knowledge of samples. Data fusion, where data from different platforms are jointed, has been known for long [88-90], and a common strategy is concatenation in different levels [2;91;92]. A prerequisite is that the samples measured at the

different platforms are identical, and it is then possible to fuse or concatenate the data. All types of data can be fused, and in Paper II, data from NMR and data from two questionnaires with information mainly concerning anthropometrics, life style habits such as smoking, alcohol intake and dietary habits have been fused. The data fusion leads to improved prediction models and known biomarkers can be confirmed. A recent study by Bro *et al.* [93] has also shown that traditional biomarkers could significantly benefit from the synergy of being fused with new chemical profiles from NMR and fluorescence.

Chapter 5

Case Studies

Throughout the work of this thesis, three metabolomics case studies have been conducted. The first one is an application of NMR spectroscopy on plasma samples from healthy females, where half of them have developed breast cancer later in life. This application is very interesting with respect to the *prediction* of the development of breast cancer *later* in life. Additionally, many of the obstacles encountered when analysing a complex metabolomic dataset are discussed (Paper II). The second application concerns analysis of plasma samples by fluorescence spectroscopy from people with symptoms of colorectal cancer. This study leans towards an early detection of cancer in the sense that none of the subjects were diagnosed with colorectal cancer at the time of blood withdraws (Paper III). The last application is a small feasibility study, where standard plasma samples were analysed using GC-MS. The focus in this application was the development of proper extraction and derivatization methods, which are challenging disciplines in GC-MS-based metabolomics.

5.1. Forecasting breast cancer by NMR

Forecasting breast cancer status from healthy individuals is a new subject in metabolomics. In order to develop meaningful and not least useful prediction models, a large sample set is needed. For this application, samples from the Danish "Diet, Cancer and Health" cohort have been analysed by NMR spectroscopy. In this section, a more thorough description of the data pre-processing and how the results from Paper II were obtained is presented.

5.1.1. The Danish "Diet, Cancer and Health" Cohort

The Danish "Diet, Cancer and Health" (DCH) cohort [94] is part of a major study called the "European Prospective Investigation into Cancer and

Nutrition" (EPIC) conducted by the Imperial College in London [95] including participation from ten European countries. The DCH cohort was established between 1993 and 1997 where 57,053 men and women were enrolled. In order to be included in the cohort, the following criteria should be met: Age between 50 and 64 years, born in Denmark and no previous cancer diagnosis registered in the Danish Cancer Registry. Hence, all participants were free of cancer and were considered to be *healthy*. A detailed food frequency questionnaire and a lifestyle questionnaire were filled in by each participant. Additionally, 99% of the participants in the cohort gave biological material at the time of enrolment. In the present case study, plasma samples have been analysed and 47 selected variables from the two questionnaires were included in the final data analysis by data fusion. The 47 questionnaire variables are described in the supplementary material in Paper II.

A subset of 3,510 samples from the DCH cohort selected by the Danish Cancer Society was analysed by NMR spectroscopy. Due to insufficient sample volumes or low data quality, the total study population was reduced to 3,419 samples. Out of these, some were diagnosed with breast cancer, colorectal cancer and/or cardio vascular disease between enrolment and 31 December, 2000 (breast cancer (N = 419), colorectal cancer (N = 414), and/or cardiovascular disease (N = 1,106)[1]). The remaining samples are a randomly selected cohort subsample (N = 1,493) not diagnosed with any of the diseases. Out of the 1,493 subsamples, 747 were women.

5.1.2. Data pre-processing

CPMG and NOESY spectra were recorded for all the 3,419 plasma samples (referred to as CPMG and NOESY in the following). The water signal was suppressed in both experiments due to the large amount of water in plasma, and for the CPMG experiment large molecules such as the proteins were suppressed as well. More details concerning the data acquisition can be found in Paper II.

Prior to any pre-treatment of data, the residue of the water peak was eliminated from the profiles. Additionally, the citrate peak was eliminated due to the use of citrate coated tubes as anti-coagulant when storing the plasma. The spectra were normalized by second order normalization.

Given the large number of variables in each set (approximately 56,000 in both), RRV was carried out to eliminate redundant variables. Since CPMG

[1] Three of the 419 breast cancer cases and ten of the 414 colorectal cancer cases had additionally developed cardiovascular disease

and NOESY complement each other, both spectra were manually inspected. Every little peak in the spectra was located resulting in two datasets only containing chemical information and leaving out pure noise. Despite the fact that CPMG and NOESY supplement each other, many identical compounds were measured in both methods with identical chemical shifts. In each of these cases, the most well-defined peak containing least baseline was included in the final data analysis meaning that each compound was only represented once by either CPMG or NOESY. This major data reduction resulted in a dataset with 189 peaks. The majority of the peaks were selected from CPMG, since this experiment generally contain a more flat baseline than NOESY. Eight peaks were selected in NOESY and 181 peaks were selected in CPMG. Figure 18 shows two selected peaks from CPMG. The left side of the figure is identified as α-glucose, whereas the peak to the right has hitherto not been identified.

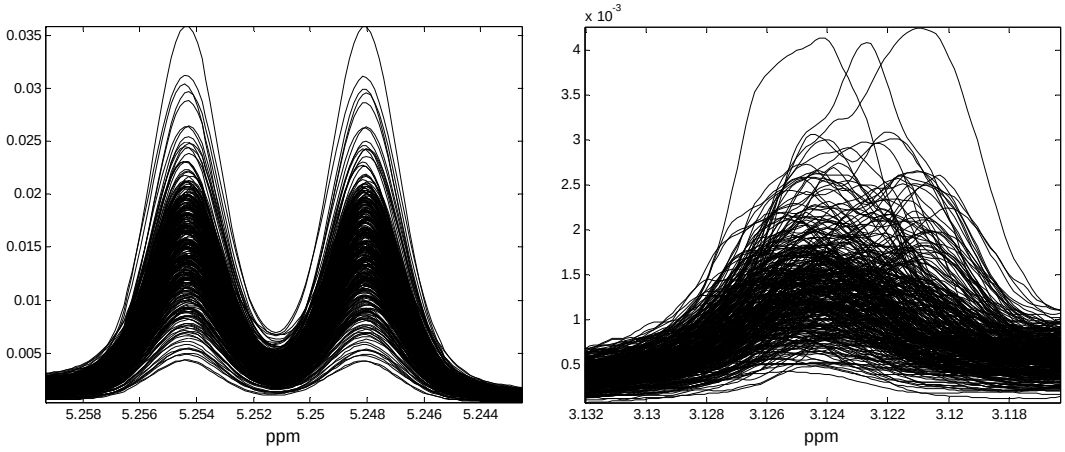


Figure 18. Two peaks selected from CPMG (only every 100th sample is shown). The peak to the left is α-glucose, whereas the peak to the right has not been identified

5.1.2.1. Baseline correction

Unfortunately, the NMR equipment broke down after measuring half of the plasma samples. This was visible in the first component in a PCA model with a clear separation between samples measured before and after the breakdown as seen in Figure 19. The problem may arise from minimal differences between the experimental settings or instrumental instability, but this was not investigated further.

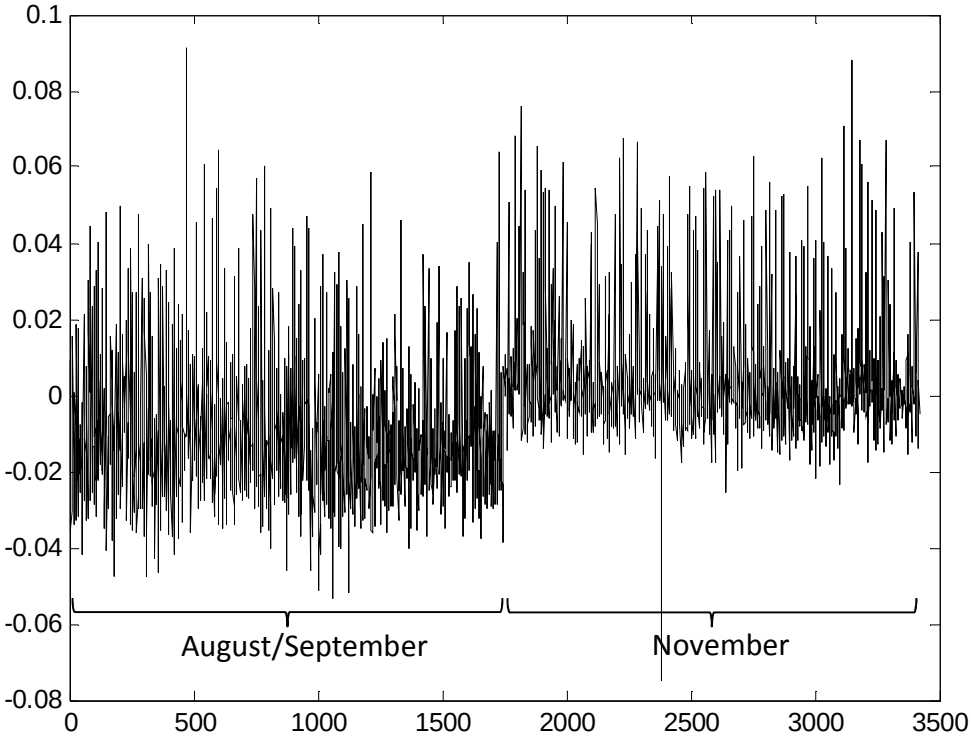


Figure 19. PCA scores from one NMR interval with 3419 samples. There is a clear jump in the score values after the equipment broke down. The two measurement periods are August/September and November

Besides the visible jump in the score values, the two measurement periods were also visible in the baseline for many peak intervals. An example is shown in Figure 20 (left), where the difference between the measurement periods is evident. Baseline correction was carried out in the peak intervals only where the shift was visible. The remaining peak intervals were left untouched. The baseline correction was specified for each interval to make sure that the polynomial used for subtraction gave the best result: elimination of baseline shift; no shape changes in the baseline ends of the interval; improvement of or retaining the peak shape. An example of an interval where baseline correction was needed is seen in Figure 20. There is a visible reduction in the baseline, the peak shape is clearer and the difference between the two measurement periods is reduced.

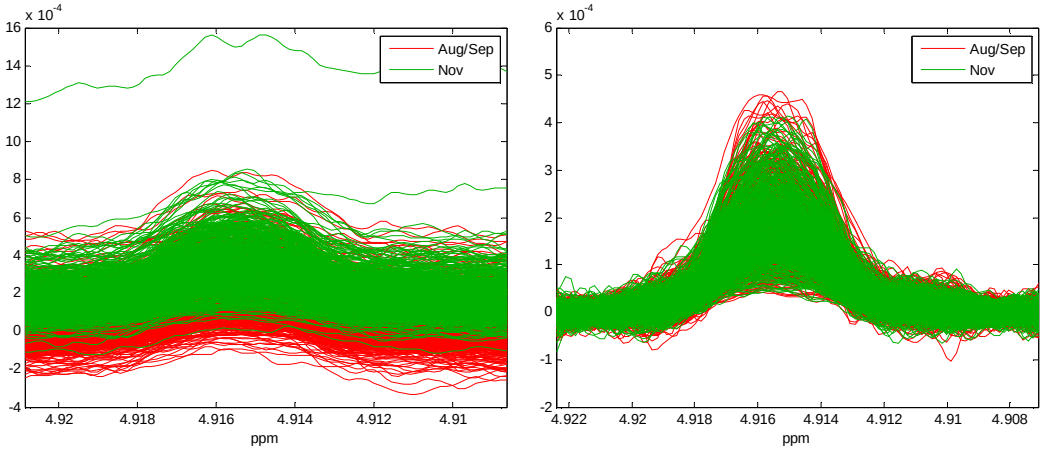


Figure 20. Interval before (left) and after baseline correction (right)

The effect of the baseline correction was immediately observed in a PCA model of the interval from Figure 20. The first score in the PCA model for the uncorrected interval is related to the measurement days, and the first period (August/September) is more spread out than the second period. The variation is completely eliminated in the PCA model when the interval has been baseline corrected. It was therefore decided that baseline correction seemed to be a proper solution for the period problem. Using baseline correction where it was necessary eliminated the variation making the dataset ready for further analysis.

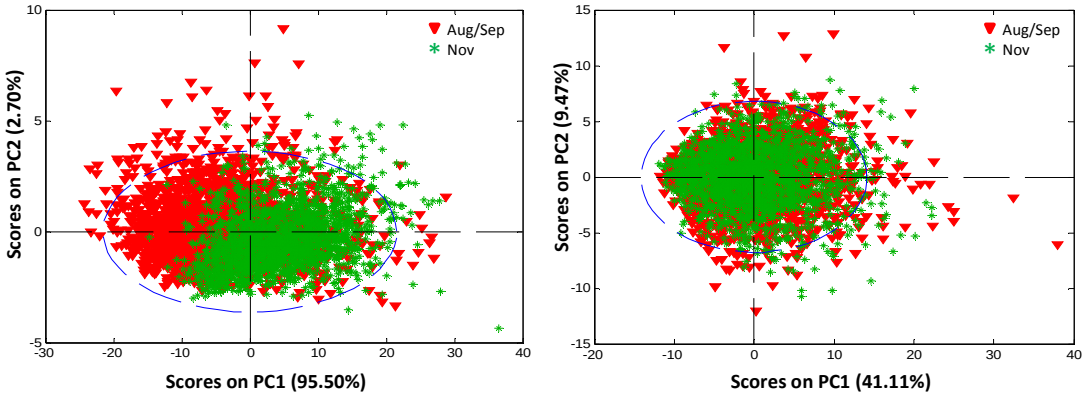


Figure 21. PCA models of the interval from Figure 20 before (left) and after (right) baseline correction.

5.1.2.2. Integration

Prior to data analysis, each peak was integrated using either the area under the peak, the height of the peak or MCR concentration profiles as described previously. The three methods were assessed for all peaks and taking the quite different peak shapes in to account, the optimal integration method was determined for each peak. One peak was integrated with a two-component MCR model, 54 peaks were integrated using a one-component MCR model, 86 peaks were integrated using the area under the curve, and 48 peaks were integrated using the height of the peak.

After selection and integration, the peaks were assigned primarily according to the findings by Nicholson et al. [96]. Several peaks were unfortunately not assigned. All the assigned regions are shown in the supplementary material in Paper II.

Some of the selected peaks were included as single peaks despite being part of a double peak (a doublet). An example is the identification of four separately selected peaks all originating from tyrosine – however, tyrosine is represented as two doublets in the NMR spectrum. A small part of the average CPMG spectrum of the 838 plasma samples is presented in Figure 22, where the four peaks identified as tyrosine are highlighted. Tyrosine is therefore represented by four variables (integration of each peak), which is not necessary.

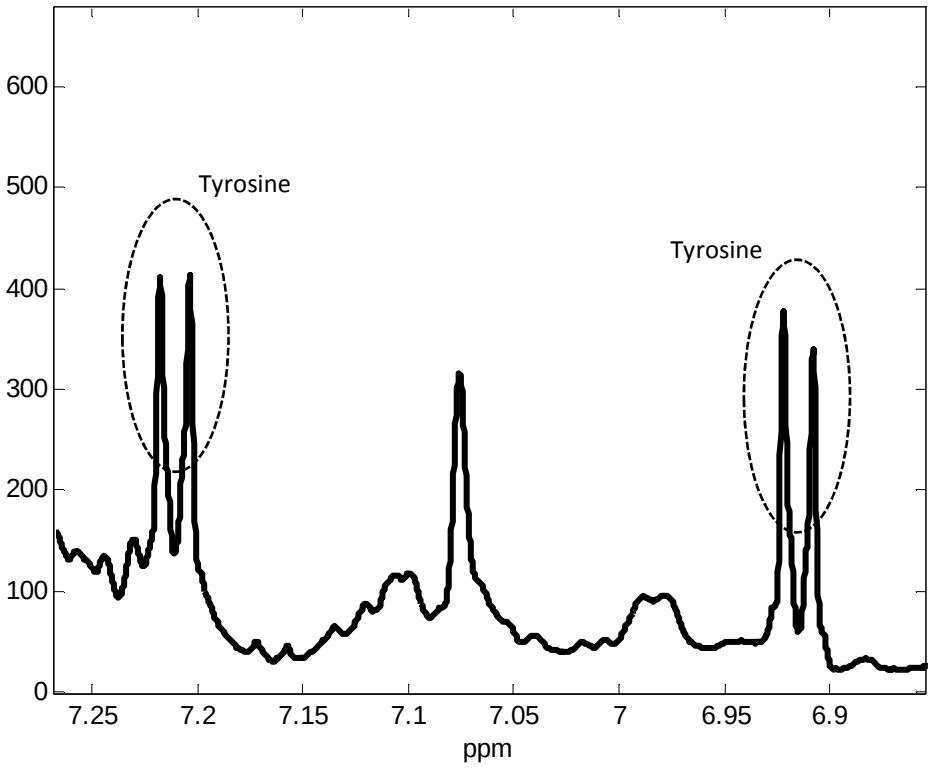


Figure 22. Average CPMG spectrum, where four peaks are highlighted (identified as tyrosine)

Peaks originating from the same molecule were summed and included in the data analysis as one single representative of the molecule. Whether peaks – represented by two or more variables – originate from the same molecule was further supported if a correlation between the integrated peaks of the molecules was present. The correlation of the four integrals of tyrosine is shown in Figure 23. Since there is a strong correlation between the integrals, it was concluded that the integrals can be merged to represent tyrosine by one variable.

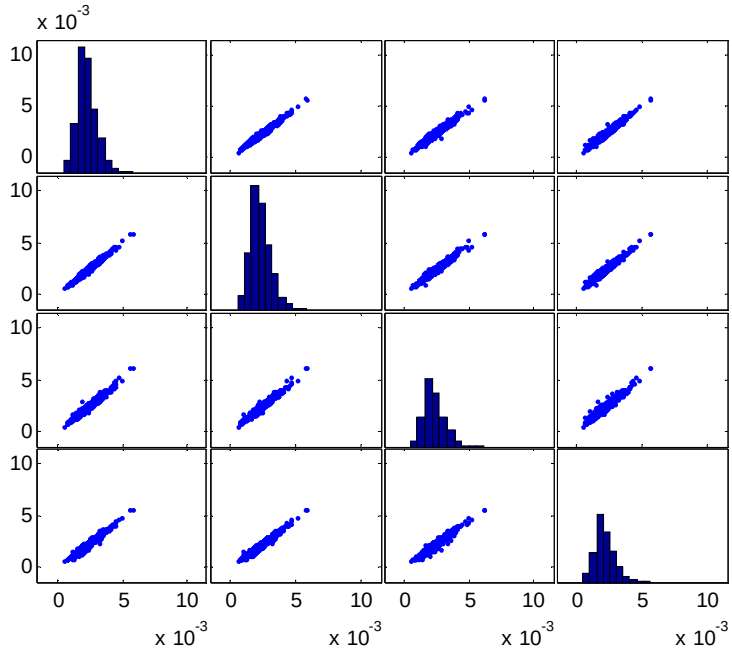


Figure 23. The four integrals identified as tyrosine. There is a strong correlation between all four integrals, and tyrosine can therefore be represented as the sum of the four integrals

It should be noted that merging integrals is only possible if the corresponding peak is identified. Otherwise, it is impossible to evaluate the specific origin of the proton. After merging all identical peaks, the number of variables was further reduced to 129 NMR variables.

5.1.3. Prediction model and bio-contours

The focus in this NMR study was the 419 women who have developed breast cancer later in life. In order to obtain a balanced dataset for the data analysis, a random female subset of 419 samples out of the 747 control samples was selected. Hence, the dataset consists of 838 samples which is a very high number compared to other metabolomics studies [19;21;48]. The variables consist of the 129 NMR variables fused with the 47 additional questionnaire variables mentioned above and the total dataset is 838 × 176.

Prior to any data analysis, the dataset was divided randomly into two sets – a calibration set and a test set. The calibration set was of size 628 × 176 and the test set of size 210 × 176 containing equally many cancer and control samples. The calibration set was used to build a classification model and the test set was applied to validate the performance of the classification model.

One of the key points in Paper II is that a complex biological problem i.e. cancer is unlikely to be described by one or few biomarkers. A complex system

must be investigated by looking at the pattern of variation in the data. Two examples of single biomarkers as classifiers for future development of breast cancer can be seen in Figure 24. Hormone replacement therapy (HRT) is a well-known risk factor for breast cancer [97;98]. However, using HRT as a single variable in an LDA model gives a sensitivity of 0.29 and a specificity of 0.85. From Figure 24 (top) it is clear that the discriminatory power is very limited. The best single variable for classification is the NMR variable at 3.12 ppm. For this variable, the sensitivity and specificity values are low (0.65 and 0.55) and as for HRT the discriminatory power is low (Figure 24 (bottom)).

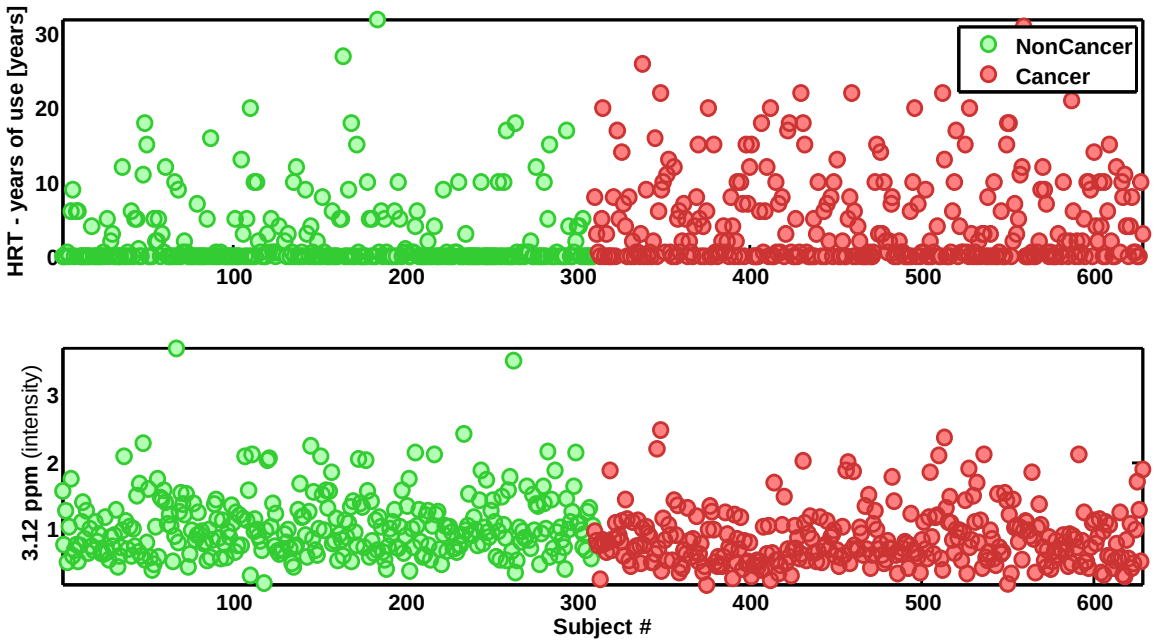


Figure 24. Examples of single biomarkers used for discrimination between cancer and non-cancer samples. The figure is from Paper II

Evidently, using single biomarkers as predictors of future cancer status is pointless. It is therefore important to develop models, where all variables are considered and irrelevant variables are carefully discarded.

The data analysis was performed using PLS-DA and variable selection by means of *i*PLS. The final model contained 28 variables out of the 176 variables and was built using eight PLS components with sensitivity and specificity values of 0.84 and 0.85, respectively. The 28 selected variables are predictors of future breast cancer status only when combined and all eight PLS components should be considered. Regarding the interpretability of the model, it is the *pattern* of the variables that reflects the biology, and picking out and interpreting single or few of the biomarkers found must be avoided,

since the study is not a design study where each variable has been varied independently. In Paper II, the underlying pattern is called a *bio-contour*, a term used in order to avoid the temptation of interpreting single biomarkers.

I should be noted that the selected 28 variables are not necessarily the only variables representative for forecasting cancer status. In order to investigate the variability between selected variables, 1000 classification models were calculated. The variables were selected by *i*PLS during resampling and in Figure 25 it can be seen that only few of the selected 28 variables are represented in almost all models.

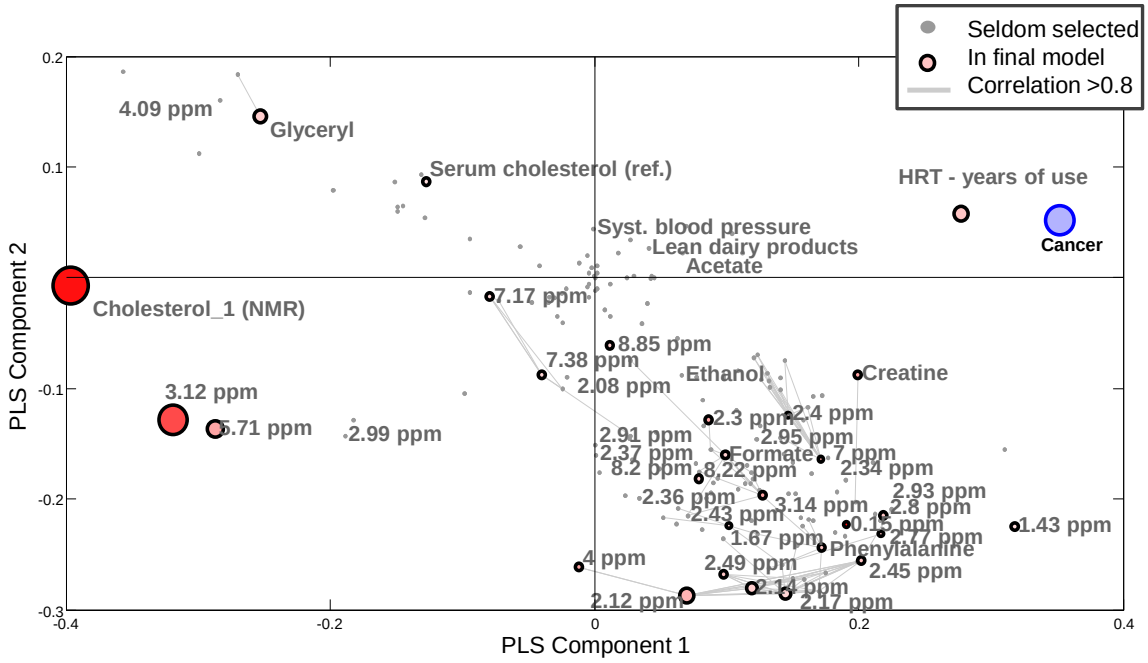


Figure 25. Loading plot of the 1000 resampled classification models. Variables encircled by a black line are the actually chosen 28 variables in the final model. Colour intensity and size of the circle indicate how often a given variable is chosen during resampling of the variable selection. The figure is from Paper II

The variables encircled by a black circle are those included in the final classification model with 28 variables in total. The larger the circle and the more intense red colour indicate that the variable has been selected in most of the 1000 resampled models. For example, the variable "Cholesterol_1 (NMR)" is selected in most models, whereas "HRT – years of use" is selected in fewer of the models. The variables marked by a grey dot are seldom selected in the models. The key point here is to understand that the selected 28 variables in the prediction model are not necessarily the best in the prediction of future breast cancer. Any combination of a selection of the variables represented in Figure 25 might predict cancer status equally well. The variable "HRT – years of use" is only included in some of the models, and this indicates that

there are other (unknown) variables present in the loading plot representing the same biology as "HRT – years of use". Additionally, it is imperative to remember that the model is composed of eight PLS components which all combined form the bio-contour. In the loading plot in Figure 25, only the two first PLS components are depicted, but there is information related to cancer status in all eight components.

