Application of PARAFAC

on Spectral Data

PhD thesis by

Åsmund Rinnan

Supervisor

Rasmus Bro

Food Technology

Department of Food Science

The Royal Veterinary and Agricultural University

Preface

This PhD thesis is part of the project "new mathematic models for modeling spectroscopic and other multi-way data" financed by the Danish Research Council (STVF). The project focuses on the investigation of new and existing models for the evaluation of spectroscopic data.

The project was carried out at the Food Technology Group, Department of Food Science, The Royal Veterinary and Agricultural University, and included a six weeks research visit in the autumn of 2003 at the Department of Chemistry and Biochemistry in Arizona State University, USA.

Special thanks go to Rasmus Bro and Frans van den Berg from Food Technology for many valuable discussions. I am thankful to Jordi Riu from Department of Analytical and Organic Chemistry at Universitat Rovira i Virgili, Spain, for good and inspiring discussions. Further, I want to thank Karl Booksh for organizing an inspiring stay at his department at Arizona State University. I would also like to thank Riikka Rinnan for proofreading the text. I would further like to thank the whole Food Technology Group for three nice years here at the Royal Veterinary and Agricultural University of Copenhagen.

Finally I would like to thank my wife Riikka and my family for support throughout my Ph.D.

Frederiksberg, May 2004
Åsmund Rinnan

Summary

The use of fluorescence and low-field nuclear magnetic resonance (LF-NMR) measurements in both research and industries such as food, petrochemical and pharmaceutics production, is increasing. The interest for and application of chemometrics to these fields has grown, and the need for more and refined applications to evaluate the data is apparent.

This thesis focuses on chemometrics, mainly applied to fluorescence and LF-NMR, to increase the informational output from spectroscopic methods. Part of the research was on food products, but the conclusions on the developed models are more general and applicable to any field where fluorescence and/ or LF-NMR might be used.

The thesis consists of 9 papers, four of which already are published in international peer-reviewed journals. The first paper establishes an automatic way to estimate the number of components in fluorescence. This work is the basis for creating a generalized method for estimation of the right number of components. Papers II and III investigate existing and new methodologies to handle the unwanted Rayleigh scatter effect in fluorescence. They are followed by a practical solution for the analyses of real samples by parallel factor analysis (PARAFAC) in order to improve convergence time and decomposition accuracy. The subsequent two papers focus on the application of PARAFAC on the decomposition of LF-NMR relaxation curves of two different food products, namely fish (Paper V) and potatoes (Paper VI). In both of these papers predictions based on the scores from PARAFAC were performed. Predictions for these two products, together with predictions on a third product (rape seeds) were compared with the more traditional two-way method partial least squares (PLS) in Paper VII. Prediction on the scores from PARAFAC is a second-order prediction, which is the focus of Paper VIII. This is an investigation into optimal performance of second order prediction on

simulated data sets. The last Paper (IX) introduces a better method to uncertainty estimates in regression models compared to conventional methods on three-way data.

Oppsummering

Fluoressens og lav-felts kjernemagnetisk ressonans (LF-NMR) målinger blir brukt i økende grad innen forskning og industri, f.eks fødevarer, petrokjemi og farmasi. Interessen for og anvendelsen av kjemometri på disse feltene har steget, og det er nødvendig med mer og bedre applikasjoner for å evaluere dataene.

Denne avhandlingen fokuserer på kjemometri, hovedsakelig brukt på fluoressens og LF-NMR, for å øke informasjonsutbytte fra spektroskopiske metodene. Deler av forskningen var på fødevarer, men konklusjonene til de utviklede modellene er generell og brukbar til andre områder hvor fluoressens og / eller LF-NMR anvendes.

Oppgaven inneholder 9 artikler, hvorav fire allerede er utgitt i internasjonale journaler. Den første artikkelen etablerer en modell for automatisk estimasjon av antall komponenter i fluoressens. Dette arbeidet er grunnlaget for utarbeidelsen av en generell metode for estimering av antall komponenter. Artikkel II og III undersøker eksisterende og nye metoder for håndtering av uønsket Rayleigh lysspredning i fluoressens. De er fulgt opp av en praktisk løsning for analysering av prøver fra industrien ved bruk av parallel faktor analyse (PARAFAC) for å forkorte konvergenstiden og øke nøyaktigheten av dekomponeringen. De to neste artiklene fokuserer på bruken av PARAFAC til dekomponeringen av LF-NMR relaksasjonskurver på to forskjellige fødevarer, nærmere bestemt fisk (Artikkel V) og poteter (Artikkel VI). I begge disse artiklene brukes skårene fra PARAFAC i regresjon. Disse to datasettene pluss et nytt datasett brukes i sammenlikningen av denne typen regresjon med den mer tradisjonelle to-veis metoden partiell lineær regresjon (PLS) i Artikkel VII. Prediksjon på skårene fra PARAFAC er en andre-ordens prediksjons metode, som er fokuset i Artikkel VIII. Dette er en undersøkelse i optimal bruk av andre-ordens prediksjon på simulerte datasett. Den siste Artikkelen (IX) introduserer en bedre metode for usikkerhetsestimering for

