Laila Stordrange

Multivariate analysis of near-infrared spectroscopic data from reaction monitoring

Department of Chemistry
University of Bergen
2003

Multivariate analysis of near-infrared spectroscopic data from reaction monitoring

Laila Stordrange

Department of Chemistry
University of Bergen
2003

Dissertation submitted for the degree of *doctor scientiarum*

Laila Stordrange
Department of Chemistry
University of Bergen
Allégaten 41
N-5007 Bergen
Laila.Stordrange@kj.uib.no

Thesis submitted in partial fulfillment of the requirements for the degree of doctor scientiarum.
October 2003

ISBN 82-7400-062-8
Bergen, Norway
2003

Preface

This thesis is submitted in partial fulfillment for the degree of Doctor Scientiarum at the University of Bergen. The thesis consists of an introduction, followed by four papers. The doctorate study was initiated in August 2000 and the work has been carried out at the Department of Chemistry, University of Bergen. Courses and contributions at conferences has been fulfilled according to requirements of the study. The Norwegian Research Council has funded the project.

Earlier studies include a bachelor degree in analytical chemistry from Agder University College (1996), and a Candidata Scientiarum degree in chemometrics from the University of Bergen (1999). In the Cand. Scient. study, chemometric techniques were used in combination with near-infrared spectroscopy to investigate chemical reactions.

Near-infrared spectroscopy is a fast and non-destructive technique, requiring little or no sample preparation. The technique is therefore suitable for automatic monitoring of chemical processes. An engagement to test near-infrared spectroscopy for monitoring a synthesis of contrast agent was carried out at Nycomed Imaging (present Amersham Health AS). The study revealed several artefacts in the spectroscopic data. A Dr. Scient. project to investigate multivariate modelling of near-infrared spectroscopy data from the complex industrial process was applied for, and was approved by the Norwegian Research Council.

High collinearity in spectroscopic data necessitates the need of multivariate techniques. Several methods to model data for quantitative purposes exist. Which method to choose depends on the chemical system. In this thesis, emphasis has been put on assessment of modelling techniques to near-infrared data. Another important issue when modelling is how to preprocess data to diminish the influence from noise or irregularities. Removal of multiplicative and additive effects, and how to model nonlinear data, have also been investigated.

The first paper is based on results achieved during the Cand. Scient. study, where a procedure to resolve near-infrared spectra was proposed. A synthesis of contrast agent at Amersham Health AS has been used as a case study in the other papers. The system is complex, making curve resolution a difficult task. If reference measurements of interesting variables are available, regression techniques that relate near infrared spectra to the responses are preferred. A feasibility study of near-infrared spectroscopy was performed in paper II. The study included testing of different preprocessing techniques. The process data were shown to have a curvature that is well described by a second order polynomial. Different techniques that account for the non-linear behaviour was tested in paper III. In the last paper, multiway techniques were tested for their ability to model near-infrared data.

Acknowledgements

Professor Olav M. Kvalheim included me as a cand. scient. student in the chemometric group at the University of Bergen. After my exam he introduced me to people at Nycomed Imaging for a project of testing near-infrared spectroscopy for process monitoring. He was also a strong promotor for applying for a Dr. Scient. project to the Norwegian Research Council. Olav is acknowledged for his enthusiasm and guidance in this and the Cand. Scient. project.

Dr. Philos Dick Malthe-Sørenssen at Amersham Health AS is thanked for being open minded for testing and adopting new techniques into the process at Lindesnes plant. His interest in chemometric methods and making it possible to carry out projects like this in the industry. Gratitude is giving to both Olav and Dick for their capability of building bridges between the academia and industry, and thereby make real process data available to the academia and let new technology be applied on industrial problems.

Dr. Scient. Fred Olav Libnau at Amersham Health AS has been my mentor in chemometrics and process analysis since my entry into Nycomed Imaging four years ago. He is acknowledge for his generous share of knowledge and for inspiring discussions. I will also express my gratitude to his encouragement to apply for a doctoral grant and taking time to discuss and see things from a new perspective when needed.

Professor Rolf Manne and associate professor Bjørn Grung are thanked for reading the draft of the thesis and for valuable comments.

Colleagues at Department of Chemistry, especially people at the rolling lounge, are thanked for pleasant company and great coffee and thereby making long days in front of the computer practicable.

I will especially express my gratitude to Kjell Olaf Halland for his encouragement during time of frustration and for bringing moments of joy into everyday life. My thanks goes also to my family for their overwhelming concern of me. Thanks to friends, making the years in Bergen full of good laugh and great moments.

I am grateful to NORFA and NICE for providing funding and scholarship to attend relevant doctorate courses at KTH, Stockholm, and at KVL, Copenhagen, and for arranging a workshop in chemometrics at Åland. The importance of providing such an instructive environment for doctoral students cannot be emphasised enough.

Last but not least, I have to thank the Norwegian Research Council for a doctoral grant.

Contents

Preface iv

Acknowledgements v

List of papers viii

1 Introduction 1
- 1.1 Outline of the thesis 2
- 1.2 Pharmaceutical process 3
- 1.3 Process analysis 4
- 1.4 Principles of near-infrared spectroscopy 5
- 1.5 Beer's law 6

2 Multivariate analysis 7
- 2.1 Data reduction by PCA and SVD 7
- 2.2 Curve resolution 8
  - 2.2.1 Rank analysis 8
  - 2.2.2 Detection of selective regions 9
  - 2.2.3 Resolution in the presence of selective regions 10
  - 2.2.4 Resolution in the absence of selective regions 10
- 2.3 Two-way regression 11
  - 2.3.1 Determination of significant components 11
  - 2.3.2 Variable selection 12
- 2.4 N-way modelling 13
  - 2.4.1 Multi-way decomposition methods 14
  - 2.4.2 Multi-way calibration 15
  - 2.4.3 Missing data 15
- 2.5 Validation of regression models 16

3 Techniques for preprocessing data 17
- 3.1 Removal of multiplicative and additive effects 17
  - 3.1.1 Additive corrections or baseline corrections 17
  - 3.1.2 Multiplicative corrections or normalization 19
  - 3.1.3 Combined methods 20
- 3.2 Modelling non-linear behaviour 20
  - 3.2.1 Transformation of data 20
  - 3.2.2 Local modelling 21

3.2.3 Quadratic PLS . . . 21

4 Effects of preprocessing NIR spectra 23

4.1 Baseline correction prior to curve resolution . . . 23

4.1.1 Inspection of data and removal of baseline . . . 24

4.1.2 A discussion of baseline correction . . . 25

4.2 Additive and multiplicative correction of process data . . . 26

4.3 A comparison of techniques for modelling non-linear structure . . . 29

5 Assessment of modelling techniques 33

5.1 Curve resolution for quantitative measurements . . . 34

5.1.1 Background of the system . . . 34

5.1.2 Rank analysis . . . 34

5.1.3 The resolution of spectra . . . 35

5.1.4 Quantitative study based on the resolved profiles . . . 36

5.1.5 A summary on resolving NIR spectra . . . 37

5.2 Performance of two-way regression methods . . . 38

5.2.1 Number of significant components . . . 38

5.2.2 Variable selection . . . 39

5.2.3 Comparison of predictive ability of models with deviation in reference values . . . 40

5.2.4 Origin of the band used to model the responses . . . 41

5.2.5 Summary of feasibility study . . . 42

5.3 Performance of three-way regression methods . . . 43

5.3.1 Inspection of data . . . 44

5.3.2 Multiway regression . . . 45

5.3.3 N-way calibration for on-line monitoring: How to handle missing values . . . 47

5.4 Summary of multivariate modelling . . . 48

6 Conclusions and future outlook 49

References 53

List of papers

This thesis is based on the following publications referred to by their Roman numerals:

I Study of the self-association of alcohols by near-infrared spectroscopy and multivariate 2D techniques.
L. Stordrange, A. A. Christy, O. M. Kvalheim, H. Shen and Y.-z. Liang.
J. Phys. Chem. A 106 (2002) 8543-8553.

II Feasibility study of NIR for surveillance of a pharmaceutical process, including a study of different preprocessing techniques.
L. Stordrange, F. O. Libnau, D. M. Sørenssen and O. M. Kvalheim.
J. Chemom. 16 (2002) 529-541.

III A comparison of techniques for modelling data with non-linear structure.
L. Stordrange, O. M. Kvalheim, P. A. Hassel, D. M. Sørenssen and F. O. Libnau.
J. Near Infrared Spectr. 11 (2003) 55-70.

IV Multiway methods to explore and model NIR data from a batch process.
L. Stordrange, T. Rajalahti and F. O. Libnau.
J. Chemom. Intell. Lab. Syst. Accepted for publication.

Notation

Scalars are indicated by italics, e.g. a and A, and vectors by bold lower-case letters, e.g. $\mathbf{x}$. Bold-face capitals, e.g. $\mathbf{X}$, are used for matrices, and underlined bold-face capitals, e.g. $\underline{\mathbf{X}}$, for three-way arrays. The size of two-way data is given as $I \times J$, i.e. I objects and J variables, while for three-way data the size is given as $I \times J \times K$, i.e. I batches, J objects and K variables. $\mathbf{X}^T$ is the transpose of the matrix. $\mathbf{X}^+$ denotes the pseudoinverse of a matrix that can be calculated by $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ or in a numerically more stable manner by using the singular value decomposition of $\mathbf{X}$. The symbol $\otimes$ is the Kronecker product;

$$\mathbf{U} \otimes \mathbf{V} = \begin{pmatrix} u_{11}\mathbf{V} & \dots & u_{1J}\mathbf{V} \\ \vdots & \ddots & \vdots \\ u_{I1}\mathbf{V} & \dots & u_{IJ}\mathbf{V} \end{pmatrix}$$

while the symbol $\odot$ is the columnwise Kronecker product (also denoted the Khatri-Rao product); $\mathbf{U} \odot \mathbf{V} = (\mathbf{u}_1 \otimes \mathbf{v}_1 | \dots | \mathbf{u}_J \otimes \mathbf{v}_J)$.

Abbreviations

EFA	Evolving factor analysis
ETA	Eigenstructure tracking analysis
HPLC	High performance liquid chromatography
ITTFA	Iterative target transformation factor analysis
LPG	Latent projection graphs
MLR	Multiple linear regression
MSC	Multiplicative scattering correction
N-PLS	Multi-way PLS regression
NIR	Near infrared
OP	Orthogonal projection
OS-2	Optimized scaling method
OSC	Orthogonal signal correction
PARAFAC	Parallel factor analysis
PCA	Principal component analysis
PCR	Principal component regression
PLS	Partial least squares regression
PV	Principal variables
RMSEP	Root mean square error of prediction
RMSECV	Root mean square error of cross validation
SEP	Standard error of prediction
SVD	Singular value decomposition
QPLS	Quadratic partial least squares regression

Chapter 1

Introduction

> ".., today significant opportunities exist for improving the efficiency of pharmaceutical manufacturing and quality assurance through the innovative application of novel product and process development, process control, and modern process analytical chemistry tools."
>
> U.S. Food and Drug Administration [1]

Many industrial processes are run under non-optimal conditions with regards to quality and yield. Hesitation towards new techniques may have hindered full understanding and control of critical process parameters. Recently, guidelines for use of process analytical technology in industry were proposed by the U.S. Food and Drug Administration [1]. The guidance document is a request to the industry to replace conventional laboratory testing methods with more efficient quality assurance techniques. Over time, the goal is to improve the quality, the safety and the efficiency of the processes.

In order to utilize new techniques, proper measuring devices and data-analytical methods providing high sensitivity are required. Near-infrared (NIR) spectroscopy is a sensitive technique that has the ability to monitor quality before and during processes, and thereby to provide useful information for process control. Processes are often carried out in hazardous environments making the lifetime of analytical instrumentation short. An advantage of the NIR spectrophotometer is that it can be equipped with fiber-optic cables sending radiation to the process while the spectrophotometer is stored in a safe environment. However, the complexity of the technique requires advanced data-analytical methods for reliable prediction of product characteristics and for chemical interpretation.

Chemometrics can be defined as the application of statistical and mathematical procedures to extract information from chemical data [2]. The consensus is that chemometrics started around the early 1970s due to increased use of instrumentation generating large data sets and the need of multivariate techniques to model these data sets [3]. Since then, new chemometric techniques and applications have increased steadily. A review of chemometrics and spectroscopy has been given in [4, 5, 6].

The major aim of this thesis is to investigate NIR spectroscopy for monitoring processes by means of chemometric techniques. Vibrational spectra acquired in the

NIR-region are characterized by broad and highly overlapping bands. In addition, they have often problems with noise from instrumental parts or are affected by physical effects from light scattering. To use the spectra for process monitoring, preprocessing of data to reduce these effects prior to modelling is often required. There are several ways to model data for quantitative purposes. In this thesis, curve resolution has been applied along with two-way and three-way regression techniques. The type of preprocessing and modelling technique to use depends on the chemical system to be analysed. Based upon results given in this thesis, strategies for robust multivariate modelling of NIR-spectra are proposed.

1.1 Outline of the thesis

This thesis includes four papers. In paper I a procedure is given for curve resolution on near-infrared spectra to investigate self-association of alcohols. By means of curve resolution the spectral and concentration profiles of the different association species are resolved. The profiles are used for qualitative and quantitative studies of the system. The results are followed by a discussion of the curve resolution problem of NIR spectra.

The strength of curve-resolution techniques is their ability to find pure spectral and concentration profiles, without any a priori information. For regression purposes the spectral profiles of all components must be known. In the future, knowing all spectral profiles, their corresponding concentrations profiles can be found directly.

In paper I it is shown that even when resolving a simple system in a non-absorbing solvent, problems due to baseline and highly overlapping bands are experienced. With access to reference methods for the components of interest, partial least squares (PLS) is easier and more reliable to use.

Paper II is a feasibility study of NIR spectroscopy for measuring the content of chemical compounds in a complex process solution, using reference data from high performance liquid chromatography. The work is focused on how to obtain calibration models with high predictive ability and includes testing of different preprocessing and variable selection methods reported in the literature.

A curvature that was well described by a second order polynomial was revealed between process data and the reference values modelled in paper II. In paper III different techniques that account for the non-linear behaviour were tested. These included use of transformations, splitting of data into few close-to-linear models and of a PLS technique taking in the second order relationship between the predictor and response. The result was simpler models with better predictive ability.

In paper IV multiway methods were tested for their ability to explore and model NIR spectra. It was revealed that blocking of data having a non-linear behaviour from two-way into three-way can improve the predictive ability. Use of N-way techniques lead to certain problems related to variable selection and how to fill in for missing values. These issues have also been discussed in paper IV.

The NIR data analysed in paper I-IV were acquired manually. During the study it was planned to use automatically acquired FT-NIR data from this process. However, automatical acquisition resulted in severe noise and artefacts in data, and reliable spectra were not acquired until recently. Chapter 6 in this thesis describes the problems and gives a further outlook for automatical acquisition of NIR spectra.

1.2 Pharmaceutical process

In this thesis the pharmaceutical production of a contrast agent for X-ray diagnostics is used as a case study. The synthesis is given in Figure 1.1. The main

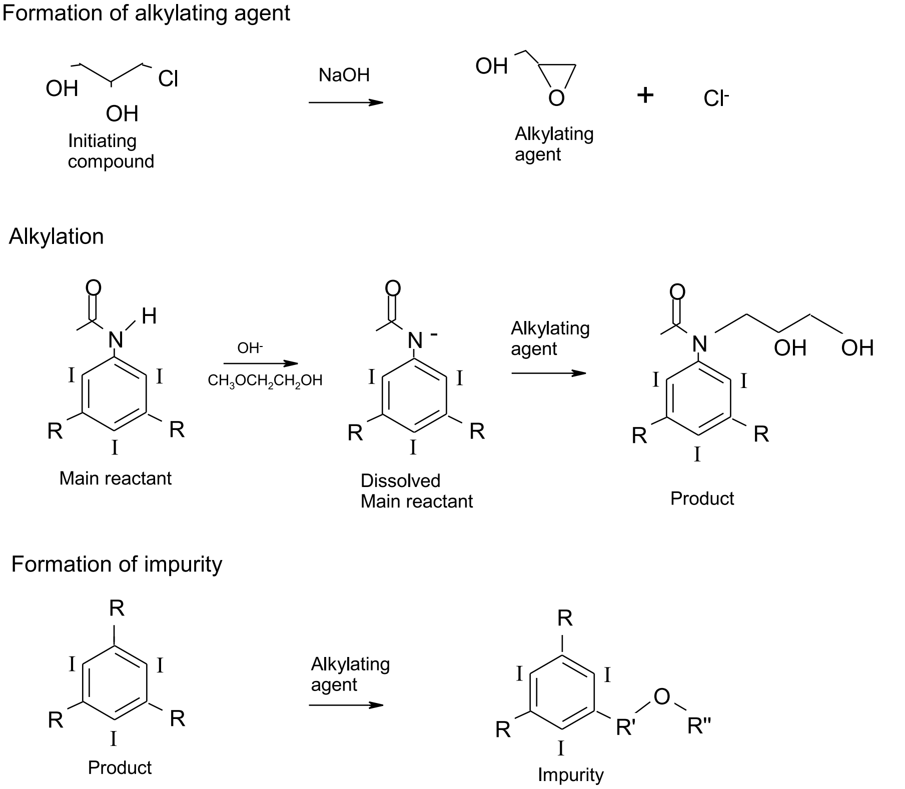

Figure 1.1: Reaction mechanisms in the process solution.

reactant is dissolved in an alkaline 2-methoxy-ethanol solution. When the initiating compound is added, the alkylating agent forms, and the transformation from the main reactant into the wanted product starts. When the product concentration gets high, it reacts with the alkylating agent forming impurity (see Figure 1.2). At termination of the reaction the goal is to have transferred as much as possible of the main reactant into product, with the formation of impurity suppressed. Therefore the synthesis is controlled empirically during reaction. The synthesis takes about 26 hours to complete. During the run there are three control points, where the area percent of the main reactant and the impurity are measured by high performance liquid chromatograpy (HPLC). If the reaction goes too fast or too slowly, analytes have to be added to adjust the reaction rate.