Finally, the test set was used to test the performance of the classification model. The sensitivity and specificity values were 0.80 and 0.79, respectively, which is a strong indicator of a reliable model. In Figure 26, the discrimination between the two groups in the test set is visualized. Compared to the univariate models shown in Figure 24, the separation between the groups is much stronger.

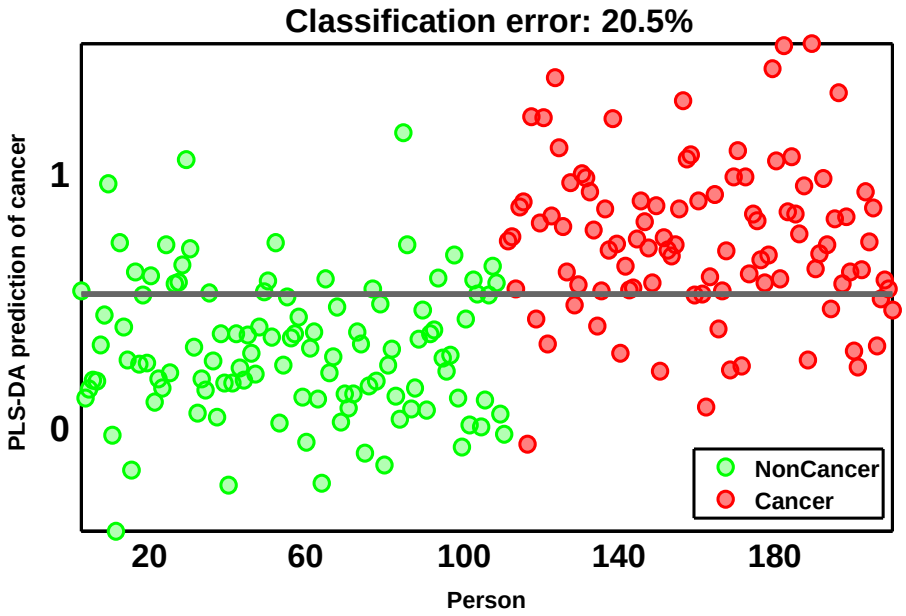


Figure 26. Prediction of the 210 samples in the test set. The sensitivity is 0.80 and the specificity is 0.79. The figure is from Paper II

5.1.4. Time to tumour

During analysis of the data, one NMR variable showed an interesting behaviour regarding the time from blood withdrawal to detection of a tumour in the individuals, who developed cancer. This timeline is here called time to tumour.

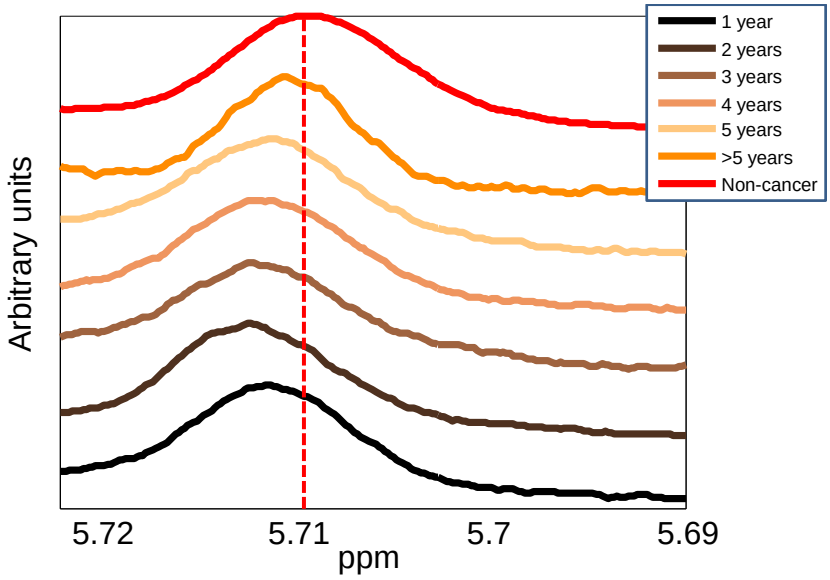


Figure 27. NMR variable "5.71 ppm" with average values of time to tumour. The cancer samples have been divided into six groups with respect to the time to tumour and one group represents the non-cancer samples. The red dotted line marks the chemical shift for the non-cancer samples

When plotting the peak, there is a slight shift towards the left side of the ppm values, when breast cancer is detected compared to the control samples. The shift is ambiguous though, since the shift seems to be minor both for the development of breast cancer within one year and after five years. The years in between are shifted more to the left. The shift indicates that there is some information related to the actual development of a breast cancer tumour. The time tendency is not directly visible in the classification model; individuals who are diagnosed with cancer later than others are not misclassified more often. The variable has a spectral range from 5.690 to 5.722 ppm, but it has up until now not been possible to identify the peak. Unfortunately, it has not been possible to even develop a slight guess of *why* the shifting behaviour between the samples is observed.

5.1.5. Conclusion

This case study is unique in the sense that the analysis has been performed on healthy individuals. The results indicate that there is a potential for forecasting breast cancer incidences years ahead in time based on a blood sample. The model uses a bio-contour consisting of eight dimensions, which should be considered as being associated with breast cancer in a combined manner.

5.2. Early detection of colorectal cancer by fluorescence

The second case study is from Paper III and concerns early detection of colorectal cancer (CRC) by fluorescence spectroscopy. Fluorescence spectroscopy is not a common data acquisition technique in metabolomics, but the current study is an attempt to show that the technique can be a valuable supplement to NMR and GC-MS. A subset of plasma samples from a multi-centre cross sectional study conducted at six Danish hospitals [99], where patients with symptoms of CRC have been undergoing large bowel endoscopy has been analysed. Early detection lies in the fact that none of the patients were diagnosed with CRC at time of enrolment. In this section, the key results from Paper III concerning the multivariate data analysis are presented.

5.2.1. The dataset

The subset consists of 308 plasma samples and the sample set is designed as a case control study. Out of the 308 samples, 77 samples were denoted as the case group with verified CRC from the endoscopy. The remaining samples were divided into three control groups representing other findings from the endoscopy, each containing 77 samples: (i) healthy patients with no findings, (ii) patients with other non-malignant findings and (iii) subjects with adenomas.

The samples were divided into smaller portions – one portion was diluted by a one-hundred fold and one portion was left undiluted. The diluted and undiluted samples were measured at excitation wavelengths ranging from 250 to 340 nm with a five nm increment and at emission wavelengths ranging from 300 to 600 nm with a one nm increment. Additionally, the undiluted samples were also measured at 385 to 425 nm (excitation) and at 585 to 680 nm (emission) in order to detect emission from porphyrins, which have proven important for the detection of CRC [100]. Additional details concerning sample preparation and measurements can be found in Paper III.

5.2.2. Classification model

Data from the fluorescence EEMs have a trilinear structure suitable for decomposition by the PARAFAC model. If the proper number of PARAFAC components is chosen for the decomposition, the scores and loadings will be estimates of the true concentrations, excitation profiles and emission profiles of the measured fluorophores. PARAFAC models were calculated for each of

the three datasets. For the undiluted samples measured at the low spectral area with excitation at 250 to 340 nm and emission at 300 to 600 nm, the PARAFAC model was calculated using ten PARAFAC components. For the diluted samples measured in the same spectral area, the PARAFAC model was calculated using six PARAFAC components. For the last dataset obtained from the undiluted samples in the higher spectral area, a three component PARAFAC model was calculated. The number of components was selected based on the core consistency diagnostic and inspection of the loadings.

In order to build a classification model, the scores from the three PARAFAC models were concatenated into one single score matrix resulting in a data matrix with 19 variables. Classification models were built for all combinations of the case group versus the three different control groups including one model with cases versus all three control groups. Additionally, the control groups were tested against each other. For all classification models, validation was performed using cross-validation and test set validation. The performance statistics of the classification models are presented as a PCA model in Figure 28. The purpose of the PCA model is to illustrate the differences and similarities between the seven PLS-DA models presented in Table 1 in Paper III based on the performance statistics: cross-validated sensitivity, cross-validated specificity, AUC, predicted sensitivity and predicted specificity.

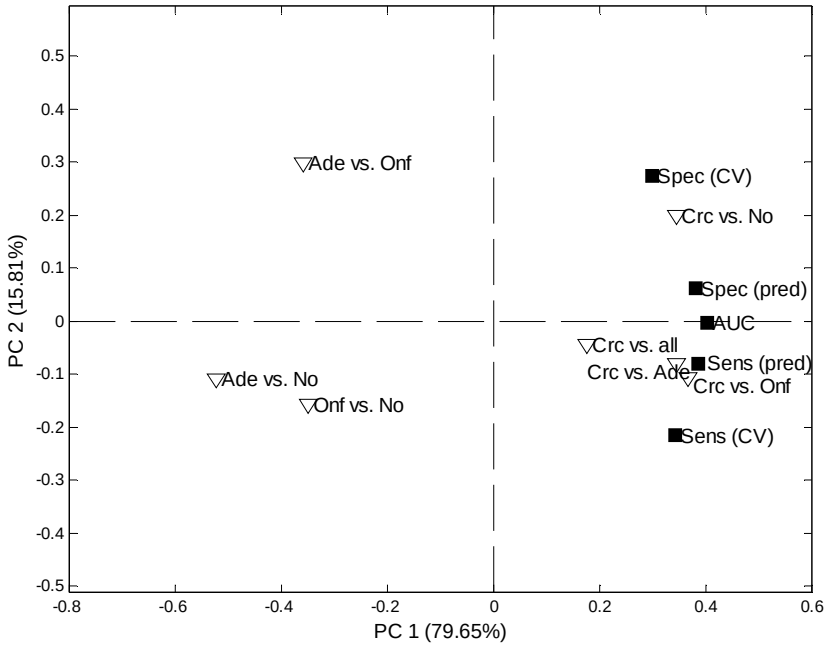


Figure 28. PCA model of the performance statistics of the seven classification models. The plot shows both scores and loadings (bi-plot), Crc = cancer, No = no findings, Onf = other non-malignant findings, Ade = adenomas, all = all three control groups, CV = cross-validated, pred = predicted. The data are adapted from the values in Table 1 in Paper III

The sensitivity and specificity values are higher for the classification models (approximately 0.75 for the test set), where CRC is separated from each control group and from all control groups merged into one. In addition, in the three models where the control groups were separated from each other, the sensitivity and specificity values are low. These models are located in the left part of Figure 28. Hence, it is possible to separate cases from all types of controls, whereas no separation between control groups is observed. In Figure 29, the ROC curves for the calibrated and cross-validated models for CRC versus other non-malignant findings are shown. The two curves are fairly similar indicating that the cross-validated model is robust.

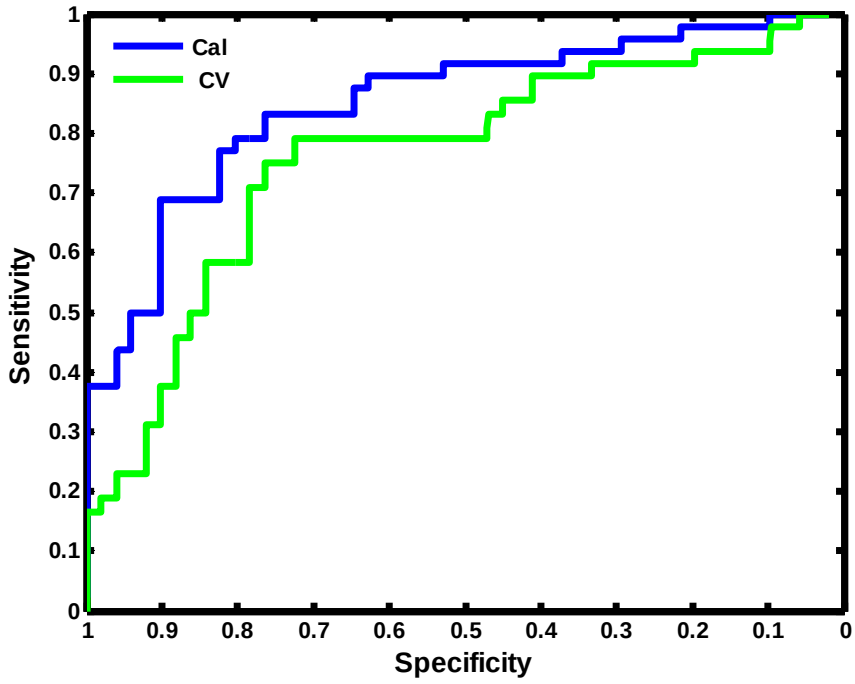


Figure 29. ROC curves for the calibrated (blue) and the cross-validated (green) models of CRC vs. other non-malignant findings

In Figure 30, the results from the four-PLS component PLS-DA model of CRC versus other non-malignant findings are presented. In the PLS-DA score plot to the left, a slight separation between cases and controls are observed. The CRC samples are the blue triangles and are positioned more to the right in the score plot, whereas the controls samples (red circles) are positioned more to the left. Hence, the cancer direction goes towards the right-hand side in the plot. This separation can be explained by the corresponding loading plot (Figure 30, right). Here, the variables are divided in two groups illustrated by the dashed diagonal line in the plot. The variables on the left side of the line are negatively correlated to the cancer direction, and the variables on the right side are positively correlated to the cancer direction. The loadings can

also be considered as a bio-contour associated to CRC and are all equally important for the separation of cancer from controls. Additionally, the model is built using four PLS components and it is the combination of the four components that form the model and provide the basis of the performance statistics. All four components should therefore be assessed in a combined manner in order to acquire the whole pattern of the bio-contour.

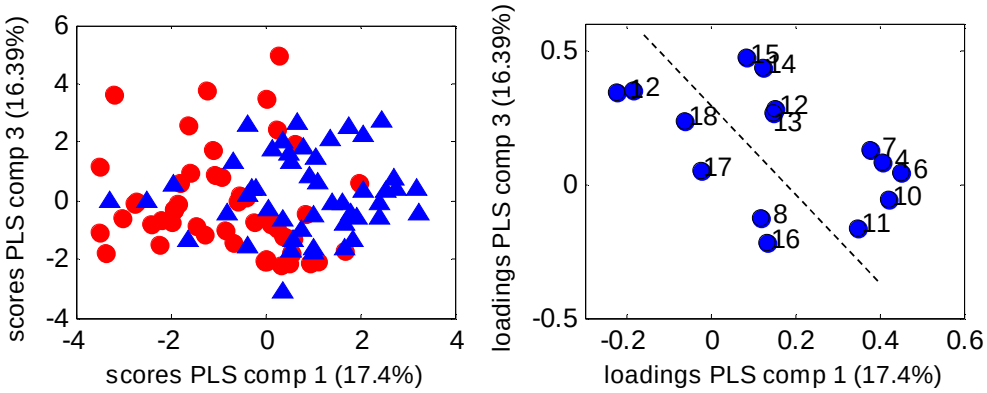


Figure 30. Four-component PLS-DA model of CRC vs. other non-malignant findings (Onf). Left: PLS-DA score plot shows a separation between CRC (blue triangles) and Onf (red circles). Right: PLS-DA loadings with 13 variables selected by forward *i*PLS. The figure is modified from Paper III

5.2.3. Conclusion

This study showed promising results for detection of CRC. The sensitivity and specificity values of approximately 0.75 for the test sets are very promising results for detection of CRC. Further studies are needed, but with the results obtained, fluorescence spectroscopy has considerable potential as a tool in the detection of CRC in blood.

5.3. Notes from a feasibility study of plasma by GC-MS

The feasibility study of plasma samples measured using GC-MS, was originally intended to be a supplement to the findings in Paper III. Hence, plasma samples from the same study were to be analysed by GC-MS in order to find more metabolites related to CRC and furthermore to combine GC-MS data and fluorescence data by data fusion. However, too many obstacles were hindering the data acquisition due to difficulties in finding an optimal method for sample preparation despite many useful publications on the subject. Therefore, the feasibility study was conducted on standardized pooled human

citrate plasma samples. Different approaches for sample preparation were tested and are described in the following. A future project is planned to analyse the samples from both the DCH cohort and the CRC plasma samples. The purpose is then to fuse the acquired GC-MS data with the already obtained NMR and fluorescence data.

5.3.1. Sample preparation

In the past few years, sample preparation in GC-MS-based metabolomics has gained more attention, and robust methods for preparation of plasma samples have been developed [101;102]. The main focus in this feasibility study concerns extraction and derivatization of plasma samples prior to analysis by GC-MS in order to detect as many metabolites as possible.

5.3.1.1. Extraction

The low molecular weight compounds which are the compounds of interest in metabolomics are usually non-covalently bound to proteins. It is therefore necessary to precipitate the proteins, commonly achieved using an organic solvent prior to further treatment of the samples. This procedure will extract the metabolites bound to the protein, and the metabolites can be treated further. Suggested solvents are methanol, ethanol, chloroform, acetone or acetonitrile, and the extraction effect of these five solvents has been tested in a study by A *et al.* in 2005 [103]. They concluded that methanol is the most suitable organic solvent for extraction of metabolites, as more metabolites were detected compared to extraction by the other solvents. Additionally, they suggest a methanol volume of 80% (the remaining 20% being water, sample material and internal standards). Internal standards can be added in order to keep track of the preparation procedure.

After extraction of the metabolites, a crucial step in the sample preparation is to dry the samples. This is necessary in order to bring the metabolites to a volatile state by derivatization. If the samples are not completely dry, the derivatization cannot be completed, and no metabolites will be detected, which will be described in the following.

5.3.1.2. Derivatization

Derivatization of the metabolites is important if the metabolites are polar, e.g. carbohydrates and amino acids, because these types of metabolites will

bind strongly to the GC column by hydrogen bonds. Lipids, on the other hand, are non-polar and do therefore not need to be derivatized [101].

Derivatization is usually carried out using a silylating agent, where the hydrogen atoms are replaced by the silyl group making the compound less prone to form hydrogen bonds. Hence, the metabolites will bind less strongly to the column, and the volatility is increased. The functional groups ketones and aldehydes are in equilibrium with their enol forms and there is a possibility of forming trimethylsilyl ethers during derivatization, which are thermally unstable. The presence of trimethylsilyl ethers will increase the risk of incomplete derivatization, resulting in multiple peaks of these compounds making quantitative interpretation a challenge. In order to prevent formation of trimethylsilyl ethers, methoxymation should be carried out in advance of derivatization. Methoxymation converts the problematic functional groups into oximes, which will prevent the undesired formation of trimethylsilyl ethers. Commonly, methoxymation is carried out using *O*-methoxylamine hydrochloride in a pyridine solution. When the methoxymation is completed, derivatization can take place using either of the silylating reagents MSTFA (*N*-methyl-*N*-(trimethylsilyl) trifluoroacetamide) or BSTFA (*N,O*-bis(trimethylsilyl) trifluoroacetamide), which are most common. The silylation reaction can be catalysed using TMCS (trimethylchlorosilane) [101]. As mentioned, it is of the utmost importance that the sample is completely dry prior to methoxymation and derivatization. If there are traces of compounds with active hydrogen atoms (water, methanol, etc.), these will also be derivatized and will hide signals from the metabolites of interest. For analysis of the plasma samples in the following, MSTFA was selected as the derivatizing agent, which is common practice for derivatizing plasma samples [101-103].

5.3.1.3. The protocol

One of the limitations in the feasibility study was the amount of sample material. Usually, the protocols suggest a sample volume of 100 µL, but for the original purpose of the present study only 50 µL of plasma was available. The following protocol in Table 3 was therefore developed with inspiration from the methods developed for pre-treatment of plasma samples, but with reduced sample material of 50 µL [101-103].

After evaluating several pre-studies, the following protocol in Table 3 was developed with three different concentrations of the methanol volume (50%, 65% and 80%), in order to find the optimal methanol/water/sample ratio. Additionally, samples with the highest and the lowest methanol concentrations were re-dissolved in the corresponding methanol

concentrations in an attempt to extract more metabolites. All samples were prepared in triplets resulting in a total of 15 prepared plasma samples.

Table 3. Protocol for pre-treatment of plasma samples prior to injection to the GC-MS. Steps 1-16 were carried out manually whereas steps 17-21 were carried out by the autosampler from the GC-MS equipment

Pre-treatment of plasma samples	
1.	Incubate the plasma samples at 37°C for 15 min to thaw them
2.	1. 150 µL methanol to 10 µL MQ water and 50 µL plasma to Eppendorf tubes (50% methanol) (**A**) 2. 195 µL methanol to 10 µL MQ water and 50 µL plasma to Eppendorf tubes (65% methanol) (**B**) 3. 240 µL methanol to 10 µL MQ water and 50 µL plasma to Eppendorf tubes (80% methanol) (**C**)
3.	Vortex the samples for 10 sec
4.	Put the samples on ice for 10 min
5.	Vortex the samples (hard!) for 2 min
6.	Put the samples on ice for 2 hours
7.	Centrifuge the samples at 19600g for 13 min at 4 °C
8.	For half of the samples: Transfer 100 µL supernatant to GC-vial-insets
9.	For the other half of the samples: Transfer 250 µL supernatant to Eppendorf tubes
10.	Dry the samples in a speed-vacuum centrifuge – they MUST be completely dry (app. 2 hours)!
11.	Cap the three samples in GC-vials with nitrogen, close the lid hard and cover with parafilm. Store in fridge (+5°C)
12.	Re-dissolve samples **A** and **C** (those in the Eppendorf tubes): (**A**): 60 µL methanol and 60 µL MQ water (**A1**) (**C**): 96 µL methanol and 24 µL MQ water (**C1**)
13.	Vortex each sample to dissolve the pellet
14.	Transfer 100 µL supernatant to GC-vial-insets
15.	Dry the samples in a speed-vacuum centrifuge – they MUST be completely dry (app. 2 hours)!

Pre-treatment of plasma samples (continued)	
16.	Cap the three samples in GC-vials with nitrogen, close the lid hard and put parafilm on. Store in fridge (+5°C)
	The following is carried out by the autosampler
17.	Add 30 µL methoxyamine hydrochloride in pyridine* (15 mg/mL)
18.	Shake the samples for 90 min at 30°C
19.	Add 30 µL MSTFA
20.	Shake the samples for 30 min at 37°C
21.	Inject to GC-MS

* Methoxyamine hydrochloride in pyridine was freshly prepared prior to the experiments

5.3.2. Data acquisition

In Table 4, an overview of the GC-MS settings is presented. The parameters have not been fully optimized, as the initial experiments focused on the samples preparation. The GC-MS instrument is an Agilent Technologies 7890A GC System with autosampler coupled to a 5975C inert XL MSD with Triple-Axis Detector (quadrupole).

Table 4. Settings for injection, the chromatograph and the mass spectrometer. The settings are not optimized

GC-MS settings	
Injection settings	Inject 1 µL sample cold in splitless mode with H_2 as carrier gas flow rate at 3 mL/min
Chromatography settings	Column: HP-5MS 5% Phenyl Methyl Silox Temperature program: Isothermal for 2 min at 80ºC, then ramped with 5ºC/min until 320ºC is reached; 5 min isothermal. Cooling for 10 min
Mass spectrometer settings	The ion source (EI) is set to 230ºC. The mass range is set to 40 to 500 Da with a scan rate of 3.15 pr. sec.

5.3.2.1. Preliminary chromatogram

Only 14 out of 15 samples were acquired, and in Figure 31 the total ion chromatogram (TIC) is presented.

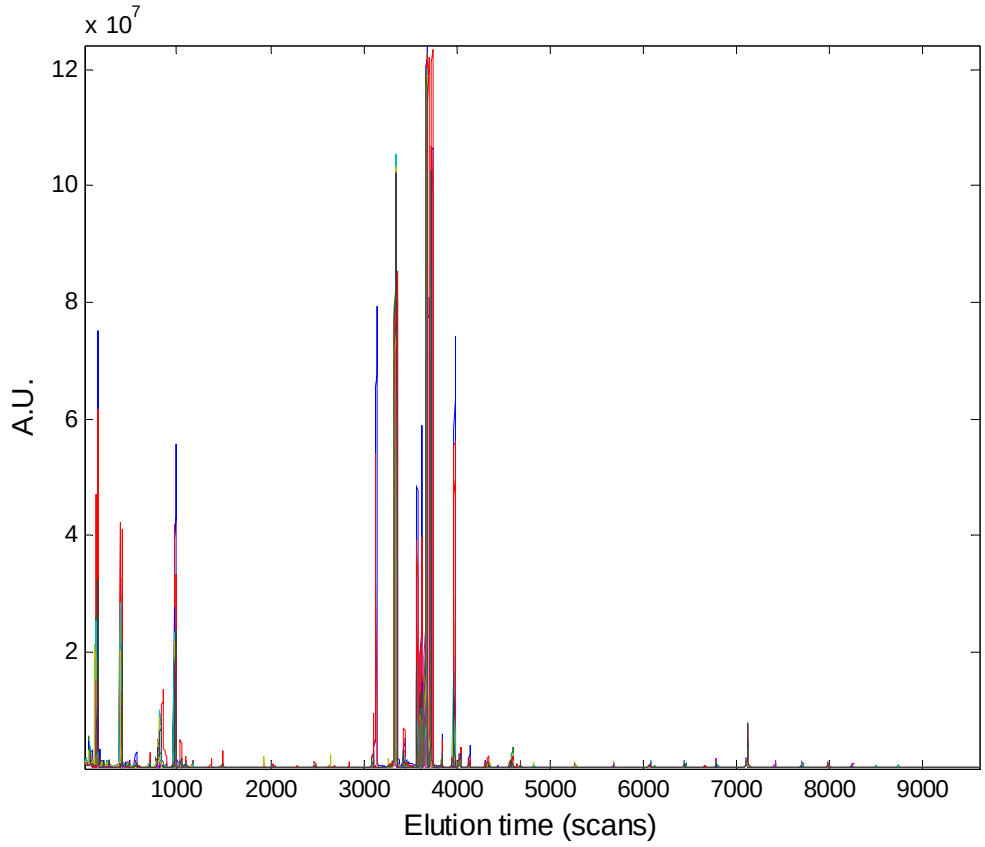


Figure 31. TIC of the 14 plasma samples

Many of the measured metabolites have a low signal-to-noise ratio and unfortunately some of the peaks were found to be column material. The low signal-to-noise ratio can probably be overcome by analysing larger amounts of sample material and more samples need to be analysed to find the optimal methanol volume. The presence of column material indicates that the settings of the GC-MS need to be adjusted or that the choice of column is not optimal. The proposed column in other protocols is a 35% phenyl methyl silox column [101;102], whereas the one in the present study is a 5% phenyl methyl silox column.

When a chromatogram is obtained, the metabolites can be identified using different databases such as the Wiley library [104] and NIST [105]. These are very useful also as initial indicators of whether the measurements are biologically meaningful and to locate column material. For the final analysis of cancer samples, the identification of metabolites from databases will aid in the determination of which biomarkers are associated with cancer. For the current state of the optimization of the analyses, little attempt has been made to identify the measured metabolites. However, one metabolite was identified as cholesterol, which is the one presented in Figure 3 in Chapter 3. This

specific metabolite has been modelled by PARAFAC2 in the following in order to briefly investigate the potential of the PARAFAC2 model in metabolomics.

5.3.2.2. PARAFAC2 models

This feasibility study is only preliminary and many optimization steps are required. However, some metabolites are real and modelling these will give an indication of whether it is possible to estimate the elution time and mass spectral profiles. Therefore, PARAFAC2 models on one selected interval (identified as cholesterol) have been calculated due to shifts in this specific peak.

Prior to the modelling of cholesterol, it was anticipated that only one component was necessary in order to estimate the true retention time profiles. After modelling the peak by PARAFAC2, it seems that three components are more appropriate than one, which is supported by the core consistency values presented in Table 5.