tre-veis PLS sammenliknet med konvensjonelle metoder brukt på fluoressensspektra.

List of papers

This work consists of a total of nine papers, all of which have been or will be submitted and/or accepted for publication. Paper I investigates automatic estimates of the number of components in a fluorescence data set. Papers II-IV discuss several ways to deal with the light scattering effects present in fluorescence spectroscopy. Papers V-VI are based on the application of LF-NMR analysis on fish and potatoes, respectively, with Paper VII comparing prediction results between two-ways and three-ways chemometrics. Paper VIII gives an into-depth discussion of how to perform second-order prediction to get optimal results, while Paper IX presents work on the estimate of the standard error of prediction on fluorescence data.

Paper I: Determining the number of components in a PARAFAC decomposition of fluorescence landscapes

Rinnan, Å., Bro, R., in prep

Paper II: Handling of first order Rayleigh scatter in PARAFAC modeling of fluorescence excitation-emission data

Rinnan, Å., Andersen, C.M., *Chemometrics and Intelligent Laboratory Systems*, submitted

Paper III: 1st Order Rayleigh scatter as a Separate Component in PARAFAC Decomposition of Fluorescence Landscapes

Rinnan, Å., Booksh, K., Bro, R., in prep

Paper IV: Stabilizing the PARAFAC decomposition of Fluorescence Spectra by insertion of zeros outside the data area

Thygesen, L.G., Rinnan, Å.,Barsberg, S., Møller, J.K.S., *Chemometrics and Intelligent Laboratory Systems*, 2004, In press

Paper V: Distribution of water in fresh cod

Andersen, C.M., Rinnan, Å., *Lebensmittel Wissenschaft und Technologie*, **35**, 2002, 687-696

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

Povlsen, V.T., Rinnan, Å., van den Berg, F., Andersen, H.J., Thybo, A.K., *Lebensmittel Wissenschaft und Technologie*, **36**, 2003, 423-432

Paper VII: Alternative regression method to PLS on NMR-relaxation curves

Rinnan, Å., *Rivista di Statistica Applicata – Italian Journal of Applied Statistics*, **15 (3)**, 2003, 393-402

Paper VIII: Multi-way prediction in the presence of uncalibrated interferents

Rinnan, Å., Riu, J., Bro, R., *Chemometrics and Intelligent Laboratory Systems*, 2003, submitted

Paper IX: Standard Error of Prediction for Multiway PLS. 2. Practical Implementation in Fluorescence Spectroscopy

Bro, R., Rinnan, Å., Faber, N.M., *Chemometrics and Intelligent Laboratory Systems*, 2004, accepted

Notation

x	Scalar
x	Vector
X	Matrix
$\underline{\mathbf{X}}$	Three mode array
a	A-scores in PARAFAC
b	B-loadings in PARAFAC
B_0	Applied magnetic field in NMR
c	C-loadings in PARAFAC
D	Nuclear spin
e, E	Residual or error
f	One component in PCA/ PARAFAC
F	Total number of components in PCA/ PARAFAC
i, j, k, n	Index
I, J, K, L, N	Max value of index
m(t)	Relaxation curve
M_0	Amplitude of an LF-NMR signal
o	Estimated value for jack-knifing or bootstrapping
$\bar{o}$	Mean estimated value for jack-knifing or bootstrapping
p	Loading in PCA
s	Estimate of uncertainty
S	Electron excitation state
τ	LF-NMR – time between the initial 90º pulse and the first 180º pulse
t	Score in PCA or time
T_2	LF-NMR – Transverse or spin-spin relaxation time constant
y	True value
$\hat{y}$	Predicted value