HPLC is a destructive and manual technique that requires operational personnel for sample collection and QC-personnel in the laboratory for the analysis. Near-infrared spectroscopy is a fast and automatic technique that is suitable for

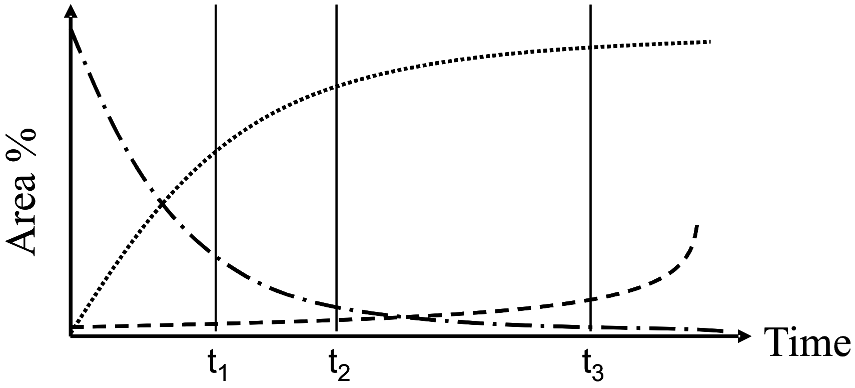

Figure 1.2: A sketch of concentration changes of the main reactant (- · - · -), the product (· · ·) and the impurity (- - -) during the synthesis of contrast agent. t_1, t_2 and t_3 refer to the three control points.

continuous monitoring. The process can be adjusted at an early stage and one may thereby prevent rejects or re-processing. The aim of using NIR spectroscopy in the process is to replace the HPLC technique with NIR technology. This is a calibration problem that is complicated due to the highly correlated variables in the NIR spectra, and due to that the spectra are noisy and show a non-linear behaviour. Different multivariate techniques and preprocessing techniques have been investigated on NIR data from the process.

NIR spectroscopy applied to the process described here is patented [7].

1.3 Process analysis

Process analysis techniques have been categorized into several groups, such as off-line, at-line, on-line, in-line, in situ, near-line, open-path, automatic, real-time and noninvasive techniques [8, 9, 10]. There are small differences between some of the categories. This makes naming the analyzer condition a confusing task. Instead a simpler division is used in this thesis. The categories used are depending upon whether the sampling is performed manually or automatically, and whether the technique is destructive or non-destructive. A non-destructive technique that acquires process data automatically is most desirable for process monitoring. Vibrational spectroscopy in the mid-IR and near-IR ranges, as well as Raman spectroscopy, fulfill these requirements.

The mid-IR region provides narrower bands, and spectra from this region are therefore more suited for quantitative and qualitative purposes compared to NIR spectra. However, in the mid-IR regions there are high transmission losses in fiber optics necessitate short fiber lengths. The use of fiber optic is highly desirable for applications of vibrational techniques to processes in corrosive environments. This restricts the use of mid-IR spectroscopy in many process applications. For NIR and Raman the radiation can be transmitted through fiber optics for long distances, and the instrumentation can be placed in a dry and temperature-controlled room. In addition to it is working well with long fiber optics, Raman spectroscopy provides a high level of spectral information with highly resolved bands. The result is robust models requiring less maintenance than models from near-IR spectroscopy. While IR absorbances are stronger for vibrations which involve a large dipole moment change, Raman bands are stronger for large changes in polarizability. Raman spectroscopy is therefore complementary to IR since materials with weak IR features often produce good Raman spectra. Raman spectroscopy is especially suitable for aqueous solutions since water is a weak Raman scatterer. A disadvantage with Raman spectroscopy is that the method has sensitivity problems in process analysis. Better instrumentation, i.e. laser diode and fiber probes, with equal sensitivity to mid-IR and near-IR is under development. For the moment NIR spectroscopy is assumed to be the best suited technique for vibrational process monitoring. For more information regarding vibrational process analysis, see [9, 10].

1.4 Principles of near-infrared spectroscopy

Infrared radiation was discovered by the astronomen William Herschel in 1800 when he investigated the distribution of heat in sunlight. He measured the temperature of the different colours in light, but it was not until he placed the thermometer below the red end of the visible spectrum that the temperature began to rise. The region was named infrared, with the Latin prefix infra meaning below [8].

Infrared spectroscopy is used to investigate the vibrational properties of a sample. Chemical bonds in a molecule vibrate with different frequencies. If the vibrations have the same frequency or energy as the light passed through a sample, and if the dipole moment of the molecule changes during the vibration, energy would be absorbed from the light and converted into vibrational energy. The absorptions at different frequencies are measured by a spectrometer, generating a spectrum.

Molecular vibrations give rise to absorption bands generally located in the mid-infrared range (2500 to 25000 nm). The near-infrared region is located in between the mid-infrared and the visible part of the electromagnetic spectrum and covers the interval between approximately 800 and 2500 nm. Absorptions in this region include overtones and combination bands of the fundamental bands observed in the mid-infrared region.

While overtones are restricted to the first, second and third overtone, there is no theoretical limit to the number of absorptions that can be involved in combining bands from the mid-IR region. The effects on near-IR spectra is that one finds absorptions at unexpected positions and broad peaks caused by the overlapping of a multitude of different absorptions. However, the absorptions due to overtones and combinations of bands are much weaker than those of the fundamentals, i.e. 10 to 100 times weaker for each higher transition. This restricts absorption to the strongly absorbing functional groups O-H, N-H and C-H.

1.5 Beer's law

Beer's law states that the absorption of light is proportional to concentration, c, if the path length, l, is kept at a constant level;

$$Abs = \epsilon l c \tag{1.1}$$

ϵ is the molar extinction coefficient. For a system following this law, regression can be performed by plotting concentration versus absorption. For near-infrared spectroscopy several species, A, absorb light at the same wavelength, λ. The light absorption is said to be additive;

$$Abs_{\lambda} = \sum_{a=1}^{A} \epsilon_a l c_a \tag{1.2}$$

Absorption spectra of a number of samples over a number of wavelengths can be arranged in a matrix, $\mathbf{X}$ $(I_{samples} \times J_{wavelengths})$, with one spectrum in each row;

$$\mathbf{X} = \sum_{a=1}^{A} \mathbf{c}_a \mathbf{s}_a^T = \mathbf{C}\mathbf{S}^T \tag{1.3}$$

The molar absorptivity, s_a, is the product of ϵ_a and l. Each column in the matrix $\mathbf{C}$ $(I \times A)$ corresponds to a concentration profile, while each column in the matrix $\mathbf{S}$ $(J \times A)$ corresponds to a spectral profile.

Chapter 2

Multivariate analysis

> ”Although NIR spectra are very complex, the fact that the same atoms are involved in many different absorptions means that these absorptions can be utilized, via complex mathematical analysis, to provide analytical information on specific functional groups.”
>
> T. Næs et al. [2]

2.1 Data reduction by PCA and SVD

An NIR spectrophotometer produces spectra of several hundred variables, and acquiring spectra over time generates large data sets. A data set can be reduced by principal component analysis (PCA) [11, 12] or singular value decomposition (SVD) [13, 14]. The centered matrix $\mathbf{X}$ is decomposed onto an A-dimensional subspace by using the least squares criterion;

$$\mathbf{X} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^T + \mathbf{E} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{2.1}$$

$$\mathbf{X} = \sum_{a=1}^{A} \mathbf{u}_a \sigma_a \mathbf{v}_a^T + \mathbf{E} = \mathbf{U}\Sigma\mathbf{V}^T + \mathbf{E} \tag{2.2}$$

For PCA the terms $\mathbf{t}_a$ are called score vectors while $\mathbf{p}_a$ are the loading vectors. In SVD, $\mathbf{u}_a$ are called the left singular vectors and $\mathbf{v}_a$ the right singular vectors. Σ is a diagonal matrix of singular values in descending order. These are the lengths of the new principal axis in the A-dimensional subspace. $\mathbf{T}$ is equal to the product of $\mathbf{U}$ and Σ, and $\mathbf{P}$ is equal to $\mathbf{V}$. PCA is therefore theoretically equivalent to SVD. The columns in matrix $\mathbf{T}$ are orthogonal, while the columns in $\mathbf{P}$, $\mathbf{U}$ and $\mathbf{V}$ are orthonormal. $\mathbf{E}$ is the variance in $\mathbf{X}$ not explained by the A-dimensional subspace, and is often referred to as noise. SVD can be computed by eigenvalue decomposition of the symmetric matrices $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$ giving $\mathbf{U}$ and $\mathbf{V}$, respectively [13]. The eigenvalue decomposition method decomposes onto one basis, called eigenvectors. For example, $\mathbf{X}^T\mathbf{X}$ can be decomposed onto the eigenvectors $\mathbf{V}$; $\mathbf{X}^T\mathbf{X} = \mathbf{V}\Lambda\mathbf{V}^T + \mathbf{E}$, Λ is the eigenvalues. The singular values are the square root of the eigenvalues.

2.2 Curve resolution

A system where no a priori information about the chemical components is available is classified as a black system [15]. A technique for resolving spectra of a black analytical two-component systems was given by Lawton and Sylvestre as early as in 1971 [16]. They named it self-modeling curve resolution (SMCR), where the self-modeling refers to that no assumptions are made concerning the shapes of the unknown. Often the curve resolution is highly supervised and the "self-modeling" term is therefore omitted here. The aim of performing curve resolution on NIR systems is twofold; (i) estimate spectra (qualitative analysis) and (ii) concentration profiles (quantitative analysis) of all the chemical species present in the system. This can be written as the decomposition of data matrix **X** into concentration profiles **C** and spectral profiles **S**;

$$\mathbf{X} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \tag{2.3}$$

Many techniques have been proposed for solving Equation 2.3. The two most important assumptions are that the profiles in **C** and **S** must be non-negative, and that there are no embedded peaks [17].

2.2.1 Rank analysis

For a black system a good start is to determine the rank of the system. A matrix **X** of size $I_{samples} \times J_{wavelengths}$ has maximum rank equal to the smaller of the numbers I and J. The chemical rank, A, is equal to number of chemical compounds causing variation in data. Due to collinearity in spectroscopic data, A is often smaller than I or J. In Equation 2.1 A latent variables are extracted. The product of **T** and $\mathbf{P}^T$ is a matrix with equal dimensionality as **X**, but the rank is equal to A. The dimension of the new matrices **T** and $\mathbf{P}^T$ is $I \times A$ and $A \times J$, respectively.

There are several ways to determine the rank of a system. In curve resolution, PCA and evolving factor analysis (EFA) [18] are frequently used methods. In PCA the rank is set equal to the number of significant eigenvalues. As the name implies, EFA is an evolving factor analysis of sub matrices of **X**. First, an eigenvalue decomposition is performed on a small sub matrix consisting of the two first objects (or variables), thereafter the sub matrix is enlarged with the third object and a new eigenvalue decomposition is performed, etc. to all the I objects are included. By plotting the eigenvalues versus objects, the number of components and when they evolve can be detected.

PCA for rank analysis can be improved by smoothing noise structure in data [19]. In PCA the eigenvalues are found by solving Equation 2.4;

$$\mathbf{X}^T\mathbf{X}\mathbf{v} = \lambda\mathbf{v} \tag{2.4}$$

v is the eigenvector and λ is the eigenvalue. This equation can be expanded with a smoothing factor $(\mathbf{I} + k\mathbf{G}^T\mathbf{G})^+$;

$$(\mathbf{I} + k\mathbf{G}^T\mathbf{G})^+\mathbf{X}^T\mathbf{X}\mathbf{v} = \lambda\mathbf{v} \tag{2.5}$$

I is the identity matrix, **G** is a second-order differentiation matrix, and k is the degree of smoothing. If the rank is equal to A, the logarithm of the ratio between eigenvalues from Equation 2.4 and 2.5 results in a significant increase at the $A+1^{th}$ eigenvalue. Plotting these values versus number of eigenvalues gives a tool to evaluate the rank visually. The idea is that the eigenvalues remain the same after smoothing while the noise residuals decrease. It is this change that is reflected in the plot. The method is in this thesis referred to as smooth PCA. To set the rank in paper I, smooth PCA was used along with PCA and EFA.

2.2.2 Detection of selective regions

Selective regions are consecutive points in the spectral or concentration direction where only one component contributes to the signal. Selective regions can be detected by latent projection graphs (LPG) [20] or eigenstructure tracking analysis (ETA) [21].

The LPG method uses the fact that vectors in a selective interval are proportional, and that their end points will map a straight line in space. Furthermore, a selective region being projected onto two arbitrarily chosen orthogonal axes, but non-orthogonal to the straight-line segment, fits as a straight line passing through the origin into the bivariate plot defined by the two axes. The result is that selective concentration and wavelength regions can be visually detected by inspection of score and loading plots (see Figure 2.1).

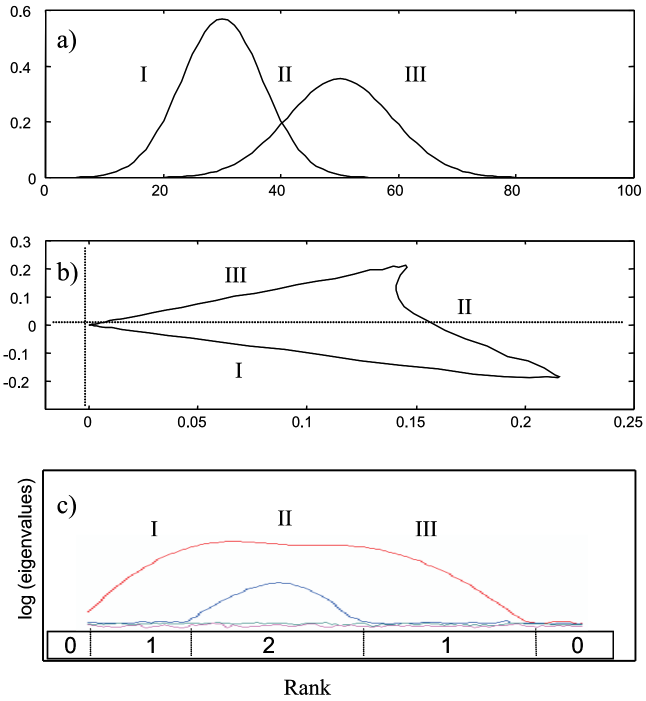

Figure 2.1: Performance of the LPG and the ETA method. a) Concentration profiles of a simulated data set, and the corresponding b) LPG plot and c) ETA plot. I and III are one-component regions, while II is a two-component region.

If the overlapping between spectral bands is severe, LPG may not be sensitive enough for the visual detection of selective regions. Instead, ETA can be used to find selective regions. To make an ETA plot, eigenvalue decomposition is performed on sub matrices of a fixed size, from start to the end of the spectra. By plotting the logarithm of the eigenvalues versus wavelengths a measure of number of components at the different wavelengths arises (see Figure 2.1). ETA is closely related to fixed-size-moving-window evolving factor analysis (FSMW-EFA) [22]. The difference is that in FSMW-EFA one chooses one window size, while ETA starts with a window-size of two up to maximum size of overlapping components, making it easier to detect when a new component evolves.

2.2.3 Resolution in the presence of selective regions

If selective regions can be found for all the significant components, for example in the concentration direction, then it is possible to calculate the concentration profiles directly;

$$\mathbf{C} = \mathbf{X}\mathbf{S}_s(\mathbf{S}_s^T\mathbf{S})^{-1} \tag{2.6}$$

The matrix $\mathbf{S}_s$ is the spectra from selective regions. In paper I, PCA performed on each selective region provided an estimate of the spectrum as the first loading vector, $\mathbf{p}$. The spectra $\mathbf{S}_s$ are then equal to $[\mathbf{p}_1\mathbf{p}_2 \ldots \mathbf{p}_A]$. In the same way, the first score vectors calculated from selective regions detected in the spectral direction, $\mathbf{C}_s$, can be used to estimate the unique spectral profiles by least squares; $\mathbf{S}^T = (\mathbf{C}_s^T\mathbf{C}_s)^{-1}\mathbf{C}_s^T\mathbf{X}$.

2.2.4 Resolution in the absence of selective regions

Frequently, not all the components have selective regions. Iterative target transformation factor analysis (ITTFA) [23, 24] is a method where selective regions are not essential. The loadings from PCA are abstract spectral profiles. In ITTFA these are transformed by rotation into the true profiles; $\mathbf{C}\mathbf{S}^T = \mathbf{T}\mathbf{R}\mathbf{R}^{-1}\mathbf{P}^T$. $\mathbf{R}$ is the invertible rotation matrix. A start vector is rotated into the eigenvector space defined by the first eigenvectors in such a way that the squared sum of the residuals between rotated start vector and start vector is minimal. A common start vector is [0 0. . . 0 1 0. . . 0 0] where the number one gives the position of the variable/target to be tested. The solution can be iterated by using constraints. For spectra the constraint of non-negativity is very useful. For every iteration the vector is approaching the true solution. Iterations are performed until a stop criterion is fulfilled, e.g. max iterations or the difference between two consecutive iterations is small.

Another useful method in the absence of selective regions is orthogonal projections (OP) introduced by Lorber [25]. He showed that the net analytical signal for a component is equal to the part of the spectrum that is orthogonal to the spectra of the other components. The part of a matrix $\mathbf{X}_A$ that is orthogonal to a reduced rank region, $\mathbf{X}_{A-1}$, can be found by;

$$\mathbf{X}_{op} = [\mathbf{I} - \mathbf{X}_{A-1}^T(\mathbf{X}_{A-1}^T)^+]\mathbf{X}_A \tag{2.7}$$

The result is a matrix $\mathbf{X}_{op}$ where each row contains the contribution from the spectrum that is not a part of the reduced matrix $\mathbf{X}_{A-1}$. By summing each row in the concentration direction an estimate of the concentration profile for the last component is obtained. OP has two functions; i) the method confirms if the reduced matrix has rank reduced by one component, and ii) a profile calculated by e.g. ITTFA can be compared with OP to check if the profile is the true solution.

2.3 Two-way regression

The purpose of regression is to relate variables in $\mathbf{X}$ to concentrations $\mathbf{y}$, i.e. estimate the regression coefficients, $\mathbf{b}$, in;

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{f} \tag{2.8}$$

$\mathbf{X}$ and $\mathbf{y}$ are mean centered, while $\mathbf{f}$ denotes a vector of noise. If the columns in $\mathbf{X}$ are linearly independent, multiple linear regression (MLR) can be performed;

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.9}$$

The variables in NIR spectra possess a high degree of collinearity and therefore ordinary least squares may not be satisfactory due to the requirement that $\mathbf{X}^T\mathbf{X}$ should be invertible. Instead, principal component regression (PCR) can be performed where $\mathbf{X}$ in Equation 2.8 is replaced with the linearly independent score vectors $\mathbf{T}$ from PCA. Another commonly used method is partial least squares (PLS) [26]. PLS decomposes $\mathbf{X}$ and $\mathbf{y}$ using the criterion of maximum covariance explained. The PLS/NIPALS [27] algorithm is given in Table 2.1. Since PLS simultaneously model $\mathbf{X}$ and $\mathbf{y}$, the LV's have usually a higher correlation with the response variable and thus provides a more parsimonious model than PCR.

