



Ph.D. Thesis by Marianne Dyrby February 2004

> Ph.D. thesis by Marianne Dyrby Cand.techn.al

> > Supervisors:

Associate professor Lars Nørgaard Associate professor Søren Balling Engelsen

Chemometrics Group, Food Technology Department of Dairy and Food Science The Royal Veterinary and Agricultural University Rolighedsvej 30, DK-1958 Frederiksberg, Denmark

Vice President Ulla Grove Sidelmann

Protein Engineering Novo Nordisk A/S Novo Nordisk Park, DK-2760 Måløv, Denmark

Ph.D. Thesis 2004 © Marianne Dyrby

ISBN 87-7611-035-4

Printed by Samfundslitteratur Grafik, Frederiksberg, Denmark

Preface

The work presented here is the result of my three years of Ph.D. studies. The work was made possible through a Ph.D. scholarship from The Royal Veterinary and Agricultural University (KVL), where I have performed a major part of the work. The Danish pharmaceutical company Novo Nordisk A/S is a partner in the project, and nearly all NMR measurements presented in this thesis have been performed at their site in Måløv, Denmark.

During my Ph.D. study an informal collaboration was established with Ph.D. student Martin Petersen from Department for Human Nutrition at KVL to pursue common interests and goals. Last but not least, a six-months stay at Centre for Magnetic Resonance (CMR) at University of Queensland, Brisbane, Australia was realised.

I am thankful to many people in the Chemometrics Group, especially to Lars Nørgaard, Søren B. Engelsen and Rasmus Bro for constructive discussions during the preparation of papers and this thesis. My fellow Ph.D. student Martin Petersen for help with getting samples, understanding human physiology and revision of papers. At Novo Nordisk, I would like to thank Ulla Grove Sidelmann, Henrik Toft Pedersen and Dorrit Baunsgaard for NMR support and project collaboration. The people at CMR should be mentioned, especially Andrew Whittaker and Ian Brereton for helping me during my stay in the group, and to my fellow Ph.D. students and post docs making my stay in Australia memorable.

1 February 2004

Summary

Multivariate data analysis (also known as chemometrics) has had significant influence on the increasing use of spectroscopic techniques in the Danish food research and production. Chemometric methods have the ability to extract important features (underlying chemical or physical phenomena) from complex data and once developed for a specific application perform rapid and stable analyses.

High-field ¹H NMR spectroscopy is an extremely versatile measurement technique. For complex biological samples like foodstuffs and biofluids, NMR has the important property of requiring minimal sample preparation thus avoiding changes in the biological matrix and yielding detailed analytical profiles from the whole biological sample. However, the practical use is limited by a number of factors including the complexity of data collection, handling and analysis. Therefore, methods to reduce the complexity and increase the information gained from NMR analyses are sought and for this purpose multivariate data analysis is a natural choice.

In this Ph.D. project, various multivariate data analytical methods were applied to one- and two-dimensional NMR spectra as well as pseudo two-dimensional data in the form of NMR time series in order to further develop methodology for the analysis of complex NMR data.

One of the simplest multivariate techniques, multivariate curve fitting, was applied to a selected region of the NMR spectra of commercial carrageenan products and the relative contents of five main carrageenan types were determined [Paper II].

Using Partial Least Squares (PLS) regression, main and subfractions of lipoproteins were quantified from the one-dimensional diffusion-edited NMR spectra of human blood plasma [Paper I]. Prediction errors were lower than those obtained in other studies, and the assessment of individual risk of coronary heart disease based on predicted concentrations was comparable to that based on reference concentrations.

Two-dimensional diffusion-edited NMR spectra of human blood plasma samples were subjected to multi-way chemometric analysis using multi-way PLS (N-PLS) and PARAllel FACtor analysis (PARAFAC) [Paper III]. N-PLS calibrations

yielded prediction errors similar to the ones obtained with PLS, however using a much smaller sample set. The analysis using PARAFAC yielded four components representing lipoproteins of different sizes with spectral contributions as well as self-diffusion coefficients that were in accordance with known properties of lipoprotein main fractions. These results clearly suggest that PARAFAC has potential for the analysis of diffusion-edited as well as other two-dimensional NMR spectra.

The multi-way Tucker model was applied to pseudo two-dimensional NMR data in the form of NMR time series [Paper IV]. The analysis of the metabolic response to three model toxins using Tucker analysis of the NMR spectra of rat urine was successful with respect to identifying major metabolites related to toxicity. Some conclusions were complementary to those reported from studies of the same substances using other data analytical techniques showing the advantage of applying and comparing different data analytical approaches.

The work presented in this Ph.D. thesis shows examples of the successful analysis of complex one- and two-dimensional NMR spectra using two-way as well as multi-way multivariate data analytical methods. The project provides material for the further development within the field of combining multivariate data analysis and high-field NMR spectroscopy.

Sammenfatning

Multivariat dataanalyse (også kaldet kemometri) har haft stor betydning for udbredelsen af spektroskopiske teknikker i den danske levnedsmiddelforskning og -industri. Kemometriske metoder har evnen til at udtrække vigtige karakteristika (underliggende kemiske eller fysiske fænomener) fra komplekse data, og når først en given applikation er udviklet, kan analyser udføres hurtigt og stabilt.

Høj-felts ¹H NMR spektroskopi er en bredt anvendelig og fleksibel målemetode. For komplekse biologiske prøver såsom mange fødevarer samt f.eks. blod og urin har NMR den specifikke fordel, at den kræver minimal prøveforberedelse. Herved undgås ændringer i den oprindelige biologiske matrix og en detaljeret analytisk profil kan opnås fra hele den biologiske prøve. I praksis er anvendeligheden begrænset af et antal faktorer, bla. af kompleksiteten af data indsamling, håndtering og analyse. Derfor er der stor interesse i udvikling af metoder, som kan reducere kompleksiteten og forøge graden af anvendelig information fra NMR analyser, og til dette formål er multivariat dataanalyse et naturligt valg.

I dette Ph.D. projekt blev forskellige metoder indenfor multivariat dataanalyse anvendt på et- og to-dimensionelle NMR spektre samt på pseudo to-dimensionelle data i form af NMR tidsserier for at videreudvikle metodikker til data analyse of komplekse NMR data.

En af de mest simple multivariate teknikker, multivariat kurve tilpasning, blev anvendt på et udvalgt område af NMR spektre af kommercielle carrageenan produkter og det relative indhold af fem hovedtyper af carrageenan blev bestemt [Paper II].

Hoved- og subfraktioner af lipoproteiner blev kvantificeret ved hjælp af Partial Least Squares (PLS) regression på et-dimensionelle diffusions-editerede NMR spektre af humant blodplasma [Paper I]. De opnåede prædiktionsfejl var lavere end i tidligere publicerede studier og den individuelle risiko for hjerte-kar sygdomme blev ikke vurderet væsentligt anderledes ved brug af prædikterede koncentrationer i forhold til brug af reference koncentrationer.

To-dimensionelle diffusions-editerede NMR spektra af humant blodplasma blev analyseret med multi-vejs metoderne N-PLS og PARAllel FACtor analysis (PARAFAC) [Paper III]. N-PLS kalibreringer gav prædiktionsfejl i samme

størrelsesorden som blev opnået vha. PLS dog baserede på et meget mindre prøvesæt. PARAFAC analysen gav fire komponenter, som kunne vises at repræsentere lipoproteiner af forskellig størrelse med spektrale bidrag og diffusions-koefficienter, som stemte godt overens med kendte værdier for lipoprotein hovedfraktioner. Disse resultater viser, at PARAFAC har potentiale for at kunne bruges til analyse af diffusions-editerede samt andre typer todimensionelle NMR spektre.

Multi-vejs metoden Tucker blev anvendt på pseudo to-dimensionelle NMR data i form af NMR tidsserier [Paper IV]. Analysen af det metaboliske respons på tre toxiner vha. Tucker analyse af NMR spektre af rotte urin var succesfuld mht. at identificere de primære metabolitter involveret i det toxiske respons. Nogle resultater var komplementære til resultater opnået med andre data analytiske metoder på studier af de samme toxiner, hvilket viser fordelen i at anvende og sammenligne forskellige analytiske tilgange til data.

I denne Ph.D. afhandling er præsenteret eksempler på succesfuld analyse af komplekse et-og to-dimensionelle NMR spektre vha. to-vejs såvel som multi-vejs multivariate dataanalyse metoder. Projektet lægger op til en videre udvikling indenfor kombinationen af områderne multivariat dataanalyse og høj-felts NMR spektroskopi.

Abbreviations

1D	one dimensional – a vector
2D	two dimensional – a matrix
3D	three dimensional – a cuboid
NMR	nuclear magnetic resonance
PARAFAC	parallel factor analysis
PCA	principal component analysis
PLS	partial least squares
RF	radio frequency
RMSECV	root mean square error of cross validation
RMSEP	root mean square error of prediction
T_1	longitudinal or spin-lattice relaxation time constant
T_2	transverse or spin-spin relaxation time constant

List of papers

Paper I

Dyrby, M., Petersen, M., Engelsen, S.B., Nørgaard, L., and Sidelmann, U.G. Quantification of lipoprotein subfractions using ¹H-NMR and chemometrics. In "Magnetic resonance in food science: Latest developments" edited by G.A. Webb, P.S. Belton, A.M. Gil and D.N. Rutledge, The Royal Society of Chemistry, 2003.

Paper II

Dyrby, M., Petersen, R.V., Larsen, J., Rudolf, B., Nørgaard, L., and Engelsen, S.B. Towards on-line monitoring of the composition of commercial carrageenan powders. *Carbohydrate Polymers, Submitted*.

Paper III

Dyrby, M., Petersen, M., Whittaker, A., Nørgaard, L., Bro, R., and Engelsen, S.B. Analysis of lipoproteins using 2D diffusion-edited NMR spectroscopy and multi-way chemometrics. *Magnetic Resonance in Medicine, Submitted*.

Paper IV

Dyrby, M., Baunsgaard, D., Bro, R., and Engelsen, S.B. Multi-way chemometric analysis of the metabolic response to toxins monitored by NMR. *Chemometrics and Intelligent Laboratory Systems, Submitted.*

Additional papers by the author

Dyrby, M.; Engelsen, S.B.; Nørgaard, L.; Bruhn, M.; and Lundsberg-Nielsen, L. Chemometric quantitation of the active substance (containing C=N) in a pharmaceutical tablet using Near-Infrared (NIR) transmittance and NIR FT-Raman spectra. *Applied Spectroscopy*, 56(5), 579-585, 2002.

Dyrby, M.; Westergaard, N.; and Stapelfeldt, H. Light and heat sensitivity of red cabbage extract in soft drink model systems. *Food Chemistry*, 72(4), 431-437, 2001.

Table of contents

Prefa	i	
Summ	ii	
Samm	iv	
Abbreviations		vi
List of papers		vii
Table	of contents	viii
1. IN	TRODUCTION	1
1.1	BACKGROUND	1
1.2	DESCRIPTION OF THE PROJECT	2
1.3	OUTLINE OF THE THESIS	2
2. MI	ULTIVARIATE DATA ANALYSIS	3
2.1	THE MULTIVARIATE ADVANTAGE	3
2.2	PROPERTIES OF DATA	4
2.3	TWO-WAY METHODS	5
2.3.1	Principal Component Analysis	6
2.3.2	Two-way calibration	7
2.4	MULTI-WAY METHODS	8
2.4.1	The Tucker model	9
2.4.2	The PARAFAC model	10
2.4.3	Multi-way calibration	11
2.5	VALIDATION	11
2.5.1	Test set validation	12
2.5.2	Cross validation	12
2.5.3	Split-half analysis	13

Multivariate Analysis of High-Field NMR Data in Food and Medicine			
3. HIGH-FIELD NMR SPECTROSCOPY			
3.1 BASIC THEORY	15		
3.2 PULSE EXPERIMENTS	23		
3.2.1 Single pulse experiment	24		
3.2.2 Diffusion-editing	24		
3.2.3 Water suppression	31		
4. APPLICATIONS OF MULTIVARIATE ANALYSIS TO NMR DATA	33		
4.1 PREPARING NMR DATA FOR MULTIVARIATE ANALYSIS	33		
4.2 HARD MODELS	35		
4.3 SOFT MODELS	39		
4.3.1 PCA on 1D NMR data	39		
4.3.2 PLS on 1D NMR data	40		
4.3.3 N-PLS and PARAFAC on 2D NMR data	41		
4.3.4 Tucker on pseudo-2D NMR data	44		
5. CONCLUSIONS			
6. REFERENCES			

1. Introduction

1.1 Background

Various spectroscopic methods are widely used both in food research and in the food industry i Denmark. Combined with multivariate data analysis (also known as chemometrics) these are very useful analytical tools for monitoring the quality of food products [1]. Industrial applications include process control in the sugar, hydrocolloid, confectionary, meat and dairy production. Applications within food research are numerous and include studies on the structural changes during processing of foods, e.g. cooking of meat and baking of bread. Within medical science, the methods are not quite as widespread as in food science, but especially in pharmaceutical production multivariate data analysis in combination with spectroscopy is gaining increasing use [2]. Examples from this area include identification of raw materials, release of finished goods and process control.