Table 5. Overview of the PARAFAC2 models of cholesterol. Four models have been calculated with one to four components, and from the core consistency values three components seem appropriate

Model	# of components	Explained variance (%)	Core consistency
PARAFAC2	1	87.00	100
	2	96.69	100
	3	99.26	98
	4	99.47	<0

The estimated retention time profiles of the three-component PARAFAC2 model are shown in Figure 32. The profiles look reasonable and all of the components seem to describe the biological variation in the peak.

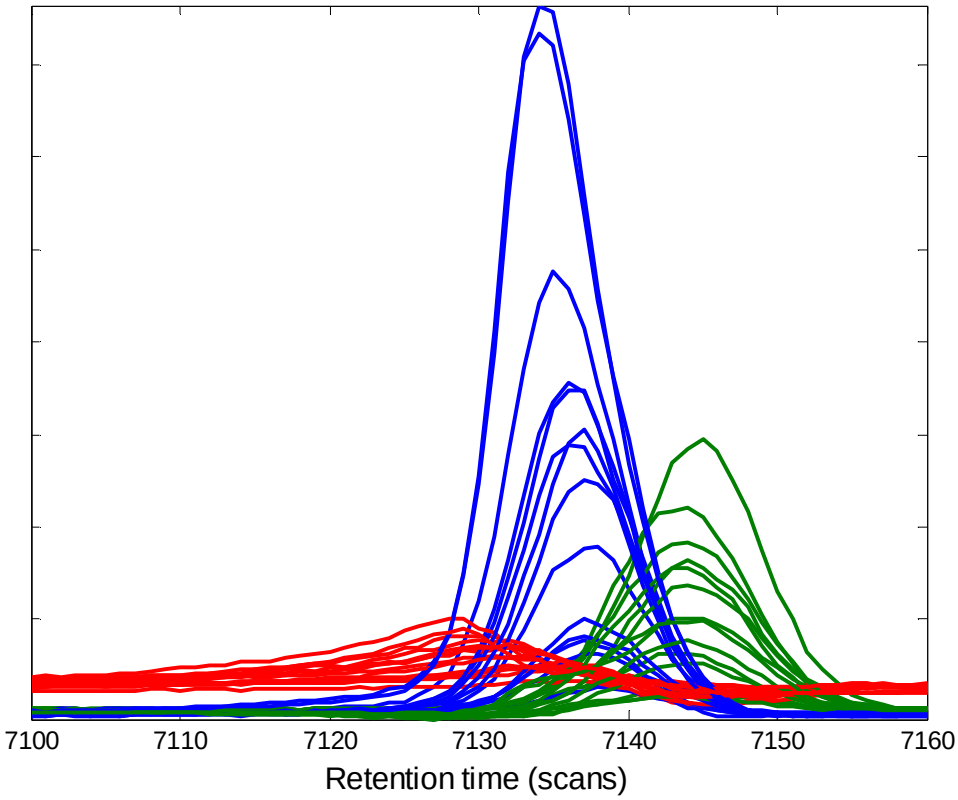


Figure 32. Estimated PARAFAC2 loadings of the retention time for cholesterol. Three components are used to model the peak

The biological interpretation of the three-component model will not be considered in detail, but the result may not be surprising since cholesterol is present in different forms in blood. Therefore, more than one component makes immediate sense biologically. The result suggests that PARAFAC2 can be used to extract the biological information from the chromatogram, and this information could form the basis of a classification model to separate cancer from controls – which was the initial purpose of the feasibility study.

5.3.3. Conclusion and perspectives

The outcome of this feasibility study shows that it is of the utmost importance to carefully prepare samples prior to data acquisition by GC-MS, and also to find the proper settings and equipment for the chromatograph and mass detector. Therefore, more thorough studies focusing of both the sample preparation step and the GC-MS should be carried out, preferably on a larger volume of sample material. However, there is considerable potential for using the GC-MS in a metabolomics context, and the perspective for this feasibility

study is especially the possibility of data fusion with measurements from fluorescence which was the original aim. Furthermore, data fusion with measurements from the NMR study on healthy individuals will most likely aid in better understanding the etiology of cancer at early stages.

Chapter 6

Conclusion

The work presented in this thesis is a clear indicator of the usefulness of multivariate models in different metabolomics studies. With three case studies, the use and importance of properly validated models have been demonstrated.

One of the main conclusions is the need for proper variable reduction in metabolomics datasets. Variable reduction, such as RRV, will exclude redundant variables and the risk of spurious groupings in a classification or prediction model is greatly reduced. Additionally, it has been shown that suitable validation of especially classification models is of the utmost importance. Models such as PLS-DA and OPLS-DA are prone to over-fit and the lack of appropriate validation will most likely result in separation of samples by chance – even if RRV has been performed in advance of modelling.

Proper use of multivariate modelling has been shown in three case studies. The first case study (Paper II) was a unique example of the development of a prediction model capable of forecasting future breast cancer status in healthy plasma samples from 838 women. The model was based on data from NMR concatenated with known risk markers from questionnaire data. The results from the study also showed that even though a model is based on carefully selected variables, these variables might not be the sole solution – other variables could have been used in the modelling with the same predictive power as the originally selected ones. The prediction model obtained consisted of eight PLS components and it is imperative to understand that all eight components must be assessed in a combined manner in order to understand the connection to the prediction of cancer status. Combining the components and the selected variables was introduced as being a bio-contour, which is a term introduced to avoid the temptation of picking out single variables positively correlated to cancer status in one of the eight components.

The second case study concerned the detection of colorectal cancer by fluorescence spectroscopy (Paper III). Here, plasma samples from patients with verified colorectal cancer were separated from three different types of

control samples with sensitivity and specificity values of approximately 0.75. The results open for a new avenue with fluorescence spectroscopy as a potential tool for detection of colorectal cancer.

The last case study (unpublished results) was a feasibility study where GC-MS was used to analyse standard plasma samples. The original aim was also to analyse the same plasma samples as those in the fluorescence study, but due to issues concerning sample preparation, only 14 standard samples were analysed. One peak was extracted and modelled by a three-component PARAFAC2 model. The core consistency (presented for PARAFAC2 models in Paper I) supported the choice of three components. The retention time profiles were estimated and the result points towards the applicability of PARAFAC2 modelling in GC-MS-based metabolomics.

Overall, the work presented throughout this thesis has proven the immense need for multivariate data analysis in metabolomics. If done properly, solid and reliable models can be built making room for increased understanding of the human metabolism. The examples have focussed on early prediction and detection of cancer, but the methods are applicable for all areas of metabolomics.

Chapter 7

Perspectives

The work and the conclusions presented in this thesis show that there are several perspectives for future work. As already mentioned, a future project has been planned for GC-MS analyses on the data from the DCH cohort analysed by NMR and from the multi-centre cross sectional study analysed by fluorescence. Additionally, data acquisition by LC-MS could also be of considerable interest. Acquisition of these data provides the foundation of multivariate data analysis based on data fusion, which will inevitably increase the understanding of the underlying etiology of cancer. Focus in this thesis has been on breast cancer and colorectal cancer, but it would be of great interest to extend the research to cover other types of cancer and other diseases in general.

The long-term perspective of this project is that cancer screening can be performed using a blood sample which will spare the patient (unpleasant) physical examinations. It will also be possible to detect a possible progress of cancer at an early stage or even anticipate the risk of cancer before the tumour is clinically detectable, which will most likely increase the five-year survival rates for different types of cancer. The current sample sets are collected from Danish individuals only, but a global perspective could be the collection of data from different parts of the world to establish a valid global prediction model. A starting point could be some of the other sample sets from the EPIC (see Chapter 5) cohort from other European countries.

The possibility of screening a blood sample for the risk of developing cancer leads to questions concerning the ethical aspects. Does the individual want to know whether there is a risk of developing cancer later in life? Will it affect the well-being and life quality of the individual if being told that there is a risk of developing cancer, which is one of the current issues in mammography screening? And finally, what about the individuals who were screened and told that the risk of developing cancer later in life was minimal, and then develops cancer? Of course questions like these will always arise when dealing with diseases where the worst case scenario is death.

On the other hand, the possibility of early screening creates room for the unique possibility of detecting the progress of cancer at a very early stage and before the cancer starts spreading to other vital parts of the body. But before this can become some sort of reality, it is essential that the prediction models developed for the purpose of detection or forecasting cancer are extremely reliable. I believe that this thesis is a step towards such a realization.

Reference list

1. Madsen R, Lundstedt T, Trygg J. Chemometrics in metabolomics – A review in human disease diagnosis. *Analytica Chimica Acta* 2010; **659**: 23-33.

2. Smolinska A, Blanchet L, Buydens LMC, Wijmenga SS. NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. *Analytica Chimica Acta* 2012; **750**: 82-97.

3. Sreekumar A, Poisson LM, Rajendiran TM *et al.* Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* 2009; **457**: 910-914.

4. Web-adress. *http://www.cancer.dk/Hjaelp+viden/kraeftformer/kraeftsygdomme/brystkraeft/statistik+brystkraeft/, accessed April* 2013.

5. Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. *The Lancet* 2013; **380**: 1778-1786.

6. Web-adress. *http://www.cancer.dk/Hjaelp+viden/kraeftformer/kraeftsygdomme/endetarmen/statistik+endetarmskraeft/, accessed April* 2013.

7. Web-adress. *http://www.cancer.dk/forebyg/screening/Screening+tarmkraeft/fordele+og+ulemper/, accessed April* 2013.

8. Fiehn O. Combining Genomics, Metabolome Analysis, and Biochemical Modelling to Understand Metabolic Networks. *Comparative and Functional Genomics* 2001; **2**: 155-168.

9. Nicholson JK, Lindon JC, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 1999; **29**: 1181-1189.

10. Oliver SG, Winson MK, Kell DB, Baganz F. Systematic functional analysis of the yeast genome. *Trends in Biotechnology* 1998; **16**: 373-378.

11. Fiehn O. Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology* 2002; **48**: 155-171.

12. Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in Biotechnology* 2004; **22**: 245-252.

13. Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L. Metabolite profiling for plant functional genomics. *Nat Biotech* 2000; **18**: 1157-1161.

14. Holmes E, Nicholls AW, Lindon JC *et al.* Chemometric Models for Toxicity Classification Based on NMR Spectra of Biofluids. *Chem.Res.Toxicol.* 2000; **13**: 471-478.

15. Nørgaard L, Sölétormos G, Harrit N *et al.* Fluorescence spectroscopy and chemometrics for classification of breast cancer samples – a feasibility study using extended canonical variates analysis. *J.Chemometrics* 2007; **21**: 451-458.

16. Asiago VM, Alvarado LZ, Shanaiah N *et al.* Early Detection of Recurrent Breast Cancer Using Metabolite Profiling. *Cancer Research* 2010; **70**: 8309-8318.

17. Chan ECY, Koh PK, Mal M *et al.* Metabolic Profiling of Human Colorectal Cancer Using High-Resolution Magic Angle Spinning Nuclear Magnetic Resonance (HR-MAS NMR) Spectroscopy and Gas Chromatography Mass Spectrometry (GC/MS). *J.Proteome Res.* 2008; **8**: 352-361.

18. Denkert C, Budczies J, Kind T *et al.* Mass Spectrometry-Based Metabolic Profiling Reveals Different Metabolite Patterns in Invasive Ovarian Carcinomas and Ovarian Borderline Tumors. *Cancer Research* 2006; **66**: 10795-10804.

19. Giskeødegård GF, Grinde MT, Sitter B *et al.* Multivariate Modeling and Prediction of Breast Cancer Prognostic Factors Using MR Metabolomics. *J.Proteome Res.* 2009; **9**: 972-979.

20. Issaq HJ, Nativ O, Waybright T *et al.* Detection of Bladder Cancer in Human Urine by Metabolomic Profiling Using High Performance Liquid Chromatography/Mass Spectrometry. *The Journal of Urology* 2008; **179**: 2422-2426.

21. Thysell E, Surowiec I, Hörnberg E *et al.* Metabolomic Characterization of Human Prostate Cancer Bone Metastases Reveals Increased Levels of Cholesterol. *PLoS ONE* 2010; **5**: 1-10.

22. Wu H, Liu T, Ma C *et al.* GC/MS-based metabolomic approach to validate the role of urinary sarcosine and target biomarkers for human prostate cancer by microwave-assisted derivatization. *Anal Bioanal Chem* 2011; **401**: 635-646.

23. Tiziani S, Lopes V, Günther UL. Early Stage Diagnosis of Oral Cancer Using 1H NMR-Based Metabolomics. *Neoplasia* 2009; **11**: 269-276.

24. van der Greef J, Stroobant P, Heijden Rvd. The role of analytical sciences in medical systems biology. *Current Opinion in Chemical Biology* 2004; **8**: 559-565.

25. Breit G, Rabi II. Measurement of Nuclear Spin. *Physical Review* 1931; **38**: 2082-2083.

26. Lenz EM, Wilson ID. Analytical Strategies in Metabonomics. *J.Proteome Res.* 2006; **6**: 443-458.

27. Beckonert O, Keun HC, Ebbels TMD *et al*. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat.Protocols* 2007; **2**: 2692-2703.

28. Claridge TDW. *High-Resolution NMR Techniques in Organic Chemistry*, 2. Edition, Oxford: Elsevier, 2013.

29. Savorani F, Tomasi G, Engelsen SB. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance* 2010; **202**: 190-202.

30. Craig A, Cloareo O, Holmes E, Nicholson JK, Lindon JC. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal.Chem.* 2006; **78**: 2262-2267.

31. Rasmussen LG, Savorani F, Larsen TM, Dragsted LO, Astrup A, Engelsen SB. Standardization of factors that influence human urine metabolomics. *Metabolomics* 2011; **7**: 71-83.

32. Herschel JFW. On a Case of Superficial Colour Presented by a Homogeneous Liquid Internally Colourless. *Philosophical Transactions of the Royal Society of London* 1845; **135**: 143-145.

33. Lakowicz JR. Introduction to Fluorescence.In: *Principles of Fluorescence Spectroscopy*, 3. Edition, New York: Springer, 2006: 1-25.

34. Wolfbeis OS, Leiner M. Mapping of the total fluorescence of human blood serum as a new method for its characterization. *Analytica Chimica Acta* 1985; **167**: 203-215.

35. Lakowicz JR. Instrumentation for Fluorescence Spectroscopy.In: *Principles of Fluorescence Spectroscopy*, 3. Edition, New York: Springer, 2006: 27-60.

36. Lawaetz AJ, Stedmon CA. Fluorescence Intensity Calibration Using the Raman Scatter Peak of Water. *Applied Spectroscopy* 2009; **63**: 936-940.

37. JAMES AT, MARTIN AJP. Gas-liquid partition chromatography; the separation and micro-estimation of volatile fatty acids from formic acid to dodecanoic acid. *Biochem.J.* 1952; **50**: 679-690.

38. Jellum E, Stokke O, Eldjarn L. Application of gas chromatography, mass spectrometry, and computer methods in clinical biochemistry. *Anal.Chem.* 1973; **45**: 1099-1106.

39. Hites RA. Gas Chromatography Mass Spectrometry.In: *Handbook of Instrumental Techniques for Analytical Chemistry*, 1. Edition, ed Settle F, Virginia: Prentice Hall, 1997: 609-626.

40. Kopka J. Gas Chromatography Mass Spectrometry.In: *Biotechnology in Agriculture and Forestry 57 - Plant Metabolomics* eds Saito K, Dixon RA, Willmitzer L, Berlin: Springer-Verlag, 2006: 3-20.

41. Want EJ, Nordstrom A, Morita H, Siuzdak G. From exogenous to endogenous: The inevitable imprint of mass spectrometry in metabolomics. *J.Proteome Res.* 2007; **6**: 459-468.

42. Hendriks MMWB, Eeuwijk FA, Jellema RH *et al.* Data-processing strategies for metabolomics studies. *TrAC Trends in Analytical Chemistry* 2011; **30**: 1685-1698.

43. Jukarainen N. NMR Metabolomics Techniques and Mathematical Tools as an Aid in Neurological Diagnosis. *Thesis* 2009; 1-157.

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

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

46. Wold S, Esbensen K, Geladi P. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* 1987; **2**: 37-52.

47. Bro R, Smilde AK. Centering and scaling in component analysis. *J.Chemometrics* 2003; **17**: 16-33.

48. Qiu YP, Zhou BS, Su MM *et al.* Mass Spectrometry-Based Quantitative Metabolomics Revealed a Distinct Lipid Profile in Breast Cancer Patients. *International Journal of Molecular Sciences* 2013; **14**: 8047-8061.

49. Tauler R, Smilde A, Kowalski B. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J.Chemometrics* 1995; **9**: 31-58.

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

51. Harshman RA. Foundations of the PARAFAC procedure: Models and Conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics* 1970; **16**: 1-84.

52. Bro R, Kiers HAL. A new efficient method for determining the number of components in PARAFAC models. *J.Chemometrics* 2003; **17**: 274-286.

53. Bro R. Multi-way analysis in the food industry, models, algorithms, and applications. *Thesis* 1998.

54. Amigo JM, Skov T, Bro R, Coello J, Maspoch S. Solving GC-MS problems with PARAFAC2. *TrAC Trends in Analytical Chemistry* 2008; **27**: 714-725.

55. Harshman RA. PARAFAC2: Mathematical and Technical Notes. *UCLA Working Papers in Phonetics* 1972; **22**: 30-44.

56. Kiers HAL, Ten Berge JMF, Bro R. PARAFAC2 - Part I. A direct fitting algorithm for the PARAFAC2 model. *J.Chemometrics* 1999; **13**: 275-294.

57. Bro R, Andersson CA, Kiers HAL. PARAFAC2 - Part II. Modeling chromatographic data with retention time shifts. *J.Chemometrics* 1999; **13**: 295-309.

58. Bro R, Kiers HAL. A new efficient method for determining the number of components in PARAFAC models. *J.Chemometrics* 2003; **17**: 274-286.

59. Barker M, Rayens W. Partial least squares for discrimination. *J.Chemometrics* 2003; **17**: 166-173.

60. Westerhuis J, Hoefsloot H, Smit S *et al*. Assessment of PLSDA cross validation. *Metabolomics* 2008; **4**: 81-89.

61. Bro R. *Håndbog i multivariabel kalibrering*, 1. Edition, Frederiksberg: Jordbrugsforlaget, 1996: 53-69.

62. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. *Analytica Chimica Acta* 1986; **185**: 1-17.

63. Wold S, Martens H, Wold H. The multivariate calibration problem in chemistry solved by the PLS method.In: *Matrix Pencils*, 973 Edition, eds Kågström B, Ruhe A, Springer Berlin Heidelberg, 1983: 286-293.

64. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* 2001; **58**: 109-130.

65. Fisher RA. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 1936; **7**: 179-188.

66. Ballabio D, Todeschini R. Chapter 4 - Multivariate Classification for Qualitative Analysis.In: *Infrared Spectroscopy for Food Quality Analysis and Control* San Diego: Academic Press, 2009: 83-104.

67. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *J.Chemometrics* 2002; **16**: 119-128.

68. Tapp HS, Kemsley EK. Notes on the practical utility of OPLS. *TrAC Trends in Analytical Chemistry* 2009; **28**: 1322-1327.

69. Wagner S, Scholz K, Sieber M, Kellert M, Voelkel W. Tools in metabonomics: An integrated validation approach for LC-MS metabolic profiling of mercapturic acids in human urine. *Anal.Chem.* 2007; **79**: 2918-2926.

70. Bellman R. *Dynamic Programming* Princeton, New Jersey, 1957.

71. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society.Series B (Methodological)* 1974; **36**: 111-147.

72. Baumann K. Cross-validation as the objective function for variable-selection techniques. *TrAC Trends in Analytical Chemistry* 2003; **22**: 395-406.

73. Xu QS, Liang YZ, Du YP. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *J.Chemometrics* 2004; **18**: 112-120.

74. Kim Y, Park YJ, Yang SO *et al.* Hypoxanthine levels in human urine serve as a screening indicator for the plasma total cholesterol and low-density lipoprotein modulation activities of fermented red pepper paste. *Nutrition Research* 2010; **30**: 455-461.

75. Altman DG, Bland JM. Diagnostic tests 1: sensitivity and specificity. *BMJ* 1994; **308**: 1552.

76. Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* 1978; **8**: 283-298.

77. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* 1993; **39**: 561-577.

78. Sharma D, Yadav UB, Sharma P. The Concept of Sensitivity and Specificity in Relation to Two Types of Errors and its Application in Medical Research. *Journal of Reliability and Statistical Studies* 2009; **2**: 53-58.

79. Lutz U, Lutz RW, Lutz WK. Metabolic Profiling of Glucuronides in Human Urine by LCΓêÆMS/MS and Partial Least-Squares Discriminant Analysis for Classification and Prediction of Gender. *Anal.Chem.* 2006; **78**: 4564-4571.

80. Wiklund S, Johansson E, Sjostrom L *et al.* Visualization of GC/TOF-MS-Based Metabolomics Data for Identification of Biochemically Interesting Compounds Using OPLS Class Models. *Anal.Chem.* 2007; **80**: 115-122.

81. Westerhuis J, Velzen E, Hoefsloot H, Smilde A. Discriminant Q2 (DQ2) for improved discrimination in PLSDA models. *Metabolomics* 2008; **4**: 293-296.

82. Szymanska E, Saccenti E, Smilde A, Westerhuis J. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics* 2012; **8**: 3-16.

83. Eriksson L, Johansson E, Kettaneh-Wold N, Trygg J, Wikström C, Wold S. *Multi- and Megavariate Data Analysis. Part I: Basic Principles and Applications*, Second Edition, Umetrics Academy, 2001.

84. Centner V, Massart D-L, de Noord OE, de Jong S, Vandeginste BM, Sterna C. Elimination of Uninformative Variables for Multivariate Calibration. *Anal.Chem.* 1996; **68**: 3851-3858.

85. Rajalahti T, Arneberg R, Berven FS, Myhr KM, Ulvik RJ, Kvalheim OM. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemometrics and Intelligent Laboratory Systems* 2009; **95**: 35-48.

86. Rajalahti T, Arneberg R, Kroksveen AC, Berle M, Myhr KM, Kvalheim OM. Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles. *Anal.Chem.* 2009; **81**: 2581-2590.

87. Nørgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB. Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. *Applied Spectroscopy* 2000; **54**: 413-419.

88. Acar E, Plopper GE, Yener B. Coupled Analysis of In Vitro and Histology Tissue Samples to Quantify Structure-Function Relationship. *PLoS ONE* 2012; **7**: 1-14.

89. Van Deun K, Smilde A, van der Werf M, Kiers H, Van Mechelen I. A structured overview of simultaneous component based data integration. *BMC Bioinformatics* 2009; **10**: 246.

90. Westerhuis JA, Kourti T, MacGregor JF. Analysis of multiblock and hierarchical PCA and PLS models. *J.Chemometrics* 1998; **12**: 301-321.

91. Smilde AK, van der Werf MJ, Bijlsma S, van der Werff-van-der Vat B, Jellema RH. Fusion of mass spectrometry-based metabolomics data. *Anal.Chem.* 2005; **77**: 6729-6736.

92. Steinmetz V, Sévila F, Bellon-Maurel V. A Methodology for Sensor Fusion Design: Application to Fruit Quality Assessment. *Journal of Agricultural Engineering Research* 1999; **74**: 21-31.

93. Bro R, Nielsen H, Savorani F *et al.* Data fusion in metabolomic cancer diagnostics. *Metabolomics* 2013; **9**: 3-8.

94. Tjønneland A, Olsen A, Boll K *et al.* Study design, exposure variables, and socioeconomic determinants of participation in Diet, Cancer and Health: A population-based prospective cohort study of 57,053 men and women in Denmark. *Scandinavian Journal of Public Health* 2007; **35**: 432-441.

95. Web-adress. *http://epic.iarc.fr/, accessed April* 2013.

96. Nicholson JK, Foxall PJD, Spraul M, Farrant RD, Lindon JC. 750-Mhz H-1 and H-1-C-13 Nmr-Spectroscopy of Human Blood-Plasma. *Anal.Chem.* 1995; **67**: 793-811.

97. Chen C, Weiss NS, Newcomb P, Barlow W, White E. Hormone replacement therapy in relation to breast cancer. *JAMA* 2002; **287**: 734-741.

98. Tjønneland A, Christensen J, Thomsen BL *et al.* Hormone replacement therapy in relation to breast carcinoma incidence rate ratios. *Cancer* 2004; **100**: 2328-2337.

99. Nielsen HJ, Brünner N, Frederiksen C *et al.* Plasma tissue inhibitor of metalloproteinases-1 (TIMP-1): A novel biological marker in the detection of primary colorectal cancer. Protocol outlines of the Danish-Australian

endoscopy study group on colorectal cancer detection. *Scandinavian Journal of Gastroenterology* 2008; **43**: 242-248.

100. Masilamani V, Al-Zhrani K, Al-Salhi M, Al-Diab A, Al-Ageily M. Cancer diagnosis by autofluorescence of blood components. *Journal of Luminescence* 2004; **109**: 143-154.

101. Fancy S-A, Rumpel K. GC-MS-Based Metabolomics.In: *Methods in Pharmacology and Toxicology: Biomarker Methods in Drug Discovery and Development* ed Wang F, Totowa, NJ: Humana Press Inc., 2008: 317-340.

102. Fiehn O, Kind T. Metabolite Profiling in Blood Plasma.In: *Methods in Molecular Biology, vol. 358: Metabolomics: Methods and Protocols* ed Weckwerth W, Totowa, NJ: Humana Press Inc., 2007: 3-17.

103. A J, Trygg J, Gullberg J *et al.* Extraction and GC/MS Analysis of the Human Blood Plasma Metabolome. *Anal.Chem.* 2005; **77**: 8086-8094.

104. Web adress. *http://www.sisweb.com/software/ms/wiley.htm, accessed July* 2013.

105. Web adress. *http://webbook.nist.gov/chemistry/, accessed in July* 2013.

Paper I

Core consistency diagnostic in PARAFAC2

Reprint from *Journal of Chemometrics*, 27(5), 99-105

Research Article

JOURNAL OF CHEMOMETRICS

Received: 17 September 2012, Revised: 14 February 2013, Accepted: 25 February 2013, Published online in Wiley Online Library: 26 March 2013

(wileyonlinelibrary.com) DOI: 10.1002/cem.2497

Core consistency diagnostic in PARAFAC2

Maja H. Kamstrup-Nielsen[a], Lea G. Johnsen[a,b] and Rasmus Bro[a]*

PARAFAC2 is applied in multiple research areas, for example, where data containing shifts are analysed, but it is a challenge to determine the appropriate number of components in the model. In this paper, it is hypothesized that the core consistency diagnostic, which is currently applied in, for example, PARAFAC1 can be used to determine model complexity in PARAFAC2. Theoretically, a PARAFAC1 model is fitted 'inside' the PARAFAC2 algorithm, and it should therefore be possible to apply the core consistency diagnostic from PARAFAC1 in PARAFAC2. To support this hypothesis, three different datasets, as well as simulated datasets, have been evaluated by means of PARAFAC2, and the core consistencies have been investigated. There is a general trend that if the core consistency is low, the model is overfitted as in PARAFAC1. Also, core consistency captures the true variation in the data, whereas small peaks are easily overlooked by visual inspection of noisy models. However, for determining the number of components in a PARAFAC2 model, we suggest usage of the core consistency in combination with other model parameters such as residuals, loadings, and split-half analysis. Copyright © 2013 John Wiley & Sons, Ltd.
Supporting information may be found in the online version of this paper

Keywords: core consistency; PARAFAC2; number of components; model complexity

1. INTRODUCTION

PARAFAC2 [1,2] has been applied in many different areas [3–5] and has for example proven to be useful for mathematical separation of overlapping chromatograms and for overcoming issues in batch data with different temporal duration and dynamics. The main reason for applying PARAFAC2 is that it can sometimes model data containing shifts and related shape changes, for example, chromatograms with shifts in retention time.