Table 2.1: Principles of the PLS/NIPALS algorithm. The regression coefficients are equal to; $\mathbf{b} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}$.

$\mathbf{G}$ is the column centered $\mathbf{X}$, and $\mathbf{h}$ is the column centered $\mathbf{y}$.	
for factor a = 1, 2, . . . , A	
$\mathbf{u}_a = \mathbf{h}$	
$\mathbf{w}_a^T = \mathbf{u}_a^T\mathbf{G}$	Loading weights
$\mathbf{w}_a = \mathbf{w}_a/(\mathbf{w}_a^T\mathbf{w}_a)^{1/2}$	Scale to length one
$\mathbf{t}_a = \mathbf{G}\mathbf{w}_a$	Scores for $\mathbf{X}$
$\mathbf{q}_a = \mathbf{h}^T\mathbf{t}_a/(\mathbf{t}_a^T\mathbf{t}_a)$	Loadings for $\mathbf{y}$
$\mathbf{p}_a = \mathbf{G}^T\mathbf{t}_a/(\mathbf{t}_a^T\mathbf{t}_a)$	Loadings for $\mathbf{X}$
$\mathbf{G} = \mathbf{G} - \mathbf{t}_a\mathbf{p}_a^T$	Remove the effect from factor a in $\mathbf{G}$
$\mathbf{h} = \mathbf{h} - \mathbf{t}_a\mathbf{q}_a^T$	Remove the effect from facor a in $\mathbf{h}$
$a = a + 1$	Start with next factor

2.3.1 Determination of significant components

The determination of the number of significant components is an important step in regression. One technique is to place all data available in one data set and perform cross validation, while another commonly used method is to split the data set into two, i.e. a calibration set to make a model and a validation set to test the predictive ability .

In cross validation [28] one or several samples are left out. Thereafter PLS models with 1, 2, ..., up to a given number of components, k, are calculated and predictions of the left-out samples are performed using these models. New samples are left out and new models calculated, etc. The prediction errors of the left out samples utilizing different number of components in the model are summarized, i.e. predicted residual error sum of squares (PRESS). The PRESS values can be plotted against number of components. The lowest PRESS value indicates number of significant components. Xu and Liang [29] showed that leave-one-out cross validation may lead to overfitting. Instead, they proposed to leave out about 50 % of the samples for validation each time. The samples are chosen randomly and the selection is performed, e.g. 500 times. The name of the method is Monte Carlo cross validation.

If the data set is split into a calibration and a validation set, explained variance of the responses of samples in an external validation set can be used to determine significant components. Models using 1, 2, ..., k components are made of samples in the calibration set. The number of significant components is based on the criterion of maximal explained variance of the responses of the validation samples. A randomized t-test is another method that can be used to evaluate the number of significant components based on the models predictive ability of samples in a validation set [30]. The method is a general distribution-free test for the equality of two distributions using paired data. The difference in the squared prediction errors of two conditions, A and B, is calculated as $d_{i,AB} = e^2_{Ai} - e^2_{Bi}$, i=1, 2, ..., I. The mean of differences between all samples is calculated, thereafter the level of significance is found by assigning random signs to d_i several times. The significance level is estimated by counting how many of the randomized signs of the differences have a higher mean than the differences between the two conditions; $p = (sum + 1)/(total + 1)$. For determination of significant values, condition A is the predicted values for the model with lowest prediction errors consisting of a components, while condition B is the predicted values obtained for models using $1, 2, \ldots, a-1$ components. If p is less than 0.05 the conditions A and B are different within a 95 % confidence interval, and a is the significant number of components to use in a model. If p is equal to or greater than 0.05 the conditions A and B cannot be said to be different.

2.3.2 Variable selection

Variable selection means finding a smaller number of variables having important information for modelling the response. Several variable selection methods exist. One possibility is to set small PLS loading values to zero [31]. A similar technique is interactive variable selection [32] and the modified interactive variable selection PLS [33, 34] that involves selective removal of single elements in the PLS weight vector. Another method proposed extends the matrix with artificial noise variables and delete the variables that are less important than the artificial variables

[35]. Stepwise deletion of variables using the uncertainty variance of the regression coefficients estimated by jack-knifing is another method [36].

In paper II and III interval PLS (iPLS) [37] and principal variables (PV) [38] were used for variable selection. For iPLS the variable region is split into many small equidistant regions, thereafter the root-mean-square error of cross validation (RMSECV) is calculated between every region and the response. The region with lowest RMSECV is chosen. PV finds the variable with greatest covariance to the response, this is called the first PV. Thereafter the data matrix is orthogonalized to the first PV, removing information related to this variable. The second PV is the variable in the orthogonalized matrix with greatest covariance with the response. The reduced data matrix is orthogonalized to the second PV. This process is continued until the data is exhausted for systematic variation.

2.4 N-way modelling

When performing standard chemometric analysis as PCA and PLS, data are arranged in a two-way structure, e.g. *samples* × *wavelengths*. Analytical data can often be structured into higher order arrays. For batch process data, the samples can be sorted into batch and time according to when the spectrum was acquired. The result is a three-way array of *batch* × *time* × *wavelength* (see Figure 2.2).

For modelling third or higher order arrays, multi-way techniques are necessary. The first techniques for decomposing N-way arrays were proposed by psychometricians already in the 1960s. Applications of N-way techniques for solving chemical problems is on the other hand relatively new. N-way techniques as Tucker3 [39] and parallel factor analysis (PARAFAC) [40] from psychometrics have been adopted to explore and model chemical systems. These are decomposition methods like PCA.

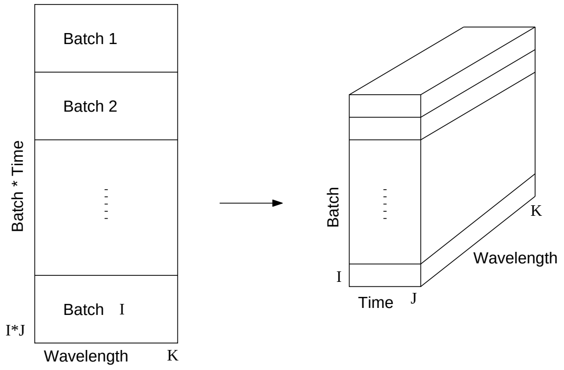

Figure 2.2: Spectra acquired at specific time intervals of a batch process can be stacked as a three-dimensional array (batch×time×wavelength).

For regression, a PLS algorithm for three-way data was developed by Ståhle [41]. Later Bro [42] developed a general multi-way PLS (N-PLS) for third or higher order arrays. During the last decade these and related techniques have received much attention for solving chemical systems. Successful applications have been given in curve resolution [43, 44], exploratory analysis [45] and calibration [46].

2.4.1 Multi-way decomposition methods

In PARAFAC the three-way array $\underline{\mathbf{X}}$ is decomposed into one score matrix, $\mathbf{A}$ $(I \times R)$, and two loading matrices, $\mathbf{B}$ $(J \times R)$ and $\mathbf{C}$ $(K \times R)$, by alternating least squares. R is the number of components (see Figure 2.3). First the three-way array, $\underline{\mathbf{X}}$ $(I \times J \times K)$, is matricized to a two-way matrix, $\mathbf{X}$ $(I \times JK)$. The loadings in $\mathbf{B}$ and $\mathbf{C}$ are set as e.g. random numbers;

$$\mathbf{X}^{I \times JK} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T + \mathbf{E}^{I \times JK} \tag{2.10}$$

where $\odot$ is the the columnwise Kronecker product [47]. Setting $(\mathbf{C} \odot \mathbf{B})^T$ equal to $\mathbf{D}$, the columnwise loadings in $\mathbf{A}$ can be estimated by alternating least squares; $\mathbf{A} = \mathbf{X}^{I \times JK}\mathbf{D}^T(\mathbf{D}\mathbf{D}^T)^{-1}$. The loadings $\mathbf{B}$ and $\mathbf{C}$ are estimated likewise, starting by unfolding the three-way array to a matrix of size $(J \times IK)$ and $(K \times IJ)$, respectively. New loadings are calculated until convergence. Since no extra constraints such as orthogonality are needed to decompose the three-way array as for PCA, the parameters can be determined uniquely. PARAFAC is therefore suitable for curve resolution.

Contrary to PARAFAC, Tucker3 allows for extraction of different number of

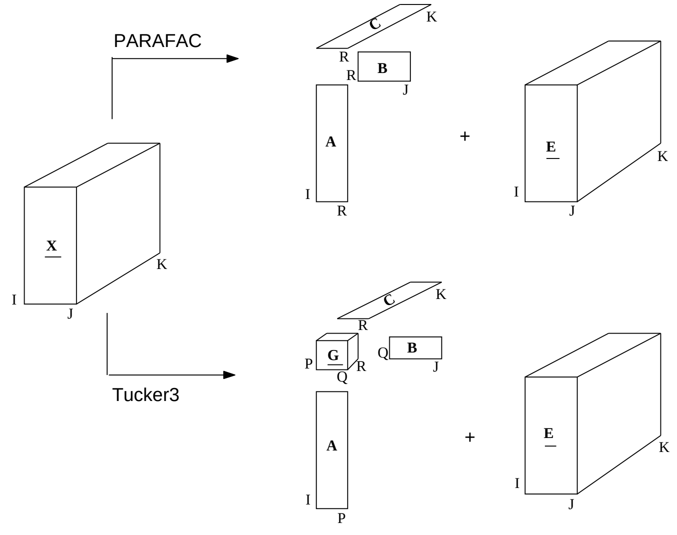

Figure 2.3: Three-way decomposition by PARAFAC and Tucker3.

factors in each direction. A core array $\underline{\mathbf{G}}$, of size $(P \times Q \times R)$, is employed (see Figure 2.3);

$$\mathbf{X}^{I \times JK} = \mathbf{A}\mathbf{G}^{P \times QR}(\mathbf{C} \otimes \mathbf{B})^T + \mathbf{E}^{I \times JK} \tag{2.11}$$

The $\otimes$ is the Kronecker product [47]. Several Tucker3 algorithms exist [48]. In paper IV, the columns in **A**, **B** and **C** are constrained to be orthogonal. **A** is initialized by; $[\mathbf{A}, \Sigma, \mathbf{V}] = SVD(\mathbf{X}^{I \times JK}(\mathbf{C} \otimes \mathbf{B}), P)$. P is the number of components to be extracted in mode **A**. **B** and **C** are determined in the same way. Iteration is performed until relative change in fit is small. Thereafter the core array $\underline{\mathbf{G}}$ can be determined by a simple regression of $\underline{\mathbf{X}}$ onto **A**, **B** and **C**. The core gives a summary of all interactions present in the three-way data and is therefore useful for exploration of data. Due to the rotational freedom the Tucker3 model is not structurally unique as the PARAFAC model.

2.4.2 Multi-way calibration

For calibration of N-way arrays there are several approaches. As for principal component regression (PCR) the data can first be decomposed by PARAFAC or Tucker3, thereafter the score matrix **A** can be related to a response **y** as in Equation 2.8.

N-PLS seeks in accordance with PLS to decompose data into a few significant components that simultaneously describe the variation in **X** and **y**. For bilinear PLS the intention is to find a weight vector, **w**, that yields a score vector, **t**, with maximal covariance with **y** (see Table 2.1). For trilinear PLS the goal is to decompose $\underline{\mathbf{X}}$ into two weight vectors, $\mathbf{w}^J$ and $\mathbf{w}^K$, corresponding to second and third mode, that produces a score vector with maximal covariance with **y**. While the weights in PLS can be determined by calculating the covariance between **X** and **y** directly, the three-way array needs to be matricized before calculating the weights. The result is a vector, $\mathbf{z} = \mathbf{X}^{(I \times JK)T}\mathbf{y}$. $\mathbf{z}^{1 \times JK}$ is matricized to **Z** $(J \times K)$. $\mathbf{w}^J$ and $\mathbf{w}^K$ are equal to the first left and right singular vectors of this matrix, respectively. The scores can now be found by; $\mathbf{t} = \mathbf{X}\mathbf{w}$, **w** is the Kronecker product between $\mathbf{w}^J$ and $\mathbf{w}^K$.

More details on theory and applications of multi-way analysis can be found in [49, 50].

2.4.3 Missing data

For on-line monitoring of a batch process, use of three-way techniques require data to be available for the entire duration of the batch. For multivariate statistical process control (MSPC) various strategies have been proposed [51]. In this thesis two approaches have been used; (i) Assume that future observations are in perfect accordance with historical data, i.e. for spectroscopic data the mean spectrum of samples in calibration set can be used to fill in for missing data, and (ii) use the ability of decomposition methods to handle missing data. For PARAFAC, the scores for a new batch up to current time point j can be calculated as;

$$\mathbf{a}_j = \mathbf{x}_{new,1:j}\mathbf{D}_{1:j}^T(\mathbf{D}_{1:j}\mathbf{D}_{1:j}^T)^{-1} \tag{2.12}$$

$\mathbf{x}_{new,1:j}$ is the vectorized $\mathbf{X}^{j\times K}$, and $\mathbf{D}_{1:j}$ is the reduced columnwise Kronecker product product of $(\mathbf{B}_{1:j}\odot\mathbf{C})^T$. To estimate the process measurements from *j+1:J* the product of the scores and the loading matrix corresponding to the time period *j+1:J* is used;

$$\mathbf{x}_{new,j+1:J}=\mathbf{a}_j\mathbf{D}_{j+1:J} \tag{2.13}$$

A combination of the already known observations up to j and the estimated values from *j+1:J* can be combined;

$$\mathbf{x}_{new}=[\mathbf{x}_{new,1:j}|\mathbf{x}_{new,j+1:J}] \tag{2.14}$$

$\mathbf{x}_{new}^{1\times JK}$ is matricized to a matrix of size $J\times K$ with no missing values. Use of projection method to fill in for missing values can also be performed for Tucker3 and N-PLS.

Meng et al. [52] used PARAFAC for MSPC. In the paper they presented a methodology where the unknown observations are calculated as a weighted combination of the scores up to the current time point, $\mathbf{a}_j$, in the new batch and those previously computed from a calibration data set, $\mathbf{a}_{mean}$;

$$\mathbf{a}_{weight,j}=\lambda_j\mathbf{a}_{mean}+(1-\lambda_j)\mathbf{a}_j,\qquad \lambda_j=\frac{J-j}{J-1} \tag{2.15}$$

Thereafter estimates of missing spectra are determined by calculating the steps in Equation 2.13 and 2.14. In the paper of Meng the method was shown to give smoother scores avoiding spurious alarms in control charts.

In paper IV the projection method is tested, both for unweighted and weighted scores along with mean spectra.

2.5 Validation of regression models

The predictive ability of a model can be evaluated by root mean square error of prediction (RMSEP) [2];

$$RMSEP=\sqrt{\frac{\sum_{i=1}^{I}(\hat{y}_i-y_i)^2}{I}} \tag{2.16}$$

I is number of samples. The same formula applies to root mean square error of cross validation (RMSECV). If a long time has passed or small differences in components between the different instruments are present, correction of the predictor offset b_o for bias may sometimes be useful [31].

$$Bias=\sum_{i=1}^{I}(\hat{y}_i-y_i)/I \tag{2.17}$$

The predicted values, $\hat{y}_i$ are bias corrected and the resulting prediction error is then expressed by the standard error of prediction (SEP);

$$SEP=\sqrt{\frac{\sum_{i=1}^{I}(\hat{y}_{i,biascorr.}-y_i)^2}{I-1}} \tag{2.18}$$

To evaluate if there is a significant difference between models, the randomized t-test [30] described in section 2.3.1 can be employed.

Chapter 3

Techniques for preprocessing data

Noise will always be present in spectra, and to reflect a real chemical system an extra noise matrix **E** should be added to Equation 1.3 ;

$$\mathbf{X} = \sum_{a=1}^{A} \mathbf{c}_a \mathbf{s}_a^T = \mathbf{C}\mathbf{S}^T + \mathbf{E} \tag{3.1}$$

Noise originates from instrument errors and from physical effects such as light scattering. For successful modelling, noise or irregular patterns in data not reflecting the chemical composition of the sample should be removed or left unmodelled. Noise and unwanted effects are often removed by mathematical transformations prior to the data analysis. The two main reasons for preprocessing spectra are either to remove noise and irrelevant features in the data, or to take care of irregularities such as non-linearity in the data [53].

3.1 Removal of multiplicative and additive effects

One classification of noise removal techniques is after their ability to remove additive or multiplicative effects [27]. Additive effects are differences in baseline between samples, while multiplicative effects occur when the pathlength differs from sample to sample.

3.1.1 Additive corrections or baseline corrections

A background is often present in near infrared spectra. It originates from reflection at the cell wall and from scattering of light by the solvent. A baseline can be expressed as a polynomial [27];

$$\mathbf{e}(j) = bo\mathbf{1} + b_1\mathbf{x} + b_2\mathbf{x}^2 + \ldots + b_n\mathbf{x}^n \tag{3.2}$$

An offset or flat baseline is expressed using $b_o\mathbf{1}$, a baseline with offset and slope with $b_o\mathbf{1} + b_1\mathbf{x}$, a curved baseline with $b_o\mathbf{1} + b_1\mathbf{x} + b_2\mathbf{x}^2$ etc. One latent variable is needed to explain each term in Equation 3.2. A correct baseline correction reduces the number of significant factors in a singular value decomposition.

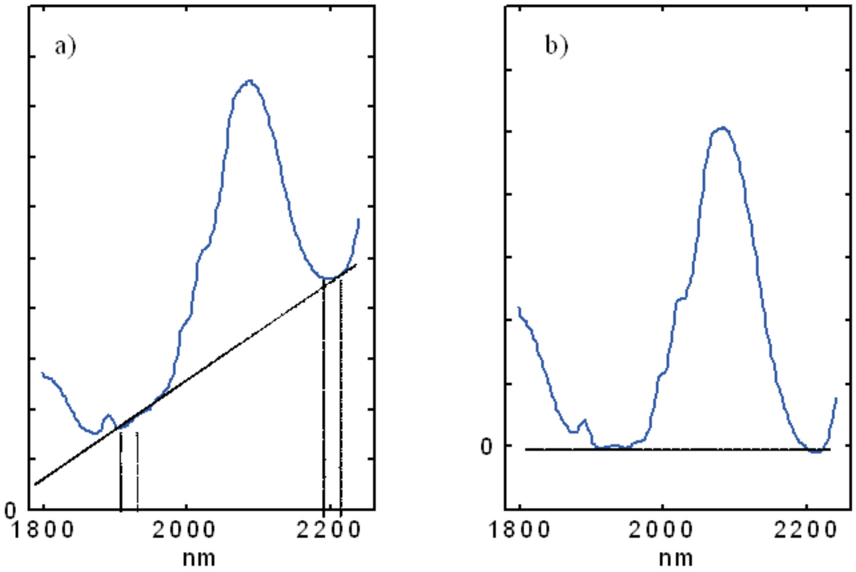

Figure 3.1: Baseline correction of a spectrum a) before and b) after removal of a straight line through two "zero-component" regions.