In the past, high-field Nuclear Magnetic Resonance (NMR) spectroscopy has mainly been concerned with the elucidation of chemical structure in solution, but today high-field NMR is emerging as a powerful exploratory tool for probing biochemical and physical processes. High-field NMR represents the most versatile and information rich spectroscopy for the analysis of foods and biological systems, yet its full potential remains to be exploited. In the following, the abbreviation NMR will refer to high-field NMR unless otherwise stated.

One of the reasons for the limited use of multivariate data analysis on NMR data is the large amount of user-interaction needed to process the data before the application of multivariate data analysis. Within food science the combination of multivariate data analytical methods and NMR spectroscopy has mainly been limited to authentication and adulteration problems [3]. In medical research, one of the major areas of application is within "metabonomics", where chemometric analysis of NMR data of biofluids has gained increasing attention and has contributed with detailed information on the effects of disease on the metabolism [4].

The chemometrics research group at Food Technology, KVL is a major Danish centre for the theoretical and applied research in multivariate data analysis. The group has collaborations within many fields, i.e. within food science, sensory science, human nutrition, pharmaceutical and medical science and thus is experienced in working with many different types of data. Apart from "traditional" two-way multivariate data analysis the group has pioneered the area of multi-way methods.

1.2 Description of the project

The aim of this research project was to improve the applicability and exploit the unique possibilities of modern high-field NMR spectroscopy to probe complex physico-chemical systems typical for foods and biological samples using chemometrics.

State-of-the-art NMR equipment and advanced chemometric methods including multi-way methods were combined in an attempt to enhance the interpretation and quantitative information of these complex systems – a combination which has enormous potential for the extraction of new and relevant information. Furthermore, some work was done in the area of preparing NMR data for data analysis.

1.3 Outline of the thesis

It is the aim of this thesis to present the work performed in the area of application of multivariate data analysis to high-field NMR spectroscopy within food science and medical science during this Ph.D. study.

The presentation is divided into three main parts: Chapter 2 gives an introduction to multivariate data analytical techniques; Chapter 3 summarises NMR basic theory including selected pulse experiments; and Chapter 4 deals with the application of multivariate data analysis to NMR data. Chapters 2 and 3 are provided as background for the work presented in this thesis, and do not cover all aspects of multivariate data analysis and NMR, respectively.

2. Multivariate data analysis

2.1 The multivariate advantage

When using spectroscopic techniques for chemical analysis, traditionally univariate data analytical methods have been used, where a single response variable is measured or selected from the whole spectrum and only this variable is analysed. Statistical methods are then applied for classification or linear regression is performed and the calculated regression line gives an indication of how well the measured variables describe the dependent variable in question.

There are a few problems that should be considered when using univariate methods. One is that out of a large number of measured variables only one or a few variables are selected for the analysis. Especially for complex samples there is a huge loss of information in using only a few variables out of a whole spectrum. Furthermore, this approach requires extensive *a priori* knowledge about the spectrum of this type of sample in order to choose the right variable and thus requires a hypothesis about the results to be found, which in turn will reduce chances for innovation.



Figure 1 Examples of problems that arise when univariate analysis is applied to spectral data.

Another problem occurs if different components in the sample give signals in the same region as the analyte or a variable baseline offset is present in the data. In cases of overlapping signals these will be inseparable when only a single variable is measured and the content will be estimated too high due to the several components adding to the signal intensity. A simplified example of some of the

problems that can occur with univariate analysis is given in Figure 1. From the figure it is obvious that if only the variable marked with a dot is recorded, spectrum A will be estimated to have a higher content of the compound in question than spectrum C, even though the two responses have equal amplitude relative to the signal baseline. If the whole spectrum is recorded, the baseline offset of spectrum A is easily detected and the necessary action can be taken to correct for this offset. For this reason univariate analysis of data including interacting or coinciding signals cannot always be expected to produce useful and trustworthy results.

Furthermore, when dealing with complex samples like foods or biological samples, univariate analysis of spectral data is often impossible. The signals are so complex, that the signal of a single chemical component cannot be identified, or the property of interest is something that cannot be described by a single peak. For this type of samples it is important to look at the sample as a whole and to analyse the spectral data taking the matrix of the sample into consideration.

An important aspect of multivariate data analysis is the possibility to detect outliers. Outliers are samples that do not conform with the majority of samples in a data set and will therefore have a detrimental effect on the calculated model. An example of this is shown in Figure 1 where one sample has a different spectral signal (spectrum D) than the others. In univariate analysis there is no way to determine that something is wrong with the sample, while in multivariate analysis, the sample will easily be detected as an outlier. The reason for a sample being an outlier can be one of many, such as wrong sample preparation or erroneous handling of the sample, but it can also be due to instrumental artefacts or just that the sample does not follow the expected sample variation. It is important to make an effort to find out why a sample is an outlier before removing it from the data set. Once a sample has been shown to be an outlier, it should be removed from the data set and the calibration model must be recalculated.

2.2 Properties of data

Several properties of data are essential for the successful application of multivariate analysis. These properties will only be briefly discussed here, whereas the practical considerations needed when applying multivariate data analysis to NMR data will be discussed in Section 4.1.

Multivariate data analysis is based on the extraction of latent components or underlying *common structures* in the data. For the analysis to make sense, the data must have common underlying structures! In the example in Figure 1, the structure, i.e. the peak, which is common for the three samples (A, B and C), is not present in the outlier (spectrum D). This could have two explanations: either this sample *is* different and contains another chemical compound than the rest, or due to some error in the spectral measurement, the *spectral axis* is different for this sample, i.e. the different position of the signal is due to an instrumental shift.

For quantitative spectral analysis, the *intensity axis* must be the same for all samples, since a basic assumption is the obedience of the law of Lambert-Beer. This is trivial for most other spectroscopic techniques, but in NMR the spectra are often scaled in the spectrometer, and this scaling must be removed before multivariate analysis. In calibration, the absence of a common linear intensity axis leads to erroneous predictions and in exploratory analysis to erroneous interpretation.

Another prerequisite for successful multivariate data analysis is that the data conform with the selected model. An assumption that applies to almost all of the methods described in the following is that the data are *low-rank bilinear* or *trilinear*. A practical explanation of the concept of bi-/trilinearity is that a given chemical compound gives one and only one spectral signal, and that a given spectral signal can come from one and only one chemical compound. Regarding the low-rank, in many cases the purpose of multivariate data analysis is to reduce the dimensionality in order to increase interpretability, so if the data are high-rank the gain in using multivariate data analysis is marginal.

2.3 Two-way methods

When analysing simple chemical mixtures or complex biological samples using spectroscopy, the data will typically consist of the spectra of a number of similar samples, which contain varying amounts of a number of chemical components each of which has a unique spectral signal. Thus the data will optimally be bilinear and two-way bilinear models can be applied.

2.3.1 Principal Component Analysis

Principal Component Analysis (PCA) [5,6] is a decomposition method where a set of corresponding scores (T) and loadings (P) are calculated for a data set (X), leaving only a residual (E):

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathrm{T}} + \mathbf{E}$$
 Equation 1

A schematic representation of the model is shown in Figure 2. The scores (T) correspond to the concentrations and the loadings (P) to the spectral signals of the common underlying structures. The outer product of a score vector and the corresponding loading vector is called a *dyad*, a *bilinear component* or in the PCA model a *principal component*.



Figure 2 A two-component Principal Component Analysis (PCA) model.

The idea of PCA is to extract principal components expressing orthogonal variation present in the data set until only random noise is left in the residual. The variation is extracted in decreasing order with the largest variation first called the first principal component and so forth. PCA is a powerful tool used to explore the variation present in the data set and often the origin of the extracted variances can be interpreted directly through plots of the scores and loadings by utilising known information about the samples.

In PCA, as in all bilinear models, there is the well-known problem of rotational freedom [7]. Thus the spectral loadings in PCA can be interpreted to yield information about the pure spectra of the analytes measured, but the loadings will not directly represent the pure spectra because of the rotation problem and the orthogonality constraint.

2.3.2 Two-way calibration

Spectral measurements can be used for multivariate predictive purposes, in a similar way as for exploratory purposes as shown for PCA. The common principle for calibration methods is that the sample spectra and the so-called dependent variable (e.g. the concentration of a specific substance) are known for a series of samples. The parameters in a mathematical model are estimated based on these samples so that the model can be used to predict the dependent variable from the spectra of samples for which this value is unknown.

A number of different possibilities exist for the calculation of multivariate prediction models [8], where the simplest approach is Classical Least Squares (CLS), the model of which is:

$$\mathbf{X} = \mathbf{Y} \cdot \mathbf{K} + \mathbf{E}$$
 Equation 2

The principle behind CLS is to view the recorded spectra as linear combinations of the pure spectra of the chemical constituents. If X (the spectra) and Y (the concentrations) are known for all constituents, including interferences, impurities etc., the pure spectra (K) can be estimated from the calibration, given that the number of samples is larger than the number of chemical components. If the pure spectra of all analytes are known, Y can be calculated directly from X and K, and the method involves no calibration. The main drawback of CLS is the need to know the concentrations or pure spectra of *all* chemical components, which is practically impossible for complex samples. Furthermore, non-linearities cannot be accounted for.

Another approach is the multivariate extension of univariate linear regression, Multiple Linear Regression (MLR). MLR is based on the assumption that the dependent variable can be calculated as a weighted sum of the spectral variables. The relationship between the spectral variables and the dependent variable is:

$$\mathbf{y} = \mathbf{X} \cdot \mathbf{b} + \mathbf{e}$$
 Equation 3

where the vector **b** represents the direct correlation between each spectral variable (X) and the dependent variable (y). As for CLS, direct and indirect interferences must be included in the calibration samples, but in MLR the concentrations need not be known. Drawbacks of MLR are that the number of variables used must be

smaller than the number of samples. Furthermore, for optimal results the spectral variables must be selective for either analyte or interference (i.e. spectra must have limited overlap between analyte and interferences) and the single spectral variables cannot be collinear with other variables.

Rather than using selected spectral variables (X) it is possible to use scores (T) for a given number of components calculated by PCA as the regression variables in MLR. This approach is called Principal Component Regression (PCR) [9,10]. The advantage of this approach compared to MLR is that the entire variation in the data set is compressed by PCA into a few new variables (represented by the scores) and the entire variation in the data set is thus used in the prediction model. PCA can easily handle spectral data sets with correlated variables, and interaction effects are taken into account in the resulting models.

Another possibility is to use Partial Least Squares (PLS) regression [11,12] where the variation in the spectra directly co-varying with the variation in the dependent variable is extracted. The variation structure calculated by PLS is not necessarily identical to that used in PCR (i.e. T is different). This is due to the fact that in PCA the spectral data matrix is decomposed independently of the dependent variable, whereas in PLS the data matrix is decomposed to obtain maximum covariance between the scores and the dependent variable.

2.4 Multi-way methods

Some data of chemical analytical problems are naturally arranged in a three-way matrix, i.e. a data cube, or even higher order data matrices. Examples of three-way data include 2D or pseudo-2D spectra measured for a series of samples, e.g. 2D diffusion-edited NMR spectra or NMR spectra measured with regular time intervals on a series of samples. It is possible to analyse these problems with two-way methods using unfolding of the three-way data matrix to a two-way matrix, where two dimensions are mixed [13]. However, ignoring the multi-way structure of the data and treating it as an ordinary two-way data set can lead to models that are less robust, less interpretable, less predictive and non-parsimonious. Furthermore, the more noisy the data, the more beneficial the use of multi-way methods will be [13]. For the analysis of multi-way data, several dedicated methods have been developed, which are described in the following.

2.4.1 The Tucker model

The Tucker3 model (henceforth referred to simply as Tucker), or multi-mode PCA, is one of the most basic multi-way models in chemometrics although originating in psychometrics [14,15,16]. In the case of three-way data the Tucker model is shown schematically in Figure 3.

The data cube \underline{X} is decomposed into a number of components, but as opposed to PCA the number of components can be different for the three modes (i.e. the dimensions or directions). If \underline{X} has dimensions I x J x K, A will have dimensions I x P, B dimensions J x Q, C dimensions K x R, and the core array \underline{G} has dimension P x Q x R, where P, Q and R are the number of components extracted in each of the three modes.



Figure 3 The Tucker model is a weighted sum of outer products between the components stored as columns in A, B and C.

The extracted components are characterised by a set of scores (**A**), and two sets of loadings (**B** and **C**). The loadings **B** and **C** can both be spectral loadings as would be the case for 2D NMR data, or can be e.g. spectral loadings and time loading as is the case for NMR time series. The model of the original data is the weighted sum of outer products between the components in **A**, **B** and **C**. The matrix <u>**G**</u> is called the core array and represents the values with which the single component products are weighed.

The interpretation of Tucker models is complicated by the presence of the core array allowing for interaction between different components. For example, in twoway PCA score one interacts with loading one only, but in Tucker it has interactions with all loadings in the other modes. The value and sign of each core element must be taken into account and Tucker models are thus more difficult to interpret than classical PCA. Furthermore, the Tucker model has rotational freedom like PCA, and thus the interpretation of loadings is not always straightforward.

The force of the Tucker model is as a generalized PCA on multi-way data with the same capabilities as PCA to compress variation, extract features, explore data, generate parsimonious models etc. There are few examples on the use of Tucker models for data that can be assumed to be generated by a process according to the Tucker model as opposed to the PARAFAC model (Section 2.4.2), which coincides with several physical models [13].