PARAFAC2 is closely related to PARAFAC. In this paper, the PARAFAC model will be denoted PARAFAC1 to distinguish it from PARAFAC2 [1]. PARAFAC1 decomposes three-way data with a low-rank trilinear structure into loading matrices that provide mostly unique estimates of the underlying variations in data. In PARAFAC2, data do not have to be low-rank trilinear—one of the directions in the data array can deviate in certain ways and still be meaningfully modelled by PARAFAC2 [2,6]. Despite the deviation from low-rank trilinearity, PARAFAC2 still provides unique estimates of the underlying latent variables under fairly mild conditions [7].

What remains a challenge in using the PARAFAC2 model is to determine the appropriate number of components. Harshman and De Sarbo [8] have proposed to use split-half analysis to determine the right number of factors. Split-half analysis can be considered as a type of resampling approach where PARAFAC2 is applied on different subsets of data. If the right number of factors is used, the result should be similar for all subsets. However, there are a number of drawbacks for split-half analysis. First of all, the subsets must be carefully selected. For instance, all compounds must be present in all subsets for the resulting models to be similar. Another inconvenience is that the computation time increases when using split-half analysis.

For the PARAFAC1 model, the core consistency diagnostic is useful when determining the number of components. The core consistency diagnostic has been described by Bro and Kiers [9]. So far, research has dealt with the determination of model complexity in PARAFAC1 by means of core consistency, but the core consistency has never been incorporated in a PARAFAC2 setting, and no similar alternative approaches to determination of model complexity have been suggested.

The objective of this paper is to develop an approach for calculating a model diagnostic similar to core consistency but for PARAFAC2 models. We will show that with some manipulations, we can define a core consistency value for a PARAFAC2 model. We will also investigate if this diagnostic can be applied to determine the number of components in PARAFAC2 models. First, the theory behind the structure of PARAFAC1 and PARAFAC2 will be outlined. Second, the theory behind and the relevance of the core consistency will be presented. Three examples on different datasets are given where core consistency is used to evaluate the model complexity. In addition, the use of core consistency in PARAFAC2 is validated using simulated data.

2. THEORY

PARAFAC1 is a multiway method used to handle three-way (or multiway in general) arrays, and the principle is outlined, for example, by Harshman [10] and Bro [11].

Let $\mathbf{X}_k$ be an $I \times J$ matrix with $k = 1, \ldots, K$ as the kth slab of an $I \times J \times K$ three-way array $\underline{\mathbf{X}}$. I is the number of observations (samples) in the first mode, J the number of variables in the second mode, and K the number of variables in the third mode [2]. With this terminology and noise disregarded for simplicity, the PARAFAC1 model has the following structure

* *Correspondence to: Rasmus Bro, Department of Food Science, University of Copenhagen, Copenhagen, Denmark.*
E-mail: rb@life.ku.dk

a *M. H. Kamstrup-Nielsen, L. G. Johnsen, R. Bro*
Department of Food Science, University of Copenhagen, Copenhagen Denmark

b *L. G. Johnsen*
Chr. Hansen A/S, Hørsholm Denmark

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^T, k = 1, \ldots, K \quad (1)$$

Here, **A** typically denotes the score matrix, and **B** is the loading matrix for the second mode, which can be considered to correspond to the loading matrix in principal component analysis (PCA). The extension from PCA then lies in the $\mathbf{D}_k$ matrix, which is a diagonal matrix of dimension $R \times R$, where R is the number of components. This matrix contains parameters from the loadings from the third mode. The loading matrix of the third mode is usually termed **C** ($K \times R$) and $\mathbf{D}_k$ holds the kth row of **C** on its diagonal. In a multiway data analysis, the component matrices **A**, **B**, and **C** are oftentimes all called loading matrices. The term score matrix can then be introduced specifically for the loadings in the sample mode.

Unlike a bilinear model, PARAFAC1 provides unique estimates of its parameters **A**, **B**, and $\mathbf{D}_1, \ldots, \mathbf{D}_K$ under certain conditions without additional abstract constraints such as orthogonality, which is used in PCA. The bilinear representation $\mathbf{A}\mathbf{B}^T$ has rotational freedom, and PCA is only uniquely identified because of the additional constraints that are imposed on the parameters.

To set the stage for PARAFAC2, the PARAFAC1 model is illustrated in the following by means of a small part of gas chromatography (GC)–mass spectrometry (MS) chromatographic data from Amigo *et al.* [3]. Instead of having samples in the first mode, as is common, these will be in the third mode for convenience of introducing PARAFAC2 subsequently. K chromatographic samples with I mass channels and J retention times have been modelled using a PARAFAC1 model. In this example, there is only one analyte present in the K samples, which is illustrated as the single peak in the second mode in Figure 1. However, in the kth sample, the retention time for this analyte is different from that of the first sample, which can also be seen in the second mode (Jth direction) in the figure. The second mode loading matrix, **B**, for a two-component PARAFAC1 model is supposed to contain estimates of the retention time profiles. Two components seem appropriate for this model, as each component in the second mode estimates the retention time profile in each sample. Hence, two components are necessary to extract the shifting information of the samples and thereby reveal the retention times of the analyte.

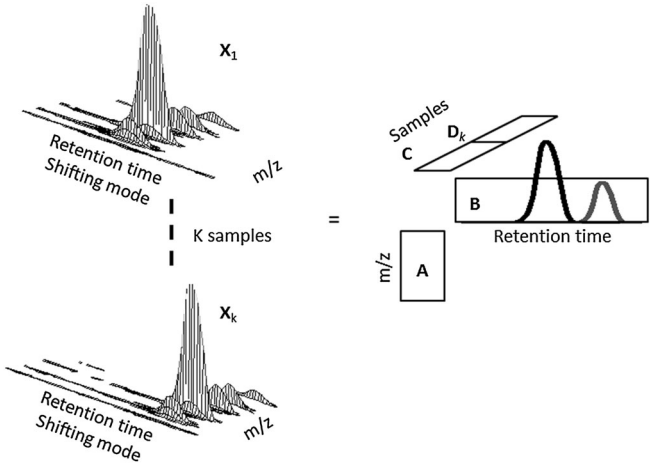


Figure 1. Chromatographic example to illustrate PARAFAC1. The data matrix $\mathbf{X}_k$ is decomposed into estimates of the parameters **A**, **B**, and $\mathbf{D}_k$ using two components.

In PARAFAC1, it is assumed that the loading matrix **B** is representative of the underlying variation in all frontal slabs, that is, that all slabs, $\mathbf{X}_k$, can be described in the row space using the same **B** ($\mathbf{A}\mathbf{D}_k\mathbf{B}^T$). This means that for chromatographic data, the underlying retention time profiles of each analyte have to have identical shapes for each sample. This is not the case in the present example, where the samples as mentioned have shifting retention times, as illustrated in the second mode in the figure. Using PARAFAC1 on such data will typically lead to including more components than underlying chemical variations as seen in the example. These subsequent components can be difficult or impossible to interpret. Using the PARAFAC2 model is one way to circumvent such problems. The PARAFAC2 model can be written as

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}_k^T, k = 1, \ldots, K \quad (2)$$

The parameters are almost identical to those of PARAFAC1. The only difference between Equations (1) and (2) is the second mode loading matrix **B**. In PARAFAC2, $\mathbf{B}_k$ is specific for every slab, k, in the third mode, whereas **B** is equal for all slabs in PARAFAC1. Note that residuals are not included in Equation (2) for simplicity.

In Figure 2, it is illustrated how the use of a sample-specific $\mathbf{B}_k$ matrix can help provide a more meaningful model of the shifting chromatographic data in Figure 1.

In a PARAFAC2 model of these data, each sample will have its own retention time loading matrix $\mathbf{B}_k$, and a one-component model is then sufficient to estimate the underlying retention time profile for the analyte present in the two samples regardless of the shift in retention time. All the information concerning the shift is extracted by this component, and the shift is modelled by the K different **B** loadings.

However, the parameter estimates in PARAFAC2 would not immediately be unique if the model was only defined through Equation (2). An additional constraint is also part of the model. The cross-product of $\mathbf{B}_k$ ($\mathbf{B}_k^T\mathbf{B}_k$) is required to be constant across k, and it can be shown that this constraint leads to the uniqueness of the model under mild conditions [7]. Constant cross-product across k is obtained by defining $\mathbf{B}_k$ as

$$\mathbf{B}_k = \mathbf{P}_k\mathbf{H}, k = 1, \ldots, K \quad (3)$$

where $\mathbf{P}_k^T\mathbf{P}_k = \mathbf{I}$; hence, $\mathbf{P}_k$ is orthogonal. The matrix $\mathbf{P}_k$ handles what is unique for each sample in the shifting mode, and **H**

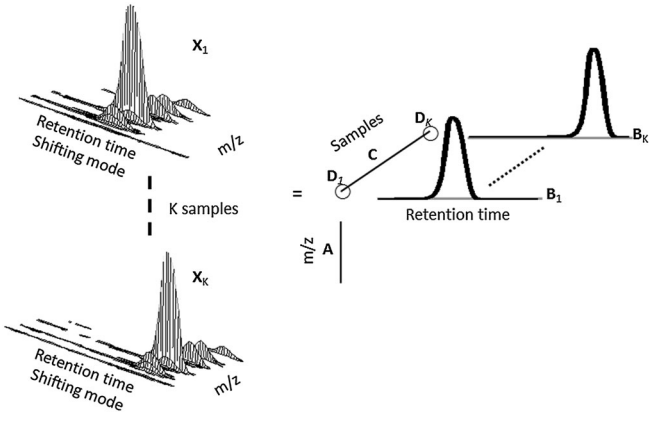


Figure 2. Same chromatographic example as illustrated in Figure 1. Here, a PARAFAC2 model is fitted to data. Only one component is necessary.

handles what is related between samples [12]. With this definition, the cross-product for $\mathbf{B}_k$ will be constant because with an orthogonal $\mathbf{P}_k$ it holds that

$$\mathbf{B}_k^T\mathbf{B}_k = \mathbf{H}^T\mathbf{P}_k^T\mathbf{P}_k\mathbf{H} = \mathbf{H}^T\mathbf{H} \tag{4}$$

If we substitute $\mathbf{B}_k$ in Equation (2) with Equation (3), we can rearrange the PARAFAC2 model in the following way:

$$\begin{aligned} \mathbf{X}_k &= \mathbf{A}\mathbf{D}_k(\mathbf{P}_k\mathbf{H})^T \quad &\Leftrightarrow \\ \mathbf{X}_k\mathbf{P}_k &= \mathbf{A}\mathbf{D}_k\mathbf{H}^T\mathbf{P}_k^T\mathbf{P}_k \quad &\Leftrightarrow \\ \mathbf{Y}_k &= \mathbf{A}\mathbf{D}_k\mathbf{H}^T, k = 1, \ldots, K \end{aligned} \tag{5}$$

Equation (5) points to an interesting approach for understanding PARAFAC2. When the orthogonal $\mathbf{P}_k$ matrices are known, we can rephrase the PARAFAC2 model as a PARAFAC1 model in terms of frontal slabs of data 'compressed' with their own specific $\mathbf{P}_k$ matrix; hence, a PARAFAC1 model can be fitted on a data array of $\mathbf{Y}_k$ slabs. This is interesting in understanding how PARAFAC2 handles changes such as retention time shifts in the second mode, and it is also useful for the purpose of developing a core consistency measure for PARAFAC2 models in this paper.

The number of components to use in a PARAFAC1 model can be estimated by means of the core consistency diagnostic [9]. PARAFAC1 can be considered as a constrained Tucker3 model [13] but where the core array has been fixed to a superdiagonal array of ones. The idea behind the core consistency diagnostic is to estimate what the core would actually have been if it was not constrained. This is estimated using the PARAFAC1 loadings as fixed loadings in a Tucker3 model, hence only estimating the core array. If this estimated core array is close to a superdiagonal of 1s, we say that the core consistency is high and that the variation described by the PARAFAC1 model is indeed low-rank trilinear. If the core is very different, for example, has high off-diagonal elements, then the core consistency is low, and this indicates that the PARAFAC1 model, which presumably should be modelling low-rank trilinear variation, is really modelling other things as well. This indicates that this particular model is not suitable.

As mentioned previously, the PARAFAC2 model can be considered a PARAFAC1 model on 'de-shifted' data with slabs $\mathbf{Y}_k$. We hypothesize that the number of components can be equally well assessed from this PARAFAC1 model and that we can therefore use the straightforward core consistency of the PARAFAC1 model 'inside' PARAFAC2 as a tool for determining model complexity. To investigate the hypothesis, obtained core consistencies have been evaluated for the three different datasets and the simulated data.

3. MATERIALS AND METHODS

All models and calculations were performed in MATLAB 2012a (MathWorks, Inc., Natick, MA, USA). PARAFAC2 models were calculated with the algorithm from the *N*-way toolbox (available from www.models.life.ku.dk, July 2012).

3.1. Fluorescence amino acid data

The first dataset consists of five samples, each containing tyrosine, tryptophan, and phenylalanine in different amounts. Each sample has been measured on a PE LS50B spectrofluorometer (excitation 240–300 nm, emission 250–450 nm). The dimensions of the dataset are 5 (samples) × 201 (emission) × 61 (excitation).

3.2. Chromatographic wine and apple data

The second dataset consists of 36 apples ripened for 5, 8, and 15 days, and the samples are analysed using headspace GC–MS. The details concerning the analysis can be found in [14]. The dataset has the dimensions 154 (masses) × 5033 (retention times) 36 (samples).

The last dataset consists of 24 samples of red wine. The aroma profiles of the samples were measured using dynamic headspace GC–MS. Details concerning the measurements can be found in the original papers [3,15]. The dimensions of the dataset are 200 (masses) × 6000 (retention times) × 69 (samples).

4. RESULTS

The use of core consistency in PARAFAC2 has been tested using three different datasets: fluorescence amino acid data [11], chromatographic apple data [14], and finally chromatographic wine data [15]. In addition, the core consistency has been tested on simulated data.

4.1. Fluorescence

In this dataset, there are no shifts, so the result from PARAFAC2 should be similar to that of PARAFAC1. This enables us to compare the core consistencies obtained from the two methods.

4.1.1. Results and discussion

The models have been calculated without any constraints and with the samples in the last mode. The core consistencies and the explained variances of the models are seen in Table I.

The core consistencies for the models from the PARAFAC1 algorithm indicate that four factors are appropriate for the dataset. Because the data are obtained from simple samples only containing three different amino acids, it would be expected that three factors would be appropriate. A visual inspection of the model (Figure 3) shows that the emission and excitation profiles for the fourth factor have some negative values. In addition, the profile for this compound seems rather noisy also, indicating that

Table I. Overview of the core consistencies and the explained variances for PARAFAC1 and PARAFAC2 models with one to five factors in the fluorescence amino acid data without shifts

Model	No. of factors	Core consistency	% fit
PARAFAC1	1	100.00	64.39
	2	100.00	86.77
	3	99.87	99.94
	4	92.49	99.95
	5	<0.00	99.96
PARAFAC2	1	100.00	67.03
	2	100.00	92.94
	3	100.00	99.96
	4	<0.00	99.97
	5	<0.00	99.98

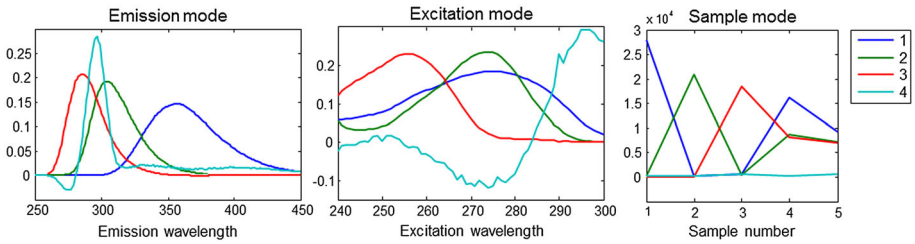


Figure 3. Illustration of the obtained PARAFAC1 model with four factors. Both the emission and excitation loadings for the fourth factor are rather noisy and have negative values, indicating that the model is overfitted.

this model is overfitted. It is likely that the fourth component is related to the small amount of Rayleigh scattering that is present in the data. In any case, for both the three-component and four-component models, the three main components come out similarly. The fourth extra component is of such a small magnitude that is does not affect the modelling of the three main components.

The core consistencies for the PARAFAC2 models indicate that three factors are appropriate for this dataset, and the visual appearance of this three-component model is also appropriate (not shown). Hence, core consistency seems to be useful for assessing the number of components for this dataset. The fact that normal PARAFAC1 and PARAFAC2 do not have the same behaviour with respect to the small and somewhat spurious fourth component is not surprising. The Rayleigh scattering that leads to the fourth PARAFAC1 component is not low-rank trilinear and hence is not expected to affect PARAFAC1 and PARAFAC2 models in a similar fashion.

Be aware that models that have not converged or have converged in a local minimum can result in a core consistency that is artificially low, and it is therefore very important to make sure that the model has converged and has reached the global minimum when core consistency is used in the evaluation of model quality. A simple ad hoc approach to this is to repeat the PARAFAC2 algorithm a number of times and make sure that the best-fitting model is obtained several times.

4.2. Chromatography

4.2.1. Results and discussion

The apple and wine datasets are very large and consist of several peak regions. Each dataset is divided into smaller parts, and PARAFAC2 models are fitted on these subsets individually. The apple data are divided manually into 26 intervals and the wine data into 50 intervals. The intervals chosen reflect a wide range of different features: overloaded peaks (e.g. wine interval 2), low signal-to-noise levels (e.g. wine intervals 31 and 32), minimal shifts in retention time (e.g. wine intervals 25 and 31), severe shifts in retention time (e.g. wine interval 42 and apple interval 3), and very complex intervals including several peaks (e.g. apple intervals 1 and 22). Intervals representing the different features are shown in Figure 4. To illustrate the features of core consistency, intervals 31 and 32 from the wine data and interval 1 from the apple data are illustrated in some detail in the following.

Core consistency was calculated for models with one to seven factors for all the 76 intervals, and all of the intervals were manually inspected to find the models with the optimal number of factors. Parts of the obtained core consistencies are shown in Figure 5. Models were evaluated on the basis of residual analysis, as well as inspection of elution profiles and spectra obtained from the models.

In agreement with the publication by Amigo *et al*. [14], we found that interval 1 in the apple dataset is best described with five factors (elution profiles not shown). As shown in Figure 5,

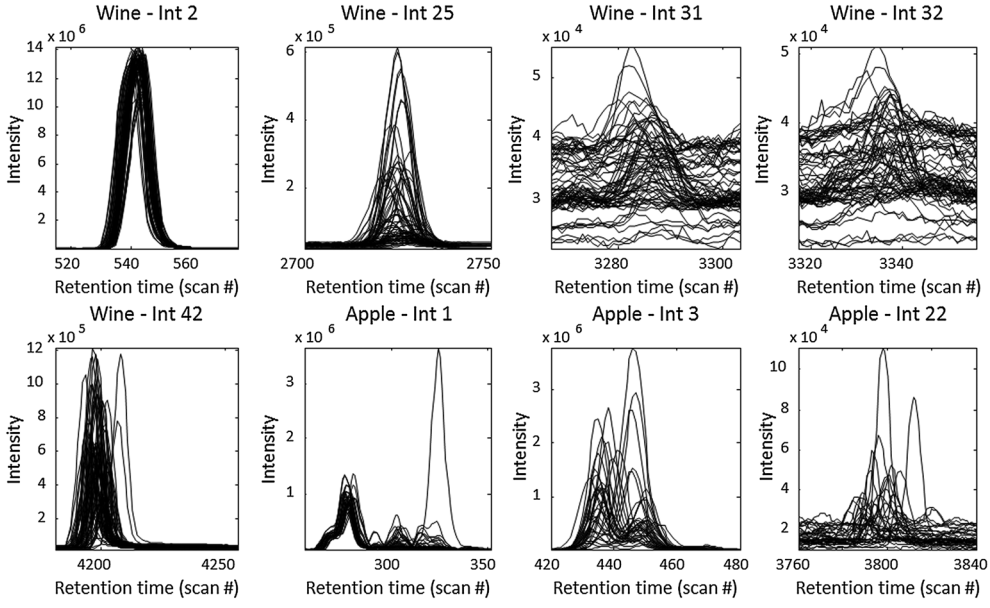


Figure 4. Examples showing a selection of the 76 intervals. The intervals cover peaks with both low and high signal-to-noise ratios, different degrees of shift, and different degrees of complexity.

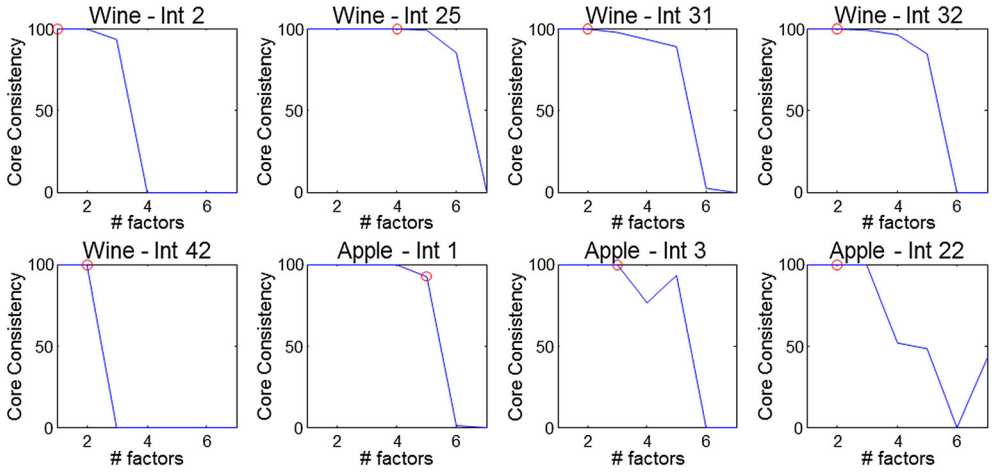


Figure 5. Examples showing a selection of the obtained core consistencies, the remaining can be found in the supplementary material. The circles indicate models with the optimal number of factors as initially decided by the authors. The line indicates the core consistency of each interval with the number of factors included in the model going from one to seven.

the core consistencies are high for the models with one to five factors and low for the models with six and seven factors. So, for this interval, it seems like the core consistency is a useful tool in the determination of the model complexity.

A manual inspection of the models calculated on interval 31 from the wine dataset suggests that a PARAFAC2 model with two factors is optimal (see elution profiles in Figure 6B). However, core consistency indicates that five factors are optimal even though the five-factor model is apparently overfitted (see elution profiles in Figure 6C). Please note that the component that does not describe a peak in the two-factor model does not indicate overfit but merely describes the baseline, which in this case is rather high compared with the height of the peak.

Figure 6D shows the estimated main peak from the two models illustrated in Figure 6B, C. Clearly, the two-factor and the five-factor models capture the same elution profile. The spectral profiles as well as the concentration profiles (plots not shown) support that it is the same chemical variation described by the two models. This tendency is also seen for other seemingly overfitted models. In the five-factor model describing interval 31, the three 'additional' components simply describe baseline. The last component seems to describe a small peak, which is only detected in the five-factor model. The same behaviour with high core consistency is observed in the models calculated on interval 32 from the wine data. The elution profiles from the models of this interval with one to six factors are shown in Figure 7.

The inspection of the seemingly overfitted models from intervals 31 and 32 with high core consistencies shows that in both models an additional factor actually appears, but it is very small and therefore difficult to locate (Figure 6C, arrow, and Figure 7, arrow). In these cases, it seems like the data contain noise and artefacts, which contribute more to the variation than the lastly described small peaks. The presence of these additional compounds is supported when the mass channels in the raw data are inspected (not shown). Nothing indicates that these peaks are not chemical compounds present in the samples, and therefore, it would be appropriate to use five factors in both intervals.

The results support that core consistency actually captures the true variation in the data, whereas a visual inspection might put too much emphasis on the noise. Thereby, small but potentially

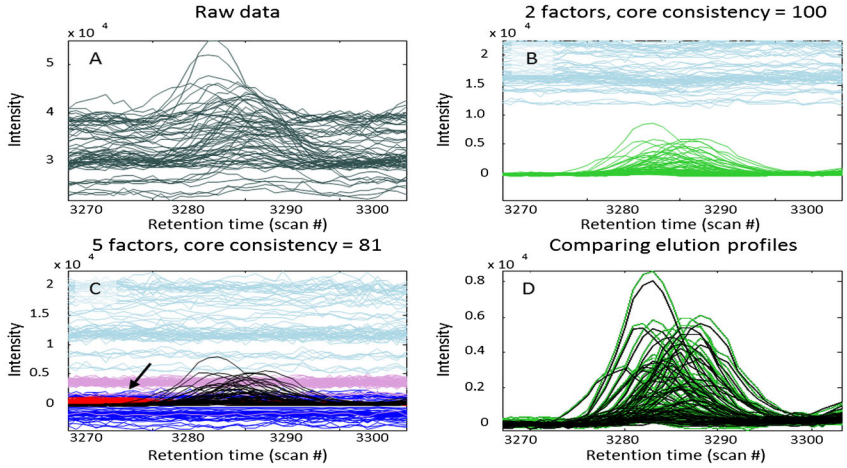


Figure 6. (A) Raw data from wine, interval 31. Elution profiles obtained with two-factor (B) and five-factor (C) PARAFAC2 models. The arrow indicates a compound that only appears in the five-factor model. (D) Illustration of the similarity between the estimated main peaks described by the models with two and five factors.

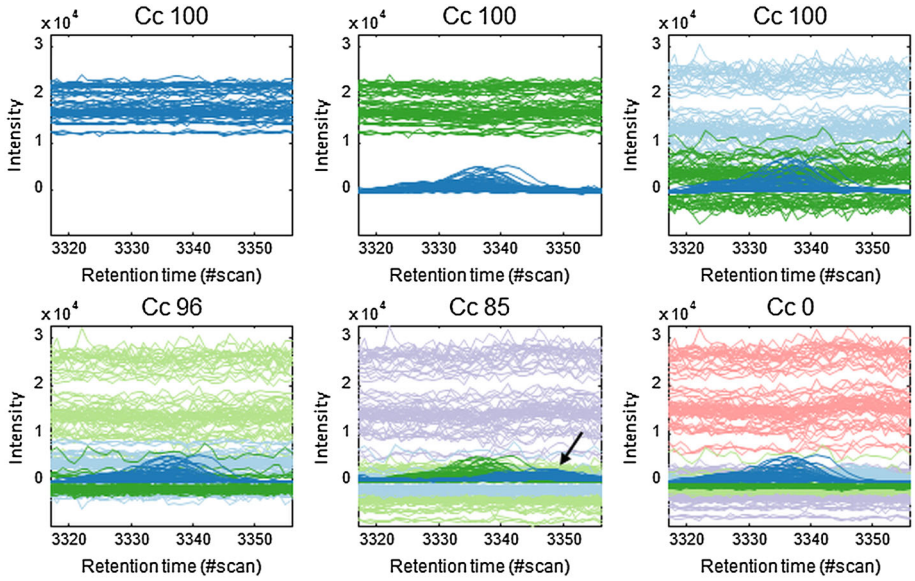


Figure 7. Interval 32, wine: elution profiles of models with one to six factors. The core consistencies (Cc) are high with the exception of the last model with six factors. Notice that in the model with five factors, an additional small peak appears (indicated with the arrow).

important peaks may be overlooked. When analysing all the intervals with low signal-to-noise ratios, the same conclusion can be made; hence, more factors than initially determined need to be included if all chemical variations are to be captured as suggested by core consistency.