The baseline is often corrected prior to curve resolution since a baseline will complicate the detection of selective and other regions. Liang et al. have developed a baseline correction method for two-way chromatography-spectroscopy data [54]. The method defines a zero-component region before and after the elution of a multi-component region. A simple least-squares fit is made of a straight line through all the elements between the two zero-component regions. An estimate of the baseline is calculated for all the variables and the baseline is removed by subtraction. For spectroscopic data, the number of bands, their shape and their width changes with concentrations. It is therefore difficult to define fixed zero-concentration regions over a large concentration scale. Baseline removal of spectra can instead be performed one spectrum at the time, choosing the minima before and after the interesting region as the "zero-concentration" regions (see Figure 3.1).

Numerical differentiation is often used for baseline corrections of near infrared spectra. First order differentiation removes the offset and second order differentiation removes the offset and slope etc. in Equation 3.2. If differentiation is performed directly, the noise increases. Therefore, the calculation of derivatives is usually combined with a smoothing procedure. The Savitzky-Golay procedure [55] combines smoothing and differentiation into one single step. A polynomial is fitted to the data in a window. The middle point of the window is interchanged with the value of the polynomial, thereafter the window is moved one step and a new value is calculated and so on through all points in the spectrum. For curve resolution, differentiation may cause problem since it introduces negative absorbances. Iterative curve resolution methods uses often non-negativity constraint, which of course cannot be used for negative data. A way to solve this is to invert the sign of second order differentiated data and thereafter set negative values equal to zero [56]. The method gives a decrease in the signal-to-noise ratio and a part of the band will be removed. For the curve resolution problem in paper I, the method of Liang et al. was chosen instead of numerical differentiation. Negative signs is not a problem when performing regression, and differentation was chosen as baseline correction method in paper II, III and IV.

3.1.2 Multiplicative corrections or normalization

Multiplicative effects occur when the pathlength varies from sample to sample. The absorbance and concentration are therefore not directly related to each other. If one is only interested in the relative amounts of the chemical components, the effect of total sample quantity can be removed from the data. The correction is called normalization or multiplicative correction.

The aim of normalization is to make spectra acquired from samples of equal composition comparable. Such samples may exhibit different spectra because different light paths were used during the acquisition. There are two ways of performing normalization, either by making use of an internal standard, or by normalization to a constant sum or length. Normalization to length one is performed by dividing each spectra by their norm. In paper II and III the spectra were normalized to length one.

Optimized scaling, introduced by Karstang and Manne [57], introduces scale factors for the individual samples. Two variants of the method were introduced, OS-1 and OS-2. While OS-1 requires all the constituent concentrations to be known, OS-2 was developed for the more common case where the reference values for only one constituent is modelled. OS-2 calculates the scaling vector, $\mathbf{c}$, and regression coefficients, $\mathbf{b}$, for object i over the variables j, in one step;

$$y_i = \frac{\sum_j x_{ij} b_j}{\sum_j x_{ij} c_j} \tag{3.3}$$

When solving this equation by least squares, the trivial solution is avoided using an auxiliary condition, setting the scale of one sample, numbered 0, equal to 1; $\sum_j x_{0j} c_j = 1$. This is introduced by the Lagrangian multiplier, λ. The equation for determination of $\mathbf{c}$ and $\mathbf{b}$ is equal to;

$$\begin{pmatrix} \mathbf{Z}^T\mathbf{Z} & \mathbf{Z}^T\mathbf{X} & \mathbf{x}_o \\ \mathbf{X}^T\mathbf{Z} & \mathbf{X}^T\mathbf{X} & 0 \\ \mathbf{x}_0^T & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{c} \\ -\mathbf{b} \\ -\lambda \end{pmatrix} = \begin{pmatrix} 0 \\ -y_0\mathbf{x}_0 \\ 1 \end{pmatrix}$$

and z_{ij} is equal to $y_i x_{ij}$. The equation is solved by least squares. Since the matrix to the left must be non-singular, it is beneficial to replace the raw data $\mathbf{X}$ by the significant principal component scores. For almost noiseless data the choice of reference sample is not important, but for noisy data the model can be tested for all possible reference samples. In paper II and III, OS-2 is used on noisy spectral data from a process and a search through all samples was carried out.

3.1.3 Combined methods

Multiplicative scattering correction (MSC) [58] combines a spectral normalization procedure with baseline correction. A linear regression is performed between a reference spectrum $\mathbf{x}_{ref}$ and the ith spectrum in the matrix. Thereafter the intercept b is subtracted and the spectrum divided by the slope a, one spectrum at the time;

$$\mathbf{x}_i = a\mathbf{x}_{ref} + b \tag{3.4}$$

$$\mathbf{x}_{i,corrected} = (\mathbf{x}_i - b)/a \tag{3.5}$$

The intention when MSC was developed was to choose ranges with no chemical information. NIR-spectra have few or no regions with no chemical information. Instead a larger spectral region can be used where the mean spectrum functions as the reference spectrum.

Orthogonal signal correction (OSC) removes information that is orthogonal, and therefore irrelevant, to the response. The method was introduced by Wold et al. [59]. Several methods based on the same idea has been introduced, e.g. a faster algorithm by Fearn [60], direct orthogonalization by Andersson [61] and a method loosening the orthogonality constraints by Westerhuis et al. [62]. The method has obtained a great deal of attention but has so far not been shown to give convincing results. Orthogonal projections have instead been found to be beneficial for interpretation purposes [63].

3.2 Modelling non-linear behaviour

The most common regression techniques assume a linear relationship between the data to be modelled and the response. For real systems this is not always the case. Sources to non-linearities in data may be [31, 64]; (i) instrumental effects, i.e. saturation of detector at high analyte concentration, or spectral shifts, (ii) physical effects from scattering of particles or temperature changes with time, (iii) chemical effects, i.e. the signal originates from another analyte not linearly related to the analyte to be modelled, and (iv) mathematical effects, i.e. spectroscopic measurements are measured in transmittance instead of absorbance [A=-log(T)].

The system investigated in paper III has a curved relation between the response and the predictors. The easiest way to model data of this kind is by adding more LVs. However, this makes the model more vulnerable towards changes and a more parsimonious model is wanted. A number of non-linear methods to deal with curvature have been developed. Reviews of these can be found in [65, 66]. In paper III several non-linear variants of PCR and PLS have been tested. The paper includes (i) transformation of data by introducing non-linear terms, both of original data and of latent variables, (ii) local modelling/splitting of data and (iii) non-linear calibration techniques that introduce a quadratic term in the PLS-algorithm.

3.2.1 Transformation of data

A nonlinear behaviour in the predictors can be accounted for by enlarging the matrix with squared variables; $\mathbf{X}_{new} = [x_j|x_j^2]$, $j = 1, 2, \ldots, J$ [64, 67, 68]. Cross-terms may also be beneficial, however, for large matrices this is not recommended since taking all possible combinations results in huge matrices requiring large computational capacity. Instead, the matrix can be decomposed by PCA or PLS and subsequently a new matrix can be formed consisting of scores and their squares and cross-terms. In order to avoid overfitting, irrelevant variables should be removed. Blanco et al. [69] employed a stepwise addition of variables, i.e. new LV's and their quadratic and cross-correlated terms are added if they result in improved correlation coefficient in an multiple linear regression model. The method is called stepwise polynomial PCR (SWP-PCR) and stepwise polynomial PLS (SWP-PLS) according to decomposition method used. Stepwise deletion of variables utilizing the uncertainty of regression coefficients estimated by jack-knifing has been proposed by Westad et al. [36]. The ±2× estimated standard deviation for each regression coefficient is compared with the regression coefficient from the global model. If the standard deviation is greater than the regression coefficient from the global model, the variable is deleted.

For data with curvature, adding squared predicted response is another way of handling the non-linear problem. Two models are needed; a PLS model of the original matrix, and a PLS model where the predicted response and the transformed response are added to the original matrix.

3.2.2 Local modelling

The problem of curvature in data can be solved by splitting the data into a few approximately linear regions. The process investigated in paper II-IV has three control points. The data set can be assumed to behave linearly close to each control point, and one model can be made for each control point. The method is referred to as time dependent splitting. Another possibility is to use locally weighted regression (LWR) [70]. In LWR the q closest samples to a sample to be predicted, $\mathbf{x}_i$, are selected and a model made. Prior to regression the samples are weighted with a constant w_k, $k = 1, 2, \ldots, q$, after increasing Mahalanobis distance (φ) to $\mathbf{x}_i$;

$$w_k(\mathbf{x}_i) = W\Big(\frac{\varphi(\mathbf{x}_k, \mathbf{x}_i)}{d(\mathbf{x}_i)}\Big) = W(u), \qquad \begin{cases} W(u) = (1-u^3)^3 & \text{if } u \leq 1 \\ W(u) = 0 & \text{if } u > 1 \end{cases}$$

The function $d(\mathbf{x}_i)$ is the maximum of $\varphi(\mathbf{x}_k, \mathbf{x}_i)$ over the q points used in each regression. As a function of k, the weight, $w_k(\mathbf{x}_i)$, is large for $\mathbf{x}_k$ close to $\mathbf{x}_i$ and small for $\mathbf{x}_k$ far from $\mathbf{x}_i$. Local regression (LR) without weighting and utilizing the q closest points in Mahalanobis distance is another possibility.

3.2.3 Quadratic PLS

A polynomial quadratic PLS (QPLS) algorithm has been proposed by Wold et al. [71]. The idea of QPLS is to perform PLS, and at the same time fulfil a quadratic inner relation between the scores $\mathbf{t}$ and $\mathbf{u}$ of $\mathbf{X}$ and $\mathbf{y}$, respectively;

$$\begin{aligned} \mathbf{u} &= c_0\mathbf{1} + c_1\mathbf{t} + c_2\mathbf{t}^2 && (3.6) \\ &= c_0\mathbf{1} + c_1\mathbf{X}\mathbf{w} + c_2\mathbf{X}\mathbf{w}^2 && (3.7) \end{aligned}$$

The start is a PLS weight vector $\mathbf{w}$ which is updated iteratively. Equation 3.6 can be rewritten as; $\hat{\mathbf{u}} = f(\mathbf{X}, \mathbf{c}, \mathbf{w})$ and differentiated with respect to the unknown parameters, $\mathbf{c}$ and $\mathbf{w}$;

$$\hat{\mathbf{u}} = f_{00} + \frac{\partial f}{\partial \mathbf{c}} |_{00} \Delta\mathbf{c} + \frac{\partial f}{\partial \mathbf{w}} |_{00} \Delta\mathbf{w} \tag{3.8}$$

The result is a vector $\Delta\mathbf{w}$ used for updating $\mathbf{w}$; $\mathbf{w}+\Delta\mathbf{w}$. Baffi et al. have developed a simpler algorithm where $\Delta\mathbf{c}$ is omitted [72]. An improvement of the error based updating procedure of Baffi et al. has been suggested in [73]. In this method $\Delta\mathbf{w}$ is calculated from the regression coefficient of a separate inner PLS algorithm where the number of latent variables is selected using cross validation. In paper III the error-based QPLS of Baffi et al. was tested along with QPLS using cross validation, later referred to as QPLScv.

Chapter 4

Effects of preprocessing NIR spectra

NIR spectra are characterized by broad and highly overlapping bands. In addition, they often have problems with noise from instrumental or physical effects. Preprocessing is performed to reduce the contribution from noise and to enhance the chemical signal of interest. The choice of preprocessing procedures depends on the noise pattern in data and on the modelling technique to be used. In the sections below some examples of preprocessing of NIR spectra for curve resolution and regression purposes are given.

The first system is a simple chemical system in a non-absorbing solvent. The spectra are intended to be used for curve resolution. A method developed for baseline removal in chromatographic data is here adapted to NIR spectra. While the chemical systems used for curve resolution in paper I are clear solutions, the process solution described in section 1.2 is a viscous yellowish solution containing particles. Light scattering due to particles in the solution results in very noisy spectra. The process is controlled by HPLC, and the intention is to model important process parameters by NIR spectroscopy instead of by the time-consuming HPLC method. The modelling technique used is regression, and NIR data is regressed on the HPLC responses. The NIR data contain multiplicative and additive effects. Different preprocessing techniques for handling these effects are tested on the process data. In addition to additive and multiplicative effects, a curvature was present in the data. Results from non-linear methods to deal with this curvature are given in section 4.3.

4.1 Baseline correction prior to curve resolution

Paper I gives a procedure for resolution of NIR spectra. Solutions of alcohols were used as a case-study to investigate the self-association phenomenon of molecules. The study included several straight-chain and branched-chain alcohol solutions in the concentration range from 0.01 to 1.00 M. Two difficulties of performing curve resolution on NIR spectra were found. Both were related to the important task to reveal selective regions. One of the disadvantages of the NIR technique is that the bands are broad and highly overlapping. This makes it difficult to find regions

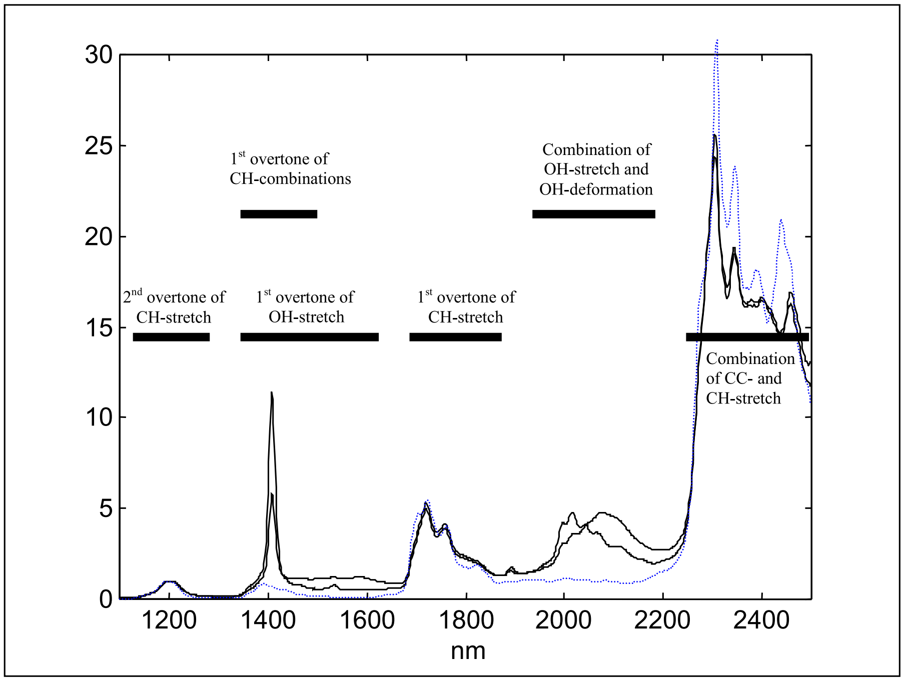

Figure 4.1: Spectra of 0.1M and 0.5M 1-pentanol (–) and n-pentane (···), normalized after the band at the 2^{nd} overtone of OH-stretch.

of low rank, and thereby diminishes the possibility of detecting selective regions. The other problem is the presence of baseline in the spectra, masking the selective regions. A linear baseline correction is performed on data investigated here [54].

4.1.1 Inspection of data and removal of baseline

The start of modelling should always be a thorough inspection of wavelength regions where interesting bands are expected to originate. The near-infrared region has two interesting regions for studying the self-association of alcohols. These are the region of the first overtone of OH-stretch, and the region of combinations of OH-stretch and OH-deformations (see Figure 4.1). Inspection of spectra of 1-pentanol and n-pentane in the near-infrared region revealed that CH-stretch absorbs strongly in the first overtone of OH-stretch. Many of the curve-resolution techniques require selective regions, and the contribution from CH-stretch must be removed before further analysis of data. Removal of CH-stretch utilizing n-alkane and general background subtraction (GBS) [74] failed.

CH-stretch does not absorb in the OH-combination region. However, upon inspection of Figure 4.1 a baseline was revealed. This was removed by linear baseline correction [54] (see section 3.1.1). The spectra and LPG plot, before and after baseline correction, are shown in Figure 4.2. The LPG plot gives the coordinates

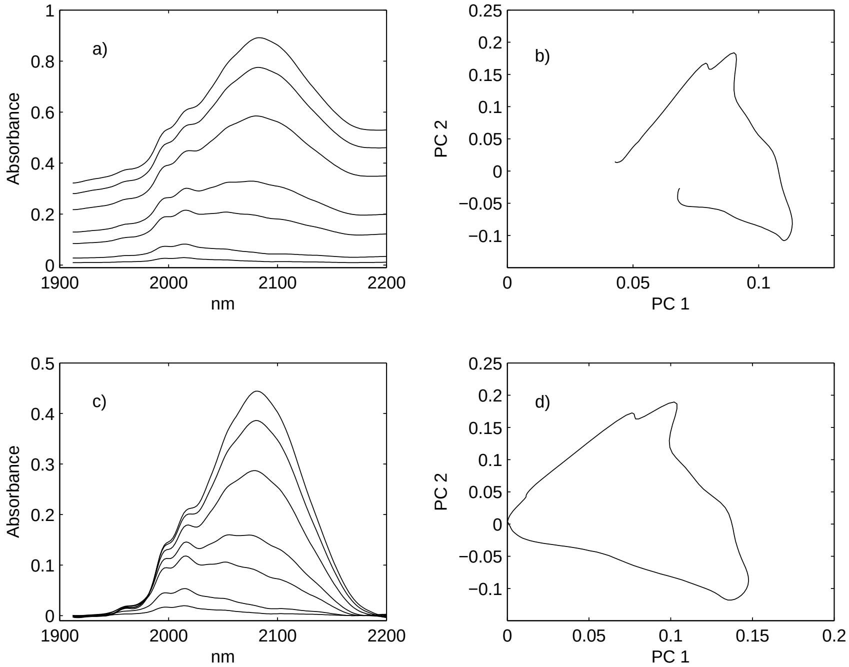

Figure 4.2: Results baseline correction of 1-heptanol; a) spectra and b) LPG-plot of data before baseline correction, and c) spectra and d) LPG-plot after baseline correction.

of the different wavelengths projected at the first and second principal components. For raw data the scores in the LPG plot start and end in different coordinates, and none of them in the origin. An offset and a slope is therefore present. After removal of baseline the scores starts and ends in the origin. Thus, the baseline is removed.