2.4.2 The PARAFAC model

The PARAllel FACtor analysis (PARAFAC) model also originates in psychometrics [17,18]. The PARAFAC model is a special case of the Tucker model, where the Tucker model is more general due to the possibility of an unequal number of loading vectors in each mode and because of a more general core array. In the PARAFAC model the number of extracted components in each dimension must be the same (P) and the core array, corresponding to the Tucker model, is a cube (P x P x P) with ones on the superdiagonal and zeros in all other positions. Therefore, the first score-vector interacts only with the first loading vector in each dimension and so forth, and the PARAFAC model can be represented as a simple three-way extension of PCA as shown in Figure 4. As for the Tucker model, PARAFAC is not limited to three-way data but can handle data sets of higher dimensionality.



Figure 4 A two-component PARAFAC model.

The PARAFAC model is based on the decomposition of the original data into trilinear components (*triads*) in a similar way to the bilinear components (*dyads*)

extracted in PCA. An obvious advantage of the PARAFAC model is the uniqueness of the solution. If the data are approximately trilinear, the components are intrinsically unique up to permutations and scaling. This means that the true underlying structures, e.g. the pure analyte spectra, will be found directly if the right number of components is used and the signal-to-noise ratio is appropriate [17]. Furthermore, the scores (or concentrations) determined by PARAFAC need not be calibrated to an external standard but only scaled according to a single reference sample in order to represent estimates of the actual concentrations of the analytes [19].

2.4.3 Multi-way calibration

For calibration on multi-way data sets, an analogue to the two-way method PCR is a possibility, using a Tucker or a PARAFAC model to decompose the data and then regressing (scaling) the scores from this model unto the dependent variable using MLR.

Furthermore, a general multi-way extension of the PLS regression model exists, that can handle multi-way data. This model is called N-PLS [8] and similar to two-way PLS, components are extracted in \mathbf{X} such that the scores have maximal covariance with the dependent variable in \mathbf{y} .

The N-PLS model is unique in the sense that it consists of successively estimated one-component models, each of which is unique. However, the uniqueness in this case will seldom imply that true underlying structures will be recovered, because the model does not reflect a physical or theoretical model [13].

2.5 Validation

Validation is an important aspect of all data analysis. In multivariate data analysis, the purpose of validation is to ensure that a suitable model is obtained, i.e. that the model is valid for future samples and is not an expression of trends in the present samples only. In practice, validation is used to determine whether the right type of model is applied, to determine the correct number of components to use, to detect outliers and for evaluation of the estimated model parameters.

Decomposition methods, such as PCA, Tucker and PARAFAC, should be validated to ensure that the interpretation is correct and based on a reasonable number of components. For calibration models, such as PLS and N-PLS, the emphasis will be more on ensuring that the estimated error of predictions is as close as possible to the true value.

Apart from validation using e.g. one of the methods listed in the following sections, all models should be chemically validated looking at how well the model reflects the physical or chemical phenomena under investigation, i.e. the appropriateness of the model. Chemical validation can include external knowledge about the shape of spectral signals, which should be reflected in the spectral loadings, or about groups within the samples, which should be reflected in the scores.

2.5.1 Test set validation

Test set validation employs the use of a set of samples that were not used in the initial model development, called an independent test set, preferably sampled at a different point in time but spanning the same variation. The validity of the model is tested applying the model to these new samples and evaluating the explained variation, spectral residuals, scores, and for calibrations the prediction error expressed as the Root Mean Squared Error of Prediction (RMSEP):

$$RMSEP = \sqrt{\frac{1}{n} \left(\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \right)}$$
Equation 4

where n is the number of samples, y_i is the reference value and \hat{y}_i is the predicted value. The problem with test set validation is merely the need for a large number of samples with the same variation as in the first set of samples, a demand that is often not practically feasible. For calibrations, the need for more reference analyses to determine the values of the dependent variable is costly and often time-consuming. If the number of original samples allows, it is possible to split these samples into two subsets – a calibration set and a test set, which is used as described above.

2.5.2 Cross validation

Cross validation can be used when sampling a new set is not practically possible and the number of original samples does not allow for splitting them into two subsets spanning the same variation. For calibration, the simplest type is leave-one-out or full cross validation. This employs leaving one sample out at a time and calculating a model on the remaining samples. The sample left out is then predicted using the model and these predictions are used for estimating the prediction error expressed as the Root Mean Squared Error of Cross Validation (RMSECV) defined equal to RMSEP.

Several other schemes than full cross validation can be used, e.g. dividing the samples into segments and leaving out one segment at a time. Which type of cross validation is optimal for a given data set depends on the number of samples, the variation in the samples etc., and must be assessed for each data set. However, full cross validation has been found optimal (with a trend towards overfit) for data sets with few samples [20].

For decomposition methods, i.e. PCA, PARAFAC and Tucker, leaving out single or several entire sample spectra is not advisable, and a different approach is used leaving out single spectral data points across samples instead [21].

2.5.3 Split-half analysis

Split-half analysis is primarily used for PARAFAC models and consists of analysing different subsets of the data independently. Due to the uniqueness of the PARAFAC model, the same loadings will be obtained in the non-splitted modes from models of any suitable subset of the data. If too many or too few components are chosen the model parameters will differ when the model is fitted to different data sets [13].

Split-half analysis can be applied to other types of models, but then the loadings will not be identical for different subsets, and instead the space spanned by the loadings must be compared, making the application less straight-forward.

3. High-Field NMR spectroscopy

The theory of Nuclear Magnetic Resonance (NMR) spectroscopy was put forth in the mid-1930's by the Dutch physicist Gorter [22], but it was not until 1945 that the American physicists Bloch and Purcell discovered NMR in the form that is known today [23,24]. The possibilities of using NMR in chemistry were not appreciated until 1950, when Proctor and Yu discovered the chemical shift which enabled elucidation of molecular structures of organic compounds [25]. Although originating in physics, it is in chemistry that NMR has provided the greatest results. The main field of use is within structure elucidation of complex molecules and especially for protein analysis in solution NMR is an indispensable tool. The use of NMR as a general analytical tool has developed especially during the last 15 years.

3.1 Basic theory

Spin and energy levels

About 2/3 of all isotopes possess a spin angular momentum, the magnitude of which is dependent on the size of the angular momentum quantum number (*I*), which is commonly referred to simply as *spin*. The spin of a nucleus depends on the mass of the isotope, and nuclei with even mass and even charge numbers possess no spin angular momentum, i.e. I = 0. Such nuclei cannot be used for NMR, since it is the nuclear spin property that enables NMR.

When a nucleus that possesses a spin different from zero is placed in a magnetic field, the nucleus will occupy one of a number of energy levels where the number of levels available depends on the value of *I*. The proton (¹H) is the most abundant NMR nucleus and has spin $I = \frac{1}{2}$. For such nuclei there are two different energy levels that the spins can occupy when placed in a magnetic field: $I = -\frac{1}{2}$ and $I = \frac{1}{2}$. This corresponds to an orientation parallel ($I = -\frac{1}{2}$) or anti-parallel ($I = \frac{1}{2}$) with the applied magnetic field. Figure 5 shows the two possible energy levels for protons and the dependence of the applied magnetic field B_0 . In the rest of this presentation of the NMR theory protons will be used as a model.

The difference in the energy levels shown in Figure 5 is equal to

$$\Delta E_0 = \mathbf{h} \cdot \mathbf{v}_0 \qquad \qquad \text{Equation 5}$$

where *h* is Planks constant and v_0 is the frequency of the excitation pulse at which transition between the two energy levels is induced. The frequency v_0 is referred to as the *resonance frequency* or the *Larmor frequency* and will depend on the type of nucleus and the magnetic field strength following the equation

$$\mathbf{v}_0 = \frac{\mathbf{\gamma} \cdot \mathbf{B}_0}{2 \cdot \pi}$$
 Equation 6

Here γ is the gyromagnetic ratio, which is a constant for a given nucleus; for protons $\gamma = 2.6752 \cdot 10^8 \text{ T}^{-1} \cdot \text{s}^{-1}$.



Figure 5 Representation of the two possible energy levels for nuclei with spin = $\frac{1}{2}$ and the influence of the strength of the applied magnetic field (B_0) on the energy difference ΔE .

When at equilibrium in a magnetic field, the protons will be distributed between the two energy states according to the equation known as the Boltzmann distribution.

$$\frac{N_{\alpha}}{N_{\beta}} = \exp\left(\frac{\Delta E}{k \cdot T}\right)$$
 Equation 7

In this equation N_{α} and N_{β} represent the populations in the parallel and antiparallel states, respectively, k is the Boltzman constant and T the absolute temperature. According to this distribution there will be a small excess of protons in the parallel state, since this is the energetically more favourable state. It is this excess that is measured in NMR and simple calculations show that at room temperature and a magnetic field strength of 14.1 T (equal to $v_0 = 600$ MHz for protons) the excess in the parallel state will be approximately 50 protons out of one million. Since it is these 50 protons that give rise to the NMR signal, a large total number of protons are required in order to generate an appreciable signal, and it should be obvious why NMR is commonly described as insensitive compared with other spectroscopic methods.

Precession and net magnetisation vector

When the sample is unaffected by an external magnetic field, the orientation of the spins will be randomly distributed in all directions. However, when the sample is placed in a magnetic field, the spins will align as described in the above equation. Under the influence of the external magnetic field the spins will start to precess about the direction of the magnetic field, as shown in Figure 6 for four spins in the parallel state and two in the anti-parallel state.



Figure 6 Precession of individual spins around the external magnetic field and the resulting net magnetisation vector.

The net magnetisation is the sum of all single spins, and since the spins are randomly distributed about B_0 , the *net magnetisation vector* is positioned exactly on the axis of the external magnetic field. The bold arrow in Figure 6 represents this net magnetisation vector, which is characterised by a small positive component

on the axis of the external magnetic field corresponding to the small excess of protons in the parallel state and no components on the other axes of the coordinate system.

In order to facilitate the description of the spin manipulation that gives rise to the NMR signal it is normal to ascribe axes to the NMR instrument as shown in Figure 7. Thus the z-axis is normally ascribed to the direction of the magnetic field (B_0) and the net magnetisation vector at equilibrium, while in modern NMR instruments, both the x- and the y-axis are equipped with coils for applying RF pulses and for detection.



Figure 7 Diagram of the coordinates normally ascribed to the NMR instrument. The drawing also depicts the RF- and detection coils and a perturbation of the equilibrium system by a RF pulse denoted a 90° pulse.

RF pulse

When the sample is at equilibrium in the magnetic field there is no observable signal since the net magnetisation vector, M, has no component in the xy-plane where the signal is to be detected ($M_x = M_y = 0$). The purpose of a RF pulse is to perturb the system to obtain an observable NMR signal, converting magnetisation on the z-axis to magnetisation in the xy-plane. A RF pulse of a duration precisely long enough to flip the net magnetisation vector into the xy-plane will generate the largest possible signal (the horizontal bold arrow in Figure 7). This pulse is referred to as a 90° pulse and a pulse with a duration long enough to flip the net magnetise z-axis is referred to as a 180° pulse. Once

in the xy-plane the magnetisation vector will continue to precess about B_0 thus inducing an oscillating signal in the detectors along the x- and y-axes.

Relaxation

After a perturbation by a 90° RF-pulse the spin system will lose coherence in the xy-plane, a process known as spin-spin or transverse relaxation. The loss of coherence is exponential and is described by a time constant called T_2 . This process is due to energy exchange between spins as well as inhomogeneities in the magnetic field, which will particularly influence molecules with mobile protons and high diffusion rates such as water. Simultaneously with the loss of coherence in the xy-plane the protons will seek to regain equilibrium orientation along the z-axis due to the influence of the magnetic field. The time it takes for the protons to regain equilibrium distribution between the two energy states depends on the probability of energy exchanges occurring between the spins and their environment (the lattice). This is characterised by a relaxation mechanism referred to as longitudinal or spin-lattice relaxation described by a time constant T_1 .

Shielding and chemical shift

Protons in different positions of a molecule do not experience exactly the same magnetic field due to the effect known as *shielding*. When brought into a magnetic field, the motion of the electrons orbiting around an atom are perturbed in such a way that a magnetic field is induced in the sample that opposes the external field. Hence the sample becomes magnetized and modifies the field. In a molecule electrons are hindered in their rotation around a particular atom by the presence of other atoms, and are therefore not capable of exerting their maximum shielding effect. Thus, differences in shielding reflect local differences in geometry and electron density in a molecule. Therefore, protons in different chemical environments experience different effective magnetic fields and thus resonate at slightly different frequencies, a phenomenon known as chemical shift. The relationship between the degree of shielding and the resulting resonance frequency is

$$v = \frac{\gamma \cdot B_0}{2 \cdot \pi} * (1 - \sigma)$$
 Equation 8

The term σ is known as the shielding constant and is a small dimensionless number usually expressed in parts per million (ppm). The existence of chemical shift is what enables the NMR spectroscopist to distinguish between nuclei in different chemical environments and makes NMR spectroscopy a powerful tool for the determination of the structure of molecules and as a general chemical analytical tool.