4.3. Simulated data

When models on real data are calculated, it can be difficult to determine the true rank of the data. Therefore, we have included results from PARAFAC2 models of simulated data as well.

In the original paper concerning core consistency in PARAFAC1 [9], calculations on simulated data were also included. The authors showed that core consistency does not work very well on perfect data, meaning data that follow the PARAFAC1 model and only has additional random identically distributed Gaussian noise. It was argued that this problem was of limited consequence as perfect data are simple to model in any case and it is very rare that such data are met in practice. This was also supported by the fact that the problems observed with ideal data were not observed for any of the quite diverse example datasets.

To assess core consistency in the original publication, a certain amount of model error was introduced in the simulated data to more adequately simulate real data. The model error introduces variance resulting in data that are not truly trilinear.

A similar approach is adopted here. Data with different ranks (three and five) and different congruence values [16] (0, 0.20, 0.50, and 0.90) were generated, to cover varying types of data, by creating a $\underline{\mathbf{Y}}$ array according to Equation (5). The components in these data were drawn from a Gaussian distribution, and in addition, independent and identically distributed noise was added to the $\underline{\mathbf{Y}}$ array in 'low' and high levels (15% and 40%, respectively). Then, three levels of model errors were introduced (5%, 10%, and 15%) to affect the trilinearity of the data. Subsequently, each slab of $\underline{\mathbf{Y}}$ was multiplied by an orthogonal $\mathbf{P}_k$ matrix to simulate PARAFAC2 data. This resulted in data arrays of size $10 \times 15 \times 30$. One hundred datasets were created for each combination of rank, Gaussian noise, model error, and congruence values. For the rank 3 data, PARAFAC2 models with one to five factors were calculated, and for each model, the core consistency was determined. Similarly, for the rank 5 data, PARAFAC2 models with one to seven factors were calculated.

Upon inspection of the obtained models, it was found that models with congruence values of 0, 0.20, and 0.50 in general fit the raw data quite accurately. However, this was, in most cases, not the case for models calculated on data with high congruence (0.90)—this was also reflected in the core consistencies. For these models, the core consistencies are in general very low for models with too few factors included (some core consistencies are below 0), regardless of the different model errors and noise levels introduced. Real data are oftentimes correlated, but a congruence value of 0.90 is quite high, and the problem has not been observed when calculating the core consistency in the three real datasets. The high-congruence data are not considered further here but may point to a limited usefulness of core consistency with highly correlated data.

The models based on the remaining data (congruence values of 0, 0.20, and 0.50) are summarized according to the core consistency in Table II. Because 100 models have been calculated for each combination of rank, congruence, noise, and model error, the core consistencies are presented as averages calculated on core consistencies where all negative values are set to 0. Otherwise, core consistencies with very high negative values would dominate the obtained mean value. Positive core consistencies are included as is.

The averaged core consistencies in the table show that there is a significant drop in core consistency when the number of factors exceeds the rank of the raw data, suggesting that core consistency indeed can be used as an indication of overfit. However, the core consistencies rarely approach 0, and in some cases, the starting point is quite low for the core consistency, that is for rank 5 data with high noise and high model error. Nevertheless, there is still a drop in the average core consistency when the number of factors included exceeds the true rank of the data.

The aforementioned observations indicate that core consistency can be used to find the true rank of data with high and

Table II. Summary of averaged core consistencies from simulated models with different properties

	Congruence	Factors	Noise level					
			Low			High		
			Model error					
			Low	Medium	High	Low	Medium	High
Rank 3	0.00	2	100	100	100	100	100	100
		3	100	100	100	100	100	100
		4	**73**	**36**	**64**	**80**	**71**	**78**
	0.20	2	100	100	100	98	100	96
		3	100	100	100	100	100	100
		4	**43**	**36**	**40**	**70**	**64**	**66**
	0.50	2	79	97	76	70	100	69
		3	78	91	74	78	96	75
		4	**35**	**13**	**31**	**65**	**43**	**61**
Rank 5	0.00	4	100	100	100	97	98	97
		5	100	100	100	99	99	99
		6	**72**	**71**	**64**	**72**	**76**	**78**
	0.20	4	99	99	100	94	93	90
		5	99	99	99	72	96	96
		6	**9**	**7**	**6**	**43**	**47**	**47**
	0.50	4	51	50	61	41	44	42
		5	58	48	58	41	49	55
		6	**12**	**13**	**7**	**22**	**27**	**19**

The bolded values mark overfitted models.

low signal-to-noise ratios and different levels of correlations within the data. When compared with the simulation results in the original publication, the results also indicate that core consistency under certain circumstances may be less effective when used for model selection with PARAFAC2 than with PARAFAC1.

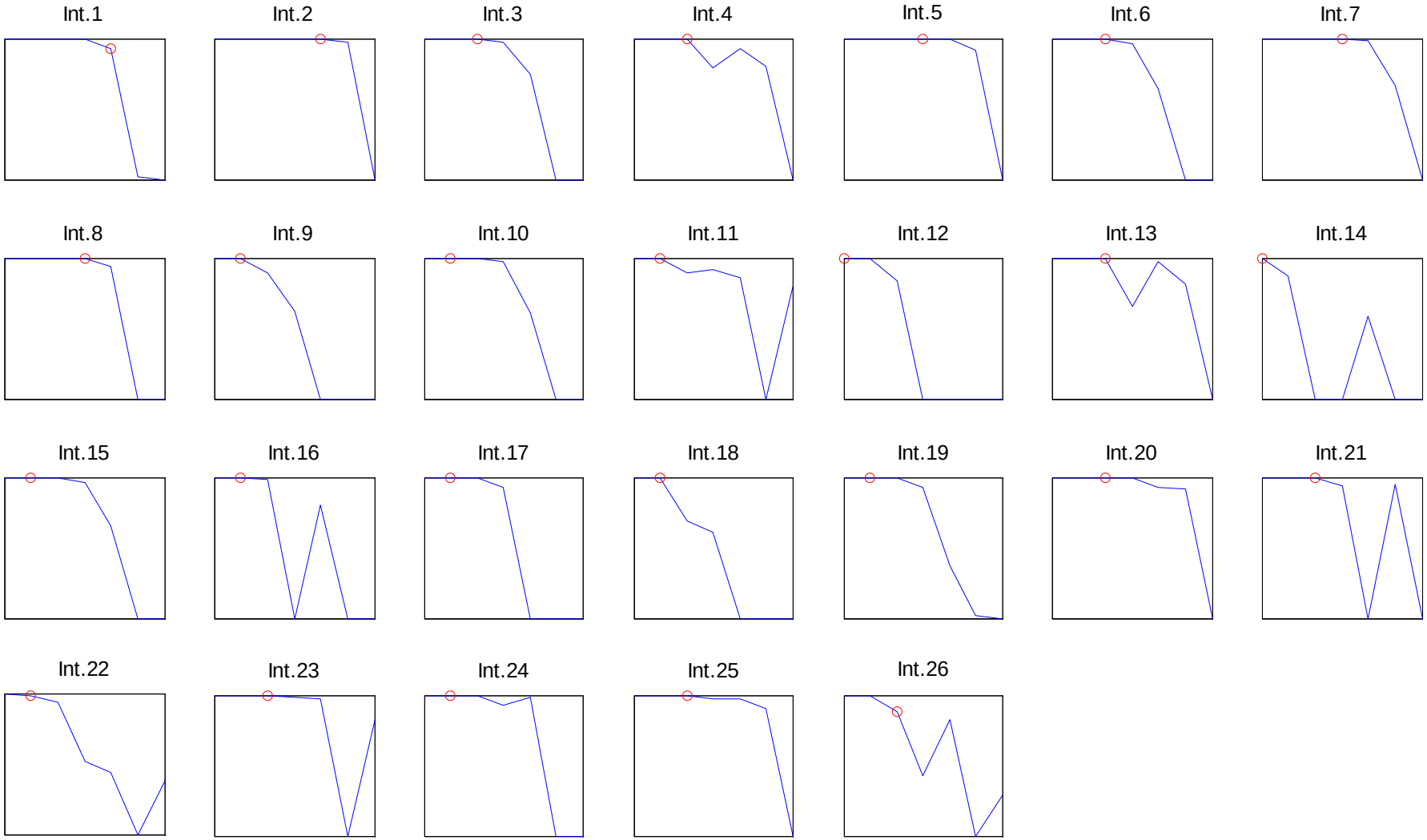
5. CONCLUSION

After evaluating the suggested core consistency diagnostic on several PARAFAC2 models from different real as well as simulated datasets, we concluded that core consistency is a helpful parameter in the evaluation of PARAFAC2 models. In some cases, usage of core consistency provides a better estimation of the underlying features than solely visual inspection. However, core consistency should not be used as the only measure of model complexity. It should be combined with additional measures or parameters such as investigation of residuals and loadings.

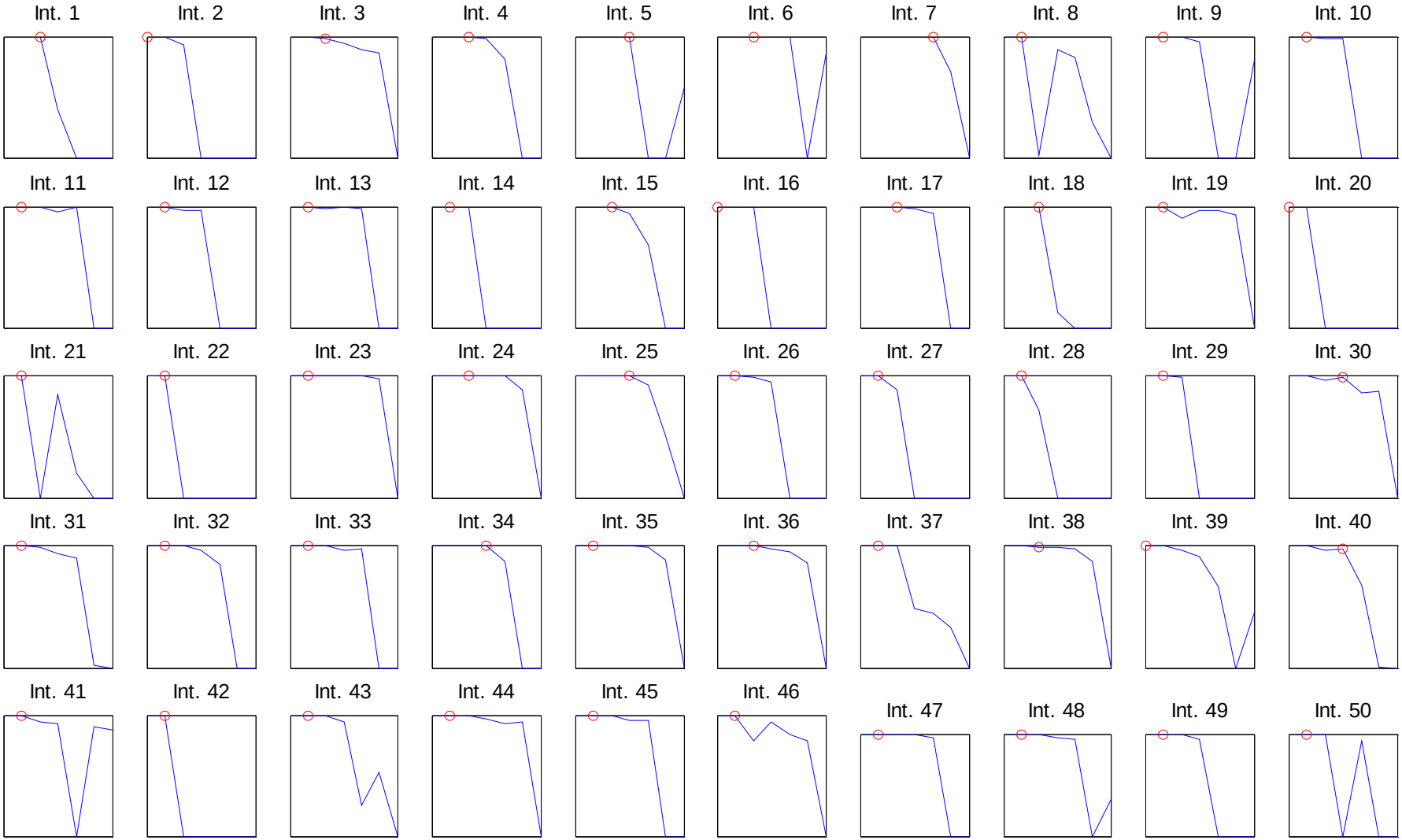
REFERENCES

1. Harshman RA. PARAFAC2: mathematical and technical notes. *UCLA Working Papers in Phonetics* 1972; **22**: 30–44.
2. Kiers HAL, Ten Berge JMF, Bro R. PARAFAC2—part I. A direct fitting algorithm for the PARAFAC2 model. *J. Chemom.* 1999; **13**: 275–294.
3. Amigo JM, Skov T, Coello J, Maspoch S, Bro R. Solving GC–MS problems with PARAFAC2. *Trends Anal. Chem.* 2008; **27**: 714–725.
4. Matero S, Poutiainen S, Leskinen J, *et al*. Monitoring the wetting phase of fluidized bed granulation process using multi-way methods: the separation of successful from unsuccessful batches. *Chemometr. Intell. Lab.* 2009; **96**: 88–93.
5. Robeyst N, Grosse CU, De Belie N. Monitoring fresh concrete by ultrasonic transmission measurements: exploratory multi-way analysis of the spectral information. *Chemometr. Intell. Lab.* 2009; **95**: 64–73.
6. de Juan A, Tauler R. Comparison of three-way resolution methods for non-trilinear chemical data sets. *J. Chemom.* 2001; **15**: 749–772.
7. ten Berge JMF, Kiers HAL. Some uniqueness results for PARAFAC2. *Psychometrika* 1996; **61**: 123–132.
8. Harshman RA, De Sarbo WS. An application of PARAFAC to a small sample problem, demonstrating preprocessing, orthogonality constraints, and split-half diagnostic techniques. In *Research Methods for Multimode Data Analysis*, Law HG, Snyder CW, Hattie JA, McDonald RP (eds.). Praeger: New York, 1984; 602–642.
9. Bro R, Kiers HAL. A new efficient method for determining the number of components in PARAFAC models. *J. Chemom.* 2003; **17**: 274–286.
10. Harshman RA. Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics* 1970; **16**: 1–84.
11. Bro R. PARAFAC. Tutorial and applications. *Chemometr. Intell. Lab.* 1997; **38**: 149–171.
12. Bro R, Andersson CA, Kiers HAL. PARAFAC2—part II. Modeling chromatographic data with retention time shifts. *J. Chemom.* 1999; **13**: 295–309.
13. Tucker LR. Some mathematical notes on 3-mode factor analysis. *Psychometrika* 1966; **31**: 279.
14. Amigo JM, Popielarz MJ, Callejon RM, *et al*. Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis. *J Chromatogr A* 2010; **1217**: 4422–4429.
15. Skov T, Ballabio D, Bro R. Multiblock variance partitioning: a new approach for comparing variation in multiple data blocks. *Anal. Chim. Acta* 2008; **615**: 18–29.
16. Lorenzo-Seva U, ten Berge JMF. Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology-Eur* 2006; **2**: 57–64.

Apple data


Int.1
Int.2
Int.3
Int.4
Int.5
Int.6
Int.7
Int.8
Int.9
Int.10
Int.11
Int.12
Int.13
Int.14
Int.15
Int.16
Int.17
Int.18
Int.19
Int.20
Int.21
Int.22
Int.23
Int.24
Int.25
Int.26

Wine data


Int. 1
Int. 2
Int. 3
Int. 4
Int. 5
Int. 6
Int. 7
Int. 8
Int. 9
Int. 10
Int. 11
Int. 12
Int. 13
Int. 14
Int. 15
Int. 16
Int. 17
Int. 18
Int. 19
Int. 20
Int. 21
Int. 22
Int. 23
Int. 24
Int. 25
Int. 26
Int. 27
Int. 28
Int. 29
Int. 30
Int. 31
Int. 32
Int. 33
Int. 34
Int. 35
Int. 36
Int. 37
Int. 38
Int. 39
Int. 40
Int. 41
Int. 42
Int. 43
Int. 44
Int. 45
Int. 46
Int. 47
Int. 48
Int. 49
Int. 50

Paper II

Forecasting breast cancer development using metabolomics and bio-contours

Submitted

FORECASTING BREAST CANCER DEVELOPMENT USING METABOLOMICS AND BIO-CONTOURS

Authors: Rasmus Bro*[1], Maja H. Kamstrup-Nielsen[1], Søren Balling Engelsen[1], Francesco Savorani[1], Flemming H. Larsen[1], Morten A. Rasmussen[1], Louise Hansen[2], Anja Olsen[2], Anne Tjønneland[2], Lars Ove Dragsted[3]

Affiliations:

1. University of Copenhagen, Department of Food Science, 30 Rolighedsvej, 1958 Frederiksberg, Denmark, *rb@life.ku.dk
2. Danish Cancer Society Research Center, Strandboulevarden 49, 2100 Copenhagen, Denmark
3. University of Copenhagen, Department of Nutrition, Exercise and Sports, 30 Rolighedsvej, 1958 Frederiksberg, Denmark

Background

Breast cancer is a major cause of death for women. To improve treatment, current oncology research focuses on discovering and validating new biomarkers for early detection of cancer, so far with limited success.

Methods

Blood plasma samples were taken from 419 women diagnosed with breast cancer up to seven years after the sample was taken. A comparative set of 419 samples of women without this diagnosis was used as control. Nuclear magnetic resonance spectroscopy measured the relative concentration of 129 plasma metabolites. Using data fusion, these metabolite concentrations were combined with 47 additional variables reflecting auxiliary information on anthropometrics, life style factors, etc. Multivariate chemometric classification models were used to build a model for discriminating the two groups.

Results

This study shows that by combining information from metabolomic measurements of plasma samples and classical risk markers from questionnaire data, it is possible to create a *bio-contour*, which we define as a complex pattern of biological and habitudial information that can be used for reliably forecasting the risk of getting breast cancer years ahead in time. While single markers have close to no predictive value, the bio-contour provides a sensitivity of 0.84 and a specificity of 0.85. This *forecast* is on par with how well most current biomarkers can *diagnose* current cancer. It is further shown that causal interpretations are not immediately possible from such models regardless of the validity of the model.

Conclusion

Metabolic forecasting of cancer by bio-contours opens new possibilities for early cancer prediction and cancer screening.

Breast cancer is the major cause of death for women in the first decade after menopause. Despite insight in to several disease risk factors, these explain only a minor fraction of the incident cases. Continuous improvements in sensitivity, resolution and precision of modern explorative technologies like metabolomics have the potential to identify additional risk factors. More importantly, they also form a basis for prediction modeling at the individual level, i.e. individual prediction of disease risk. This translational aspect has not been exploited to any large extent until now, primarily due to the inherent difficulties associated with the technology. Omics-based biomarker profiling is a complex and truly multi-disciplinary subject.

Proliferation of the tumor at time of diagnosis is probably the factor with the greatest effect on survival rates among cancer patients. Consequently, an important focus in cancer research is to improve our ability to detect malignancy prior to the stage where the tumor has evolved into a clinically detectable disease. Breast cancer is the most common type of cancer diagnosed among women in the Western part of the world. In Europe, 425,417 women were diagnosed with breast cancer in 2008 and 128,737 women died. Worldwide, close to 1.4 million women are diagnosed with breast cancer each year and approximately 450,000 die from breast cancer. To facilitate detection of breast cancer prior to the occurrence of clinical symptoms, many Western countries have introduced mammography screening programs that are broadly aimed at middle-aged women. The risk of too many false positives in mammography screening, that is, detection of tumors that never progress to a stage that will affect the wellbeing of the patient has, however, been heavily discussed [1]. A method of early breast cancer detection will have substantial implications.

METHODS

Cancer cohort

In the current project, a subset of 838 women from the Danish Diet, Cancer and Health (DCH) cohort have been analyzed. The cohort was established in 1993-1997 and consists of a total of 57,053 men and women free of cancer at the time of recruitment [2]. The DCH cohort is part of the European Prospective Investigation into Cancer and Nutrition (EPIC) study including cohort participants from ten European countries. In the part of the cohort investigated here, 419 women were diagnosed with breast cancer between time of enrolment and the chosen follow-up date (December 31, 2000). For the current study, an equal number of randomly selected women (419) free of cancer during the same timespan, were selected as controls in an evenly balanced dataset. Several baseline characteristics were recorded for all individuals and a standardized questionnaire was used to collect subjective information on life style factors, including dietary habits, smoking, alcohol intake, etc.[2]. A total of 99 percent of the participants in the DCH cohort gave biological material at their time of recruitment, and for this project, plasma samples were used. All blood samples were withdrawn in a non-fasting state, and citrate was used as anticoagulant. The samples were stored at -150/-80 degrees until analysis. Further description of the study subjects can be found in the supplementary material.

Metabolomic data

The 838 plasma samples were analyzed by proton Nuclear Magnetic Resonance (^{1}H NMR). The ^{1}H NMR analytical platform[3] has several advantages compared to other common metabolomics analytical platforms. In particular, it is inherently quantitative and provides an unbiased and highly reproducible simultaneous observation of many metabolites. The so-called 'curse of dimensionality'[4] poses a practical hindrance for how much information can be obtained when few samples and many variables are

measured. In the ^{1}H NMR data, there are resonances from each hydrogen atom in hundreds of molecules sampled in several thousand variables. The high number of variables increases the risk of *spurious correlations* and this is a fundamental problem in non-targeted and comprehensive analyses[5]. A way to counter the curse of dimensionality is to have a sufficient number of samples compared to the number of variables and to avoid inflating the number of variables if possible. In this case, the NMR spectra of 838 subjects' blood samples were transformed into a less redundant representation by using integrals of 189 identified spectral regions. These peak regions were further reduced to 129 variables as some regions contained resonances from the same chemical compounds (see Supplementary Appendix). Each individual region was carefully selected and assessed and, in order to avoid selection bias, the best approach for integrating was decided in a blinded way, i.e. without any knowledge of the outcome (cancer status).

In addition to the NMR data, 47 variables were included which contain information about the lifestyle and phenotype of the subjects, resulting in a final dataset of 176 variables. A complete list of these additional parameters, which mainly relate to anthropometrics, life style habits such as smoking, alcohol intake and dietary habits, can be found in the Supplementary Appendix.

RESULTS

Forecasting cancer status

Prior to data analysis, the samples were randomly split into two groups – one group containing 628 samples and another group containing 210 samples. Each group contained equally many controls and future breast cancer patients. The larger group – the calibration set – was used for building a prediction

model. The smaller set was only used subsequently as a test set, to validate the predictive quality of the resulting model.

A complex biological problem such as future cancer development is unlikely to be well described by a single or a few biomarkers. It is necessary to investigate the multivariate complex *pattern* of variation in the current data in order to extract all the relevant information.

This is also evident when examining the performance of individual variables for the given data. The best single variable for classification was identified as an NMR spectral region ranging from 3.116 to 3.132 ppm (called "3.12 ppm" in the following). Using this single variable in linear discriminant analysis, yields a fairly low sensitivity of 0.65 and a specificity of 0.55 on the test set (Figure 1 bottom shows "3.12 ppm" for calibration). As another example of a univariate approach, consider the number of years using hormone replacement therapy (variable "HRT – years of use"). This is an established risk factor for breast cancer[6]. A linear discriminant analysis using just hormone replacement therapy yields a specificity of 0.85 but a sensitivity of 0.29 and a plot of this variable (Figure 1 top) clearly shows the limited discriminatory power of this risk marker. The present dataset is rather high in the number of samples and therefore also in statistical power. Null hypothesis testing of "HRT – years of use" and "3.12 ppm" reveals apparent strong results ($p_{\text{HRT – years of use}} = 0.00001$ and $p_{3.12\ \text{ppm}} = 0.0000001$). Although these results suggest real differences between cases and controls on the population level, it is clear from Figure 1 that these variables offer close to *no* power in terms of predicting the status of an individual.

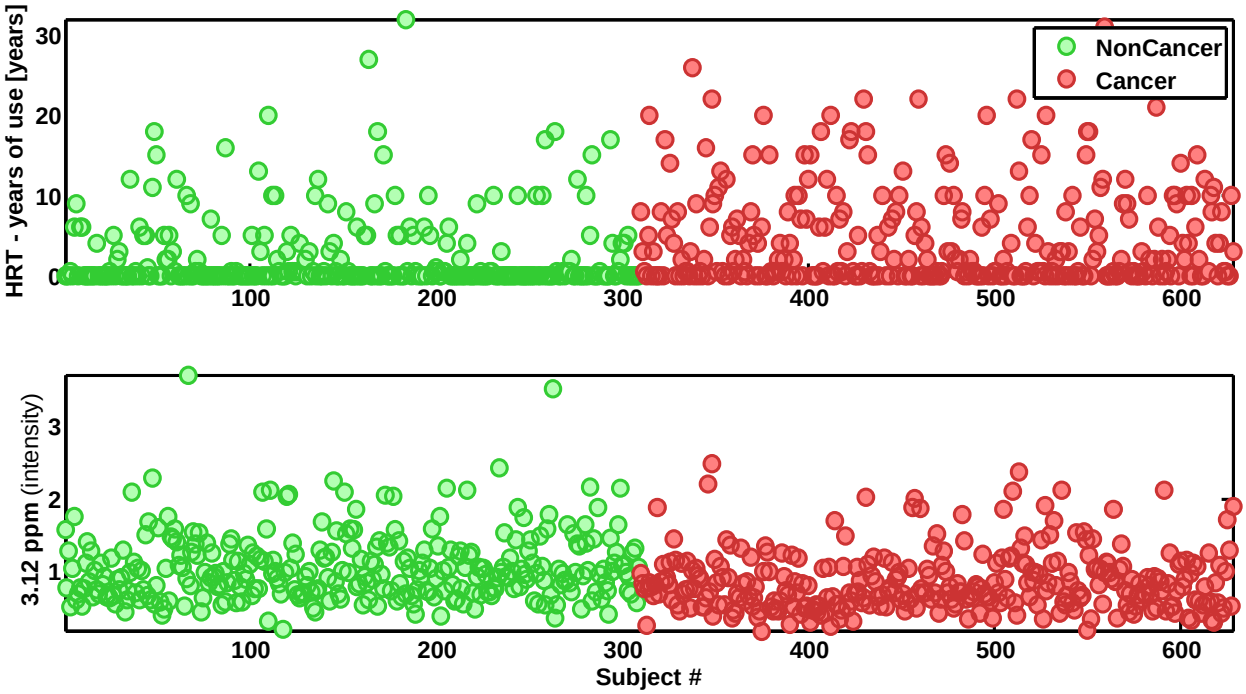


Figure 1. (Top) Years of hormone replacement therapy for the 628 persons in the calibration set. It is clear that there is little discriminatory power in this measure. Using it for classifying, the test set will yield a classification error of approximately forty percent. (Bottom) Similar plot of the best discriminatory variable (3.12 ppm).

Rather than using single variables, it is imperative to use a sufficient number of relevant variables to reflect the *biological patterns* that relate to the given endpoint. The chemometric classification model Partial Least Squares Discriminant Analysis (PLS-DA)[7, 8] allows building multivariate classification models with correlated variables. Thereby data fusion of the NMR and additional variables is possible[9]. By using cross-validated variable selection[10, 11], a classification model was determined using 28 of the original variables. The model provides a hitherto unseen effective means for forecasting breast cancer and is visualized in Figure 2. The model is built using eight PLS components, only two of which are depicted in the figure. Variables close to the position of 'Cancer' are positively correlated to cancer incidence and those opposite to cancer are negatively correlated within these PLS components. Note, that variables may be positively correlated in one dimension but negatively correlated in others.