4.1.2 A discussion of baseline correction

The linear baseline correction of Liang et al. [54] requires wavelength regions where no compounds absorb radiation, both before and after interesting region. This is not likely to happen since the bands in the near infrared region are broad. The baseline correction is therefore expected to remove some absorbance due to OH-stretch vibrations.

General background subtraction [74] was tested for removal of the CH-stretch contribution in the first overtone of OH-stretch region, by utilizing *n*-pentane for 1-pentanol. After the subtraction negative regions were observed in the CH-stretch region. The reason for this may be that the number of methyl-, methylene and methin groups are not exactly the same, and that the presence of hydroxyl groups influence on the force constant to neighbouring atoms. A shift can therefore be expected comparing the absorbance of CH-stretch origin from alcohols and similar *n*-alkanes.

A possibility is to use compounds in which the hydrogen of the hydroxyl group is replaced by deuterium [75]. The oxygen-deuterium bond is not expected to give absorptions in the first overtone of the OH-stretch region, whereas the CH-stretch in the deuterated alcohol is expected to have force constants closely related to the CH-stretch in the alcohol having the same carbon skeleton. This could be tested using the subtraction method for the first overtone of OH-stretch, and for the OH-combination region.

A summary of the performance of the baseline correction methods tested on NIR spectra is given in Table 4.1.

Table 4.1: A summary of baseline correction methods tested on NIR spectra.

Linear baseline removal	No contribution from other compounds than the one to be modelled must be present. ÷ Method is brutal. Fitting of a polynomial may improve the results [76].
General background subtraction	Removal of contribution from other compounds. ÷ A spectrum having the same physical properties and all bands, except for the one of interest, must be available.

4.2 Additive and multiplicative correction of process data

There are several techniques for reducing additive and multiplicative effects. A few techniques are broadly used, while others techniques proposed have gained little attention. In paper II a feasibility study of NIR for monitoring a pharmaceutical process was performed. The study included testing of different preprocessing techniques. The results are presented here.

Samples from 20 batches were scanned manually on a FOSS NIRSystems 6500 spectrophotometer. Spectra were recorded in the region from 1100 to 2500 nm, with a resolution of 2 nm, using a transflectance probe attached by fiber optics to the spectrophotometer. For each sample an HPLC analysis was carried out to determine the content of main reactant and impurity. The samples were divided into a calibration set of 45 samples to make calibration models, and a validation set of 43 samples to test the performance of the models.

Inspecting the raw data in Figure 4.3 reveals a severe baseline problem. In addition, particles are present in solution which results in differences in light path from sample to sample. Therefore, both additive and multiplicative effects can be expected. Normalization and optimized scaling remove multiplicative effects, while differentiation removes additive effects. MSC and OSC are expected to deal with both additive effects and multiplicative effects. Since the data must be closed before the removal of multiplicative effects [27], only the results obtained performing multiplicative correction after additive correction, along with only additive correction and only multiplicative correction, are given in Table 4.2.

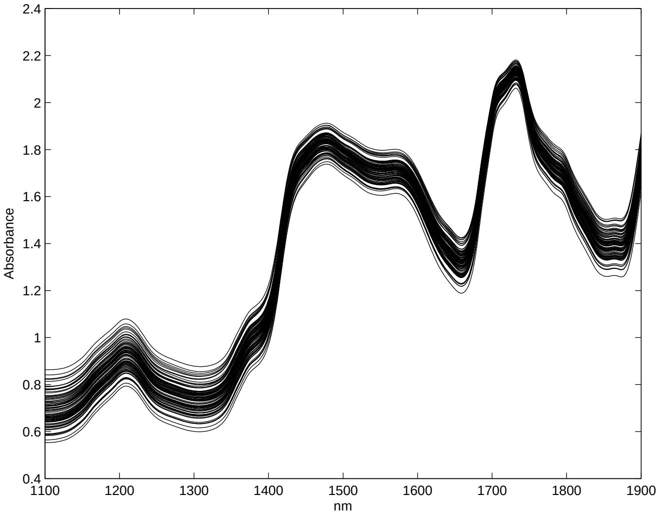

Figure 4.3: Spectra of raw data.

For the main reactant first-order differentiation provides better models than second-order differentiation. Second-order differentiation adds more noise than first-order differentiation. The spectra consist of highly overlapping bands, and higher-order differentiation increases the selectivity. Second-order differentiation provides more parsimonious models compared to other preprocessing techniques. Performing differentiation is therefore a trade-off between increased selectivity and increased noise level.

Small improvements were observed when combining differentiation with normalization. These result support that both additive and multiplicative effects are present in spectra. In a paper of de Noord [77], MSC on differentiated data was shown to perform better than using only MSC. This was not obtained here, in fact combining MSC with differentiation performed worse than using only MSC.

Table 4.2: Predictive ability of models using different preprocessing techniques for main reactant and impurity. The p-values were found using the predictions from optimized scaling on first order differentiated data.

	MAIN REACTANT			IMPURITY		
Preprocessing + followed by	No. comp.	RMSEP	p-value	No. comp.	RMSEP	p-value
None	9	0.62	0.005	5	0.041	0.005
Normalization	8	0.67	0.010	6	0.040	0.005
1st derivative	7	0.54	0.075	4	0.043	0.005
1st derivative + norm.	7	0.50	0.095	5	0.043	0.005
2nd derivative	5	0.67	0.010	2	0.045	0.005
2nd derivative + norm.	5	0.59	0.005	2	0.046	0.005
OS-2	9	0.51	0.075	7	0.045	0.155
1st derivative + OS-2	10	0.41	1.000	6	0.034	1.000
2nd derivative + OS-2	8	0.50	0.035	5	0.037	0.045
MSC	8	0.45	0.005	5	0.043	0.005
2nd derivative + MSC	5	0.59	0.010	2	0.046	0.015
OSC (1 OSC)	8	0.58	0.010	6	0.041	0.070

The best model was obtained with optimized scaling on first-order differentiated data. For MSC preprocessed data the RMSEP value is close to the value obtained for the best model. However, inspecting the p-value reveals that the two methods differ in precision. Since the p-values are small, the residuals obtained for first-order differentiated data followed by optimized scaling are expected to be smaller than for MSC. A histogram showed that MSC had a few very large residuals. This method is therefore less reliable for the data investigated here. According to the p-value, first-order differentiation followed by normalization is the second best alternative. Also for the impurity, optimized scaling on first-order differentiated data gave the model with lowest prediction errors. However, OS-2 and orthogonal signal correction (OSC) cannot be said to be different from the best model. Due to the low prediction error using OS-2 on first order differentiated data, this method is favoured. Note that second order differentiation gave a model of low complexity and moderate error for impurity.

Two different methods of OSC were tested in paper II. The methods of Wold et al. [59] and Fearn [60] gave the same results. It was observed that removal of one OSC component decreases the number of significant components in the PLS model by one; one OSC component approximates one LV. OSC removes systematic information in $\mathbf{X}$ orthogonal to $\mathbf{y}$. The scores, loadings and weights orthogonal to $\mathbf{y}$ in the calibration set it used to remove irrelevant information in the validation set. The problem is that OSC removes systematic patterns, not random noise. Regression methods such as PLS, model also systematic patterns in data. Utilizing OSC seems only to be a detour to the same solution.

The predictive ability using OS-2 on differentiated data is significantly better than no preprocessing of data. For most of the other techniques this is not so. It is therefore appropriate to ask whether preprocessing is necessary. When modelling a new chemical system, different preprocessing techniques should always be tested along with no preprocessing.

A summary of the performance of the different preprocessing methods is given in Table 4.3. In the process data investigated here both multiplicative and additive effects are present. For preprocessing data where these effects are expected to be

Table 4.3: Performance of multiplicative and additive correction methods.

Differentiation	+ Performs well. ÷ Requires testing of numbers of consecutive points to utilise.
Normalization	+ Simple to use, performs well.
OS-2	+ Best results. ÷ The method combines poorly with internal validation since a reference spectrum is needed. Gives a complex model.
MSC	+ Simple to use. ÷ Results in a few large residuals, i.e. unstable.
OSC	Performs equally to no preprocessing. ÷ 1 OSC ≈ 1 LV.

present, differentiation and normalization should be tested along with OS-2 and raw spectra. If there are small differences between the best model and a model of lower complexity, the more parsimonious model should be chosen for future modelling.

4.3 A comparison of techniques for modelling non-linear structure

The aim of paper III was to investigate different techniques for modelling data having a slight non-linear relationship between predictors and the responses. The alkylation process described in section 1.2 was used as a case study. An ANOVA test [78] revealed the process data to have a curvature that is well described by a second degree polynomial (see Figure 4.4).

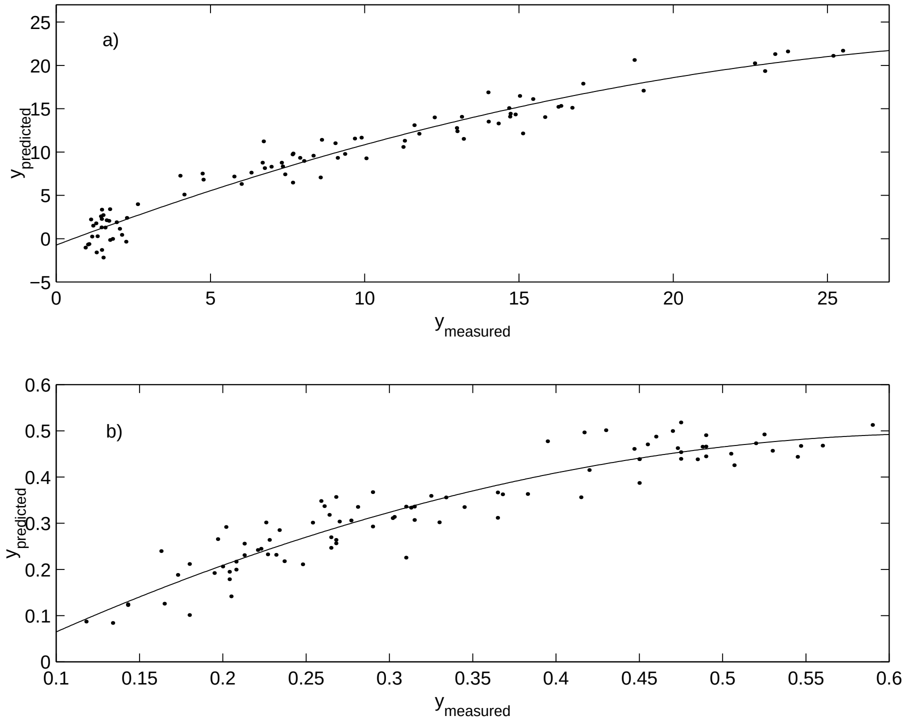

Figure 4.4: Measured vs. predicted response. PLS1 model, two LV's, using the spectral region from 1100-1900 nm; a) main reactant, 1^{st} order differentiated data followed by normalisation, b) impurity, 2^{nd} order differentiated data. Solid line is fitted 2^{nd} degree polynomial.

The reason for the curvature was investigated. Spectra of pure compounds in process solution were acquired along with spectra of process solution before and after the addition of the initiating agent (see Figure 1.2). The alkylating agent was found to have the highest covariance with the responses for main reactant and impurity. The alkylating agent takes part in at least two reactions that differ in the reaction rates. In addition, the synthesis is controlled emprically by adding initiating compound or NaOH. A non-linear relationship between the predictor and the main reactant or impurity can therefore be expected.

Different techniques were tested for managing the curvature. They include (i) transformation of original variables [64, 67] and scores (both stepwise addition [69] and deletion [36]) and incorporation of squared predicted response, and (ii) two different approaches of local modelling; one model per control point and locally weighted regression (LWR) [70], and (iii) quadratic PLS (QPLS) [71, 72, 73].

The data set used in paper III is the same as in paper II. In addition, an extended test was performed of the models predictive abilitities on data recorded one year later on another NIRSystems spectrophotometer. This test set included 34 samples. For preprocessing, some of the non-linear techniques combine poorly with optimized scaling due to the requirement of finding a reference object. Instead, first derivative followed by normalization to length one was chosen for main reactant, while for impurity the data was second order differentiated. A summary of the best results are given in Table 4.4.

Transformation of original variables, i.e. adding squared variables, using principal variables (PV) and the region from 1616 to 1656 nm gave models of better predictive ability compared to original data. Including interaction terms in addition to squared terms may improve the model further. All possible combinations of

Table 4.4: Summary of the best results obtained for the different nonlinear techniques. SEP is the Standard Error of Prediction of a test set acquired one year after. The p-values are calculated to compare the residuals of the method with lowest SEP ($p = 1.00$) with the residuals obtained from the other methods.

	# LV	SEP	p-value
	orig.var.*	test set	rand. t-test
MAIN REACTANT			
Transformed principal variables	12*	0.97	< 0.05
Transformed variables, 1616-1656 nm	3	0.76	0.23
Incorporation of $\hat{y}$, $\hat{y}^2$	6	0.73	0.30
SWP-PLS	9	1.04	< 0.05
LR, 1616-1656 nm	3	0.84	0.11
QPLScv	1	0.67	1.00
Linear PLS	7	1.04	< 0.05
Linear PCR, OS-2	10	0.91	< 0.05
IMPURITY			
Tansformed principal variables	2*	0.057	1.00
Tansformed variables, 1616-1656 nm	2	0.060	< 0.05
Incorporation of $\hat{y}$, $\hat{y}^2$	2	0.059	0.26
SWP-PLS	2	0.068	< 0.05
LR, 1616-1656 nm	2	0.075	< 0.05
QPLScv	1	0.077	< 0.05
Linear PLS	2	0.075	< 0.05
Linear PCR, OS-2	6	0.059	0.36

interaction terms between original variables would result in an enormous matrix. Therefore, data was decomposed into a few LVs. Scores, squared scores and interaction terms of the scores were used for modelling. Stepwise deletion of variables by jack-knifing gave complex models with poor predictive ability. Stepwise addition of variables by polynomial PCR and PLS performed better. SWP-PLS gave better or equally performing models to SWP-PCR. SWP-PLS was therefore chosen as the best way for modelling with transformed scores. Incorporating predicted response and its squared, $[\mathbf{x}|\hat{y}|\hat{y}^2]$, gave models with good predictive ability.

Splitting data according to the three control points into a few approximately linear models resulted in complex models of equal or lower predictive ability compared to linear PLS. In local regression (LR) and locally weighted regression (LWR), the closest samples in Mahalanobis distance of the first important scores are used to model the new sample. LR without weighting of samples provided better results than LWR. When weighting, samples closest to the sample to be predicted are weighted up while the samples furthest away are weighted down. The assumption is based on that the absorbance is proportional to concentration. The score plot of a sample taken at control point two showed that the 20 closest samples could cover the concentration region from the first to the the third control point. Weighting of data having a non-linear structure is therefore expected to introduce more disturbances into data. The results in Table 4.4 show that other techniques gave better models than LR. The reason is expected to originates from the methods possibility of selecting a few samples covering a large concentration region. By so doing, fewer samples are used to model systematic variation in data.

The last technique tested was quadratic PLS (QPLS). The error-based QPLS technique of Baffi et al. [72] gave one LV that captured 100 % variance of the response. The QPLS-algorithm without cross-validation re-estimates the weights until it explains all information in $\mathbf{y}$, and the result is an enormous overfit. Instead

Table 4.5: Performance of non-linear methods.

Transformation	
Orig. variables; 1616-1656 nm	+ Simple to use, improve model.
Orig. variables; PV	+ Simple to use, improve model.
Scores, Jack-knifing	÷ Complex models of poor predictive ability.
Scores, SWP-PLS/PCR	+ Best way to choose variables of transformed scores.
$[x\|\hat{y}\|\hat{y}^2]$	+ Simple to use. Best non-linear method.
Splitting data	
Time dependent splitting	÷ Complex models of equal predictive ability to linear models.
LR/LWR	LWR; ÷ Weighting may introduce more disturbance into data. LR is therefore favoured over LWR. ÷ Risk of finding few samples covering a large concentration region.
Quadratic PLS	
QPLS	÷ Danger in overfitting.
QPLScv	÷ Unstable? Perform best for main reactant and worst for impurity. More testing is required.

QPLS with cross-validation (QPLScv) [73] was used. Comparing the results from QPLScv with PLS shows that while the linear PLS explains close to 100 % variance in both $\mathbf{X}$ and $\mathbf{y}$, QPLScv explains less than 40 % variance of $\mathbf{X}$, but close to 100 % variance of $\mathbf{y}$. For QPLScv only one LV is needed in the model. One of the reason for incorporating the curvature of the predictor having a non-linear behaviour was the demand for a more parsimonious model. This requirement is fulfilled using QPLScv. Inspecting the score $\mathbf{u}_1$ vs. $\mathbf{t}_1$ showed a close to perfect fit for QPLScv, while the deviations were large for linear PLS. The disadvantage is the danger of overfitting, and it was doubted whether this would give problems when modelling future samples. The predictive ability of the model showed that the QPLScv was robust for the main reactant. However, for the impurity QPLScv gave the worst result of all the methods tested. The differences in predictive ability between the two compounds my indicate that the method is unstable. Further testing is necessary.

A summary of the performance of different non-linear methods is given in Table 4.5. Incorporation of a squared predicted response is the best way to incorporate curvature into the model. It gives low prediction errors and it is also a simple mathematically transformations so that it can be employed in commercial software without additional implementations of algorithms.

Chapter 5

Assessment of modelling techniques

The variables in near-infrared data are highly collinear. This necessitates use of multivariate techniques. Different ways to model data for quantitative purpose exist. In this chapter, curve resolution has been applied along with two-way and three-way regression techniques for modelling NIR spectra.

In a dynamic system species are transformed into one or several other species. Acquiring spectra at different stages of the process provide a matrix of concentration composite in one direction and spectra of the different species in the other direction. The idea of performing curve resolution on a matrix of an evolving system is to obtain the pure spectral profiles of each species and their corresponding concentration profiles without any a priori information. A procedure to resolve near-infrared data from a chemical system is given below.