Subtracting the carrier frequency

The net magnetisation vector thus consists of a number of spins that, once in the xy-plane, will precess with slightly different frequencies close to the Larmor frequency. The signal detected in NMR is the current induced in the coils by the precessing magnetization. Since we are not interested in the Larmor frequency but in the difference in shielding and thus chemical shift, in the receiver the Larmor frequency (some MHz) is "subtracted" from the detected frequency for each spin, resulting in what is called the audio signal (chemical shift frequencies, some kHz). In practice, a frequency called the carrier frequency positioned in the middle of the range of frequencies of interest is subtracted instead of the Larmor frequency, so that we get both positive and negative frequencies of the different spins in the sample.



Figure 8 Example of the evolution of two spins relative to the carrier frequency. One has a frequency lower and one higher than the carrier frequency.

Quadrature detection

This is illustrated in Figure 8, where one spin has a frequency of -v and one a frequency of +v with respect to the carrier frequency. To the receiver, this

corresponds to having spins moving both clock-wise and counter-clock-wise in the xy-plane, and this is the reason for using a detection system with two simultaneous channels, which are mutually 90° out of phase.

This system is called quadrature detection and enables the discrimination between a spin with a frequency -v and one with frequency +v, something which is not possible when detection is only done on one axis. This is illustrated in Figure 9 where the signals from the two spins with frequency -v and +v in each of the detectors is depicted. For simplicity, the signals are shown as if no relaxation occurred. As can be seen, the signals from the detector situated on the y-axis are identical for the two spins while the signals on the x-axis are different and thus enables the differentiation of the two frequencies.



Figure 9 The signals of spins with frequencies -v and +v using quadrature detection.

Fourier Transformation

During a modern NMR experiment, the signal is measured in the time domain, i.e. as a function of time. This signal, being a superposition of sinusoids damped due to relaxation, is Fourier transformed to obtain the spectrum in the frequency domain. The signals from the two detectors are handled as the real and imaginary parts of a complex signal, enabling the use of complex Fourier transformation. Prior to, and after, Fourier transformation several cosmetic operations are performed that influence the signal-to-noise ratio of the spectrum, the resolution between peaks, and the correction of artefacts such as deviations from a flat horizontal baseline. These include zero filling, application of window function and baseline corrections, topics that will not be discussed further here.

Phase correction

The result of a complex Fourier transformation is also complex and thus consists of two parts, the real and imaginary part, $S(v) = R(v) + i \cdot I(v)$. Optimally, R(v) is equal to the absorption spectrum, A(v), and I(v) is equal to the dispersion spectrum, D(v), the shapes of which are shown in Figure 10.



Figure 10 The absorption- and dispersion-mode of a single resonance.

However, due to a difference in phase between the receiver and the carrier frequency, this is not exactly the case, and in practice R(v) and I(v) are linear combinations of A(v) and D(v):

$$R(v) = A(v)\cos(\phi) - D(v)\sin(\phi)$$

$$I(v) = A(v)\sin(\phi) - D(v)\cos(\phi)$$

Equation 9

where the term ϕ is called the phase error. The pure absorption and dispersion spectra, which are preferred for analysis, can be obtained though a linear combination of R(v) and I(v), a process known as phase correction:

$$A(v) = R(v)\cos(\phi) + I(v)\sin(\phi)$$

Equation 10
$$D(v) = -R(v)\sin(\phi) + I(v)\cos(\phi)$$

Since the phase error is not constant over the entire spectrum the correction is made frequency dependent:

$$\phi(\mathbf{v}) = \phi_0 + \mathbf{v} \cdot \phi_1 \qquad \qquad \text{Equation 11}$$

where ϕ_0 and ϕ_1 are called the 0th and 1st order phase correction values. In practice, phase correction is done using the instrument software where the operator tries to make the real spectrum look as 'absorption-like' as possible by adjusting the values of ϕ_0 and ϕ_1 .

Referencing

The frequencies in the Fourier transformed spectrum are dependent on the size of the applied field, and thus an alternative axis is defined relative to a reference compound, resulting in the chemical shift axis (in units ppm) on which a given spin always has the same value independent of the size of the applied magnetic field. The most widely used reference substance is tetramethyl silane, Si(CH₃)₄ (abbreviated TMS) or in aqueous solutions the water-soluble sodium salt of trimethylsilyl proprionic acid (abbreviated TSP). TMS/TSP is a suitable reference compound because it exhibits almost maximum shielding so that most sample NMR peaks have a smaller shielding constant (σ) and thus a positive value on the axis compared with the reference compound (per definition 0 ppm). The chemical shift of a sample spin can be calculated as

$$\delta = \frac{v_{\text{sample}} - v_{\text{reference}}}{v_{\text{reference}}} \cdot 10^6 \text{ ppm}$$
Equation 12

In cases where the addition of TMS/TSP to the sample for some reason is not an option, other reference compounds can be used given that the chemical shift value is well-known.

3.2 Pulse experiments

The possibility to design pulse experiments for the selection of specific signals is the basis for the enormous diversity of NMR applications. Here only two types of pulse experiments will be presented, the most simple unselective pulse experiment and a group of pulse experiments designed to measure self-diffusion of molecules in solution. Finally, a short review of solvent suppression techniques will be given, which can be implemented in both the simple and the more complex pulse experiments described.

3.2.1 Single pulse experiment

This pulse experiment is used when the purpose of the experiment is to view all nuclei of a given type in the sample and is schematically described in Figure 11. It employs the application of a 90-degree RF-pulse followed by the measurement of the signal, which is exponentially decreasing due to relaxation. The decreasing signal is generally known as the Free Induction Decay (FID), a name also commonly used as a name for the experiment. The notation 90_x implies that the RF pulse is applied on the x-axis. Note that the timing diagram is not produced to scale, which applies to the pulse experiments described in the following text as well. The duration of a 90-degree pulse is typically in the order of a few microseconds, a 180-degree pulse twice as long, and the data acquisition time may be as long as seconds.



Figure 11 The single pulse experiment

3.2.2 Diffusion-editing

The purpose of diffusion-edited NMR experiments (also known as Diffusion Ordered SpectroscopY, DOSY) can be many including to determine self-diffusion coefficients of molecules in solution, to monitor intermolecular binding, to separate signals from molecules with similar or overlapping NMR signals but different size or to edit the NMR spectrum to represent molecules of a specific size only. Diffusion-editing is usually acquired as a 2D spectrum consisting of a series of 1D spectra, through which the intensity of signals decrease exponentially due to diffusion. The relationship between signal intensity, diffusion coefficient and gradient strength is:

$$\mathbf{I} = \mathbf{q} \cdot \mathbf{M}_0 \cdot \exp\left[-\mathbf{D} \cdot \gamma^2 \cdot \delta^2 \cdot \mathbf{g}^2 \cdot \Delta' - \mathbf{R}\right]$$
 Equation 13

where I is the measured signal intensity, M_0 is the equilibrium magnetisation, D is the diffusion coefficient (m²·s⁻¹), γ is the gyromagnetic ratio (s⁻¹·Tesla⁻¹), δ is the time duration of the applied gradient pulse (s) and g is the applied gradient strength (Tesla·m⁻¹). The term Δ ' is equal to $\Delta - \delta \cdot k$, where Δ is the diffusion time (the time between the two gradient pulses) and k a correction factor, which depends on the shape of the gradient pulses. The term R is a constant that accounts for relaxation and the constant q is a correction factor accounting for the loss of signal when using pulse experiments based on the stimulated echo sequence (see below).

Since small molecules have large self-diffusion coefficients, the signals of these will decrease more in intensity than those of large molecules at a given gradient strength. This is exploited when acquiring a 1D diffusion-edited spectrum, which is especially used for blood plasma, where the signals from small metabolites, e.g. sugars and acids, are fully attenuated so that only the signals from macromolecules, e.g. lipoproteins and other proteins, are seen.

The gradient spin echo

The gradient spin echo pulse experiment shown in Figure 12 is a modification of the original Hahn echo [26] and is the simplest pulse experiment for measuring diffusion.



Figure 12 The gradient spin echo.

After excitation by the 90° RF pulse, the spins are labelled with a positiondependent phase angle using a gradient pulse, thus sensitising the sample to diffusion and flow. A gradient pulse is a pulse which results in different magnetic fields in different parts of the sample. Because the spins are always undergoing random translational motion in solution, some will change position along the zdirection during the successive evolution period or diffusion time. Then the spin magnetization is rotated 180° by a RF pulse, and another gradient pulse identical to the first refocuses the spins that have not changed position during evolution thus forming an echo. In the gradient spin echo experiment the magnetization is transverse, i.e. in the xy-plane, during the entire evolution period. Here it is subject to transverse relaxation (governed by T_2) and signals from especially macromolecules, which generally have small T_2 , will lose intensity due to relaxation during the pulse experiment, leading to low signal-to-noise rations when using the gradient spin echo experiment.

The stimulated echo

To reduce the loss of signal from especially macromolecules due to relaxation, the Stimulated Echo (STE) pulse experiment shown in Figure 13 can be used. The STE sequence essentially replaces the 180° pulse of the gradient spin echo with two 90° pulses, the first of which is placed just after the first gradient pulse thus keeping the time spent as transverse magnetisation (the time period 2τ) short. This pulse flips the magnetization back on the z-axis, where the relaxation is governed by T_1 , and hence minimizes the time the magnetization spends in the xy-plane, where it is governed by T_2 . This is generally more favourable because the ratio T_1/T_2 is greater than or equal to one for ¹H nuclei [27].



Figure 13 The stimulated echo (STE) pulse sequence.

The main drawback of this sequence is that half of the total signal intensity is lost, since only y-magnetisation is flipped onto the z-axis by the second 90° pulse and the x-magnetisation is destroyed. Because the T_1/T_2 ratio is usually greater than one, in most cases the benefits of the STE sequence in terms of less attenuation of the signal due to relaxation generally outweigh the halving of the signal [27].

The Longitudinal Eddy current Delay

In the STE experiment, the time τ will usually be kept as short as possible to minimise loss of signal due to transverse relaxation. The reduction of τ to a value only slightly larger than the gradient pulse duration means that the echo and hence detection comes almost immediately after the second gradient pulse, which can result in gradient pulse induced eddy current effects that will distort the signal.

Eddy currents are electrical currents caused by the gradient pulses. Whenever the magnetic field is changed, eddy currents are induced within the metal structures of the probe and the magnet. The result is slowly decaying magnetic fields opposing the applied gradient, which can be experienced by the sample and therefore lead to distortions resulting from time dependent phase changes. Moreover, the extent of the distortion is dependent upon the strength, and rate of change, of the applied gradient pulse and thus will produce a systematic change in the spectra. The presence of eddy currents during acquisition can be seen as lineshape and baseline errors in the spectra, as well as a too fast decay in signal intensity with increasing gradient strength.

There are several ways to avoid, or at least reduce, eddy currents [28,27], including actively shielded gradients, shaped gradient pulses, and bipolar pulses, which have been included in most modern analysis. Furthermore, if these measures are not sufficient, a delay called the Longitudinal Eddy current Delay (LED) can be added in the pulse sequence before acquisition to allow the eddy currents to decay before acquisition is started [29].

The Double Stimulated Echo

When measurements are done at temperatures above or below room temperature, uneven heating or cooling of the sample can result in the development of convection currents inside the sample due to temperature gradients. Most modern NMR systems introduce air of variable temperature through the base of the sample region and this can easily create a situation where the bottom of the sample tube experiences warmer (or colder when cooling) gas than the top. Especially in nonviscous samples, this results in a temperature gradient along the tube and hence in the formation of convection currents.

Even small temperature gradients can lead to significant loss of magnetization in diffusion experiments and sometimes to its complete extinction or partial inversion. The reason for this is that the different velocities in the sample result in phase distortions and superimposes an oscillating behaviour in the signal decay, which can even create negative signals. The relative effects of convection and diffusion depend in a complex way on a number of parameters, such as sample viscosity, temperature, molecular shape and size, sample geometry, probe design, and filling height, which makes the occurrence of convection seemingly random [30].

An example of the effect of convection currents is shown in Figure 14, in which the attenuation of a specific peak in the spectra of several samples of human blood plasma is shown. The spectra were acquired at 40 °C using the STE sequence with LED. If the attenuation were due to diffusion only, the curves should be linear (cf. Equation 13). For one sample (black line) this is clearly the case, while the decaying oscillatory effect of convection currents is seen on the curves of the other three samples.



Figure 14 Illustration of the detrimental effect of convection currents using a LED sequence. See text for details.

There are several ways to reduce convection currents [27,28]: Using a well designed temperature control system, using a high flow rate of the heating/cooling air, using sample tubes with smaller diameters, reducing the sample height, and spinning the sample tube [31]. In some cases many of these measures might not be

feasible or even when using them the problem might persist. In that case, a convection current compensated experiment [32] may be the final option, and the most commonly used sequence is shown in Figure 15.



Figure 15 The double stimulated echo with LED. All RF-pulses are 90-degree pulses.

This pulse sequence is based on a double stimulated echo (DSTE) configuration, where the gradient direction is reversed in the second echo sequence with respect to the flow, thereby cancelling its effect. The sequence shown incorporates a LED, characterised by the time T_e . Using the double stimulated echo, the signal is again halved compared to the single STE and furthermore the longer pulse sequence makes the effect of relaxation due to T_1 and T_2 more severe. Hence the sensitivity is seriously impaired, but for some samples the effect of convection currents is so strong that there is no choice.