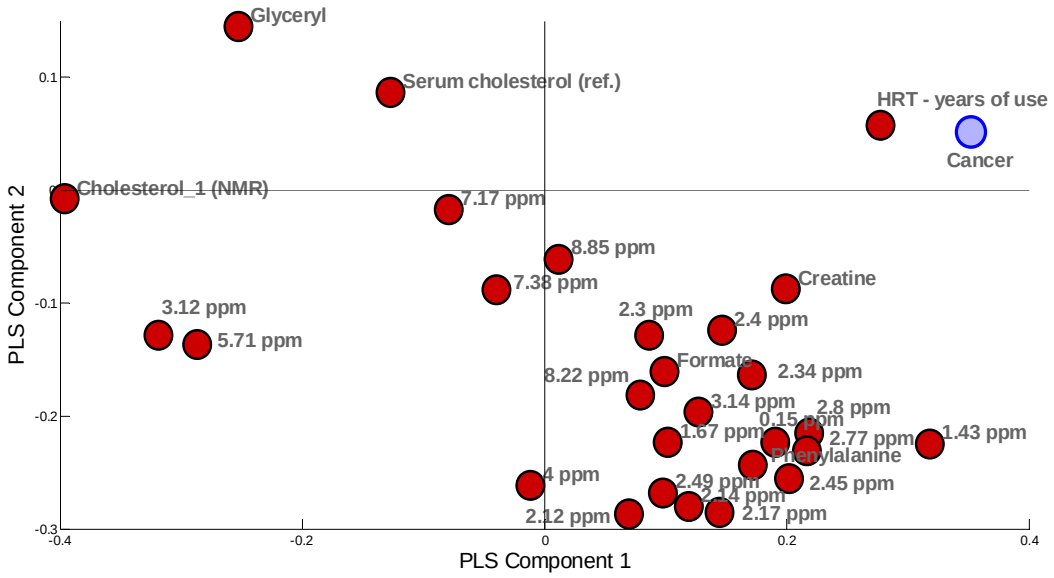


Figure 2. Loadings of a PLS-DA model. Variables in the direction of Cancer are positively related to cancer incidence whereas variables in the opposite direction are negatively related within the two components. Note, that the plot only shows two of a total of eight PLS components that are used in the PLS-DA model.

Validation is of utmost importance, especially when the variable to sample ratio is high and the relevant signals are deeply buried in the data. Oftentimes, the so-called score plot is used as an indication of group separation, but this can easily lead to overly optimistic interpretations[5]. The situation can become even more misleading when orthogonal signal correction is used as the model quality remains the same, while possible score plots will show even better apparent separation regardless that the actual classification ability has not improved. Instead of relying on score plots, the classification power must be assessed using independent samples. In the present study, the high number of samples, the reduction in number of variables and the high-quality experimental data lead to a robust model as evidenced by the similarity of the receiver operating characteristic (ROC) curves [12] from the calibration model, the cross-validated model and even from the test set (Figure 3).

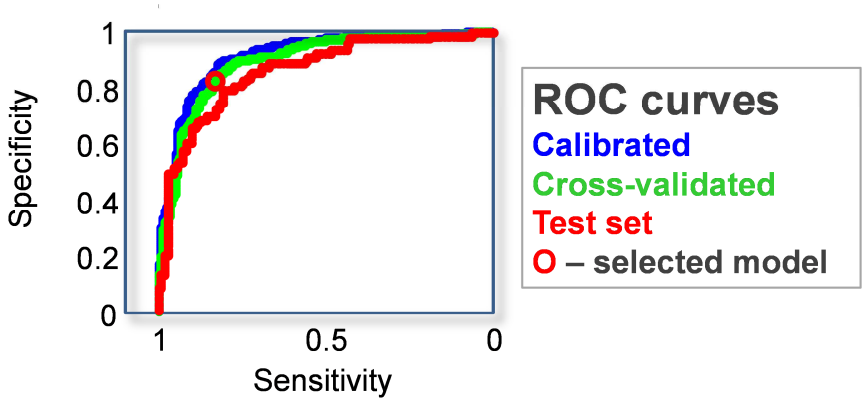


Figure 3. Receiver operating characteristic curves for the PLS-DA model of 28 selected variables during calibration (blue), cross-validation (green) and test set (red).

The sensitivity/specificity obtained is 0.84/0.85 for the calibration set, 0.82/0.83 for cross-validation and 0.80/0.79 for the 210 samples in the test set that have not taken part in the model building. The predicted cancer class membership of the test set samples is seen in Figure 4. In comparison to the univariate relations shown in Figure 1, it is clear that the multivariate discriminatory power is much stronger.

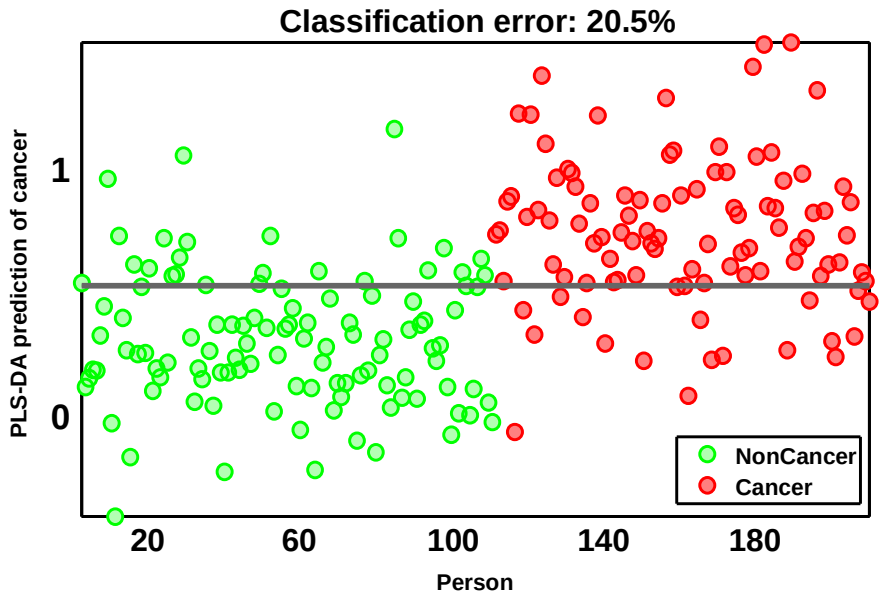


Figure 4. Cancer model based on 28 selected variables. Prediction of cancer status of test set. Horizontal line is the selected threshold for assigning a sample to either class.

DISCUSSION

Understanding the classification model

While variable selection can lead to improved predictions[10], this must not be confused with subsequent interpretations of how variables are causally related to cancer development. The classification model is based on an eight component PLS-DA model. This means that there are eight underlying 'variations' – an eight-dimensional subspace – that, *when* combined, can predict cancer status. The 28 selected variables shown in Figure 2 are merely indicators of this eight-dimensional subspace. It is important to note that other variables could have been selected and serve as a probe of the subspace as well.

With respect to interpretation, all variables that are correlated *must* be assessed in a combined manner. It is the pattern of variables rather than the individual variables that is biologically meaningful for

interpretations. We call the underlying pattern reflected by the relevant variables a *bio-contour*. We use this term deliberately to enforce an understanding that the temptation to pick out individual biomarkers and elevate these to be causal markers for explaining this complex biological phenomenon is a misleading avenue and has, in fact, led to very little progress so far[13, 14].

For example, a statistically significant predictive variable may be years of HRT use. However, this variable may merely be *indirectly correlated* because it is associated with the actual causal factor, in this case possibly the activation of cell division by estrogen signaling. Selection of significant variables has at best only some indirect relation to causality. This is especially true for untargeted analyses.

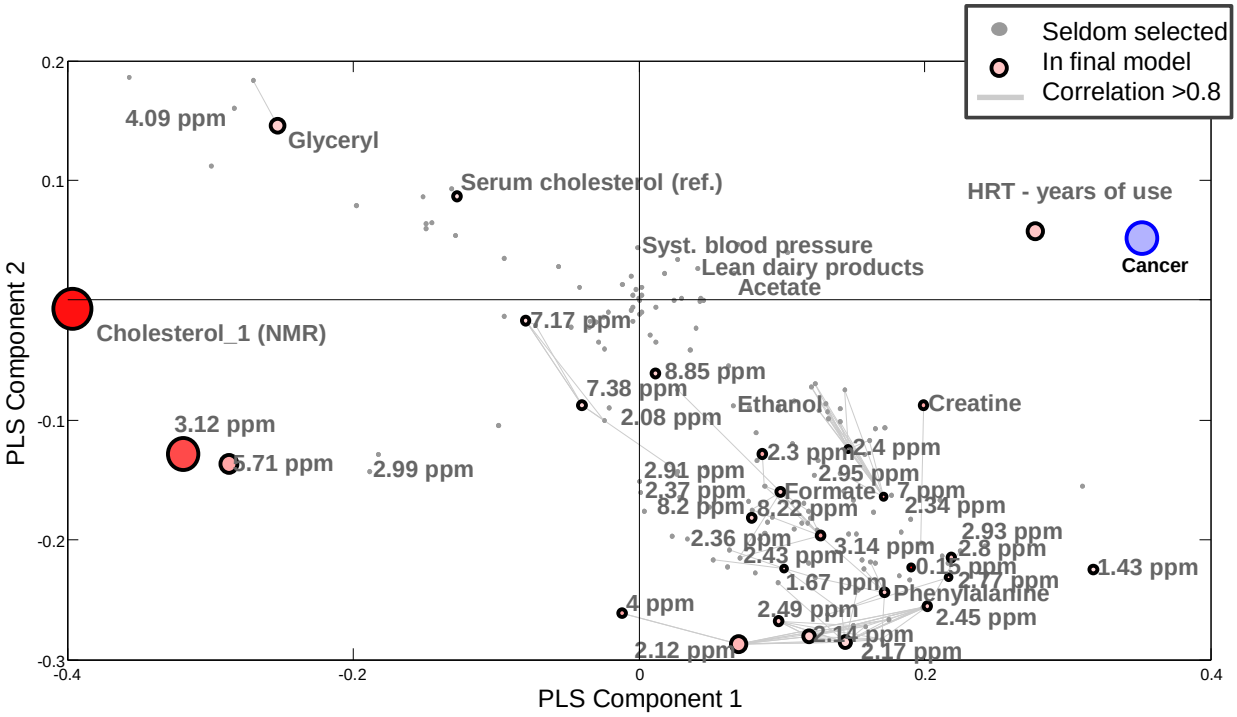


Figure 5. All variables relevant for predicting cancer status. Variables encircled by a black line are the actually chosen variables in the final model and color intensity and size of the circle indicate how often a given variable is chosen during resampling (bootstrapping[15]) of the variable selection. This allows assessing how unique a certain selection is. The plot is similar to Figure 2 but now represents the complete bio-contour projected onto the two-dimensional space. Note that the known marker hormone replacement therapy ("HRT – years of use") is not selected as often as for example, the variable "3.12 ppm" (lower middle plot), representing a 'preventive' factor related to the second disease component.

The shape of the bio-contour is indicated in Figure 5, where it is illustrated how the eight-dimensional space is related to all the measured variables shown for the two first dimensions of the model. Each dimension of the bio-contour may be interpreted as representative of a 'pseudo-etiology' containing complex biological information associated with – but *not* representing – causality. The link to causality will, in most cases, not be direct, and causal claims will necessarily have to come from *other* theories than the empirically observed correlations.

Cholesterol is seen in the plot as a variable that is often selected and which is in a 'protective' position in the first component of the biocontour but neutral in the second component. In the remaining components, cholesterol has opposing effects and overall only a moderate but positive total effect on the prediction of cancer. Note, though, that observed correlations and effects in a biocontour can even be opposite to the actual biological effect. This can happen e.g. if cholesterol is also affected by other aspects that has an opposing effect on cancer. This can be termed a *luring* correlation. Cholesterol is well known to have a complex relationship with breast cancer, which is supported by the observations here[16].

It needs to be considered that the model forecasts diagnosis of breast cancer and not necessarily aggressively progressing disease. Some of the women diagnosed with breast cancer may have a slow growing tumor that would never have affected lifespan if not discovered; on the other hand, some of the included controls may have an undiagnosed tumor.

In this study, we have described a bio-contour that can forecast diagnosis of breast cancer several years ahead. It has been exposed to strong internal validation, but its applicability for other populations of women with other diets, lifestyles, medications and habits remains unproven. The global validity of the bio-contour needs to be tested on future datasets. In addition, the method should be tested against traditional screening tools.

The perspectives in early detection of cancer by use of bio-contours from blood samples and life style factors from apparently healthy persons are of worldwide importance. We advocate that bio-contours get a much more prominent role in cancer prediction and disease diagnostics. At the same time, it is stressed that predictive models from empirical data can never form the actual basis of a causal interpretation. Empirically observed correlations or effects can be misleading because they are spurious, indirect or luring (affected by other variations that also have an effect). These different problems all lead to different limitations that must be considered when interpreting models. Observed model parameters, though, may be used as a lead for generation of new hypotheses in an exploratory fashion.

Reference List

1. Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. Lancet 2013;380:1778-1786.
2. Tjønneland A, Olsen A, Boll K et al. Study design, exposure variables, and socioeconomic determinants of participation in Diet, Cancer and Health: A population-based prospective cohort study of 57,053 men and women in Denmark. Scandinavian Journal of Public Health 2007;35:432-441.
3. Beckonert O, Keun HC, Ebbels TMD et al. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. Nature Protocols 2007;2:2692-2703.
4. Bellman R. Dynamic Programming. Princeton, NJ: Princeton University Press; 1957.
5. Kjeldahl K, Bro R. Some common misunderstandings in chemometrics. J Chemom 2010;24:558-564.
6. Tjønneland A, Christensen J, Thomsen BL et al. Hormone replacement therapy in relation to breast carcinoma incidence rate ratios - A prospective Danish cohort study. Cancer 2004;100(11):2328-2337.
7. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. Anal Chim Acta 1986;185:1-17.

8. Næs T, Indahl U. A unified description of classical classification methods for multicollinear data. J Chemom 1998;12:205-220.

9. Bro R, Nielsen HJ, Savorani F et al. Data fusion in metabolomic cancer diagnostics. Metabolomics 2013;9(1):3-8.

10. Andersen CM, Bro R. Variable selection in regression—a tutorial. J Chemom 2010;24:728-737.

11. Ståhle L, Wold S. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. J Chemom 1987;1:185-196.

12. Zweig MH, Campbell G. Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. Clin Chem 1993;39(4):561-577.

13. Khleif SN, Doroshow JH, Hait WN. AACR-FDA-NCI Cancer Biomarkers Collaborative Consensus Report: Advancing the Use of Biomarkers in Cancer Drug Development. Clinical Cancer Research 2010;16:3299-3318.

14. Drucker E, Krapfenbauer K. Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. The EPMA Journal 2013;4(7):2-10.

15. Efron B. The Jackknife, the Bootstrap and other resampling plans. Philadelphia: Society for industrial and applied mathematics; 1982.

16. Fagherazzi G, Fabre A, Boutron-Ruault MC, Clavel-Charelon FO. Serum cholesterol level, use of a cholesterol-lowering drug, and breast cancer: results from the prospective E3N cohort. European Journal of Cancer Prevention 2010;19(2):120-125.

17. Xie YL, Hopke PK, Paatero P. Positive matrix factorization applied to a curve resolution problem. J Chemom 1998;12(6):357-364.

18. Nicholson JK, Foxall PJD, Spraul M, Farrant RD, Lindon JC. 750 MHz ^{1}H and ^{1}H-^{13}C NMR spectroscopy of human blood plasma. Anal Chem 1995;67:793-811.

19. Nørgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. Appl Spectrosc 2000;54(3):413-419.

Acknowledgements: The study was carried out as a part of the research program Metabonomic Cancer Diagnostics supported by The Villum Foundation (www.veluxfoundations.dk) as well as the research program of the Danish Obesity Research Centre (DanORC, www.danorc.dk), supported by the Danish Council for Strategic Research (grant 2101-06-0005). Both the Diet, Cancer and Health study and the

current sub-study were approved by the regional ethics committees on human studies in Copenhagen and Aarhus and by the Danish Data Protection Agency. Informed consent was obtained from all participants.

Supplementary Appendix

Study population

The present study is based on data from the prospective Danish Diet, Cancer and Health (DCH) cohort study. A total of 57,053 men and women were enrolled into the cohort in 1993-1997 and were included if they fulfilled the following criteria: age between 50-64 years, born in Denmark and no previous cancer diagnosis in the Danish Cancer Registry. A detailed food frequency questionnaire (FFQ) and lifestyle questionnaire were filled in by each participant. Development and validation of the FFQ is described elsewhere. The FFQ contained questions regarding 192 food and beverage items and was developed to obtain information on the participants' habitual diet during the preceding year. Biological and anthropometric measurements were taken including height (m) and weight (kg), from which body mass index (BMI) was calculated (kg/m^2) as well as a non-fasting 30 mL blood sample. Citrate was used as the anticoagulant. The blood samples were spun and divided into fractions of citrate plasma, serum, red blood cells, and buffy coat and stored in 1 mL tubes. All samples were processed and frozen within two hours at -20°C and were ultimately transferred to liquid nitrogen vapor (max. -150°C). Shortly before NMR analysis, samples were retrieved from the bio-bank and stored at -80°C. For the current project, the plasma fraction was used. A thorough description of data collection has been published elsewhere [2].

From the DCH cohort, a combined study population of 3510 persons was created, consisting of three case groups: breast cancer (433), colon cancer (428) and coronary heart disease (1149) as well as a sub-cohort (1500 persons) to be used as a reference group for all. After exclusions due to insufficient sample volume or low data quality, the total study population consisted of 3,419 persons, of which 419 were breast cancer cases. As only the breast cancer case group was used for this project, no further mention will be made of the two remaining case groups. The sub-cohort consists of 1500 persons (750 men, 750 women)

who were chosen randomly in the DCH cohort, and only the women in the sub-cohort were used in this investigation.

Follow-up and ascertainment of cases

Information on breast cancer incidence was obtained by linkage of the personal identification number of each participant to the Danish Cancer Registry, and follow-up was nearly complete (99.8%). All cohort members were followed up for primary breast cancer diagnosis from the date of their visit to the study clinic until the date of diagnosis of any cancer (except for non-melanoma skin cancer), date of death or emigration, or December 31, 2000, whichever came first. Furthermore, the Danish Breast Cancer Co-operative Group (DBCG) has information on estrogen receptor (ER) status (ER+ or ER-) for approximately 90% of all Danish breast cancer cases, and ER status was available for 384 of the 419 breast cancer cases in this study.

NMR analysis

The plasma samples were measured using ^{1}H CPMG-presat and ^{1}H NOESY-presat NMR spectra according to a standardized and highly automated procedure for sample preparation, handling and analysis on the basis of the guidelines introduced by Bechonert et al [3]. In the NOESY-presat data, the water-resonance is suppressed, but otherwise all components are observed in a quantitative manner. In the CPMG-presat data, also resonances from large molecules such as proteins are suppressed, which facilitates easier identification of small molecules.

The NMR spectra have been subjectively evaluated by spectroscopists and data analysts in order to exclude as many noise regions from the data as possible and to include all peaks present. The

spectroscopists and data analysts were blinded to the case/control status. Many of the peaks were present in both the CPMG-presat and NOESY-presat spectra and were only selected in one of these. The majority of the peaks were selected from CPMG-presat as there was generally a more well defined baseline in these spectra due to the suppression of protein resonances. In total 181 intervals from CPMG-presat and eight intervals from NOESY-presat were selected. The 189 intervals were individually baseline corrected and integrated by means of either Multivariate Curve Resolution (MCR) [17], the area under the peak or the height of the peak. Peak height provides a reasonable estimate of concentration (up to a scaling) when the line shapes of the peak in different samples are similar. Peak height has been used for integrating when e.g. baseline resolution was difficult to achieve. Integrating peaks leads to one value instead of a lineshape for each peak. For one peak, two MCR components were used instead of one. Many intervals represent the same compounds. For the intervals which have been assigned [18], those originating from the same molecule have been summed, meaning that all assigned compounds are only represented once. Consequently, the NMR data contribute with a total of 129 discrete integrated variables. The assigned peaks and areas in the NMR spectrum can be seen in Figure 6.

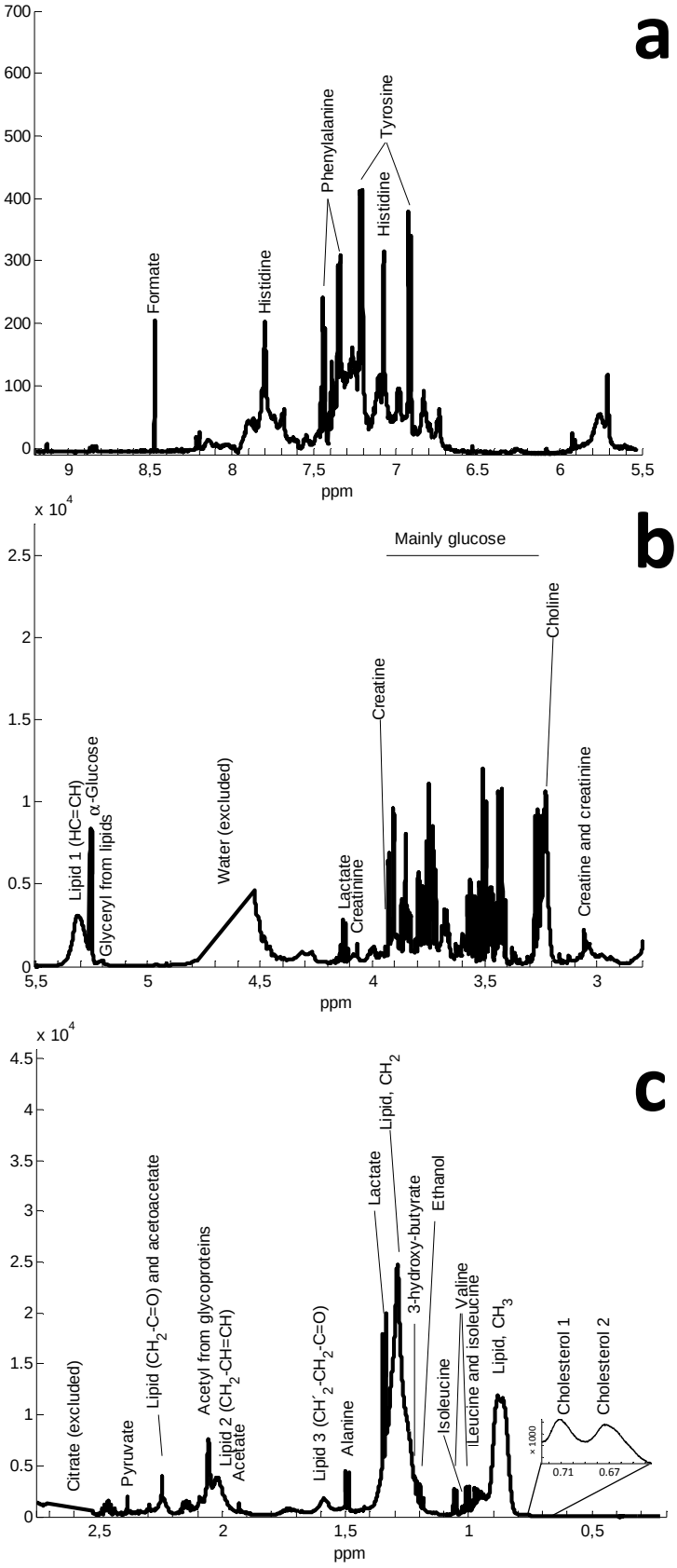

a
Formate
Histidine
Phenylalanine
Histidine
Tyrosine
ppm
b
x 10^4
Mainly glucose
Lipid 1 (HC=CH)
α-Glucose
Glyceryl from lipids
Water (excluded)
Lactate
Creatinine
Creatine
Choline
Creatine and creatinine
ppm
c
x 10^4
Citrate (excluded)
Pyruvate
Lipid (CH2-C=O) and acetoacetate
Acetyl from glycoproteins
Lipid 2 (CH2-CH=CH)
Acetate
Lipid 3 (CH´2-CH2-C=O)
Alanine
Lactate
Lipid, CH2
3-hydroxy-butyrate
Ethanol
Isoleucine
Valine
Leucine and isoleucine
Lipid, CH3
Cholesterol 1
Cholesterol 2
x 1000
0.71
0.67
ppm

Figure 6. ^{1}H CPMG-presat NMR spectra with annotation for selected spectral regions: a) 5.6-9.2 ppm, b) 2.7-5.6 ppm, and c) 0.3-2.7 ppm. In each region, the sub-spectrum is vertically scaled according to the most intense resonance. The average spectrum is shown.

Additional data

The additional variables included in the data analysis stem from the dietary and lifestyle questionnaires that were collected at baseline in the DCH cohort and from clinical markers obtained initially. More than 1000 variables were potentially available; however, of these variables, some are derived directly from the questionnaire, while others are summed or cumulated variables. We chose the ones that were deemed most relevant through a careful selection process prior to the data analysis. A total of 47 dietary and lifestyle variables were chosen based on knowledge of factors known or speculated to be important for cancer or heart disease development. The additional data included in the data analysis contribute with additional information regarding the life styles of the persons and are shown below with a short description (Table 1).

Table 1. Additional variables included in the classification model.

Variable	Explanation
Age >35 at first birth or no births	Older than 35 when having the first child OR have not given birth at all
Age at birth of first child	Age at first child birth
Alcohol intake	Alcohol (grams/day)
Alcohol intake (cumulated)	Cumulative alcohol, excluding pauses (units/week)

Amount of body fat	Fat weight (amount of fat mass in the body) (kg)
BMI	Body mass index (kg/m^2)
Carbohydrate intake	Carbohydrates (grams/day)
Coffee intake	Coffee (grams/day)
Dietary fibre intake	Dietary fibre (grams/day)
Energy intake	Total energy intake, incl. alcohol (kJ/day)
Fat intake	Fat (grams/day)
Fat pct.	Fat percent in diet
Fatty dairy products	Dairy products, fatty (grams/day)
Fruit intake (no juices)	Fruits (excl. juices) (grams/day)
Fruits (incl. juice)	All fruits (incl. juices) (grams/day)
Glycemic index (carbohydrates)	Overall glycemic index (based on all carbohydrates) (grams/day)
Glycemic load (carbohydrates)	Overall glycemic load (based on all carbohydrates) (grams/day)
Height	Standing height (cm)
High level school	Highest level school education
Hip circumference	Hip circumference (cm)
HRT – years of use	Years of HRT use
Intake of rye bread	Rye bread (grams/day)
Juice intake (fruit and vegs)	Juices (fruit and vegetable) (grams/day)

Lean dairy products	Dairy products, lean (grams/day)
Level of serum cholesterol	Level of serum Cholesterol (mmol/L)
Low level school	Low level school education
Marine omega-3	Marine fats (n-3) in diet (grams/day)
Medium level school	Medium level school education
Mono unsaturated fatty acids	Monounsaturated fat (grams/day)
Number of births	Number of births
Poly unsaturated fatty acids	Polyunsaturated fat (grams/day)
Potato intake	All potatoes (grams/day)
Processed meat	Processed meat (grams/day)
Protein intake	Protein (grams/day)
Red meat	Red meat (grams/day)
Saturated fatty acids	Saturated fat (grams/day)
Sugar intake	Total, all sugars (grams/day)
Syst. Blood pressure	Blood pressure, systolic
Total exercise (MET score)	MET-score, hours/week
User of nsaid	Current user of NSAIDS, including aspirin (self-reported)
Vegetables (incl. juice)	All vegetables (incl. juices) (grams/day)

Water intake	Water (grams/day)
Waist circumference	Waist circumference (cm)
Waist/hip ratio	Ratio of waist to hip
Weight	Weight (kg)
Years of quitting smoking	Time since smoking cessation (years)
Years of smoking	Smoking duration (years)

Data analysis

All data preprocessing and data analysis was carried out in MATLAB 2013A, The MathWorks Inc., Natick, MA, 2013 and with the chemometric toolbox PLS_Toolbox 7.0.3, Eigenvector Research, Inc., Wenatchee, WA, 2013. After integration of peaks, the 129 peak variables and the 47 additional variables were kept in a single dataset. Before further analysis, the data was randomly split into a calibration set of 628 samples and a test set of 210 samples. The fraction of cancer and non-cancer diagnosed samples were kept the same in the two sets. The particular split size was rather arbitrarily made as a compromise of having as many samples as possible for building the calibration model, yet retain enough samples in the test set to be able to get sufficiently robust test set results.