The system used as a case study for curve resolution is simple. Even then the curve resolution techniques showed sensitivity problems due to broad overlapping bands and presence of baseline. If the purpose is process monitoring and a reference method is available, methods as PLS is expected to be easier to carry out. A feasibility study of NIR spectroscopy for modelling process variables has been performed. The goal is to find robust calibration models of process variables that are used to control a chemical process empirically. Another important task is to obtain an estimate of the error in a model based on NIR spectroscopy compared to the reference technique. This is to obtain more information of the model and to check if the introduction of a new and faster technique can be supported.

NIR data acquired from a batch process can be arranged in a three-way structure (see Figure 2.2). To decompose data of three- or higher order arrays N-way techniques are necessary. In this thesis N-way decomposition techniques are tested for their ability to decompose data and to model NIR data for quantitative purposes. A problem using three-way regression for process variables is that all data from a batch must be available. Different ways to fill in for missing values have been tested.

Which modelling technique to choose depends on the chemical system. The requirements for using these techniques on NIR data are summed at the end of this chapter.

5.1 Curve resolution for quantitative measurements

Paper I gives a procedure for resolving NIR spectra of alcohols. The aim is twofold; (i) to perform a qualitative and quantitative study of the self-association phenomena, and (ii) to test the performance of curve-resolution of NIR spectra and thereby emphasize the benefits and obstacles on use of curve resolution for the purpose of reaction monitoring.

5.1.1 Background of the system

The self-association of alcohols has been investigated for several decades by techniques such as dielectric [79, 80], mid-infrared [81, 82, 83] and near-infrared spectroscopy [75, 84, 85]. The studies suggest that the alcohols exist in a ”free” unbound form at very low concentrations, and that they form larger aggregates of different sizes and shapes at higher concentrations. Controversy exists about the sizes and shapes of the aggregates. Curve resolution studies of alcohols solutions in the mid-infrared region [82, 83] suggested a three-component system of monomers, linear and cyclic aggregates.

Most of the near-infrared studies of self-association of alcohols have used 2D-correlation techniques for band assignments. However, these techniques provide no quantitative information. A quantitative NIR study was performed by Iwahashi et al. [75]. They suggested use of one frequency under the monomer peak to quantify the association degree of alcohols. A disadvantage is that the method is only able to quantify the monomer and polymer species. The literature study above states that more than two generic structures of alcohol species exist. Curve resolution techniques may find the concentration and spectral profiles of all different species in the systems.

5.1.2 Rank analysis

The curve resolution was carried out on baseline-corrected data (see section 4.1.1). A thorough rank analysis on each of the alcohol systems was performed prior to curve resolution. This gave important information for resolving the spectral data. In paper I, the results from 1-heptanol is used as an example. Inspecting the eigenvalues gave no unique rank of the system. Inspecting scores from PCA in concentration direction indicated a two- or three-component system, while EFA indicated a three-component system. Correlated noise was removed prior to rank analysis by differentiation in spectral direction. Inspecting the loading vectors revealed structure in the first three components. The performance of smooth PCA (see section 2.2.1) in the concentration and in the spectral direction is shown in Figure 5.1. A jump is observed from the third to the fourth eigenvalue in both directions. Based on these results the rank was set to three.

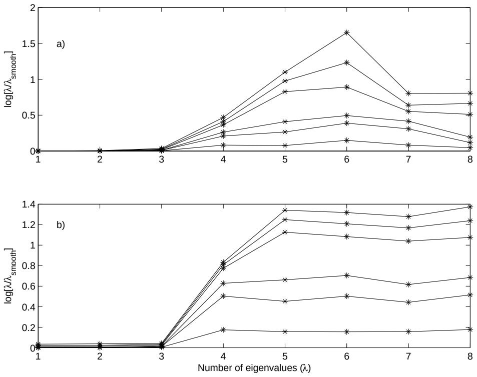

Figure 5.1: Smooth PCA in a) the concentration and in b) spectral direction of differentiated data of 1-heptanol from 0.01 to 0.92 M. ($*$) is the degree of smoothing.

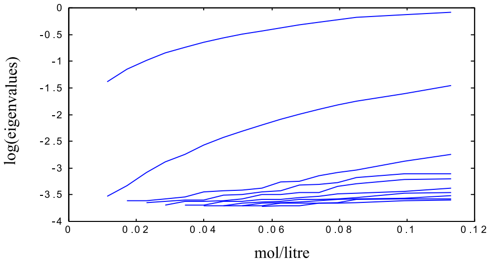

Figure 5.2: Two-component region revealed by EFA of 1-heptanol solutions from 0.006 to 0.113 M.

5.1.3 The resolution of spectra

EFA revealed a two-component region at low concentrations (see Figure 5.2). Six selective regions were found in the ETA plot in spectral direction of this region. Cross-correlating the scores from the selective regions gave two groups. These are estimates of the concentration profiles of the two species at low concentration. The spectral profiles of component one and two were calculated by least squares. The resolved profiles are shown in Figure 5.3.

The monomers are expected to dominate at low concentration. In addition, the free OH-stretch has the strongest force constant, therefore the monomer band appears at the highest frequency. The profiles having a dotted line is therefore the monomer species. The spectral profile of component two is asymmetric. This is expected to be due to that at higher frequencies many small, linear species exist, whereas a lower intensity at lower frequency is caused by the presence of fewer but longer linear aggregates [82]. Component two is therefore assigned to the linear species, i.e. the profiles having a dotted line.

The curve resolution was expanded to the concentration range from 0.01 to 1.00 M, i.e. the three-component region. The third component is expected to be from cyclic aggregates. The OH-stretch of cyclic aggregates absorbs at a lower frequency than the linear aggregates. This information was used in combination with ITTFA to resolve the third component. Several initial estimates were chosen at a frequency slightly lower than the band maximum of the second component. For every start estimate, a least-squares procedure was performed 15 times, setting the regions of negative intensity to zero after each iteration. The different start estimates gave resolved spectra of the same shape and of almost same maximum. The shape of the third resolved profile was checked by means of OP. Having the spectral profiles of all three components, the concentration profiles of the large concentration region were estimated by least squares, i.e. $\mathbf{C} = \mathbf{XS}(\mathbf{S}^T\mathbf{S})^{-1}$. The profiles are given in Figure 5.3.

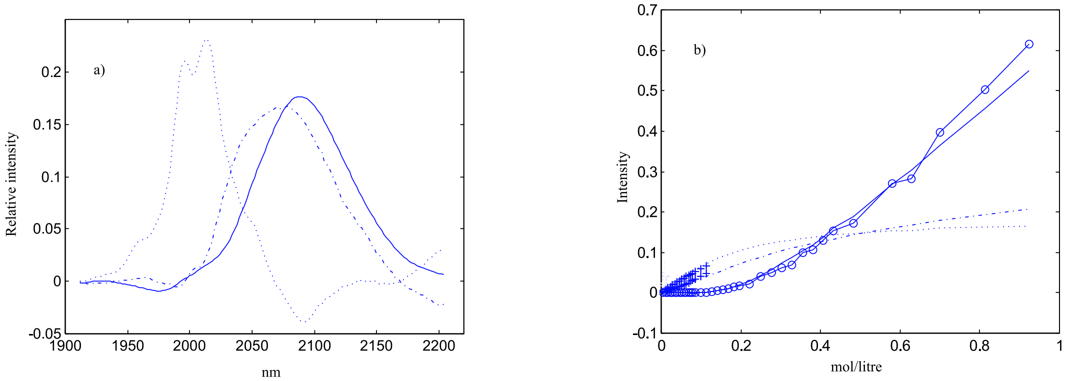

Figure 5.3: Resolved a) spectral and b) concentration profiles of 1-heptanol. First $(\cdots)$, second $(- \cdot -)$ and third component $(—)$. $(++)$ are the concentration profiles estimated from the two-component region, $(- \circ -)$ is the third component checked by OP.

5.1.4 Quantitative study based on the resolved profiles

In paper I it has been shown that it is possible to resolve spectra in the near infrared region. The resolved profiles obtained were used to increase the understanding of the self-association of alcohols. The spectral profiles confirm earlier results that free alcohol species are present in majority at low concentrations and that they tend to associate into linear and thereafter cyclic or higher associated species with increasing concentration.

In order to perform quantitative studies, i.e. to find estimates of association numbers and equilibrium constants, the absolute concentration profiles have to be calculated. When the total concentration is known, absolute concentration profiles can be calculated using the assumption that the sum of the concentrations of different species is equal to total concentration;

$$\mathbf{c}_{total} = \sum_{a=1}^{A} \mathbf{c}_a b_a = \mathbf{C}\mathbf{b} \tag{5.1}$$

$\mathbf{b}$ is the scaling coefficient of the resolved concentration profiles. The association number was determined by plotting $\mathbf{c}_{associated}^{1/n}$ versus $\mathbf{c}_{free}$ and the association number, n, is found when the fitted straight line passes through the origin [82, 83]. The association numbers and the equilibrium constants are given in Table 5.1.

The linear aggregates have smaller aggregates sizes and equilibrium constants than the cyclic aggregates. In addition it is observed that the branched alcohols have smaller equilibrium constants than the linear alcohols. This is expected since branched alcohols have greater difficulty in aggregating compared to linear alcohols.

The results in Table 5.1 differentiate well between linear and sterical hindered alcohols. However, the method was not sensitive enough to differentiate within these two groups. For example, the linear alcohols show a relatively large range of association numbers that cannot be explained by structural differences. The method used to determine the aggregate size is expected to partly contribute to the differences observed, since a strong extrapolation makes one data point influence a

Table 5.1: Association numbers and equilibrium constants calculated for the given concentration regions.

	linear aggregates		cyclic aggregates		
	size	equilibrium	size	equilibrium	concentration
alcohol	(n)	constant (K)	(n)	constant (K)	(M)
1-propanol	3	109.9	8	8.1×10^{6}	0.30-0.70
1-butanol	3	13.5	6	7.7×10^{3}	0.40-1.03
1-pentanol	3	9.6	10	2.5×10^{8}	0.20-1.10
1-hexanol	2	13.4	11	3.4×10^{10}	0.23-0.65
1-heptanol	3	44.9	8	9.0×10^{5}	0.41-0.92
2-propanol	2	0.9	6	2.6×10^{3}	0.38-1.04
2-methyl-1-propanol	3	1.7	6	2.1×10^{3}	0.42-1.17
2-methyl-2-propanol	2	1.1	5	211	0.40-1.11

lot on the physical constants given in Table 5.1. The main reason for the variation in the constants is however expected to be caused by the rough baseline correction and the absence of pure one component regions.

5.1.5 A summary on resolving NIR spectra

A way to perform curve resolution of highly overlapped NIR-spectra has been provided in this thesis. The performance of the different methods is given in Table 5.2.

Table 5.2: A summation of methods used to resolve NIR-spectra.

Rank analysis	
Eigenvalues	÷ Problems of detecting rank.
Score/loadings	Manual inspection of noisy pattern. Can be difficult to detect rank since some noise may be present even in the significant scores/loadings.
smooth PCA	+ Visual interpretation.
EFA	+ A map over evolving species. ÷ Not as clear in rank determination as smooth PCA.
Curve resolution	
ETA	+ reveal selective regions in spectral direction.
ITTFA	+ No selective region is required.
OP	+ Indirect rank map since only noise is present after projection if a too large region embracing all species are used. Indicate also the form of the profile. ÷ profile is to uncertain and is only used to check profile from ITTFA.

Several applications of curve resolution in the mid-IR region have been given [86, 87, 88, 89]. The applications have two things in common. Prior to curve resolution a baseline correction is performed. In addition the system is of low rank, or is divided into several low-rank regions (two or three). The important issues when performing curve resolution of spectral data can be summarized as; (i) baseline correction, and (ii) rank mapping, that is finding regions of low rank. The greatest difference between spectra acquired in the mid-IR and near-IR region is that bands in the near-IR region are broader, making it difficult to find regions of low rank. A possibility could be to use a subtraction method like general background subtraction [74]. The problem of finding low rank regions and baseline removal can therefore be reduced to removing the spectrum of a compound/solution that is very similar to the system investigated, except for the functional group of interest. Furthermore, the problem can be reduced to finding the proper component/solution to use for each chemical system to be resolved.

5.2 Performance of two-way regression methods

Paper II includes a feasibility study of NIR for monitoring a pharmaceutical process. The process solution is chemically complex, and the study focuses on how to improve the predictive ability of regression models. The paper includes testing of different preprocessing techniques. These results were given in section 4.2. The purpose of the current chapter is to discuss important issues related to regression and includes; (i) how to select significant components, (ii) an investigation of how models of small, informative wavelength regions perform compared to models using all wavelengths available, and (iii) an inspection of the difference between prediction errors in the models and the deviations in HPLC-values.

5.2.1 Number of significant components

The optimal number of components to use in a model can be found using cross validation or explained variance of responses in a validation set. Leave-one-out and Monte Carlo cross validation were examined, along with explained variance of validation set. In addition, manual inspection versus statistical tests were investigated. The results from the determination of significant components in models using different preprocessing techniques are given in Table 5.3.

Leave-one-out and Monte Carlo cross validation give similar numbers of significant components. The correspondence between cross validation and the use of explained variance of responses in a validation set was also high. For the data inspected here, choosing between cross validation and validation set is therefore a matter of taste. However, some methods may have problems with cross validation. Optimized scaling 2 (OS-2) is difficult to combine with cross validation owing to its calculation intensity, and the difficulty of finding an internal automatic criterion for determining the object to use as reference.

The results from the F-tests differ in several cases from manual inspection of cross-validation plots. Better control of the number of components is obtained with manual inspection. Manual inspection is therefore favoured. Explained variance of responses and random t-test use the validation set to determine the number

Table 5.3: Determination of significant components for the impurity.

	Cross-validation		Test set		Number of
	Leave-one-out	Monte Carlo	Explained	Random	significant
Preprocessing	(F-test)	(F-test)	variance	t-test	components
None	6 (5)	6 (3)	5	10	5
Normalization	6 (6)	6 (3)	7	6	6
1st derivative	3 (2)	2 (2)	4 or 10	10	4
1st derivative + norm.	3 (2)	3 (2)	5	10	5
2nd derivative	2 (2)	2 (2)	2	2	2
2nd derivative + norm.	2 (1)	2 (1)	2	2	2
OS-2	-	-	7	7	7
1st derivative + OS-2	-	-	6	6	6
2nd derivative + OS-2	-	-	5	5	5
MSC	5 (5)	5 or 6 (2)	5	9	5
2nd derivative + MSC	2 (1)	2 (1)	2	2	2
OSC (1 OSC)	6 (6)	6 (3)	6	9	6

of significant components. The t-test gives too many significant components for several cases, and is therefore not reliable.

5.2.2 Variable selection

The predictive ability of models based on a few principal variables (PVs) is compared to models using larger informative regions, and the large wavelength regions from 1100 to 1900 nm. For variable selection, interval PLS (iPLS) and PVs were used. Matlab files for iPLS are given at the KVL homepage [90]. Interval PLS [37] splits the data into several equidistant regions. The RMSECV is calculated for every region. The region with the lowest RMSECV is optimized by expanding and contracting it, symmetrically and asymmetrically. This is time-consuming. Often, the combination of several regions may lead to better results than what one achieves with any single region. This complicates the search for the optimal region(s) even more. A solution to this problem, called synergy iPLS, has also been given at the KVL homepage. The method examines all possible combinations of regions. Because of the increasing amount of calculations that has to be performed, synergy iPLS is limited to combinations of two, three and four regions. Since synergy iPLS makes it difficult to establish narrow variable regions, small modifications of iPLS were performed. The variables were split into many small intervals, and the intervals were ranked according to increasing RMSECV. The region of lowest RMSECV was expanded with the region of second lowest RMSEV. PLS was performed on the increased matrix, and the prediction error of the expanded region was compared to the prediction error of the region of lowest RMSECV. If the prediction error has decreased when expanding the region, further expanding with the region of third lowest RMSECV is performed and a new PLS model made, etc. The optimal variable region is obtained when no or only insignificant improvement in the model is observed. Interval PLS and optimized scaling, here referred to as iOS-2,

Table 5.4: Results variable selection.

Method	Preprocessing	λ (nm)	No. of var.	No. of LVs	RMSEP
MAIN REACTANT					
PLS	None	1100-1900	401	9	0.62
iPLS 40 int.	None	1124-1170 1628-1674	48	5	0.66
OS2	1st der. + scaling	1100-1900	401	10	0.41
iOS2 40 int.	1st der. + scaling	1148-1170 1604-1722	84	9	0.48
PV	2nd der. + norm.	1644, 1878, 1424, 1734, 1888, 1704	6	MLR	0.70
IMPURITY					
PLS	None	1100-1900	401	5	0.041
iPLS 40 int.	None	1628-1650	12	3	0.044
OS2	1st der. + scaling	1100-1900	401	6	0.034
iOS2 40 int.	1st der. + scaling	1604-1674 1700-1722	48	5	0.035
PV	2nd derivative	1644	1	MLR	0.047

was carried out using a validation set. A new reference object was found for each variable region.

The results from variable selection for the model of the main reactant and the impurity are given in Table 5.4. For iPLS only the results from no preprocessing and optimized scaling on first order differentiated data are given. For principal variables the best preprocessing method was second-order differentiation. Interval PLS found two interesting regions, 1100 to 1200 nm and 1600 to 1720 nm, for both compounds. These regions correspond to the second and first overtone of CH-stretch. Six variables were detected as PVs for the main reactant, while only one was detected for the impurity. Use of one or few variables detected as PVs gave the highest prediction error. Utilizing the larger region detected by interval PLS improved the model. However, the best models were obtained using the large region from 1100 to 1900 nm.

Prior to the data analysis, a manual inspection of the spectra revealed a noisy region from 1900 to 2500 nm. This region was removed. When testing different variable selection techniques of the remaining region, it was found that the large region from 1100 to 1900 nm gave a more stable model than using a narrower region. It seems that as long as noisy regions are removed before modelling, further variable selection is not required to improve models.

5.2.3 Comparison of predictive ability of models with deviation in reference values

The models with lowest prediction errors were obtained using optimized scaling on first order differentiated data in the region from 1100 to 1900 nm. These models gave an RMSEP-value of 0.41 and 0.034 area percent (HPLC-values) for the main reactant and impurity, respectively. Note that the RMSEP in a relative sense is higher for the impurity, relative deviation (RD) ≈ 9, than for the main reactant, RD ≈ 5. RD is equal to; (deviation/mean$_{HPLC}$)×100.