To get a deeper understanding of how convection compensation works, it is necessary to look into the effect of molecular motion in spin echo experiments, which has been extensively described by Callaghan [33]. It can be shown that the motion of spins during a gradient spin echo experiment causes an average phase distortion of

$$\overline{\phi(t)} = r_0 \cdot \int_0^t g^*(t) dt + v \cdot \int_0^t t \cdot g^*(t) dt +$$
Equation 14
$$\frac{1}{2}a \cdot \int_0^t t^2 \cdot g^*(t) dt$$

where r_0 is the average spin position in the sample, v is the average velocity, a the average acceleration, and $g^*(t)$ is the effective gradient defined as $g^*(t) = p(t) \cdot g(t)$, where p(t) is the coherence order and g(t) the applied gradient at time t. An explanation of coherence and coherence order can be found in standard NMR text

books [34,35] and will not be covered here. The three integral terms are equal to the 0th, the 1st and the 2nd moment of the effective gradient in time, $m_k = \int t^k g^*(t) dt$.

From this equation it can be concluded that if the 0^{th} moment is zero, the signal will be insensitive to the mean position of the spins, which is the condition for echo formation and is obeyed in all pulse experiments designed to measure motion and not position. If the 1^{st} moment is zero, the experiment will be insensitive to constant velocity. Nulling the 2^{nd} moment as well will make the experiment insensitive to acceleration.

A simple example of nulling of the 0^{th} and 1^{st} moments is shown in Figure 16. The practical meaning of nulling the 0^{th} moment is simply that the area of the effective gradient taken over the entire pulse sequence must be zero. Nulling of the 1^{st} moment can be described as the effective gradient having to be symmetric in time. In these simple experiments, the effective gradient, $g^*(t)$, is equal to the applied gradient, g(t), since the coherence order is one during the entire experiment. In Figure 16, the 0^{th} moment is clearly zero in both (a) and (b), since the positive and negative parts of the effective gradient, $g^*(t)$, have equal area, and hence an echo will form in both experiments. However, in (a) the 1^{st} moment is not zero, since the positive and negative parts of the effective gradient weighed by the time, $t \cdot g^*(t)$, do not have equal area, while this is the case in (b).



Figure 16 Simple gradient spin echo experiments, which are (a) velocity sensitive and (b) velocity insensitive. Both RF-pulses are 90-degree pulses.

For more complicated pulse sequences as the double stimulated echo, nulling of the 1st moment is done by appropriate selection of coherence-transfer pathways by phase cycling. In the pulse experiment shown in Figure 15 the 0th and 1st moment are zero and hence the position and velocity terms are nulled but not the acceleration term, which is supposedly negligible for most experiments [32].

3.2.3 Water suppression

Many biological samples contain large amounts of water, which gives a very strong signal in NMR being 110 molar in protons. The purpose of water suppression is to obtain agreement between, on one side, the dynamic range of the interesting NMR signals and on the other side the dynamic range of the analogue-to-digital converter and of the receiver. In some cases, it can also be a purpose to avoid the disappearance of small signals on the huge water peak. Various means of suppressing the water signal exist, here only two methods will be discussed, presaturation and WATERGATE.



Figure 17 The general pulse sequence for presaturation.

One of the simplest, most robust and widely used methods is presaturation, where the water signal is dephased through a long weak RF pulse, a pulse which affects only the water peak. The presaturation pulse is applied in the beginning of a pulse sequence (shown in Figure 17 for a single pulse experiment) and can thus be implemented into practically any pulse experiment. The method is simple and reasonably effective, but has the disadvantage of also partly suppressing other signals close to the water peak. The effectiveness is impaired in long pulse sequences where the water may regain some phase coherence and thus detectable signal. A special presaturation scheme is the NOESY-presat sequence, which consists of a non-selective 1D NOESY with zero mixing time, or in other words, the first increment of a 2D NOESY experiment [35].



Figure 18 The pulse sequence for WATERGATE solvent suppression.

A more complex – but also very effective – water suppression technique is the WATERGATE (WATER suppression by GrAdient Tailored Excitation) sequence [36], which is usually implemented as the last part of a pulse sequence. It employs two gradient pulses of equal strength and polarity around a binomial-type hard pulse sequence consisting of three pairs of symmetric pulses. The hard pulse sequence has no effect on the water resonance but the effect of a 180° pulse for all other resonances. Thus the water signal is further dephased by the second gradient pulse, while all other resonances are refocused. The conventional WATERGATE sequence shown in Figure 18 is also known as W3, while a further development is the W5 sequence [37], where the hard pulse consists of five pairs of short pulses.

4. Applications of multivariate analysis to NMR data

4.1 Preparing NMR data for multivariate analysis

Most types of spectroscopic data, e.g. IR, NIR, Raman can be subjected to multivariate analysis almost directly from the spectrometer. Usually, only spectral preprocessing with one of many methods is applied in order to correct for path differences, scatter effects due to different particle sizes etc.

NMR data on the other hand need to be processed in several ways in order to conform with the prerequisites for multivariate data analysis (cf. Section 2.2). The first problem one encounters is the existence of phase errors (cf. Section 3.1). Manual phase correction is usually implemented in the instrument software, but this process is very time-consuming, especially for the large data sets that are often analysed using chemometrics. Furthermore, manually phase correcting a series of spectra using the instrument software may yield sub-optimal results due to the subjective evaluation of the correction necessary for individual spectra. During this project, an in-house routine (written in Matlab, The Mathworks, Inc., Natick, MA) was developed, where individual spectra are phase corrected to resemble a reference spectrum, thereby reducing the errors due to subjective evaluation of the spectra.

Several attempts have been made to develop automatic procedures for phase correction, eliminating the need for extensive user-interactions in the processing of NMR spectra. The first automatic phase correction method was suggested in 1969 [38], and since then numerous other methods have been presented based on various principles [39,40,41,42]. In a series of papers, PCA is introduced for simultaneous phase correction of single resonances in a series of high-field NMR spectra [43,44,45,46,47], an approach which all has shown to perform very well. However as the correction requires a single isolated peak, the approach is not directly suited for complex spectra like ¹H NMR spectra of biological matrices. Recently, a method called Principal Phase Correction (PPC) based on PCA was shown to perform perfectly on low-field NMR spectra correcting one spectrum at a time [48]. Based on the success of this simple method, we have worked on developing an automatic phase correction method that will enable phase correction of single complex high-field NMR spectra using the same principle. Our work so far has

dealt with developing an automatic phase correction method for single spectra based on baseline optimisation (as suggested in [49]) using PCA to improve starting values for the optimisation procedure, but presently work is being conducted in the field of modifying the previously mentioned methods based on PCA to enable handling of single entire NMR spectra of complex matrices (work not yet published).

All NMR spectra are referenced to a standard in order to correct for small differences between samples in effective magnetic field, which result in a global shift of the spectrum. When referencing the spectra, an internal standard is most often used, however in 1D diffusion-edited experiments, typical internal standards cannot be used, due to their small size and hence high degree of attenuation in the diffusion-edited spectrum. If EDTA is used as anticoagulant in blood plasma, the Ca-EDTA peak at 2.56 ppm can be used as reference compound, since it is not fully attenuated in a typical 1D diffusion-edited experiment of blood plasma. In 2D diffusion-edited NMR spectra, the Ca-EDTA peak in the first (weak gradient) spectrum is used as reference for the entire 2D spectrum.

Yet another type of correction can be necessary due to variance in pH between samples, which results in local shifts, i.e. shifts that are different from peak to peak and from sample to sample. This can represent a major problem when analysing urine samples, since urine from different subjects and at different times can have very different pH values. Several methods exist for the general problem of shifts in spectra. For NMR spectra the methods Partial Linear Fit [50] and genetic algorithm peak alignment [51] have been suggested among others. A method for shifting a single resonance peak across a series of in vivo ³¹P NMR spectra has been presented, but this type of method is of little use when dealing with ¹H NMR spectra of biofluids with hundreds of resonances [45,46]. Dynamic Time Warping (DTW) and Correlation Optimised Warping (COW) have previously been used for the correction of peak shifts on chromatographic data with good results [52,53] and the application to NMR data should be fairly straight-forward. However, initial attempts to correct the data analysed in Paper IV using the methods of genetic algorithm peak alignment and COW were not successful due to the complexity of the spectra and the large range of shifts for some resonances (results not published). Instead a simple but effective method was used, which employs integration over small chemical shift ranges (usually 0.01 - 0.04 ppm) resulting in a number of NMR regions, which are analysed instead of the original spectra. Typically, a given peak will fall into the same NMR region for all samples thus eliminating the pH dependence of the peak positions. The disadvantage of this method is that resonances from different compounds may end up in the same NMR region, thus complicating the interpretation of models built on spectra reduced in this manner.

Finally, NMR spectra acquired with the Bruker XWIN–NMR software have been scaled to a common maximum peak height, a scaling that must be removed before multivariate analysis. The scaling factor is specified in the acquisition parameter file accompanying the spectrum and is thus relatively easily corrected for. If this scaling is removed, other types of scaling should generally not be necessary. However, sometimes in NMR spectra of biofluids a few very large peaks dominate the spectra making it harder for the multivariate model to extract the variation in smaller peaks. In cases like this, VAriable STability (VAST) scaling [54] or pareto scaling [55] can be of use.

4.2 Hard models

Hard models are based on the main principles of chemical physics. For every new system under investigation, a specific model must be chosen based on assumptions or known properties of the system. The analysis results in estimates of the parameters of the model, e.g. the concentrations or decay rates. In the following, three examples of hard modelling of NMR spectra are given using curve fitting applied to 1D and 2D spectra.

Manual curve fitting

A simple way of analysing 1D NMR spectra in a multivariate manner is to apply curve fitting to selected peaks. This approach can only be used if the NMR signal of the substance of interest is known, and if the overlap with peaks from other substances is limited. In Paper II five carrageenan types are quantified in commercial carrageenan products from the NMR spectra of carrageenan solutions using this approach. Manual curve fitting is used, since the five carrageenans give unique ¹H NMR signals in the anomeric proton range, 5.0 - 5.6 ppm, and using more sophisticated data analysis in this case was not necessary. An introduction to carrageenans is found in Box 1.

Box 1. Carrageenans are sulfated polygalactans extracted from algae. There are several types of carrageenans, the major ones being iota (1), kappa (κ) and lambda (λ). Minor types are mu (μ) and nu (ν), which are kappa- and iota-precursors, respectively, and are never found pure but always in coexistence with kappa and iota. In the food industry, carrageenans are widely used as stabilizing, thickening and gelling agents. The different carrageenan types have different functional properties, and therefore it is of great importance for the carrageenan industry to obtain detailed knowledge about the composition of their products in order to be able to design the functionality of the final product.

At the field strength used, 600 MHz for ¹H, the NMR peaks were slightly overlapping, and the five carrageenans as well as Floridean starch and four unknown substances were therefore quantified using *simultaneous* manual fitting with Lorentzian functions to specific NMR peaks.



Figure 19 Simultaneous fitting of the peaks from five carrageenan types and five other peaks yielding the relative composition of the sample.

The graphical user interface of the in-house built program for the simultaneous fitting of the carrageenan peaks is show in Figure 19. With this tool, it is far easier to accurately determine the areas of slightly overlapping peaks compared to using standard integration.

While the quantification of carrageenans from NMR spectra is not new, the use of simultaneous fitting of the carrageenan peaks in ¹H NMR spectra has not been published before, since normally a combination of ¹H and ¹³C NMR spectra are used to compensate for the overlap between specific peaks [56].

Automatic curve fitting

If the peaks of interest are heavily overlapping, the simple approach cannot be used but Classical Least Squares (CLS) can be applied (cf. Section 2.3.2). CLS is also a type of hard modelling, as it requires that the concentrations or spectra of all chemical constituents in the analysed spectral region to be known. As opposed to the manual fitting described above, in CLS the fitting is done in an automated way, optimising the fit in a least-squares sense. This approach is used in several publications and in a patented method for the quantification of lipoproteins in human blood plasma using ¹H NMR spectra [57,58,59,60,61]. Box 2 gives an introduction to lipoproteins and their importance.

The patented method by Dr. J.D. Otvos is described in the patent [60] and details can also be found on the website of the company performing the analysis, www.liposcience.com. It is based on CLS with non-negativity constraints on a selected region of the ¹H NMR spectra using known spectra of the lipoprotein subfractions. The method quantifies six subfractions of VLDL, IDL, three subfractions of LDL, and five subfractions of HDL. In reality, the 15 reference spectra have not been measured for all these subfractions, as only one VLDL, two LDL and two HDL subfractions were measured. The other subfraction reference spectra were constructed mathematically by simply shifting the spectra one or two variables to the left or right, hence producing the desired number of subfractions. Although empirical, this approach has shown in several studies to yield reasonable concentrations for the lipoprotein subfractions compared to reference methods such as ultracentrifugation. Furthermore, the method gives a measure of the risk of CHD, which has been shown to be highly correlated to the actual occurrence of CHD [62,63].

Box 2. People with high concentrations of plasma cholesterol and triglyceride have an increased risk of coronary heart disease (CHD). To assess the risk of CHD it is of great importance to be able to measure the content of lipoproteins – the lipid-transporters – in blood plasma. Lipoproteins can be divided into subgroups based on their density. The main fractions are very low density lipoproteins (VLDL), intermediate density lipoproteins (IDL), low density lipoproteins (LDL) and high density lipoproteins (HDL). The definition of the main fractions is partly empirical, as they do not represent four strictly distinct types of particles but rather particles with densities within certain ranges. Each of these main fractions can be further divided into a number of subfractions simply by dividing the density range of the main fraction into smaller ranges.