Variables for the calibration model were selected based on interval Partial Least Squares regression – iPLS[19] using only the calibration data and scaling each variable to unit standard deviation within the cross-validation. Variables were added one by one in a forward selection as long as the model improved. As a criterion for selection, the cross-validated classification error was used in an automated fashion.

Cross-validation was performed using ten randomly selected segments and this cross-validation was repeated ten times for each step. This quite conservative repetition of randomly selected segments provides a practical means for avoiding spurious correlations but has no effect on luring or indirect correlations. The assessment of how often each variable was selected was based on bootstrapping this approach 100 times. Note, that the bootstrapping has no effect on the actually selected variables and the quality of the model. It merely provides means for assessing the variability in the selection method.

Upon selecting variables, a final PLS discriminant model was made on the selected variables using autoscaling and selecting the number of PLS components based on the same cross-validation as above and on minimum cross-validated classification error. This model was used for predicting the class assignments on the test set with the prior selected variables.

Paper III

Fluorescence spectroscopy as a potential metabonomic tool for early detection of colorectal cancer

Reprint from *Metabolomics*, 8, 111-121

Metabolomics (2012) 8:S111–S121
DOI 10.1007/s11306-011-0310-7

ORIGINAL ARTICLE

Fluorescence spectroscopy as a potential metabonomic tool for early detection of colorectal cancer

Anders Juul Lawaetz · Rasmus Bro · Maja Kamstrup-Nielsen · Ib Jarle Christensen · Lars N. Jørgensen · Hans J. Nielsen

Received: 12 January 2011 / Accepted: 6 April 2011 / Published online: 21 April 2011

Abstract Fluorescence spectroscopy Excitation Emission Matrix (EEM) measurements were applied on human blood plasma samples from a case control study on colorectal cancer. Samples were collected before large bowel endoscopy and included patients with colorectal cancer or with adenomas, and from individuals with other non malignant findings or no findings ($N = 308$). The objective of the study was to explore the possibilities for applying fluorescence spectroscopy as a tool for detection of colorectal cancer. Parallel Factor Analysis (PARAFAC) was applied to decompose the fluorescence EEMs into estimates of the underlying fluorophores in the sample. Both the pooled score matrix from PARAFAC, holding the relative concentrations of the derived components, and the raw unfolded spectra were used as basis for discrimination models between cancer and the various controls. Both methods gave test set validated sensitivity and specificity values around 0.75 between cancer and controls, and poor discriminations between the various controls. The PARAFAC solution gave better options for analyzing the chemical mechanisms behind the discrimination, and revealed a blue shift in tryptophan emission in the cancer patients, a result that supports previous findings. The present findings show how fluorescence spectroscopy and chemometrics can help in cancer diagnostics, and with PARAFAC fluorescence spectroscopy can be a potential metabonomic tool.

Keywords Fluorescence spectroscopy · Colorectal cancer · Chemometrics · PARAFAC · Metabonomics

A. J. Lawaetz (✉) · R. Bro · M. Kamstrup-Nielsen
Department of Food Science, Quality and Technology, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark
e-mail: ajla@life.ku.dk

I. J. Christensen
Finsen Laboratory, Copenhagen, Denmark

L. N. Jørgensen
Department of Surgery, Bispebjerg Hospital, University of Copenhagen, Copenhagen, Denmark

H. J. Nielsen
Department of Surgical Gastroenterology, Copenhagen University Hospital Hvidovre, Hvidovre, Denmark

1 Introduction

The idea of using autofluorescence measurements of blood to discriminate people with cancer from non-cancer was first presented by Leiner, Wolbeis and co-workers in the 1980s. They considered the fluorescence excitation emission matrix (EEM) of a diluted blood serum sample as a base for pattern recognition to monitor the health status of a person. The hypothesis was that, due to the high sensitivity of fluorescence spectroscopy, it would be possible to observe even small deviations in the fluorescence spectrum from "normal" healthy subjects to diseased subjects (Leiner et al. 1983, 1986; Wolfbeis and Leiner 1985). This hypothesis actually fits well into the present theories of metabonomic based diagnostics. Metabonomic based diagnostics explores metabolites in a biological system and its response to a stress situation such as disease. Metabonomics is often based on non-targeted quantitative and qualitative measurements using nuclear magnetic resonance spectroscopy (NMR) or chromatography [liquid (LC) or gas (GC)] combined with mass spectroscopy (MS) (Nordström and Lewensohn 2010; Zhang et al. 2007). In the present study we explore the possibilities for

introducing fluorescence spectroscopy of blood plasma samples as an alternative metabonomic tool for detection of cancer.

Other publications have followed up on the work from Leiner and co-workers or applied other strategies in using autofluorescence on blood to detect cancer (Hubmann et al. 1990; Kalaivani et al. 2008; Leineret al. 1986; Madhuri et al. 1997, 1999, 2003; Masilamani et al. 2004; Nørgaard et al. 2007; Uppal et al. 2005; Xu et al. 1988). Different approaches have been used; some use extracts or controlled fractions of the plasma, whereas others use the plasma or serum merely diluted or with no sample treatment at all. The studies by Madhuri et al. (1999, 2003) and by Masilamani et al. (2004) use an acetone extract of blood plasma in order to reduce spectral interference in their attempt to measure emission from porphyrins. The results from these studies show elevated levels of porphyrins in cancer patients compared to healthy patients. In the present study we will therefore also have a focus on emission from porphyrins.

Common for almost all of the previous studies was the use of only few or single specific wavelength pairs as opposed to the whole spectral approach combined with chemometrics used in the present study. Only the study from Nørgaard et al. (2007) applied chemometrics in their data analysis, and they got promising results on serum samples from breast cancer patients. The use of chemometrics allows us to use the whole spectrum instead of focusing on single wavelength pairs. Multivariate data analysis/chemometrics is a cornerstone in metabonomics used to extract important information from the complex data output, and hereby hopefully identify specific metabolites with discriminatory or predictive ability (biomarkers) that can be used e.g. for a diagnostic purpose (Ragazzi et al. 2006; Ward et al. 2006). The lack of methods to extract the useful information from the EEMs was exactly a problem for Leiner and co-workers and hence, despite the rather complex EEM measurements, the outcome of their analysis was a simple ratio between two wavelength pairs. In the present study, we apply chemometrics on the fluorescence spectra to discriminate between blood plasma samples from colorectal cancer (CRC) patients and healthy individuals. We apply two different methods of data analysis; one which has been applied previously using the raw spectra as input to the classification model, and one where we extract underlying chemical information from the spectra by Parallel Factor Analysis (PARAFAC) (see materials and methods for a description of PARAFAC). The combination of fluorescence spectroscopy and PARAFAC has not previously been applied in a diagnostic test approach. The combination of PARAFAC and three-way fluorescence data (the EEMs) is especially fruitful, as the parameters of the PARAFAC model can be seen as estimates of the relative concentrations (scores) and the emission and excitation spectra (loadings) of the fluorophores in the sample (Andersen and Bro 2003; Bro 1997). As for conventional NMR and LC–MS this chemical identification opens for fluorescence spectroscopy as a metabonomic tool.

Fluorescence spectroscopy is widely applied in biomarker research though almost solely in the field of labeled fluorescence, where designed fluorescence probes are used to detect the presence of specific biomarkers (Hamdan 2007). In autofluorescence or intrinsic fluorescence, naturally occurring fluorophores are measured with or without minimal sample preparation (Lakowicz 2006). The number of fluorophores in a blood sample is limited compared to the number of compounds detectable by MS and NMR, though among the fluorophores, biologically important compounds are found. In blood for example, the amino acids tryptophan, tyrosine and phenylalanine and also some cofactors and flavonoids NAP, NAD(P)H, FAD are among the fluorophores (Wolfbeis and Leiner 1985). Compared to MS and NMR, fluorescence spectroscopy is highly sensitive and can thus measure concentrations down to parts per billion (Lakowicz 2006). The fluorescent signal from a fluorophore is dependent on the surrounding environment. For example, tryptophan groups in different proteins or on different positions in the same protein can have different excitation and emission maxima, and can thus be distinguished from each other (Abugo et al. 2000). In fact Leiner et al. (1986) showed a difference in the fluorescence from the amino acid tryptophan in human serum from healthy individuals and patients with gynaecological malignancies.

In the practical data acquisition, fluorescence spectroscopy has some advantages compared to both traditional metabonomic techniques. Sample preparation is limited to a minimum of only diluting the sample, and the time of acquisition can be down to few minutes, depending on the spectral area covered and the integration time. A spectrofluorometer can be small and compact compared to MS and NMR, and the price is often much lower. Compared to standard diagnostic tools such as X-ray, MR and CT scanning, fluorescence spectroscopy is very cheap, but at the present stage not a viable alternative. Compared to targeted methods for single biomarkers based on immunochemical tests the onetime investment in fluorescence spectroscopy is, like in MS and NMR, relatively high, but the running costs are much lower, and fluorescence spectroscopy is faster and easy to use.

Some drawbacks of fluorescence spectroscopy are the instrument dependent results that call for spectral correction before they are globally comparable (DeRose and Resch-Genger 2010). The fluorescence intensity is also highly dependent on the overall absorbance of the sample. At low concentrations of fluorophores (and/or low absorbance), the

linear relation between concentration and intensity known from Lambert-Beers law is also valid in fluorescence spectroscopy. At higher concentrations/high absorbance this relation is broken. This phenomenon is called concentration quenching or the inner filter effect (Lakowicz 2006). Blood plasma is highly absorbent, and thus precautions must be taken to avoid or reduce inner filter effects. In the present study the samples are both diluted and undiluted. For the undiluted samples the pathway of the exciting light is reduced to reduce absorbance.

Colorectal cancer is one of the most frequent malignant diseases for both women and men in the western world. In Denmark in 2008, 4194 cases of CRC were diagnosed, which accounted for more than 12% of all malignant diseases (The Danish Cancer Society 2010; The Danish National Board of Health 2010). The 5-year survival rate of CRC patients is approximately 50%, only ovarian, lung, and pancreas cancers have lower rates (UK, national statistics, 2010). The low rate is primarily due to high recurrence frequencies in some patients undergoing intended curative resection and disseminated disease at the time of diagnosis in other patients. At present fecal occult blood test (FOBT) combined with subsequent colonoscopy in those with positive tests is the method of choice for early detection of colorectal cancer. In recent years national screening programs based on FOBT have been introduced in several countries. The FOBT has been criticized for limited compliance rates, which reduce the advantage of the test, and therefore new, improved screening modalities with high compliance rates are urgently needed (Jenkinson and Steele 2010). The only accepted serum biomarker for CRC is carcinoembryonic antigen (CEA), but with sensitivity and specificity values of 0.34/0.93, this is only accepted for prognosis after detection. Other biomarkers have been suggested with similar or better performance, for example free DNA (Flamini et al. 2006) and plasma lysophosphatidylcholine levels (Zhao et al. 2007). None of these biomarkers have yet been clinically accepted. In search for alternative methods with improved detection rates, and/or better compliance rates in screening for CRC, a metabonomic approach with broad unbiased search for changes in the metabolic profile is a possible solution. Interesting results have been published by Ward et al. (2006) by use of MALDI MS. The present paper will explore whether a solution with fluorescence spectroscopy could be an interesting approach.

2 Materials and methods

2.1 Samples

Human plasma samples (sodium citrate anticoagulant) from 308 individuals were used for the experiment. The samples are a part of a larger sample set from a multicentre cross sectional study conducted at six Danish hospitals of patients undergoing large bowel endoscopy due to symptoms associated with CRC (Nielsen et al. 2008). The present sample set is designed as a case control study with one case group (verified CRC) and three different control groups. The three control groups are (1) healthy subjects with no findings at endoscopy, (2) subjects with other, non malignant findings and (3) subjects with pathologically verified adenomas (Lomholt et al. 2009). Each of the groups, case and controls, consisted of samples from 77 individuals. Additional control samples, standardized pooled human citrate plasma, were purchased from 3H-Biomedical AB, Sweden.

2.2 Sample handling and data acquisition

Before measurements, the samples were defrosted on wet ice (0°C) for app. one hour, or until thawed, and each sample was divided in four aliquots of 200 µL to 1 mL for different analytical methods. The divided samples were immediately refrozen at −80°C. The standardized plasma samples were received in 50 mL aliquots, and stored at −80°C. Before use they were thawed at 0°C and divided into aliquots of 300 µL, and refrozen at −80°C. For fluorescence measurements, the samples were defrosted on wet ice (0°C) for app. 40 min.

The samples were measured both undiluted and in a hundred fold dilution in Phosphate Buffered Saline (PBS) (pH 7.4). The diluted samples were prepared immediately after the samples were thawed, and then stored on wet ice (0°C) until measured (app. 20 min). The non diluted fractions of the samples were measured as fast as possible after thawing. Fluorescence spectra were acquired on an FS920 spectrometer (Edinburgh Instruments) with double monochromators and a red sensitive photomultiplier (R928P, Hamamatsu) in a cooled detector house. The EEMs were acquired for the samples using the following settings. Diluted and undiluted samples were measured with excitation from 250 to 450 nm with a 5 nm increment, and emission from 300 to 600 nm with a 1 nm increment. Integration time was 0.05 s. This spectral area consists of light in both the ultra violet and visual area. The ultra violet area is dominated by excitation and emission from the aromatic aminoacids tyrosine and tryptophan hence the fluorescence from proteins. The visual area covers among other things excitation and emission from vitamins and cofactors (for example riboflavin and NAD(P)H) (Wolfbeis and Leiner 1985). In an attempt to capture emission from porphyrins, additional EEMs were acquired from the undiluted samples with excitation wavelengths from 385 to 425 nm with a 5 nm increment and emission wavelengths from 585 to 680 nm with a 1 nm increment, and an

integration time of 0.2 s. Every day a spectrum of the PBS used for dilution was measured with the same settings as the diluted samples. Excitation and emission slit widths were set at 4 nm for all measurements. The fluorescence data were corrected for the wavelength dependent excitation intensity by an internal reference detector in the spectrometer. Likewise the spectra were corrected for instrument dependent emission spectral biases by a correction factor supplied with the instrument. Total time spent for measuring all three EEMs was app. 40 min.

Diluted samples were measured in a 10 × 10 mm quartz cuvette. To reduce inner filter effect in the undiluted samples, these were measured in a 2 × 10 mm quartz cuvette with the 2 mm in the emission direction.

An external cooling system was mounted on the spectrometer keeping the measurement temperature constant at 15°C. To monitor the performance of the fluorescence instrument, a standard plasma sample was measured every day. All spectra were saved as ASCII and exported to Matlab® by an in-house routine. The raw spectra are available for download at http://www.models.life.ku.dk/.

2.3 Data analysis

Some samples were discarded due to either obviously erroneous measurements, or too little sample material. From the three different EEMs acquired, the numbers of samples ready for data analysis were then 301, 295 and 300 from low wavelength undiluted, high wavelength undiluted and diluted, respectively. Before the actual data analysis, the data were subjected to certain signal processing steps meant to appropriately handle and minimize the influence from non-relevant artifacts. When measuring fluorescence EEMs, non-chemical phenomena such as Rayleigh scatter and second order fluorescence may be present (Lakowicz 2006). These were removed and replaced with missing data and zeros using in-house software (Andersen and Bro 2003). For the diluted samples, a background spectrum of the solute PBS, measured the same day as the sample, was subtracted from each sample in order to remove possible Raman scatter (McKnight et al. 2001). All samples were intensity calibrated by normalizing to the integrated area of the water Raman peak of a sealed water sample measured each day prior to the measurements. This converts the scale into Raman units and allows comparison of intensity of samples measured on other fluorescence spectrometers (Lawaetz and Stedmon 2009).

A data reduction/decomposition of the fluorescence EEMs to less complex features was performed using the multi-way decomposition method called PARAFAC. A set of fluorescence EEMs can be seen as a three-way data array ($I \times J \times K$), where I is the number of samples measured (objects), J the number of emission wavelengths, and K the number of excitation wavelengths. Just as PCA is decomposing a two-way data matrix, a three-way data structure can be decomposed by PARAFAC into a number of latent PARAFAC components, by minimizing the sum of squared residuals e in the PARAFAC model (equation below).

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

a_{if} is the ith element of the score vector, b_{jf} the jth element of the loading vector of the emission mode and c_{kf} the kth element of the loading vector for the excitation mode, for the fth PARAFAC component. If the correct number of PARAFAC components is used to decompose data with an approximately true trilinear structure and an appropriate signal to noise value, the solution from the PARAFAC model will give estimates of the true underlying profiles of the variables. This makes PARAFAC perfect for fluorescence spectroscopy when applied on EEMs. The loadings and scores can be treated as estimates of the excitation and emission spectra, and relative concentrations of the fluorophores in the samples respectively (Andersen and Bro 2003; Bro 1997).

PARAFAC models were fitted applying nonnegativity constraints on all parameters in the model (Andersen and Bro 2003); hence the estimated parameters were found in such a way that they would not be negative. Models were validated by split-half analysis (Harshman and DeSarbo 1984) combined with trained judgment of the loadings. PARAFAC models were fitted separately to each of the three sets of EEMs. The score matrices from the PARAFAC analyses were pooled to one matrix with 19 variables, which were subjected to further data analysis. PCA was fitted to get a preliminary overview of the data. Classification models were built using PLS-DA, a PLS regression with the pooled PARAFAC scores as independent X variables and a dummy matrix as the dependent Y variable with ones for samples belonging to the class, and zeros for samples not belonging to the class (Wold et al. 2001). Forward selection was applied for variable selection. For all classifications, the data sets were divided into training and test sets (10–30% in test set). The training sets were used for model building, and the test samples were used for validating the models. During model building of the training sets, the models were cross validated with 10% of the samples randomly removed in each segment and averaging over ten repetitions for each cross-validation run. The test sets for subsequent model validation were randomly selected from the data with the same relative number of samples removed from each class.

As an alternative to building classification models on the three combined PARAFAC score matrices, classification was tried directly with the raw spectra as the independent

variables. Variable selection was applied using Interval PLS (iPLS) (Nørgaard et al. 2000). Before the direct classification the three-way array of EEMs were unfolded to a two-way matrix.

All data analyses were performed in Matlab R2010® (The Mathworks Inc.) and chemometric analyses were performed in PLS_Toolbox v.5.8.2 (Eigenvector Research, Inc).

3 Results and discussion

Spectra from the three setups are seen in Fig. 1. Comparing the spectra from one undiluted sample and a sample diluted 100 times (leftmost and rightmost spectra respectively in the figure) the effect of dilution is clear. In both the raw undiluted sample and in the diluted, the major peak is in the region with fluorescence from the aromatic amino acids tryptophan and tyrosine (phenylalanine is also among the fluorescing amino acids, but it has excitation/emission maximum outside the measured area). For the undiluted sample there are two distinct peaks in that area, whereas in the diluted sample there is only one distinct peak. Furthermore in the undiluted sample a distinct peak is seen with emission maximum at a higher wavelength. The complex peak structure indicates that it is a mixture of several peaks, which could reflect analytes such as NAD(P)H, FAD, Riboflavin etc. (Wolfbeis and Leiner 1985). This peak structure is not apparently visible in the diluted sample.

It is also worth noticing that the intensity of the diluted sample is higher than the raw. This shows that even though the raw sample is measured in a micro cuvette, it still suffers from inner filter effect. Though it was also observed that the dilution in PBS buffer had an effect besides the reduced inner filter effect, a slight blue shift was observed in emission following excitation at 295 nm in the diluted samples. This might be explained by a slight change in the configuration of the proteins, which can change the emission profile.

The high wavelength area of the undiluted samples was measured separately in order to capture possible fluorescence from porphyrins. In the diluted samples this area gave no signal and was therefore not measured. In Fig. 1, middle plot, the high wavelength area primarily shows the descending tail of a peak with maximum outside the measured area, but a closer inspection of the EEM reveals a little bump at app. 405/610 nm which is in accordance with literature values of porphyrin fluorescence (Madhuri et al. 2003).

In order to monitor the performance of the fluorescence spectrometer, a standard plasma sample was measured every day. The standard deviation among these standard samples was up to five times lower than the standard deviation for the real samples, indicating good performance of the instrument and consistent sample handling, and at the same time revealing a large biological variation among the real samples.

On each of the three measured areas, a PARAFAC model was fitted. Due to the high complexity of the plasma matrix and the large biological variation in the samples, a large number of PARAFAC components was expected, which makes modelling more challenging. For the undiluted samples in the main spectral area (excitation from 250 to 450 nm, emission from 300 to 600 nm), ten PARAFAC components were chosen. To the spectra from the diluted samples, a model of six PARAFAC components was fitted. Only a reduced area of the spectra from the diluted samples was used, as the highest emission and excitation wavelengths did not contribute positively to the model. To the last selected area, the high wavelength area of the undiluted samples, a three component PARAFAC model was fitted. The number of PARAFAC components reflects the chemical rank of the system. For each component we get a set of loadings and scores, which are estimates of the excitation and emission profiles for the underlying chemical compounds. The excitation and emission loadings for the three models are seen in Fig. 2. Many of the components can be identified chemically but some are more difficult and even impossible to assign to specific chemical analytes. Despite the large number of PARAFAC components it is possible that some of these peaks reflect more than one chemical compound and the non-Gaussian peak shape of some of the loadings supports this.

In case of "just" making a model to discriminate between cancer and non cancer the issue would be to; objectively and in an unsupervised manner reflect the underlying variation, and then chemical assignment is of secondary concern. On the other hand if we at the same time want to gain knowledge about the reason for the discrimination and hereby move fluorescence spectroscopy into the world of metabonomics, chemical identification is an important parameter. A perfect PARAFAC model will give loadings which are estimates of the underlying excitation and emission spectra, and therefore we expected more unambiguous loadings with better options for chemical assignment. The reason for such non-ideal behaviour can be a low signal of some analytes, correlation between different compounds or non-linear behaviour due to quenching and similar phenomena. Given the relatively low number of samples and that some of the samples are not diluted, it is actually impressive that the PARAFAC models come out as chemically interpretable as they do. Still, we anticipate that the interpretability would be possible to improve if many more samples were included in

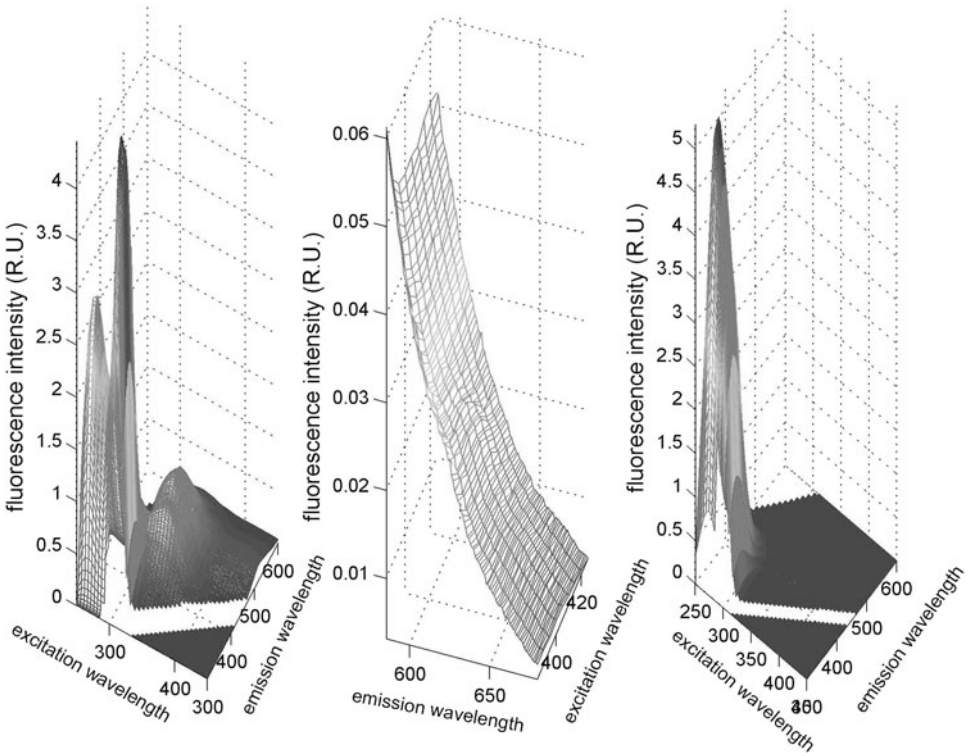


Fig. 1 Different EEMs recorded on one sample. *Left*: undiluted sample in main spectral area. *Middle*: undiluted sample in high wavelength area (notice the *axes* are different from the two other). *Right*: sample diluted 100 times in PBS

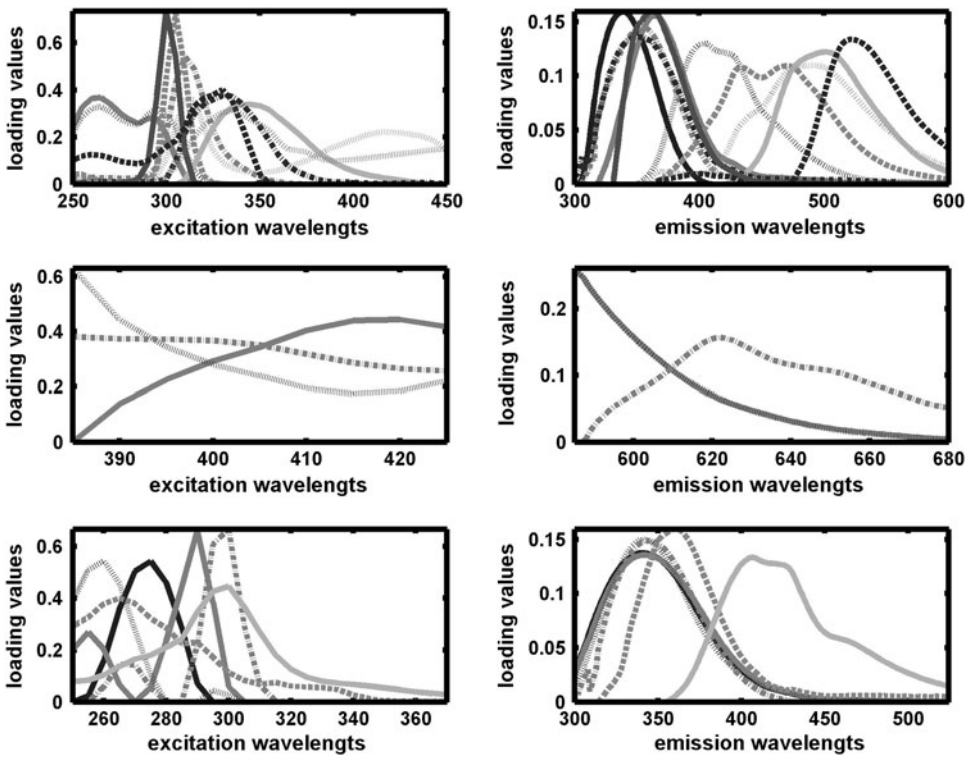


Fig. 2 PARAFAC excitation and emission loadings from the three datasets. *Upper*: undiluted main area. *Middle*: undiluted high wavelength area. *Lower*: diluted main area

the model and possibly also by using targeted standard addition of hypothesized analytes in the modelling phase.