The contribution to the prediction errors in a partial least squares originates from the uncertainty in estimated model parameters, the unmodelled part of the response and from the uncertainty in the measurement in the predictor vector for the unknown object [91]. Here an inspection of the uncertainty in the response measurements has been carried out. Three different HPLC-instruments were used to determine the area percent of the main reactant and the impurity in the process. The average deviation in replicate runs on HPLC-instrumentation is given in Table 5.5.

Table 5.5: Deviation of highest and lowest value of replicate runs on different HPLC-instruments.

	Control point 1		Control point 2		Control point 3	
Compound	Deviation	RD	Deviation	RD	Deviation	RD
Main reactant	0.22	1.55	0.09	1.17	0.06	4.07
Impurity	0.03	12.54	0.03	8.37	0.03	5.67

There is a decrease in deviation from control point one to control point three. The amount of main reactant in the process solution decreases from about 10-15 area percent to 1-2 area percent at the last control point. The decrease is therefore expected, since deviation between replicates is related to concentration. For impurity the average deviation is the same at all control points since the concentration is in the same order of magnitude throughout the process.

For the impurity the highest deviation is 0.03 area percent. The RMSEP value is close to this value, and a reliable model for surveillance of the impurity in the process using NIR is feasible. For the main reactant the highest deviation is equal to 0.22 area percent, while the lowest RMSEP achieved is equal to 0.41 area percent. These deviations cannot be attributed only to the HPLC-measurements. Some noise will always be present in spectra due to spectrometer hardware, light scatter in solution or interaction of compounds, e.g. hydrogen bonding [92]. However, the impurity deviation agrees well with the HPLC deviations, and the reason for the higher deviations between modelled and HPLC values must be attributed to NIR data having greater problems of modelling this compound. In order to improve the model the cause for the high prediction error for main reactant should be investigated.

5.2.4 Origin of the band used to model the responses

The difference between the predicted and measured responses were too large to be attributed to the variation in the HPLC measurements alone. An investigation of observed bands in the near infrared region was therefore performed.

Bands from O-H, N-H and C-H stretches are strongly absorbing in the near infrared region. These are all present in the reactants involved. The bands from the solvent, $CH_3OCH_2CH_2OH$, absorb strongly in the entire near-infrared region. The result is that the solvent seems to mask interesting bands. However, upon inspection of differentiated spectra a narrow region in the region around 1644 nm was observed to relate well to the responses. This band was shown in paper III to originate from the alkylating agent. This observation explains the errors between predicted values and the HPLC-measurements. The impurity has low concentration and the expectation for successful modelling of this compound was low. However, the results in paper II showed that this compound is well modelled by NIR spectroscopy. This indicates strongly that the content of alkylating agent is highly correlated to the content of impurity. For the main reactant a greater difficulty in relating the NIR-spectra and the HPLC-responses is obtained, requiring more LVs.

Why do we not observe changes in the concentration of other bands? The initiating compound is expected to react almost completely when caustic soda is added. When the main reactant dissolves in the alkaline environment, the hydrogen atom of the N-H stretch reacts with OH^-, forming water. The strongly absorbing N-H bond is converted to N^-, and absorptions due to the N-H bond is therefore not expected. The R-group of the main reactant and of the product is expected to give close to the same absorptions, and the changes in the spectra are too small to be observed. For the impurity, a C-H stretch frequency different from the main reactant and the product is expected. Due to small concentration changes in the process solution, the absorption differences cannot be observed. Because of water formation, band shifting is expected. In addition, the strongly absorbing solvent masks severely the bands from other absorbing species in the spectra.

5.2.5 Summary of feasibility study

A summary of the different methods used in the feasibility study is given in Table 5.2.5. There has been a disagreement between chemometricians of whether to use internal cross validation (CV) or an external validation set when choosing the optimal number of components. For the data investigated here they perform equally, and the choice between internal or external validation is more a matter of taste. For variable selection, PV and iPLS gave information of spectral region. However, for linear modelling use of the large region from 1100 to 1900 nm gave models of lowest prediction errors.

In a feasibility study, an estimate of the errors from response-measurements should be found. A large difference between errors originating from the response-measurements and the obtained RMSEP value indicates that NIR spectra obey some problems of modelling this compound. To be able to improve the model, more research is needed to find the cause.

Table 5.6: A summary of methods used in the feasibility study.

Determination of significant components	
Leave-one-out CV	Performs well.
Monte Carlo CV	Performs equally to leave-one-out CV.
F-test on results from CV	+ Automatic selection. ÷ Deviates somewhat from manual selection.
Explained variance of test set	Performs equally to CV. + Able to test model on an independent data set. ÷ Requires more samples to be available than CV.
Random t-test, test set	÷ Risk for overfitting.
Variable selection	
PV	+ Ranking of important variables. ÷ Similar variables that could stabilise model, will not be detected.
iPLS	+ Visual method for detecting the most informative regions.
Large region, 1100-1900 nm	+ Use of the large region gives models of lowest prediction errors.

5.3 Performance of three-way regression methods

Little research has been carried out combining multiway techniques with NIR spectroscopy. For process analysis there are two main problems. While PARAFAC has been shown to be successful to resolve profiles from fluorescence data, this is difficult for NIR spectra due to broad overlapping bands and baseline variations [93, 94]. Another difficulty is that when predicting responses from a new spectrum, data from the entire duration of the batch must be available . For on-line analysis quick and reliable methods dealing with missing data are therefore needed.

Despite these difficulties, decomposing NIR data by N-way techniques has been shown to be suitable, both for classification purposes [95], and to increase understanding of chemical reactions [93, 94]. No applications of NIR spectra in combination with N-way regression techniques are found in the literature. However, three-way PLS have been tested on fluorescence and UV/Visible-spectroscopy with better results than two-way PLS [46, 96].

The use of N-way techniques on the process data investigated in this thesis is especially appealing since a curvature was observed between predictor and responses, and that scatter effects are observed in the NIR spectra. In paper IV it was tested if these problems could be diminished by stacking samples into higher order arrays. The objectives included testing of N-way decomposition techniques for exploration and regression of NIR-data, and to find a suitable technique to fill in for missing values.

5.3.1 Inspection of data

This study is based on manually collected NIR spectra. The data set comprises 17 batches where a spectrum has been collected at each control point. To investigate the data, the wavelength region from 1100 to 1900 nm was utilized. The data was arranged in a three-way array of size 17 batches × 3 time points × 401 wavelengths. 9 batches were included in a calibration set, while the remaining 8 batches were included in a validation set. An additional 16 batches were used to test the performance of the regression models. The data were preprocessed by second-order differentiation. Data analysis was carried out using MATLAB (The MathWorks, Inc., version 6.5) and the N-way Toolbox [97] for MATLAB, version 2.10.

N-PLS and two-way PLS were tested for interesting wavelength regions. The results showed the region from 1638 to 1650 nm to be important for modelling the main reactant (see Table 5.7). Poor models were obtained for N-PLS making use of the large wavelength region from 1100 to 1900 nm. This is different from two-way PLS. Inspecting the weights from N-PLS revealed that they were small in the important region around 1644 nm, while for two-way PLS the regression coefficients were large. Since N-PLS blocks data, centering of data removes distinct concentration differences between control points. When calculating the covariance between the spectra and the response, N-PLS is more vulnerable to wavelengths regions changing in intensity when the process evolves, and the response is erroneously registered as beeing dependent on these variables. The result is poor models that fail to predict new samples. This is not obtained for PLS1 where all control points are centered simultaneously, giving three distinct bands. Variable selection is therefore more important for N-PLS than for two-way PLS.

Table 5.7: Predictive ability of main reactant for different wavelength regions using N-PLS and PLS1, full cross-validation.

Wavelength	**N-PLS**		**PLS1**	
(nm)	No. LVs	RMSEP	No. LVs	RMSEP
1100-1900 nm	9	1.63	6	0.72
1622-1660 nm	3	0.71	3	0.76
1638-1650 nm	3	0.59	1	0.78

The region from 1638 to 1650 nm was chosen for a further analysis of data. Prior to multiway analysis, a test is required to see if there is enough variation to perform multiway analysis. For the process data, the correlations between the first score of the NIR spectra and the belonging responses were equal to 0.9, 0.8 and 0.6 for control point one, two and three, respectively. It is therefore appropriate to run multiway decomposition of the process data.

PARAFAC and Tucker3 were used to explore the process data. Dimensionality tests for PARAFAC, i.e. sum of squares (SSQ) and CORCONDIA [98], gave two significant factors. A split-half test [40] using two components gave different profiles. The rank of the system is therefore different from two. While PARAFAC is restricted to decompose data into same number of factors in all modes, Tucker3 does not have this limitation. A dimensionality test, plotting SSQ vs. number of factors, gave [2 2 1] factors to be significant. In paper III it was revealed that the band used to model the main reactant and the impurity most likely stems from the alkylating agent. This explains a rank equal to one in the spectral direction. In addition, the profiles obtained in the concentration direction resembled the concentration profiles of the main reactant and the impurity. One may ask whether the two significant factors in the time direction correspond to the main reactant and impurity. To check this, simulated data resembling the real data set were made. Three data sets were made of different concentration profiles multiplied with the alkylating band. The results of the dimensionality tests are given in Table 5.8. They show that only one component contributes to the signal. Since the system investigated is complex, the rank is not expected to be [2 2 1] as obtained for Tucker3. The profiles obtained from PARAFAC or other N-way decomposition techniques are therefore not expected to be pure. N-way decomposition techniques provides instead information of number of significant components needed to explain most of the variance in the system. This information can be used for regression purposes.

Table 5.8: Dimensionality test of simulated data set.

Response multiplied with "alkylating band"	**Tucker3**
Main reactant + impurity	[2 2 1]
Main reactant	[2 2 1]
Impurity	[0 0 0]

5.3.2 Multiway regression

The second issue addressed in paper IV was to test the performance of N-way calibration techniques for the modelling of the main reactant. The results from N-way techniques are given along with two-way regression techniques in Table 5.9. PARAFAC, Tucker3 and N-PLS perform equally well. In addition, N-way regression techniques were shown to model process data better than two-way techniques. This applies especially to control point three where two-way regression fails. The predictions of new samples using N-PLS and PLS1 are shown in Figure 5.4. The residuals locate close to the perfect fit line for N-PLS, while for two-way PLS the

Table 5.9: Results from off-line monitoring. RMSEP of validation set using two-way and three-way regression methods.

Control point	PARAFAC	Tucker3	N-PLS	PLS1	PLS1 per control point
1	0.32	0.33	0.33	0.90	0.74
2	0.47	0.47	0.47	0.77	0.78
3	0.23	0.24	0.23	1.15	1.19

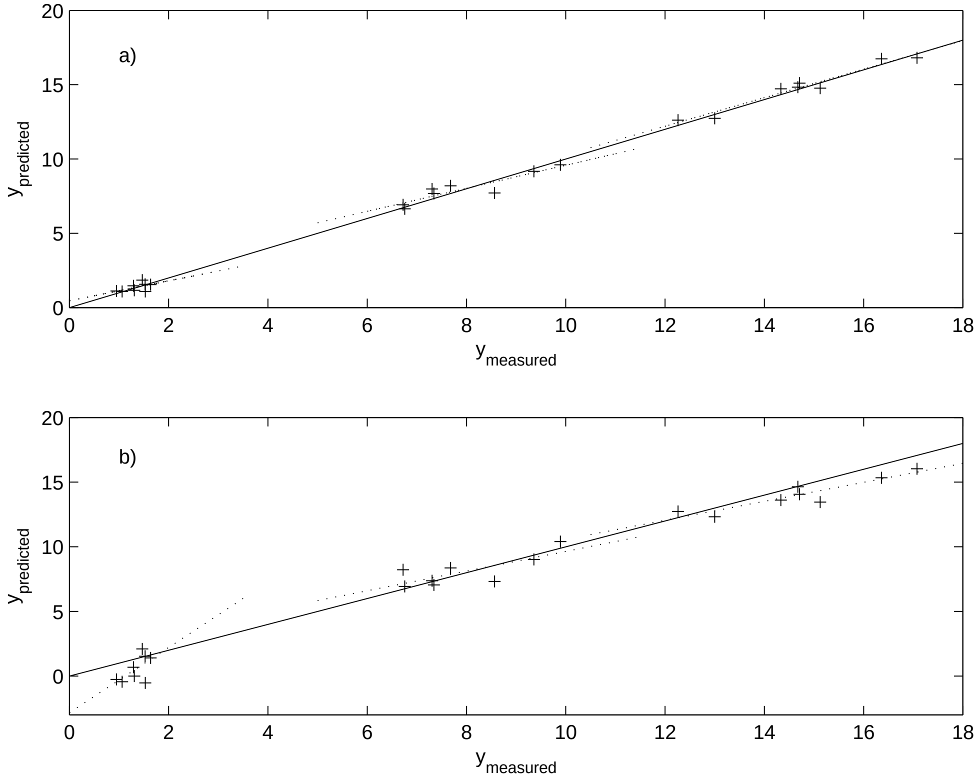

Figure 5.4: Predicted vs. measured for validation samples calculated using a) N-PLS, and b) two-way PLS. Dashed line is 1^{st} order polynomial fit of residuals at each control point.

residuals are large and obey a systematic pattern. For control point one and three most of the objects plot beneath the perfect line of fit. In addition it is observed that the slope of fit line at each control point differ more from perfect fit for the two-way PLS results. The behaviour in Figure 5.4 b) is due to the non-linear behaviour in data discussed in paper III.

When decomposing three-way data the data array is projected onto the loadings in the time and spectral directions. The result is scores that explain the behaviour of the batch. The score from one batch is used to model the response at all control points for the given batch. For N-PLS the regression coefficients are used to weight scores for a batch. The same regression coefficients are used for all control points;

$$\hat{\mathbf{Y}} = \mathbf{TBQ}^T \tag{5.2}$$

The differences in concentration are given by the loadings for the responses, $\mathbf{Q}$. By organizing data in a box, i.e. stacking data per control point, the size of the signal variation at each control point is removed due to centering of data. The result is that the signals at the different control points do not need to be linearly related, however, they must be linearly related within the control point. This is opposite to two-way PLS where the signals need to be linearly correlated at all control points. The results is that N-way regression may be a good alternative for other calibration techniques handling data with a slight nonlinear behaviour such as for the process data here. Note: Division of data into too narrow concentration regions may result in false positive results, since the response weight, **Q**, forces the score to have a value within or close to the concentration range. Therefore, it is very important that the data are checked for correlation before modelling, e.g. that one performs a correlation test as described in section 5.3.1.

5.3.3 N-way calibration for on-line monitoring: How to handle missing values

For on-line prediction of new batches, data are missing from the control points not yet monitored. Filling in with random variables gave terrible predictions (see Table 5.10). The N-way models predictive ability do therefore rely upon the values chosen. Different ways to fill in for the missing values were tested. These included mean spectra, scores and weighted scores. The results are given in Table 5.10.

Table 5.10: Results on-line monitoring. Data for the entire duration of the batch process is missing in control point I and II. The predictive ability (RMSEP) is given for different ways to in-fill yet unknown data.

	Validation set		**Test set**	
Control point	**1**	**2**	**1**	**2**
PARAFAC				
Random	8.64	5.23	8.63	5.27
Mean	0.77	0.59	0.93	0.73
Scores	0.74	0.58	0.88	0.72
Weighted	-	0.59	-	0.72
Tucker				
Random	9.05	4.60	10.18	5.49
Mean	0.77	0.59	0.92	0.73
Scores	0.74	0.61	0.93	0.87
Weighted	-	0.58	-	0.76
N-PLS				
Random	9.56	5.87	9.86	5.97
Mean	0.78	0.60	0.93	0.73
Scores	0.73	0.59	0.92	0.73
Weighted	-	0.59	-	0.73

Small differences are observed between the different techniques. However, a few things should be noted. Use of scores performed best for all N-way techniques in control point one for the batches in the validation set. This indicates that scores might be a good way to fill in for missing values. Weighted scores in control point one gives results equal to utilizing the mean spectrum of calibration samples, while in control point two a half weight is given to the mean spectrum and another half to spectra estimated from scores. Therefore only control point two is inspected. The results are close to the results obtained using the mean spectrum of calibration samples. In the future, automatical acquisition of spectra by fiber optic cables makes it possible to scan spectra more frequently. More reliable estimates of scores (and thereby missing spectra) are then expected. More tests employing scores and weighted scores should be performed in the future.

The prediction errors of the samples acquired on a later moment have in average increased with 0.15 area percent. Updating of calibration models with new samples is therefore recommended.

5.4 Summary of multivariate modelling

In this chapter, curve resolution, two-way and three-way regression techniques have been tested on NIR data for the purpose of quantitative analysis. Based on the results, a scheme giving the requirements of the different techniques is proposed in Table 5.4.

Table 5.11: Requirements of multivariate techniques for process monitoring.

Curve resolution	+ No response needs to be available. ÷ Baseline correction is required. ÷ The system must be simple, rank 2-3 is desirable.
Two-way regression	+ Simple to use. Works well even for highly scattering solutions.
Three-way regression	+ Non-linear data. ÷ Methods to fill in for missing values are required for on-line monitoring.

No reference method is required for curve resolution techniques. This makes it possible to bring out studies of chemical systems as in paper I. The disadvantages is that a proper way to perform baseline correction must be found, and that the system must be simple.

The synthesis of contrast agent used as a case study in papers II-IV is a complex system. The rank is expected to be high, since several of the compounds in the process solution have functional groups that are highly absorbing in the near infrared region. Since a technique to measure the required responses is available, regression techniques relating NIR-spectra to those responses are preferred.

Two-way regression techniques were used in paper II. The methods are simple to use and work well even for highly scattering solutions. In paper IV process data was stacked in a box of $batch \times control\ point \times wavelength$ and N-way regression was carried out. The N-way decomposition techniques model samples in each control point separately, making it suitable for non-linear data. A problem using the N-way techniques is that a way to fill in for values not yet monitored is needed.

Chapter 6

Conclusions and future outlook

In this thesis various strategies for handling noise and irregularities in NIR-spectroscopy data have been proposed and tested. In addition, different ways to model data for quantitative purposes have been assessed.