For the assessment of individual risk of CHD, the US National Cholesterol Education Program (NCEP) guidelines can be used. These guidelines present a set of lipoprotein main and subfraction contents that are known to be associated with low or high risk of CHD. Thus, measuring the concentrations of lipoprotein subfractions in blood plasma of a subject, the individual risk can be assessed.

The established standard reference method for the separation and analysis of lipoproteins is ultracentrifugation (UC). The disadvantage of lipoprotein quantification using ultracentrifugation is that the method is time and labour consuming and requires a large amount of plasma. Furthermore, the method quantifies the lipoprotein fractions on their cholesterol and triglyceride content, measures that are not always linearly correlated to the number of particles.

Exponential curve fitting

Certain types of 2D NMR data are characterised by exponentiality in the second dimension. This is valid for 2D relaxation-edited NMR spectra using the Inversion Recovery (INVREC) pulse experiment for T_1 edition [64] or the Carr-Purcell-Meiboum-Gill (CPMG) pulse experiment for T_2 edition [65] as well as for 2D diffusion-edited NMR (cf. Section 3.2.2).

This type of data is traditionally analysed using exponential fitting on each single resonance peak, resulting in an second dimension representing the relaxation time constants T_1 or T_2 or the self-diffusion coefficient, D. However, the robustness of

exponential fitting is limited especially in the case of spectral overlap, due to the high number of parameters to be estimated in the model.

For complex samples like blood plasma, exponential fitting of 2D diffusion-edited NMR spectra give reasonable estimates of the self-diffusion coefficients for small molecules with peaks separated from the peaks of other substances, while the determination of diffusion coefficients for lipoproteins is not possible using exponential fitting due to broad peaks and extensive overlap [28]. In a recent study, the lipoprotein region of a 2D diffusion-edited NMR spectrum was first deconvoluted in the spectral dimension using curve fitting with six Lorentzian functions [66]. Subsequently, the areas of the standard curves in the diffusion dimension were calculated and the diffusion coefficients were calculated from the areas using exponential fitting. However, the analysis did not yield the diffusion coefficients of any pure lipoprotein fractions, since the Lorentzian functions were found to represent mixtures of the main lipoprotein fractions.

4.3 Soft models

Soft models are characterised by general mathematical models, which are not specific for the particular system and thus do not require knowledge about the chemical or physical properties of the system. In this context, the only assumptions of most of the models are that the data are bi- or trilinear.

4.3.1 PCA on 1D NMR data

PCA has been applied to NMR data for many different purposes in the last decade. The applications include investigation of differences in lipoprotein subfraction NMR spectra between healthy and CHD patients [67], metabolic response to flavonoids from tea [68], characterisation of beer [69], investigation of structural changes during a pulping process [70], and compositional analysis of cellulose types [71]. Furthermore, PCA has been used for the classification of neural cell types [72], Slovenian wines [73], olive oils [74,75], apple juices [76], instant coffee [77], organic unsaturated compounds [78], and of naphtalene isomers [79]. Another area where PCA is commonly used on NMR data is in metabonomics, which is discussed in Section 4.3.4.

The uses of PLS for quantitative analysis is described in the following section. However, PLS can also be used for classification purposes as an alternative to PCA. The application of PLS discriminant analysis (PLS-DA) to NMR data has been limited in the last decade, but a few references exist. PLS-DA has been used for metabolic phenotyping [80], classification of saponin structures [81], and classification of olive oils [74].

4.3.2 PLS on 1D NMR data

In the last decade, PLS has been applied to NMR data within biological, pharmaceutical, and chemical research. Examples include the quantification or prediction of the active substances in tablets [82], biological activity of progestagens [83], composition of heparin type mixtures [84,85], crystallinity of lactose and cellulose [86,87], composition of cellulose samples [88], and kappanumber and processability for softwood and viscose pulps [89,90].

In Paper I, PLS is used for the quantification of lipoprotein main and subfractions from the 1D diffusion-edited ¹H NMR spectra of human blood plasma. These spectra are free from signals from small molecules, which disturb the extraction of lipoprotein information from the NMR spectra. The relative RMSECV (RMSECV/mean reference value) for main fractions were between 10 and 19 % with correlations 0.87 - 0.98, and for subfractions between 17 and 39 % with correlations 0.54 - 0.97. Plots of PLS predicted concentrations against reference values are shown in Figure 20 for the four main lipoprotein fractions. Despite the large relative error in some prediction models, it was shown that the placement in risk categories according to the NCEP guidelines was successful. The agreement in individual risk based on NMR derived values and on values determined by ultracentrifugation was between 76 and 100 %. This result emphasizes the ability to assess individual CHD risk from NMR data.

The only other reference for the quantification of lipoproteins using PLS is Bathen et al. [91]. However, this reference only attempted the quantification of lipoprotein main fractions and obtained relative prediction errors of 19 - 46 % with correlations of 0.74 - 0.97. Some attempts to use other methods for quantification of lipoprotein main fractions from NMR spectra have also been made, e.g. using an Artificial Neural Network approach [92] yielding correlations of 0.74 - 0.99 and Wavelet transform [93] yielding correlations of 0.79 - 0.92. In the case of quantification of lipoprotein main fractions, the results presented in Paper I are better than these references.



Figure 20 Predicted versus reference values for the four main lipoproteins fractions [Paper I].

4.3.3 N-PLS and PARAFAC on 2D NMR data

Many types of 2D NMR data exist – e.g. COSY, NOESY, TOCSY, HMQC, as well as relaxation- and diffusion-edited spectra. Of these, only the last two are optimally bilinear, and the one most suited for the application of chemometric methods is 2D diffusion-edited NMR since the individual resonance peaks are edited according to the self-diffusion coefficient of the entire *molecule*, while in 2D relaxation-edited NMR the resonance peaks are edited according the T_1 or T_2 of the single *atoms*. While relaxation-editing leads to data that are usually high-rank bilinear, diffusion-editing generally leads to low-rank data, where the rank is equal to the number of chemical constituents in the sample. If bilinear 2D NMR spectra are measured for a series of samples, the resulting 3D data set will optimally be trilinear and thus multi-way methods assuming trilinearity can be applied. The trilinear character of this type of NMR data sets have not previously been exploited.

Quantification of lipoproteins using N-PLS and PARAFAC

Due to the different sizes of lipoprotein subfractions they have different diffusion properties, i.e. self-diffusion coefficients. An additional "axis of separation" is therefore present in 2D diffusion-edited NMR spectra compared to regular 1D spectra and the application of multi-way methods for the quantification of lipoproteins is possible. This is done in Paper III, where 2D diffusion-edited NMR spectra of 17 plasma samples are analysed using N-PLS and PARAFAC.

The application of N-PLS to 2D diffusion-edited NMR data is straight-forward. Different spectral ranges are tested for prediction abilities regarding the concentrations of lipoprotein main and subfractions as determined by ultracentrifugation. N-PLS calibration models with three to nine components yielded relative prediction errors of 12 - 18 % and correlations between 0.82 and 0.96 for main fractions, and relative errors of 12 - 36 % and correlations 0.58 - 0.97 for subfractions. The relative prediction errors obtained here are slightly higher than the ones obtained in Paper I, but using only 17 samples. This result shows the potential of using 2D diffusion-edited NMR in order to get a further separation of the different lipoprotein signals.

The application of PARAFAC to 2D diffusion-edited NMR data is more subtle. Due to the uniqueness of the PARAFAC model, the analysis optimally results in the pure components, represented by the 1D spectra and the diffusion curves of the pure components, as well as the concentrations of each pure component. In the case of lipoproteins the number of pure components is really infinite, since all lipoprotein fractions are distributions in density and size. The data *are* theoretically trilinear, since each specific size of lipoprotein has a specific self-diffusion coefficient and thus diffusion curve, but the data are not *low-rank* trilinear! However, since PLS on 1D NMR spectra and N-PLS on 2D NMR spectra give reasonable results, it is expected that the data nevertheless can be approximated by a low-rank trilinear model. In order to be able to quantify the lipoprotein subfractions using PARAFAC, the analysis should optimally result in 12 components, corresponding to the contents determined by ultracentrifugation, VLDL 1-2, IDL, LDL 1-6 and HDL 1-3. The main fractions, VLDL, LDL and HDL could then be quantified as the sum of the subfractions. Given the low number of samples in the study, the extraction of 12 components is not realistic,

therefore the analysis is conducted only to examine the potential for analysing 2D NMR data using PARAFAC. Also due to the low number of samples, the analysis is conducted on the CH_2 peak only, as any signal from other constituents of the blood would require the extraction of PARAFAC components for these constituents also.



Figure 21 The 2D diffusion-edited NMR spectrum (only the CH_2 peak is shown) of a single sample (a) and the result of the PARAFAC analysis of 17 such spectra (b and c). Plot (b) shows the spectral loadings and plot (c) the diffusion loadings [Paper III].

The analysis resulted in a PARAFAC model with four components (an example of the original data as well as the model are shown in Figure 21). Validation using split-half analysis was not possible due to the low number of samples and the large variation between them, so the result was validated chemically. The four spectral loadings have similar shape only shifted on the chemical shift axis. The diffusion loadings are seemingly exponential and the order of the decay rates correspond to the order of shift in the spectral loadings. The calculation of self-diffusion coefficients from the diffusion loadings using exponential fitting yielded spherical hydrodynamic radii of 29, 18, 12 and 6 nm, which is within the reference values for the sizes of VLDL, IDL, LDL and HDL. However, the corresponding concentrations did not correlate well with the concentrations determined by ultracentrifugation, and the analysis using PARAFAC was therefore not successful with respect to the quantification of the lipoproteins. This is easily explained by the fact that the PARAFAC model did not explain all important variance in the data (i.e. the residual spectra contained structured information) and by the fact that the true number of components is much higher than the four that were resolved in this study. Nevertheless, the analysis shows the potential of using PARAFAC for analysing this type of data, and the possibilities of interpretation that the method offers.

4.3.4 Tucker on pseudo-2D NMR data

Pseudo-2D data in the form of NMR time series data could be trilinear but cannot a priori be guaranteed to be so, since the same spectral component can have different time evolution for different samples. Thus the application of the Tucker model is a possibility since this model does not require trilinearity in the sense that PARAFAC does. Tucker has never before been applied to NMR data.

In Paper IV, Tucker is applied to NMR time series data to evaluate the potential of using multi-way methods for the analysis of this type of data. The purpose of the present analysis is the investigation of the metabolic response to model toxins (see Box 3 for background).

Box 3. Metabonomics is defined as "the quantitative measurement of the multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification" [94]. Metabonomics is most often based on NMR measurements analysed using chemometrics and can e.g. be used to study the metabolite profile in humans with a particular disease or the metabolic response to a toxin administered to animals.

In the toxicological studies subjected to multi-way analysis in Paper IV, three model toxins were administered orally to rats, and urine samples were collected over time from 24 hours pre-dose to 168 hours post-dose. Each study comprised 30 animals of which ten received a high dose of the toxin, ten a lower dose and ten were controls.

Numerous toxic substances have been investigated in the past using NMR and chemometrics, a review of which can be found in [95]. The analysis of the urinary metabolic model toxins hydrazine response to the dihydrochloride, α -naphthylisothiocyanate (ANIT) and butylated hydroxytoluene (BHT) has previously been made using PCA on the unfolded data matrix [96,97,98,99,100]. The unfolding was made such that the sample and time modes were mixed into one common mode while the spectral mode remained unchanged. The disadvantage of this data analytical approach is that using the highly flexible PCA, all metabolic perturbations are modelled including minor ones pertaining to only a few samples at specific times. Furthermore, information about the time evolution of important perturbations must be extracted from the score plots, since the sample and time modes are mixed. Another method, batch analysis (based on PCA and PLS), has been applied to the study of hydrazine and ANIT toxicity [101,102]. Batch analysis has the disadvantage of forcing the time evolutions to be linear and with a quantitative difference between each time point, due to the use of PLS against a constructed Y consisting of the numbers from one to ten, thus ruling out reversible metabolic effects.

The advantage of Tucker analysis over the analysis using PCA and PLS is that the three-way structure of the data is retained and used in the analysis. Therefore, only the metabolic perturbations that have common time profiles will be modelled and random or less general changes in metabolite levels, which do not have common time profiles, will be considered as noise. No constraints are put on time profiles, and thus the models can handle effects with any kind of time evolution.

In Paper IV, data from the toxicological studies of these three model toxins were analysed using Tucker models. The models were fitted to the three-way data sets with dimensions 30 (rats) \times 10 (time points) \times 202 (reduced NMR spectra). The datasets were centered across the rats mode using the mean of control rats only, thus facilitating the interpretation as all perturbations are viewed as changes with respect to the control rats. The Tucker analysis showed the same major trends in terms of metabolic response to the toxins as were found in the previous studies although minor differences were seen in individual metabolites.