Qualified presumptions on the chemical origin of some of the loadings are made. In both the undiluted and the diluted samples, several loadings are seen with excitation maximum from 250 to 305 nm, and emission maximum from app. 330 to 350 nm. In this region, fluorescence from protein-bound tryptophan is strong. The emission from

tryptophan can shift when the polarity of the microenvironment changes, hence tryptophan which is bound to different proteins, or at the internal or external parts of a protein, can give rise to different emission maxima. In fact, literature values are reported for tryptophan emissions from 307 to 355 nm (Vivian and Callis 2001). This can explain the numerous peaks for tryptophan emission. Some of the excitation loadings fit well with excitation of tyrosine (app. 265 nm) whereas there is no emission loading supporting the presence of tyrosine emission (app. 300 nm). Energy transfer from excited state tyrosine to tryptophan is a known phenomenon and a reasonable explanation of the absent emission from tyrosine (Lakowicz 2006).

The peaks with maximum at higher wavelengths in both the undiluted and diluted samples can possibly be assigned to compounds such as NAD(P)H, FAD and FMN. In the model from the high wavelength region, it is worth noticing that the little, hardly visible "bump" in the pure spectra gives a clear component with excitation/emission maximum at 400/620 which is in agreement with literature values for porphyrins. There are two other components in this model. One has excitation maximum at 420 nm, but emission maximum outside the measured area, and the other has both excitation and emission maxima outside the measured area. The loadings are in agreement with some of the peaks in the undiluted "main" area (two rightmost peaks in Fig. 2 upper right), and could be tentatively assigned to compounds such as NADH or flavins.

The score matrices from the three PARAFAC models are "pooled" into one common score matrix. This matrix now contains all the quantitative information extracted from the fluorescence measurements. Thus we have reduced the complex spectra with several thousand variables to a matrix with 19 variables consisting of estimated relative concentrations of the underlying chemical compounds of the plasma samples. This matrix is now the input to a classification analysis. Note that absolutely no information about the health status of the patients has been used for building the PARAFAC models. This is important from a validation point of view, as it ensures that the matrix is simply an unbiased representation of the raw data.

3.1 Classification

The combined score matrix is used for building classification models. An initial exploratory PCA analysis of the score matrix explains 52% of the variation in the first three components and needs more than 12 components to explain 95% of the variation. The somewhat low explained variation is most likely due to the biological variation in the data and shows that the 19 PARAFAC scores are not overly redundant. No clear separation of cancer and control samples is found by the PCA analysis. There is thus no unsupervised direction in the variable space directly separating cancer from controls and hence the major part of the variation in the data is not related to the cancer/non cancer issue at all. Supplementary information such as age, gender, smoking habits, and co-morbidity could not explain further of this variation either. It is most likely just individual differences.

The score matrix with 19 variables was used as input to a PLS-DA classification model. During model building, some samples were removed as outliers based on evaluation of residuals and Hotellings T^2 (Jackson 1991). Classification models were built for all combinations of cancer and control and also control/control. Models are cross validated and the models are tested on a set of samples left out during model building. The huge biological variation from the raw data is still reflected in the extracted 19 variables in the score matrix. Therefore it makes sense to apply variable selection to select those variables of the 19 that reflect the variation relevant for discriminating cancer and non-cancer. We applied forward selection on the calibration data to find the optimal variables for classification. In the different models the number of variables was reduced from 19 variables to between five and 15 variables.

Results from the different models with sensitivity and specificity values for the cross validated and the tested models as well as area under the receiver operating characteristic (ROC) curve are seen in Table 1. A PLS-DA model with all the three control groups pooled to a common control versus the cancer patients gives an area under the ROC curve of 0.69 with optimal sensitivity and specificity values of 0.70 in the cross validated model, and similar values of 0.73 and 0.77 validated on new samples. Similar values are obtained on models with cancer vs. controls from the group of healthy individuals with no findings, and cancer vs. other non malignant findings. These models give areas under the ROC curves of 0.75 and 0.77, and sensitivity and specificity values between 0.73 and 0.80. In the models of cancer vs. adenomas, the area under the curve, sensitivity and specificity values are at the same level as the model with all controls. The results are to some extent surprising as one would expect it to be easier to discriminate between individuals with no findings and cancer, than between individuals with adenomas and cancer. Models of the different controls against each other give poor models with area under the curve values of 0.5–0.6. Even though they have different imbalances (adenomas or other non malignant findings), the controls are thus not much different from a fluorescence point of view. This result is important for future work of building better diagnosis models, as it underlines that the essential differences found in this study are related to cancer, non-cancer. In a different study on the same samples searching for differences in plasma levels of soluble urokinase

Table 1 PLS-DA models for classification of different classes based on the PARAFAC scores

Groups	Sensitivity CV	Specificity CV	AUC CV	Sensitivity predict	Specificity predict
Crc vs. no	0.68	0.84	0.75	0.73	0.77
Crc vs. onf	0.79	0.73	0.76	0.79	0.73
Crc vs. ade	0.73	0.74	0.77	0.92	0.63
Ade vs. no	0.57	0.55	0.50	0.45	0.43
Ade vs. onf	0.47	0.75	0.57	0.47	0.47
Onf vs. no	0.63	0.58	0.59	0.53	0.40
Crc vs. all controls	0.70	0.70	0.69	0.74	0.71

Crc cancer, *No* no findings, *Onf* other non malignant findings, *Ade* adenomas, *All* all three control groups, *CV* cross validated

plasminogen activator receptor (suPAR), the level of discrimination between cancer and other non malignant findings was better than between cancer and no findings. The discrimination between cancer and adenomas was less significant in this study (Lomholt et al. 2009).

The sensitivity and specificity values in Table 1 are found as the optimal value (maximizing the sum of the two). In diagnostic models, a high specificity value is often preferred as this reduces the number of false positives. For the models cancer vs. other non malignant findings and cancer vs. no findings we get sensitivity values of 0.48 and 0.43 at specificity values of 0.9. The result achieved by use of fluorescence spectroscopy and PARAFAC is thus comparable to the performance of the known biomarkers for CRC; CEA that has sensitivity and specificity values of 0.34 and 0.93.

The table above shows the results of the different classification models. The different models are based on different data, and thus use different variables for classification. A score and a loading plot for the classification model of cancer vs. other non malignant findings based on the PARAFAC scores are seen in Fig. 3. As expected from the sensitivity and specificity values, there is not a perfect separation between the two classes. However, there is a tendency towards separation along the diagonal from the second to forth quadrant in the score plot of the first vs. third PLS-DA component. From the loading plot we can see which variables are important for this separation. The loadings are likewise separated along a diagonal, with samples that are positively correlated to the "cancer direction" and samples negatively correlated to the "cancer direction" or positively correlated to the control samples; in this case the samples with other non malignant findings. A similar exercise can be done for all models.

Common for the models with cancer vs. one or all groups of controls is that the variables 1, 2, 8, 16 and 19 for several of the models are negatively correlated to the cancer direction, and likewise variables 6, 7 and 10 are positively correlated to the cancer direction. These variables are thus important in the discrimination between cancer and controls, though a model based on only those variables does not perform as well as models with more variables. The excitation and emission loadings from components seven and 10 which are positively correlated to cancer and likewise from components eight and 17 which are positively correlated to the controls are shown in Fig. 3 (lower plot). From the excitation and emission loadings these variables can most likely be assigned to tryptophan (variables 7 and 17) or tyrosine, with energy transfer to tryptophan (variables 1 and 4). They have pair wise similar excitation loadings, but the tryptophan emissions in the "cancer variables" are all shifted to shorter wavelengths (blue shift) compared to the "control variables". This confirms the findings from Leiner et al. (1986) who also experienced a blue shift in tryptophan emission in blood from cancer patients.

As opposed to what was expected, variable 3 (excitation/emission at 400/620), which corresponds to porphyrin, was not correlated to cancer. Several studies have shown elevated porphyrin levels in the blood from cancer patients (Madhuri et al. 2003; Masilamani et al. 2004; Xu et al. 1988). In this study all the subjects were included due to symptoms associated with CRC, and thus, even though three of four do not have cancer, some cellular biochemical imbalance might be expected, and therefore elevated levels could be expected in some of these controls. Additionally, the studies showing porphyrin to be important used acetone extracts of either blood plasma or cells, and not pure blood plasma as in the present study.

In the above models, PARAFAC scores were included from measurements on both diluted and undiluted samples, and as explained earlier there are some important effects of dilution. Fluorescence measurements on the undiluted samples may suffer from inner filter effect due to the high absorbance from the plasma samples. Diluting the samples induce physical/chemical changes in the plasma causing blue shift in the spectra. We found that variables from both the diluted and undiluted measurements were important for detecting cancer. Modelling only on scores from the diluted or undiluted samples gave similar but slightly worse

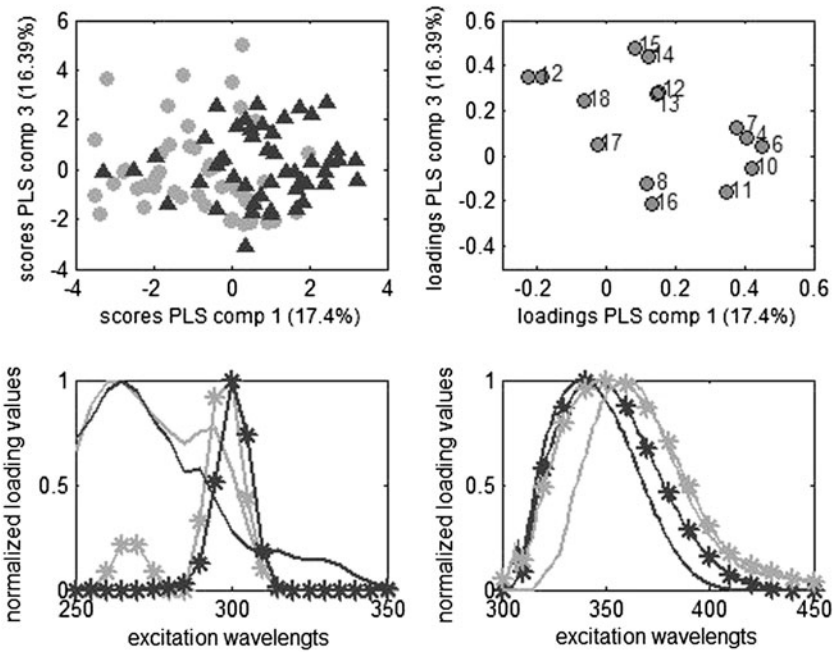


Fig. 3 *Upper left*: PLS-DA score plot of the first vs. third PLS-DA component from the model cancer vs. other non malignant findings on PARAFAC loadings. *Triangles* are cancers and *circles* are controls. *Upper right*: corresponding loading plot. *Lower*: selected PARAFAC excitation (*left*) and emission (*right*) loadings. *Dark gray line* (loading #7) and *dark grey* with *asterisk* (loading #10) are correlated with cancer, *light gray* (#8) and *light gray* with *asterisk* (#17) are correlated with control samples

models compared to the combination of scores from the diluted and undiluted samples, thus predictive power is gained by including both. From an analytical point of view, measuring only on the undiluted samples would be preferred as it makes the measurements faster and simpler to perform. Additionally there is a risk that the changes in sample matrix due to dilution could break some of the cancer specific correlations/interactions and thus make discrimination more difficult. A more thorough study addressing this could be interesting. In fact in analysis of the raw spectra (see below) better models were obtained using only the undiluted samples.

3.2 Classification on the raw data

A study similar to this on breast cancer by Nørgaard et al. (2007) applied discrimination only on the raw spectra. The authors did recommend applying more advanced techniques such as PARAFAC on the spectra but did not pursue this. Recall that we have used PARAFAC here, in order to provide more direct chemical information on how a possible classification can come about. Nevertheless, it is interesting to see whether we have gained anything from a quantitative point of view by applying PARAFAC on the data. Hence, classification models were built directly on the raw spectra as well. We have analyzed both diluted and undiluted samples individually and combined, and achieved similar results. However, the results from the undiluted measurements were slightly better than the alternative results, and are thus the only ones presented below. In Table 2 the results from the classifications based on the raw spectra are shown. Compared to the results based on the PARAFAC scores, these classification models perform equally well and these results are thus also comparable to the performance of CEA. Again the models on control vs. control perform worse than the cancer vs. control models. As for the models based on the PARAFAC scores we have applied variable selection on the models. Different variables are used for the models, but some of the same areas are represented in all four models.

Although it is possible to trace the original wavelengths behind the variables, these do not give the same intuitive information compared to the PARAFAC loadings. The scores and loadings for the model classifying cancer and other non malignant findings (Fig. 4) show a fairly good separation between the two groups in the first and fifth components. The loadings can be traced back to wavelengths around maxima for tryptophan, and the loading for the fifth component has a second derivative-like shape, which can be connected to the shift in the spectra from control to cancer that was shown above in the models based on PARAFAC scores. The results are thus similar, which was expected as it is originally the same data. Still, the extracted features by PARAFAC make the interpretation more straight forward and more comprehensive.

4 Conclusion

We have introduced excitation emission matrix fluorescence measurements on human blood plasma combined with multivariate data analysis as a potential alternative method to discriminate CRC patients from healthy controls, and controls with other cellular imbalances than cancer. With

Table 2 Results from the PLS-DA on the raw unfolded spectra

Groups	Sensitivity CV	Specificity CV	AUC CV	Sensitivity predict	Specificity predict
Crc vs. no	0.64	0.79	0.73	0.73	0.67
Crc vs. onf	0.73	0.79	0.75	0.73	0.73
Crc vs. ade	0.78	0.71	0.74	0.64	0.87
Ade vs. no	0.68	0.61	0.63	0.33	0.63
Ade vs. onf	0.84	0.34	0.55	0.70	0.33
Onf vs. no	0.45	0.82	0.62	0.20	0.82
Crc vs. all controls	0.69	0.7	0.73	0.67	0.83

Crc cancer, *No* no findings, *Onf* other non malignant findings, *Ade* adenomas, *All* all three control groups

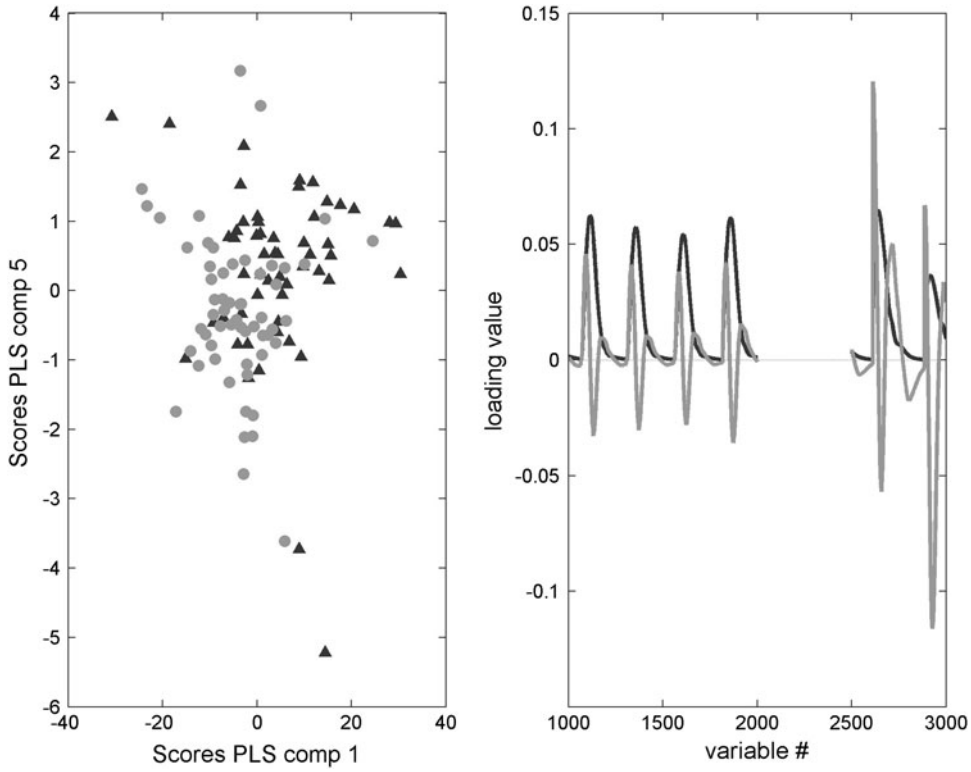


Fig. 4 *Left*: score plot of the first component vs. the fifth component for the PLS-DA model on cancer (*triangles*) vs. other non malignant findings (*circles*) on the raw spectra. *Right*: loadings from the first component (*dark gray*) and the fifth component (*light gray*)

sensitivity and specificity values of app 0.75 on a test set, the results are comparable to the known biomarker CEA. Previous studies with fluorescence spectroscopy have obtained similar results on other types of cancer but with a smaller number of samples. We obtained similar results in regards to discrimination whether we applied classification directly on the raw unfolded spectra or extracted estimates of the underlying fluorophores by use of PARAFAC. By the latter method, however, we obtained better conditions for a chemical interpretation/understanding of the results. We could see a blue shift in the tryptophan emission from cancer patients as one of the reasons for discrimination, a phenomenon described earlier in the literature. The use of PARAFAC on the fluorescence data to extract qualitative and quantitative chemical information from the human blood plasma samples, and base classification on this information is an example on how fluorescence spectroscopy can be used as a tool for metabonomic research. Compared to biomarker tests, fluorescence spectroscopy is an inexpensive alternative, and with minor sample preparation it is easy to perform the analysis. Further research is needed but we believe that there is room for fluorescence spectroscopy as metabonomic tool in cancer research.

Acknowledgments The VILLUM FOUNDATION is thanked for funding Anders Juul Lawaetz. Abdelrhani Mourhib is thanked for his laboratory assistance. Knud Nielsen, Randers Hospital, Søren Laurberg, Aarhus Hospital, Jesper Olsen, Glostrup Hospital and Hans B Rahr, Odense Hospital, are acknowledged for their contribution to the original protocol.

References

Abugo, O. O., Nair, R., & Lakowicz, J. R. (2000). Fluorescence properties of rhodamine 800 in whole blood and plasma. *Analytical Biochemistry, 279*, 142–150.

Andersen, C. M., & Bro, R. (2003). Practical aspects of PARAFAC modeling of fluorescence excitation-emission data 1. *Journal of Chemometrics, 17*, 200–215.

Bro, R. (1997). PARAFAC. Tutorial and applications 1. *Chemometrics and Intelligent Laboratory Systems, 38*, 149–171.

DeRose, P. C., & Resch-Genger, U. (2010). Recommendations for fluorescence instrument qualification: The new ASTM standard guide. *Analytical Chemistry, 82*, 2129–2133.

Flamini, E., Mercatali, L., Nanni, O., Calistri, D., Nunziatini, R., Zoli, W., et al. (2006). Free DNA and carcinoembryonic antigen serum levels: An important combination for diagnosis of colorectal cancer. *Clinical Cancer Research, 12*, 6985–6988.

Hamdan, M. H. (2007). *Cancer biomarkers*. Hoboken: John Wiley and sons.

Harshman, R. A., & DeSarbo, W. S. (1984). An application of PARAFAC to a small sample problem, demonstrating preprocessing, orthogonality constraints, and split-half diagnostic techniques. In H. G. Law, et al. (Eds.), *Research methods for multimode data analysis* (pp. 602–642). New York: Praeger.

Hubmann, M. R., Leiner, M. J. P., & Schaur, R. J. (1990). Ultraviolet fluorescence of human sera.1. Sources of characteristic differences in the ultraviolet fluorescence-spectra of sera from normal and cancer-bearing humans 1. *Clinical Chemistry, 36*, 1880–1883.

Jackson, J. E. (1991). *Operations with group data*. Hoboken: John Wiley & Sons, Inc.

Jenkinson, F., & Steele, R. J. C. (2010). Colorectal cancer screening—methodology. *Surgeon-Journal of the Royal Colleges of Surgeons of Edinburgh and Ireland, 8*, 164–171.

Kalaivani, R., Masilamani, V., Sivaji, K., Elangovan, M., Selvaraj, V., Balamurugan, S. G., et al. (2008). Fluorescence spectra of blood components for breast cancer diagnosis. *Photomedicine and Laser Surgery, 26*, 251–256.

Lakowicz, J. R. (2006). *Principles of Fluorescence Spectroscopy*. New York: Springer.

Lawaetz, A. J., & Stedmon, C. A. (2009). Fluorescence intensity calibration using the Raman scatter peak of water. *Applied Spectroscopy, 63*, 936–940.

Leiner, M. J., Schaur, R. J., Desoye, G., & Wolfbeis, O. S. (1986). Fluorescence topography in biology. III: Characteristic deviations of tryptophan fluorescence in sera of patients with gynecological tumors. *Clinical Chemistry, 32*, 1974–1978.

Leiner, M., Schaur, R. J., Wolfbeis, O. S., & Tillian, H. M. (1983). Fluorescence topography in biology. 2. Visible fluorescence topograms of rat sera and cluster-analysis of fluorescence parameters of sera of Yoshida ascites hepatoma-bearing rats. *IRCS Medical Science-Biochemistry, 11*, 841–842.

Lomholt, A. F., Hoyer-Hansen, G., Nielsen, H. J., & Christensen, I. J. (2009). Intact and cleaved forms of the urokinase receptor enhance discrimination of cancer from non-malignant conditions in patients presenting with symptoms related to colorectal cancer. *British Journal of Cancer, 101*, 992–997.

Madhuri, S., Aruna, P., Summiya Bibi, M. I., Gowri, V. S., Koteeswaran, D., Schaur, R. J., et al. (1997). Ultraviolet fluorescence spectroscopy of blood plasma in the discrimination of cancer from normal. *Proceedings of SPIE, 2982*, 41–45.

Madhuri, S., Suchitra, S., Aruna, P., Srinivasan, T. G., & Ganesan, S. (1999). Native fluorescence characteristics of blood plasma of normal and liver diseased subjects. *Medical Science Research, 27*, 635–639.

Madhuri, S., Vengadesan, N., Aruna, P., Koteeswaran, D., Venkatesan, P., & Ganesan, S. (2003). Native fluorescence spectroscopy of blood plasma in the characterization of oral malignancy. *Photochemistry and Photobiology, 78*, 197–204.

Masilamani, V., Al-Zhrani, K., Al-Salhi, M., Al-Diab, A., & Al-Ageily, M. (2004). Cancer diagnosis by autofluorescence of blood components. *Journal of Luminescence, 109*, 143–154.

McKnight, D. M., Boyer, E. W., Westerhoff, P. K., Doran, P. T., Kulbe, T., & Andersen, D. T. (2001). Spectrofluorometric characterization of dissolved organic matter for indication of precursor organic material and aromaticity. *Limnology and Oceanography, 46*, 38–48.

Nielsen, H. J., Brunner, N., Frederiksen, C., Lomholt, A. F., King, D., Jorgensen, L. N., et al. (2008). Plasma tissue inhibitor of metalloproteinases-1 (TIMP-1): a novel biological marker in the detection of primary colorectal cancer. Protocol outlines of the Danish-Australian endoscopy study group on colorectal cancer detection. *Scandinavian Journal of Gastroenterology, 43*, 242–248.

Nordström, A., & Lewensohn, R. (2010). Metabolomics: Moving to the Clinic. *Journal of Neuroimmune Pharmacology, 5*, 4–17.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., & Engelsen, S. B. (2000). Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy, 54*, 413–419.

Nørgaard, L., Soletormos, G., Harrit, N., Albrechtsen, M., Olsen, O., Nielsen, D., et al. (2007). Fluorescence spectroscopy and chemometrics for classification of breast cancer samples—a feasibility study using extended canonical variates analysis. *Journal of Chemometrics, 21*, 451–458.

Ragazzi, E., Pucciarelli, S., Seraglia, R., Molin, L., Agostini, M., Lise, M., et al. (2006). Multivariate analysis approach to the plasma protein profile of patients with advanced colorectal cancer. *Journal of Mass Spectrometry, 41*, 1546–1553.

The Danish Cancer Society. (2010). http://www.cancer.dk.

The Danish National Board of Health. (2010). *National screening for tyk- og endetarmskræft*. The Danish National Board of Health, 1.

Uppal, A., Ghosh, N., Datta, A., & Gupta, P. K. (2005). Fluorimetric estimation of the concentration of NADH from human blood samples 1. *Biotechnology and Applied Biochemistry, 41*, 43–47.

Vivian, J. T., & Callis, P. R. (2001). Mechanisms of tryptophan fluorescence shifts in proteins. *Biophysical Journal, 80*, 2093–2109.

Ward, D. G., Suggett, N., Cheng, Y., Wei, W., Johnson, H., Billingham, L. J., et al. (2006). Identification of serum biomarkers for colon cancer by proteomic analysis. *British Journal of Cancer, 94*, 1898–1905.

Wold, S., Sjostrom, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*, 109–130.

Wolfbeis, O. S., & Leiner, M. (1985). Mapping of the total fluorescence of human-blood serum as a new method for its characterization. *Analytica Chimica Acta, 167*, 203–215.

Xu, X. R., Meng, J. W., Hou, S. G., Ma, H. P., & Wang, D. S. (1988). The characteristic fluorescence of the serum of cancer-patients. *Journal of Luminescence, 40–1*, 219–220.

Zhang, Xuewu, Li, Lin, Wei, Dong, Yap, Yeeleng, & Chen, Feng. (2007). Moving cancer diagnostics from bench to bedside. *Trends in Biotechnology, 25*, 166–173.

Zhao, Z., Xiao, Y., Elson, P., Tan, H., Plummer, S. J., Berk, M., et al. (2007). Plasma Lysophosphatidylcholine Levels: Potential Biomarkers for Colorectal Cancer. *Journal of Clinical Oncology, 25*, 2696–2701.

Metabolomics (2012) 8:S122
DOI 10.1007/s11306-011-0321-4

ERRATUM

Erratum to: Fluorescence spectroscopy as a potential metabonomic tool for early detection of colorectal cancer

Anders Juul Lawaetz · Rasmus Bro · Maja Kamstrup-Nielsen · Ib Jarle Christensen · Lars N. Jørgensen · Hans J. Nielsen

Published online: 4 June 2011

Erratum to: Metabolomics
DOI 10.1007/s11306-011-0310-7

The original version of this article unfortunately contained a mistake. In the Materials and methods section the number of samples ready for data analysis is incorrectly given as 301, 295 and 300 for the three groups. The correct numbers are 299, 299 and 289. The incorrect numbers come from an intermediate step in the analysis where some irrelevant standard samples were included. Similar quality of results was obtained on datasets of the size incorrectly given in the paper. The data from the paper can be downloaded from our homepage. http://www.models.life.ku.dk/datasets.

The online version of the original article can be found under doi:10.1007/s11306-011-0310-7.

A. J. Lawaetz (✉) · R. Bro · M. Kamstrup-Nielsen
Department of Food Science, Quality and Technology, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark
e-mail: ajla@life.ku.dk

I. J. Christensen
Finsen Laboratory, Copenhagen Biocenter, Copenhagen, Denmark

L. N. Jørgensen
Department of Surgery, Bispebjerg Hospital, University of Copenhagen, Copenhagen, Denmark

H. J. Nielsen
Department of Surgical Gastroenterology, Copenhagen University Hospital Hvidovre, Hvidovre, Denmark