A procedure for resolving NIR spectra to investigate the self-association phenomenon of alcohols was presented in paper I. The resolved spectral profiles confirm earlier results. Free species dominate at low concentrations. They associate into linear and cyclic species with increasing concentration. Physical constants were calculated from the resolved concentration profiles. These discriminated well between linear and sterical hindered alcohols. However, the method was not sensitive enough to discriminate between the linear alcohols of different molecular size and of different sterially hindered alcohols tested. This is expected to be due to the problems of finding low-rank regions and a sensitive baseline correction method. For future research, subtraction methods providing regions of low rank (selective regions) should be investigated. For nonhazardous and simple chemical processes other techniques, such as mid-IR region combined with an ATR cell, are expected to provide narrower bands. This makes it easier to perform curve resolution for quantitative purposes.

The advantage of curve resolution techniques is the possibility of performing quantitative and qualitative studies in the absence of reference methods. When the purpose is process monitoring of a chemically complex system of high rank and a technique to measure interesting responses is available, regression methods such as PLS are easier and more reliable to use.

Paper II is a feasibility study of NIR spectroscopy for surveillance of the production of contrast agent. Different ways to determine model complexity and different variable selection methods were tested. Automatical determination of model complexity by statistical tests such as t-tests and F-tests are unstable. Instead, manual inspection of cross-validation plots or explained-variance plots of responses in an external validation set, is recommended. Variable selection showed that use of a large wavelength region provided the best models. Errors in the response measurements were compared to errors in the NIR models. A greater deviation than can be attributed only to the response measurements was observed. It remains to be seen if this knowledge can be used to obtain better models. The feasibility study also included testing of different techniques to handle additive and multiplicative effects in the NIR data. The most surprising result was that the little used method of optimized scaling, proposed by Karstang and Manne in 1992 [57], gave the best results. The more popular technique orthogonal signal correction (OSC) [59, 60] gave little or no improvements from the use of raw data. An OSC component removes systematic noise, in a way which is similar to the addition of one extra LV using raw data. For process data where multiplicative and additive effects are expected, differentiation and normalization should be tested along with OS-2 and no preprocessing.

The process data in paper II showed a curvature well described by a second-order polynomial. In paper III several methods for handling the curvature when building calibration models were tested. These included transformations, local regression and quadratic PLS. Simple mathematical transformations were shown to be the best way to deal with non-linearities in data. Quadratic PLS is a new interesting technique, but contradicting results were obtained. More studies of the method are therefore required.

NIR data acquired from a batch process can be arranged in a three-way structure. No applications of N-way regression in combination with NIR spectra have so far been found in the literature. For monitoring chemical processes, the N-way techniques are appealing in the presence of curvature and scatter effects in data. In paper IV the performance of N-way decomposition techniques was tested. Variable selection was shown to be of the utmost importance for N-way modelling when many irrelevant variables are measured. N-way regression techniques provided calibration models of better predictive ability than two-way PLS. The residuals were found to be nonlinear for two-way PLS while for N-PLS they were random and small. The process data are nonlinear and when stacking data in a box, each time point and thereby concentration range is treated separately. If the data are close to linear within each time point, N-way regression seems to be a good option for handling non-linear data. To be able to use N-way techniques for on-line monitoring, data missing at the control points not yet monitored must be filled in. The results obtained in paper IV indicated that the use of scores could be a good way. However, frequent measurements are required to find a score describing the present monitored batch in the best possible way. How to deal with these missing values needs further investigations. Continuous monitoring, e.g. frequent acquisition of a spectrum, may be requested. Note that for two-way regression one model can be used to predict models from the start to the end of the process. On the contrary, for N-way techniques only samples acquired at the three control points are shown to work. Further investigations of how a slice in a box can be extrapolated and how to deal with this problem are required.

The process data investigated in papers II-IV were scanned manually using a monochromator instrument. An FT-NIR spectrophotometer enabling automatical acquisition of spectra has been installed in the process. A multiplexer is connected to the instrument. This makes it possible to scan spectra at different sites in the production line.

Several reactors are used to alkylate the main reactant. A score plot of the preprocessed data revealed that spectra scanned at one of the reactors do not group with the spectra scanned at the other reactors. The differences cannot be removed by techniques handling multiplicative and additive effects. The differences between reactors have been assigned to the fiber optics. A possibility is to replace the fiber with a new one. However, finding fibers of equal quality may be a problem. Instead, two models can be made; one for the reactors giving similar spectra and one for the other reactor. Another possibility is to use calibration transfer techniques [99] to convert spectra in the reactors to behave in a similar manner (shifts and peak intensity). In the future, having suitable calibration transfer techniques available may also be required if a new lamp or other spare parts replaced alter the spectra in a minor way. The issue of calibration transfer is therefore important. Calibration techniques requiring no standards is especially requested. For the process used as a case study in this thesis no stable reference sample of process solution is available, complicating the calibration transfer even more.

The correlation between each control point and corresponding HPLC value of main reactant revealed the data to be very noisy. The problems were assigned to instrumentation and physical effects. While mathematical transformations were shown to deal with noise and irregularities well in Chapter 4, the noise in these data was too severe to be handled simply by preprocessing techniques. Improvements in the sampling construction and technical devices were required. Noise is generated from turbulent flow in the pipe [8]. The noise was reduced by stopping the flow when acquiring data. Technical devices diminishing noise should be looked for in the future, e.g. use of bundle vs. single fiber, testing of fiber lengths and sampling devices, among others.

Continuous process monitoring is the goal. Surveillance of a process using parameters such as temperature, pressure and flow rate, has been used for multivariate statistical process control [51, 100, 101]. In the future it would be interesting to relate vibrational spectroscopy data to other process parameters, including raw-material quality, to detect which parameters that disturb the process. Being able to control this and thereby obtain same high yield and good quality in every batch, is the ultimate goal.

Bibliography

[1] Center for drug evaluation and research (CDER), *Guidance for industry. PAT - a framework for innovative pharmaceutical manufacturing and quality assurance* (U.S. Food and Drug Administration, Rockville, MD, USA, 2003).

[2] T. Næs, T. Isaksson, T. Fearn, and T. Davies, *A user-friendly guide to multivariate calibration and classification* (NIR Publications, Chichester, UK, 2002).

[3] K. Esbensen and P. Geladi, J. Chemom. **4**, 389 (1990).

[4] J. J. Workman, Jr, P. R. Mobley, B. R. Kowalski, and R. Bro, Appl. Spectrosc. Reviews **31**, 73 (1996).

[5] P. R. Mobley, B. R. Kowalski, J. J. Workman, Jr., and R. Bro, Appl. Spectrosc. Reviews **31**, 347 (1996).

[6] R. Bro, J. J. Workman, Jr., P. R. Mobley, and B. R. Kowalski, Appl. Spectrosc. Reviews **32**, 237 (1997).

[7] D. Malthe-Sørenssen, A. C. Schelver Hyni, A. Aabye, H. R. Bjørsvik, G. Brekke, and C. E. Sjøgren, Patent no. US 6,232,499 B1 (2001).

[8] D. Burns and E. Ciurczak, *Handbook of near infrared analysis* (Marcel Dekker, Inc, New York, 1992).

[9] J. M. Chalmers, *Spectroscopy in process analysis* (Sheffield Academic Press, 2000).

[10] J. J. Workman, Jr, D. J. Veltkamp, S. Doherty, B. B. Anderson, K. E. Creasy, M. Koch, J. F. Tatera, A. L. Robinson, L. Bond, L. W. Burgess, et al., Anal. Chem. **71**, 121R (1999).

[11] O. M. Kvalheim, Chemom. Intell. Lab. Syst. **2**, 283 (1987).

[12] S. Wold, K. Esbensen, and P. Geladi, Chemom. Intell. Lab. Syst. **2**, 37 (1987).

[13] L. Trefethen and D. I. Bau, *Numerical linear algebra* (SIAM, 1997).

[14] B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. de Jong, P. J. Lewi, and J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics: Part B* (Elsevier, 1998).

[15] Y.-z. Liang, O. M. Kvalheim, and R. Manne, Chemom. Intell. Lab. Syst. **18**, 235 (1993).

[16] W. H. Lawton and E. A. Sylvestre, Technometrics **13**, 617 (1971).

[17] R. Manne, Chemom. Intell. Lab. Syst. **27**, 89 (1995).

[18] M. Maeder and A. D. Zuberbuehler, Anal. Chim. Acta. **181**, 287 (1986).

[19] B. W. Silverman, Ann. Stat. **24**, 1 (1996).

[20] O. M. Kvalheim and Y.-z. Liang, Anal. Chem. **64**, 936 (1992).

[21] J. Toft and O. M. Kvalheim, Chemom. Intell. Lab. Syst. **19**, 65 (1993).

[22] H. R. Keller and D. L. Massart, Anal. Chim. Acta **246**, 379 (1991).

[23] P. J. Gemperline, J. Chem. Inf. Comput. Sci. **24**, 206 (1984).

[24] B. G. M. Vandeginste, W. Derks, and G. Kateman, Anal. Chim. Acta **173**, 253 (1985).

[25] A. Lorber, Anal. Chem. **58**, 1167 (1986).

[26] A. Höskuldsson, J. Chemom. **2**, 211 (1988).

[27] R. Nortvedt, F. Brakstad, O. M. Kvalheim, and T. Lundstedt, *Anvendelse av kjemometri innen forskning og industri*, ISBN 82-91294-01-1 (Tidsskriftforlaget kjemi AS, Norway, 1996).

[28] S. Wold, Technometrics **20**, 397 (1978).

[29] Q.-s. Xu and Y.-z. Liang, Chemom. Int. Lab. Syst. **56**, 1 (2001).

[30] H. van der Vöet, Chemom. Int. Lab. Syst. **25**, 313 (1994).

[31] H. Martens and T. Næs, *Multivariate calibration* (John Wiley and Sons, Inc., Chichester, 1989).

[32] I. E. Frank, Chemom. Int. Lab. Syst. **1**, 233 (1987).

[33] F. Lindgren, P. Geladi, S. Rännar, and S. Wold, J. Chemom. **8**, 349 (1994).

[34] F. Lindgren, P. Geladi, A. Berglund, M. Sjöström, and S. Wold, J. Chemom. **9**, 331 (1995).

[35] V. Centner, D.-L. Massart, O. E. de Noord, S. de Jong, B. M. Vandeginste, and C. Sterna, Anal. Chem. **68**, 3851 (1996).

[36] F. Westad and H. Martens, J. Near Infrared Spectr. **8**, 117 (2000).

[37] L. Nørgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, and S. B. Engelsen, Appl. Spectr. **54**, 413 (2000).

[38] A. Höskuldsson, Chemom. Int. Lab. Syst. **23**, 1 (1994).

[39] H. A. L. Kiers and I. Van Mechelen, Psychological Methods **6**, 84 (2001).

[40] R. Bro, Chemom. Int. Lab. Syst. **38**, 149 (1997).

[41] L. Ståhle, Chemom. Intell. Lab. Syst. **7**, 95 (1989).

[42] R. Bro, J. Chemom. **10**, 47 (1996).

[43] R. Bro, Chemom. Int. Lab. Syst. **46**, 133 (1999).

[44] M. M. Reis, D. N. Biloti, M. M. C. Ferreira, F. B. T. Pessine, and G. M. Teixeira, Appl. Spectr. **55**, 847 (2001).

[45] V. Pravdova, C. Boucon, S. de Jong, B. Walczak, and D. L. Massart, Anal. Chim. Acta **462**, 133 (2002).

[46] R. Bro and H. Heimdal, Chemom. Int. Lab. Syst. **34**, 85 (1996).

[47] H. A. L. Kiers, J. Chemom. **14**, 105 (2000).

[48] C. A. Andersson and R. Bro, Chemom. Int. Lab. Syst. **42**, 93 (1998).

[49] P. M. Kroonenberg, *Three-mode principal component analysis. Theory and applications*, vol. 2 (DSWO Press, Leiden, 1983).

[50] R. Bro, Ph.D. thesis, Universiteit van Amsterdam (1998).

[51] P. Nomikos and J. F. MacGregor, Technometrics **37**, 41 (1995).

[52] X. Meng, A. J. Morris, and E. B. Martin, J. Chemom. **17**, 65 (2003).

[53] G. R. Flåten and A. D. Walmsley, Analyst **128**, 935 (2003).

[54] Y.-z. Liang, O. M. Kvalheim, A. Rahmani, and R. G. Breereton, Chemom. Intell. Lab. Syst. **18**, 265 (1993).

[55] A. Savitzky and M. J. E. Golay, Anal. Chem. **36**, 1627 (1964).

[56] W. Windig, Chemom. Intell. Lab. Syst. **23**, 71 (1994).

[57] T. Karstang and R. Manne, Chemom. Intell. Lab. Syst. **14**, 165 (1992).

[58] P. Geladi, D. MacDougall, and H. Martens, Appl. Spectrosc. **3**, 491 (1985).

[59] S. Wold, H. Antti, F. Lindgren, and J. Öhman, Chemom. Intell. Lab. Syst. **44**, 175 (1998).

[60] T. Fearn, Chemom. Intell. Lab. Syst. **50**, 47 (2000).

[61] C. A. Andersson, Chemom. Intell. Lab. Syst. **47**, 51 (1999).

[62] J. A. Westerhuis, S. de Jong, and A. K. Smilde, Chemom. Intell. Lab. Syst. **56**, 13 (2001).

[63] J. Trygg and S. Wold, J. Chemom. **17**, 53 (2003).

[64] J. Verdú-Andrès, D. L. Massart, C. Menardo, and C. Sterna, Anal. Chim. Acta **349**, 271 (1997).

[65] I. E. Frank, Chemom. Intell. Lab. Syst. **27**, 1 (1995).

[66] S. Sekulic, M. B. Seasholtz, Z. Wang, B. R. Kowalski, S. E. Lee, and B. E. Holt, Anal. Chem. **65**, 835 A (1993).

[67] J. Verdú-Andrès, D. L. Massart, C. Menardo, and C. Sterna, Anal. Chim. Acta **389**, 115 (1999).

[68] O. M. Kvalheim, Chemom. Intell. Lab. Syst. **8**, 59 (1990).

[69] M. Blanco, J. Coello, H. Iturriaga, S. Maspoch, and J. Pags, Anal. Chim. Acta **384**, 207 (1999).

[70] T. Næs, T. Isaksson, and B. Kowalski, Anal. Chem. **62**, 664 (1990).

[71] S. Wold, N. Kettaneh-Wold, and B. Skageberg, Chemom. Int. Lab. Syst. **7**, 53 (1989).

[72] G. Baffi, E. B. Martin, and A. J. Morris, Comp. Chem. Eng. **23**, 395 (1999).

[73] B. Li, P. A. Hassel, E. B. Martin, and A. J. Morris, in *Proceedings of the PLS'02 international symposium*, edited by V. Vinzi, C. Lauro, A. Morineau, and M. Tenenhaus (CICIA-CERESTA, Montreuil, France, 2001), p. 163.

[74] A. Lorber, Z. Goldbart, and A. Halon, Anal. Chem. **57**, 2537 (1985).

[75] M. Iwahashi, M. Suzuki, N. Katayama, M. A. Matsuzawa, H. Czarnecki, Y. Ozaki, and A. Wakisaka, Appl. Spectr. **54**, 268 (2000).

[76] P. Gemperline, J. Whang Cho, and B. Archer, J. Chemom. **13**, 153 (1999).

[77] O. E. de Noord, Chemom. Int. Lab. Syst. **23**, 65 (1994).

[78] J. C. Miller and J. N. Miller, *Statistics for analytical chemistry* (Ellis Horwood, PTR Prentice Hall, UK, 1993), 3rd ed.

[79] G. Brink and L. Glasser, J. Phys. Chem. **82**, 1000 (1978).

[80] K. Shinomiya and T. Shinomiya, Bull. Chem. Soc. Jpn. **63**, 1093 (1990).

[81] G. M. Førland, F. O. Libnau, O. M. Kvalheim, and H. Høiland, Appl. Spectr. **10**, 1264 (1996).

[82] G. M. Førland, O. M. Kvalheim, Y.-z. Liang, H. Høiland, and A. Chazy, J. Phys. Chem. B **101**, 6960 (1997).

[83] E. Nodland, Appl. Spectr. **54**, 1339 (2000).

[84] A. N. Fletcher, J. Phys. Chem **76**, 2562 (1972).

[85] M. A. Czarnecki, J. Phys. Chem. A **104**, 6356 (2000).

[86] F. O. Libnau, A. A. Christy, and O. M. Kvalheim, Vibr. Spectr. **7**, 139 (1994).

[87] B. V. Grande, H. Kallevik, F. O. Libnau, and O. M. Kvalheim, Chemom. Intell. Lab. Syst. **45**, 7 (1999).

[88] E. Furusjö, L.-G. Danielsson, E. Könberg, M. Rentsch-Jonas, and G. Skagerberg, Anal. Chem. **70**, 1724 (1998).

[89] R. Tauler, B. Kowalski, and S. Fleming, Anal. Chem. **65**, 2040 (1993).

[90] J. Wagner, *The Graphical iPLS Toolbox for MATLAB, ver. 2.1: http://www.models.kvl.dk/source/ipls/* (KVL, Denmark, 2000).

[91] K. Faber and B. R. Kowalski, Chemom. Int. Lab. Syst. **34**, 283 (1996).

[92] H. W. Siesler, Y. Ozaki, S. Kawata, and H. M. Heise, *Near-infrared spectroscopy; principles, instruments, applications* (Wiley-VCH Verlag GmbH, Weinheim, Germany, 2002).

[93] P. Geladi and P. Åberg, J. Near Infrared Spectr. **9**, 1 (2001).

[94] P. Geladi and J. Forsström, J. Chemom. **16**, 329 (2002).

[95] N. Allosio, P. Boivin, D. Bertrand, and P. Courcoux, J. Near Infrared Spectr. **5**, 157 (1997).

[96] J. L. Beltrán, R. Ferrer, and J. Guiteras, Anal. Chim. Acta **373**, 311 (1998).

[97] C. A. Andersson and R. Bro, Chemom. Int. Lab. Syst. **52**, 1 (2000).

[98] R. Bro and H. A. L. Kiers, J. Chemom. **17**, 274 (2003).

[99] R. N. Feudal, N. A. Woody, H. Tan, A. J. Myles, S. D. Brown, and J. Ferre, Chemom. Int. Lab. Syst. **64**, 181 (2002).

[100] E. N. M. van Sprang, H.-J. Ramaker, J. A. Westerhuis, S. P. Gurden, and A. K. Smilde, Chem. Engng. Sci. **57**, 3979 (2002).

[101] D. J. Louwerse and A. K. Smilde, Chem. Engng. Sci. **55**, 1225 (2000).