As an example, Figure 22 shows the Tucker model of the hydrazine dataset. A Tucker model with two components in each of the three modes, i.e. the

rats/treatment mode, the time mode and the spectral mode, was found to provide good interpretation of the data explaining 79 % of the total variance. Detailed interpretation of the score and loading plots can be found in Paper IV, and will not be given here. An interesting feature that was extracted by the Tucker analysis and was not reported in the papers using unfold PCA or batch analysis [97,99,101], is that the low-dose rats experience a metabolic perturbation, which is less severe and with an earlier onset and a faster recovery than the high-dose rats, but otherwise similar to that of the high-dose rats and described by the same spectral loading.



Figure 22 Overview of the hydrazine Tucker (2 2 2) model. Plot (a) shows scores, plot (b) time loadings and plot (c) spectral loadings. Symbols and metabolite abbreviations can be seen in Paper IV.

5. Conclusions

Multivariate data analysis was applied to different types of high-field ¹H NMR data ranging from simple 1D spectra to complex 2D diffusion-edited spectra and NMR time series. Several new approaches to data analysis of NMR data have been studied and compared.

Multivariate curve fitting applied to a selected region of the NMR spectra of carrageenan mixtures gave reliable relative concentrations for the five major carrageenan types and yielded excellent PLS models on Raman, IR and NIR spectra using the NMR determined carrageenan concentrations as the dependent variable.

Lipoprotein main and subfractions were quantified from their 1D and 2D diffusionedited NMR spectra using PLS and N-PLS. The correlations as well as relative errors were improved compared to previous attempts to use chemometrics and NMR to quantify lipoprotein fractions. An intrinsic problem of these data, which limits the application of both PLS and N-PLS, is the fact that lipoprotein subfractions are *distributions* in density and thus the data are basically not lowrank bi- or tri-linear, but high-rank. The high-rank bi- or tri-linearity of the NMR data limits the appropriateness of fitting low-rank models and thus sets a limit for the performance of calibration models of lipoprotein main and subfraction contents. Further work is being done on improving the calibrations using more samples.

Extraction of spectral and diffusion profiles of lipoprotein fractions from 2D diffusion-edited spectra were achieved using PARAFAC. The attempt to use the extracted components to determine concentrations of lipoprotein main and subfractions was not successful. However, the results are promising with regard to analysing 2D diffusion-edited NMR spectra of other data where the requirement for low-rank trilinearity is met.

It was furthermore shown that Tucker analysis of NMR time series is feasible and yields interpretable models. Tucker models were applied to the data from three toxicological studies consisting of NMR spectra of rat urine collected over time, and the analysis gave results that were complementary to results obtained using unfold PCA and batch analysis on the same toxicological substances.

The work presented in this thesis shows that application of two-way as well as multi-way multivariate data analysis to complex high-field NMR data is feasible and may yield models that are informative and with good predictive power. Many problems are still to be solved regarding the preprocessing of 1D NMR spectra, e.g. phase correction and shift alignment, as NMR spectra thus better suited for multivariate data analysis may give more accurate and detailed multivariate models. The possibility of analysing 2D NMR spectra measured on series of samples with PARAFAC yielding the pure spectra and concentrations of all analytes is very promising. However, the analysis of a relatively simple system needs to be performed to investigate the possibilities and limitations of this application.

6. References

- 1. Munck, L., Nørgaard, L., Engelsen, S.B., Bro, R., and Andersson, C.A. Chemometrics in food science a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance. *Chemometrics and Intelligent Laboratory Systems* 44(1-2): 31-60, 1998.
- 2. Gabrielsson, J., Lindberg, N.O., and Lundstedt, T. Multivariate methods in pharmaceutical applications. *Journal of Chemometrics* 16(3): 141-160, 2002.
- 3. Guillou, C., Remaud, G., and Martin, G.J. Applications of NMR to the characterization and authentication of foods and beverages. *Trends in Food Science and Technology* 3: 197-201, 1992.
- 4. Nicholson, J.K., Lindon, J.C., and Holmes, E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29(11): 1181-1189, 1999.
- 5. Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psycology* 24: 417-441,498-520, 1933.
- 6. Wold, S., Esbensen, K., and Geladi, P. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* 2(1-3): 37-52, 1987.
- 7. Martens, H. and Næs, T. "Multivariate Calibration", John Wiley & Sons, Inc., New York, 1989.
- 8. Bro, R. Multiway calibration. Multilinear PLS. *Journal of Chemometrics* 10(1): 47-61, 1996.
- 9. Hotelling, H. The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology* 10: 69-79, 1957.
- 10. Kendall, M.G. "A Course in Multivariate Analysis", Griffin, London, 1957.
- 11. Wold, S., Martens, H., and Wold, H. The Multivariate Calibration-Problem in Chemistry Solved by the PLS Method. *Lecture Notes in Mathematics* 973: 286-293, 1983.
- 12. Geladi, P. and Kowalski, B.R. Partial Least-Squares Regression A Tutorial. *Analytica Chimica Acta* 185: 1-17, 1986.
- 13. Bro, R. "Multi-way analysis in the Food Industry. Models, Algorithms, and Applications", Doctoral Thesis, University of Amsterdam, 1998.
- 14. Tucker, L. Some mathematical notes on three-model factor analysis. *Psychometrika* 31: 279-311, 1966.
- 15. Kroonenberg, P.M. and de Leeuw, J. Principal components analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika* 45: 69-97, 1980.

- 16. ten Berge, J.M.F., de Leeuw, J., and Kroonenberg, P.M. Some additional results on principal components analysis of three-mode data by means of alternating least squares. *Psychometrika* 52: 183-191, 1987.
- 17. Harshman, R.A. Foundations fo the PARAFAC procedure: model and conditions for an 'explanatory' multi-mode factor analysis. *UCLA Working Papers in Phonetics* 16: 1, 1970.
- Carroll, J.D. and Chang, J. Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. *Psychometrika* 35: 283, 1970.
- 19. Bro, R. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems* 38(2): 149-171, 1997.
- 20. Martens, H. and Dardenne, P. Validation and verification of regression in small data sets. *Chemometrics and Intelligent Laboratory Systems* 44(1-2): 99-121, 1998.
- 21. Wold, S. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics* 20(4): 397-405, 1978.
- 22. Gorter, C.J. Negative results of an attempt to direct nuclear magnetic spins. *Physica* 3: 995-998, 1936.
- 23. Bloch, F., Hansen, W.W., and Packard, M. Nuclear Induction. *Physical Review* 69: 127, 1946.
- 24. Purcell, E.M., Torrey, H.C., and Pound, R.V. Resonance absorption by nuclear magnetic moments in a solid. *Physical Review* 69: 37-38, 1946.
- 25. Proctor, W.G. and Yu, F.C. The dependence of a nuclear magnetic resonance frequency upon chemical compounds. *Physical Review* 77: 717-717, 1950.
- 26. Hahn, E.L. Spin echoes. Physical Review 80: 580-594, 1950.
- 27. Antalek, B. Using pulsed gradient spin echo NMR for chemical mixture analysis: how to obtain optimum results. *Concepts in Magnetic Resonance* 14(4): 225-258, 2002.
- 28. Johnson, C.S. Diffusion ordered nuclear magnetic resonance spectroscopy: principles and applications. *Progress in Nuclear Magnetic Resonance Spectroscopy* 34(3-4): 203-256, 1999.
- 29. Gibbs, S.J. and Johnson, C.S. A PFG NMR experiment for accurate diffusion and flow studies in the presence of eddy currents. *Journal of Magnetic Resonance* 93(2): 395-402, 1991.
- 30. Jerschow, A. and Müller, N. Convection compensation in gradient enhanced nuclear magnetic resonance spectroscopy. *Journal of Magnetic Resonance* 132(1): 13-18, 1998.
- 31. Esturau, N., Sanchez-Ferrando, F., Gavin, J.A., Roumestand, C., Delsuc, M.A., and Parella, T. The use of sample rotation for minimizing convection effects in self-diffusion NMR measurements. *Journal of Magnetic Resonance* 153(1): 48-55, 2001.

- 32. Jerschow, A. and Müller, N. Suppression of convection artifacts in stimulated-echo diffusion experiments. Double-stimulated-echo experiments. *Journal of Magnetic Resonance* 125(2): 372-375, 1997.
- 33. Callaghan, P.T. "Principles of Nuclear Magnetic Resonance Microscopy", Oxford University Press, Oxford, 1993, pp. 353-367.
- 34. Freeman, R. "Spin Choreography. Basic Steps in High Resolution NMR", Oxford University Press, Oxford, 1997.
- 35. Claridge, T.D.W. "High-resolution NMR Techniques in Organic Chemistry", Pergamon Press, Oxford, 1999.
- Piotto, M., Saudek, V., and Sklenar, V. Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *Journal of biomolecular NMR* 2(6): 661-665, 1992.
- 37. Liu, M.L., Mao, X., Ye, C., Huang, H., Nicholson, J.K., and Lindon, J.C. Improved WATERGATE pulse sequences for solvent suppression in NMR spectroscopy. *Journal of Magnetic Resonance* 132: 125-129, 1998.
- 38. Ernst, R.R. Numerical Hilbert Transform and Automatic Phase Correction in Magnetic Resonance Spectroscopy. *Journal of Magnetic Resonance* 1(1): 7-26, 1969.
- 39. Sotak, C.H., Dumoulin, C.L., and Newsham, M.D. Automatic Phase Correction of Fourier-Transform Nmr-Spectra Based on the Dispersion Versus Absorption (Dispa) Lineshape Analysis. *Journal of Magnetic Resonance* 57(3): 453-462, 1984.
- 40. Craig, E.C. and Marshall, A.G. Automated phase correction of FT NMR spectra by means of phase measurement based on Dispersion Versus Absorption Relation (DISPA). *Journal of Magnetic Resonance* 76(3): 458-475, 1988.
- 41. van Vaals, J.J. and van Gerwen, P.H.J. Novel methods for automatic phase correction of NMR spectra. *Journal of Magnetic Resonance* 86(1): 127-147, 1990.
- 42. Dzakula, Z. Phase angle measurement from peak areas (PAMPAS). *Journal of Magnetic Resonance* 146(1): 20-32, 2000.
- 43. Ober, R.J. and Ward, E.S. Correcting for phase-distortion of NMR spectra analyzed using singular value decomposition of Hankel-matrices. *Journal of Magnetic Resonance Series* A 114(1): 120-123, 1995.
- 44. Brown, T.R. and Stoyanova, R. NMR spectral quantitation by principal component analysis. II. Determination of frequency and phase shifts. *Journal of Magnetic Resonance Series B* 112(1): 32-43, 1996.
- 45. Witjes, H., Melssen, W.J., Zandt, H.J.A.I., van der Graaf, M., Heerschap, A., and Buydens, L.M.C. Automatic correction for phase shifts, frequency shifts, and lineshape distortions across a series of single resonance lines in large spectral data sets. *Journal of Magnetic Resonance* 144(1): 35-44, 2000.

- 46. Witjes, H., Pepers, M., Melssen, W.J., and Buydens, L.M.C. Modelling phase shifts, peak shifts and peak width variations in spectral data sets: its value in multivariate data analysis. *Analytica Chimica Acta* 432: 113-124, 2001.
- 47. Stoyanova, R. and Brown, T.R. NMR spectral quantitation by principal component analysis. III. A generalized procedure for determination of lineshape variations. *Journal of Magnetic Resonance* 154: 163-175, 2002.
- 48. Pedersen, H.T., Bro, R., and Engelsen, S.B. Towards rapid and unique curve resolution of low-field NMR relaxation data: Trilinear SLICING versus two-dimensional curve fitting. *Journal of Magnetic Resonance* 157(1): 141-155, 2002.
- 49. Brown, D.E., Campbell, T.W., and Moore, R.N. Automated phase correction of FT NMR spectra by baseline optimization. *Journal of Magnetic Resonance* 85(1): 15-23, 1989.
- 50. Vogels, J.T.W.E., Tas, A.C., Venekamp, J., and van der Greef, J. Partial linear fit: A new NMR spectroscopy preprocessing tool for pattern recognition applications. *Journal of Chemometrics* 10: 425-438, 1996.
- 51. Forshed, J., Schuppe-Koistinen, I., and Jacobsson, S.P. Peak alignment of NMR signals by means of a genetic algorithm. *Analytica Chimica Acta* 487(2): 189-199, 2003.
- 52. Nielsen, N.P.V., Carstensen, J.M., and Smedsgaard, J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A* 805(1-2): 17-35, 1998.
- 53. Tomasi, G., van den Berg, F., and Andersson, C.A. Correlation Optimized Warping and Dynamic Time Warping as preprocessing methods for chromatographic data. *Submitted*.
- Keun, H.C., Ebbels, T.M.D., Antti, H., Bollard, M.E., Beckonert, O., Holmes, E., Lindon, J.C., and Nicholson, J.K. Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling. *Analytica Chimica Acta* 490: 265-276, 2003.
- 55. Wold, S., Johansson, E., and Cocchi, M. "PLS". In *3D-QSAR in Drug Design, Theory, Methods, and Applications*, Kubinyi, H. (Ed.). ESCOM Science, Leiden, The Netherlands, 1993.
- 56. van de Velde, F., Knutsen, S., Usov, A.I., Rollema, H.S., and Cerezo, A.S. ¹H and ¹³C high resolution NMR spectroscopy of carrageenans: application in research and industry. *Trends in Food Science & Technology* 13: 73-92, 2002.
- Hiltunen, Y., Ala-Korpela, M., Jokisaari, J., Eskelinen, S., Kiviniitty, K., Savolainen, M.J., and Kesäniemi, Y.A. A lineshape fitting model for ¹H NMR spectra of human blood plasma. *Magnetic Resonance in Medicine* 21(2): 222-232, 1991.
- 58. Ala-Korpela, M., Korhonen, A., Keisala, J., Horkko, S., Korpi, P., Ingman, L.P., Jokisaari, J., Savolainen, M.J., and Kesäniemi, Y.A. ¹H NMR-based absolute quantitation of human lipoproteins and their lipid contents directly from plasma. *Journal of Lipid Research* 35(12): 2292-2304, 1994.

- 59. Otvos, J.D., Jeyarajah, E.J., and Bennett, D.W. Quantification of plasma lipoproteins by proton nuclear magnetic resonance spectroscopy. *Clinical Chemistry* 37(3): 377-386, 1991.
- 60. Otvos, J.D. Method and apparatus for measuring classes and subclasses of lipoproteins. *United States Patent*, 5 343 389, 1994.
- Soedamah-Muthu, S.S., Chang, Y.-F., Otvos, J.D., Evans, R.W., and Orchard, T.J. Lipoprotein subclass measurements by nuclear magnetic resonance spectroscopy improve the prediction of coronary artery disease in Type 1 Diabetes. A prospective report from Pittsburgh Epidemiology of Diabetes Complications Study. *Diabetologia* 46(5): 674-682, 2003.
- 62. Blake, G.J., Otvos, J.D., Rifai, N., and Ridker, P.M. Low-density lipoprotein particle concentration and size as determined by nuclear magnetic resonance spectroscopy as predictors of cardiovascular disease in women. *Circulation* 106: 1930-1937, 2002.
- 63. Kuller, L.H., Arnold, A., and Tracy, R. Nuclear Magnetic Resonance spectroscopy of Lipoproteins and risk of coronary heart disease in the cardiovascular health study. *Arteriosclerosis Thrombosis and Vascular Biology* 22: 1175-1180, 2002.
- 64. Vold, R.L., Waugh, J.S., Klein, M.P., and Phelps, D.E. Measurement of spin relaxation in complex systems. *Journal of Chemical Physics* 48(4): 3831-3832, 1968.
- 65. Meiboom, S. and Gill, D. Modified spin-echo method for measuring nuclear relaxation times. *Review of Scientific Instruments* 29(2): 688-691, 1958.
- 66. Liu, M.L., Tang, H., Nicholson, J.K., and Lindon, J.C. Use of ¹H NMR-determined diffusion coefficients to characterize lipoprotein fractions in human blood plasma. *Magnetic Resonance in Chemistry* 40(13): S83-S88, 2002.
- 67. Bathen, T.F., Engan, T., and Krane, J. Principal component analysis of proton nuclear magnetic resonance spectra of lipoprotein fractions from patients with coronary heart disease and healthy subjects. *Scandinavian Journal of Clinical and Laboratory Investigation* 59(5): 349-360, 1999.
- 68. Solanky, K.S., Bailey, N.J.C., Holmes, E., Lindon, J.C., Davis, A.L., Mulder, T.P.J., Van Duynhoven, J.P.M., and Nicholson, J.K. NMR-based metabonomic studies on the biochemical effects of epicatechin in the rat. *Journal of Agricultural and Food Chemistry* 51(14): 4139-4145, 2003.
- 69. Duarte, I., Barros, A., Belton, P.S., Righelato, R., Spraul, M., Humpfer, E., and Gil, A.M. High-Resolution Nuclear Magnetic Resonance Spectroscopy and multivariate analysis for the characterization of beer. *Journal of Agricultural and Food Chemistry* 50: 2475-2481, 2002.
- 70. Malkavaara, P., Alen, R., and Kolehmainen, E. Chemometrics: An important tool for the modern chemist, an example from wood-processing chemistry. *Journal of Chemical Information and Computer Sciences* 40(2): 438-441, 2000.

- Lennholm, H. and Iversen, T. Estimation of Cellulose-I and Cellulose-Ii in Cellulosic Samples by Principal Component Analysis of C-13-Cp/Mas-Nmr-Spectra. *Holzforschung* 49(2): 119-126, 1995.
- 72. Griffin, J.L., Bollard, M., Nicholson, J.K., and Bhakoo, K. Spectral profiles of cultured neuronal and glial cells derived from HRMAS H-1 NMR spectroscopy. *NMR in Biomedicine* 15(6): 375-384, 2002.
- 73. Kosir, I.J., Kocjancic, M., Ogrinc, N., and Kidric, J. Use of SNIF-NMR and IRMS in combination with chemometric methods for the determination of chaptalisation and geographical origin of wines (the example of Slovenian wines). *Analytica Chimica Acta* 429(2): 195-206, 2001.
- Shaw, A.D., diCamillo, A., Vlahov, G., Jones, A., Bianchi, G., Rowland, J., and Kell, D.B. Discrimination of the variety and region of origin of extra virgin olive oils using C-13 NMR and multivariate calibration with variable reduction. *Analytica Chimica Acta* 348(1-3): 357-374, 1997.
- Sacchi, R., Mannina, L., Fiordiponti, P., Barone, P., Paolillo, L., Patumi, M., and Segre, A. Characterization of Italian extra virgin olive oils using H-1-NMR spectroscopy. *Journal of Agricultural and Food Chemistry* 46(10): 3947-3951, 1998.
- 76. Belton, P.S., Colquhoun, I.J., Kemsley, E.K., Delgadillo, I., Roma, P., Dennis, M.J., Sharman, M., Holmes, E., Nicholson, J.K., and Spraul, M. Application of chemometrics to the H-1 NMR spectra of apple juices: discrimination between apple varieties. *Food Chemistry* 61(1-2): 207-213, 1998.
- 77. Charlton, A.J., Farrington, W.H.H., and Brereton, P. Application of H-1 NMR and multivariate statistics for screening complex mixtures: Quality control and authenticity of instant coffee. *Journal of Agricultural and Food Chemistry* 50(11): 3098-3103, 2002.
- 78. Tasic, L. and Rittner, R. Principal component analysis of carbon-13 substituent-induced chemical shifts of some unsaturated compounds. *Journal of Molecular Structure* 616(1-3): 49-54, 2002.
- 79. Airiau, C.Y., Shen, H.L., and Brereton, R.G. Principal component analysis in liquid chromatography proton nuclear magnetic resonance: differentiation of three regioisomers. *Analytica Chimica Acta* 447(1-2): 199-210, 2001.
- Gavaghan, C.L., Wilson, I.D., and Nicholson, J.K. Physiological variation in metabolic phenotyping and functional genomic studies: use of orthogonal signal correction and PLS-DA. *FEBS Letters* 530(1-3): 191-196, 2002.
- Nord, L.I., Kenne, L., and Jacobsson, S.P. Multivariate analysis of H-1 NMR spectra for saponins from Quillaja saponaria Molina. *Analytica Chimica Acta* 446(1-2): 199-209, 2001.
- 82. Talebpour, Z., Maesum, S., Jalali-Heravi, M., and Shamsipur, M. Simultaneous determination of theophylline and caffeine by proton magnetic resonance spectroscopy using partial least squares regression techniques. *Analytical Sciences* 19(7): 1079-1082, 2003.

- 83. Bursi, R., Dao, T., van Wijk, T., de Gooyer, M., Kellenbach, E., and Verwer, P. Comparative spectra analysis (CoSA): Spectra as three-dimensional molecular descriptors for the prediction of biological activities. *Journal of Chemical Information and Computer Sciences* 39(5): 861-867, 1999.
- Ruiz-Calero, V., Saurina, J., Galceran, M.T., Hernandez-Cassou, S., and Puignou, L. Estimation of the composition of heparin mixtures from various origins using proton nuclear magnetic resonance and multivariate calibration methods. *Analytical and Bioanalytical Chemistry* 373(4-5): 259-265, 2002.
- 85. Ruiz-Calero, V., Saurina, J., Galceran, M.T., Hernandez-Cassou, S., and Puignou, L. Potentiality of proton nuclear magnetic resonance and multivariate calibration methods for the determination of dermatan sulfate contamination in heparin samples. *Analyst* 125(5): 933-938, 2000.
- 86. Gustafsson, C., Lennholm, H., Iversen, T., and Nystrom, C. Comparison of solid-state NMR and isothermal microcalorimetry in the assessment of the amorphous component of lactose. *International Journal of Pharmaceutics* 174(1-2): 243-252, 1998.
- 87. Lennholm, H., Larsson, T., and Iversen, T. Determination of Cellulose I-Alpha and I-Beta in Lignocellulosic Materials. *Carbohydrate research* 261(1): 119-131, 1994.
- Wickholm, K., Hult, E.L., Larsson, P.T., Iversen, T., and Lennholm, H. Quantification of cellulose forms in complex cellulose materials: a chemometric model. *Cellulose* 8(2): 139-148, 2001.
- Hauksson, J.B., Edlund, U., and Trygg, J. NMR processing techniques based on multivariate data analysis and orthogonal signal correction. C-13 CP/MAS NMR spectroscopic characterization of softwood kraft pulp. *Magnetic Resonance in Chemistry* 39(5): 267-275, 2001.
- 90. Elg-Christofferson, K., Hauksson, J., Edlund, U., Sjostrom, M., and Dolk, M. Characterisation of dissolving pulp using designed process variables, NIR and NMR spectroscopy, and multivariate data analysis. *Cellulose* 6(3): 233-249, 1999.
- 91. Bathen, T.F., Krane, J., Engan, T., Bjerve, K.S., and Axelson, D. Quantification of plasma lipids and apolipoproteins by use of proton NMR spectroscopy, multivariate and neural network analysis. *NMR in Biomedicine* 13(5): 271-288, 2000.
- 92. Hiltunen, Y., Heiniemi, E., and Ala-Korpela, M. Lipoprotein-lipid quantification by neural-network analysis of 1H NMR data from human blood plasma. *Journal of Magnetic Resonance* Series B 106(2): 191-194, 1995.
- 93. Serrai, H., Nadal, L., Leray, G., Leroy, B., Delplanque, B., and de Certaines, J.D. Quantification of plasma lipoprotein fractions by wavelet transform time-domain data processing of the proton nuclear magnetic resonance methylene spectral region. *NMR in Biomedicine* 11(6): 273-280, 1998.
- 94. Lindon, J.C., Nicholson, J.K., Holmes, E., and Everett, J.R. Metabonomics: Metabolic Processes Studied by NMR Spectroscopy of Biofluids. *Concepts in Magnetic Resonance* 125(5): 289-320, 2000.

- Beckonert, O., Bollard, M.E., Ebbels, T.M.D., Keun, H.C., Antti, H., Holmes, E., Lindon, J.C., and Nicholson, J.K. NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches. *Analytica Chimica Acta* 490(1-2): 3-15, 2003.
- 96. Beckwith-Hall, B.M., Nicholson, J.K., Nicholls, A.W., Foxall, P.J.D., Lindon, J.C., Connor, S.C., Abdi, M., Connelly, J.C., and Holmes, E. Nuclear magnetic resonance spectroscopic and principal components analysis investigations into biochemical effects of three model hepatotoxins. *Chemical Research in Toxicology* 11(4): 260-272, 1998.
- Holmes, E., Nicholls, A.W., Lindon, J.C., Connor, S.C., Connelly, J.C., Haselden, J.N., Damment, S.J., Spraul, M., Neidig, P., and Nicholson, J.K. Chemometric models for toxicity classification based on NMR spectra of biofluids. *Chemical Research in Toxicology* 13(6): 471-478, 2000.
- Robertson, D.G., Reily, M.D., Sigler, R.E., Wells, D.F., Paterson, D.A., and Braden, T.K. Metabonomics: Evaluation of nuclear magnetic resonance (NMR) and pattern recognition technology for rapid in vivo screening of liver and kidney toxicants. *Toxicological Sciences* 57(2): 326-337, 2000.
- 99. Nicholls, A.W., Holmes, E., Lindon, J.C., Shockcor, J.P., Farrant, R.D., Haselden, J.N., Damment, S.J.P., Waterfield, C.J., and Nicholson, J.K. Metabonomic investigations into hydrazine toxicity in the rat. *Chemical Research in Toxicology* 14(8): 975-987, 2001.
- 100. Waters, N.J., Holmes, E., Williams, A., Waterfield, C.J., Farrant, R.D., and Nicholson, J.K. NMR and pattern recognition studies on the time-related metabolic effects of alphanaphthylisothiocyanate on liver, urine, and plasma in the rat: An integrative metabonomic approach. *Chemical Research in Toxicology* 14(10): 1401-1412, 2001.
- 101. Antti, H., Bollard, M.E., Ebbels, T., Keun, H., Lindon, J.C., Nicholson, J.K., and Holmes, E. Batch statistical processing of H-1 NMR-derived urinary spectral data. *Journal of Chemometrics* 16(8-10): 461-468, 2002.
- 102. Azmi, J., Griffin, J.L., Antti, H., Shore, R.F., Johansson, E., Nicholson, J.K., and Holmes, E. Metabolic trajectory characterisation of xenobiotic-induced hepatotoxic lesions using statistical batch processing of NMR data. *Analyst* 127(2): 271-276, 2002.