Abbreviations

CMPG	Carr-Purcell-Meiboom-Gill
DECRA	Direct Exponential Curve Resolution Algorithm
DTLD	Direct Tri-Linear Decomposition
EEM	Excitation-emission-matrix
GRAM	General Rank Annihilation Method
LF-NMR	Low Field NMR
MLR	Multi-linear regression
NMR	Nuclear Magnetic Resonance
NPLS	Multi-linear PLS
OLS	Ordinary least squares
PARAFAC	Parallel Factor Analysis
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Squares
RMSECV	Root Mean Square Error of Cross Validation

Table of contents

1 Introduction

This thesis is divided into five main parts. The first Chapter gives a short introduction on the overall contents of the thesis. Chapter 2 concerns basic theory of fluorescence spectroscopy, while Chapter 3 describes the basic principals of LF-NMR. Chapter 4 explains the basis of the chemometric tools used throughout the thesis and presents the major results. In the last chapter the importance of this thesis is placed into a more holistic perspective.

During the last decades, the use of spectroscopic measurements has grown considerably. There are two main reasons for this trend. First of all, the general knowledge of spectroscopy has increased as the amount of research on the techniques has developed both through thorough understanding of the spectroscopic methods and by applying mathematics. This has led to an increase in applications of spectroscopic methods, both in industry and in new research areas. The subsequent demand for more precise and affordable instruments has again called for more research and better methods to handle the data of varying quality. Secondly, spectroscopic methods are in general fast and have the potential to be non-invasive, and thus they have a broader use than the slower traditional methods (e.g. extraction and titration), that often are invasive. The replacement of (or addition to) these older methods has shown valuable in many fields, e.g. process industry [He *et al.* 2003], pharmaceutical research [Kauppinen *et al.* 1993], food science [Munck *et al.* 1998], environmental science [Silverman 1993] and biochemistry [Plugge and van der Vlies 1993].

Historically, spectroscopy was mainly used for establishing the chemical features of pure samples. It was early realized that there was a linear trend between the spectroscopic signal of the sample and the concentration of the dissolved analytes. This observation made it possible to make prediction models based on the change in amplitude of the signal from one single wavelength/-number. However,

recording the amplitude change only at one point in the spectrum will not give any insight into the underlying complexity of the signal, and hence it will result in bad predictions if there are several phenomena contributing to the signal. The technique can be extended into recording the height of several peaks, which increases the amount of extracted information. Further increasing this to include all the data from the spectroscopic measurement is an obvious extension. The development in spectroscopic methods has also been to separate and identify the different analytes in a mixture. This has partly been achieved through the use of hyphenated systems (combing methods like gas chromatography and infrared spectroscopy). The aim is to separate the analytes in the first step (often chromatography) and subsequently identify them by a spectroscopic method (e.g. infrared). An approach which has proven to be a robust visualizing and modeling tool, and which includes most information available in the data, is chemometric methods. Chemometric methods are capable of e.g. performing predictions, curve resolution (separating the signal of each analyte in the solution), handling large datasets and exploring the space the data spans [Martens and Næs 1989, Grung and Kvalheim 1995].

In spectroscopy the energetic level of a sample is increased by the absorption of some external enforced energy beam. The spectroscopic methods are divided into categories depending on the wavelength range used for the energy beam, e.g. infrared (IR, ca. 2.5-15μm), ultraviolet/visible (UV-VIS, ca. 200-1000nm), and X-ray (ca. 10-250pm). Most spectroscopic methods measure the amount of energy absorbed by the sample. However, in fluorescence spectroscopy, not only the amount of absorbed energy, but also the amount of energy subsequently released from the sample is recorded.

In this thesis two spectroscopic techniques were studied: fluorescence spectroscopy, and low-field nuclear magnetic resonance (LF-NMR). In fluorescence spectroscopy the sample is excited by visible and near-visible light, and in LF-NMR by a radio frequency signal. Fluorescence spectroscopy is a

sensitive technique, and can measure concentrations at ppb or even at ppt levels [Li *et al.* 2003]. It has been used in various scientific fields, e.g. archaeology [Lyons *et al.* 2003], chemistry [Greetham and Ellis 2003] , food science [Moshou *et al.* 2003], environmental science [Claret *et al.* 2003], and psychology [Dmitrieva *et al.* 2004]. LF-NMR is a valuable spectroscopic technique thanks to its rapid measurements and the availability of small portable instruments. It has proven useful and reliable especially in the analysis of food products [Hills and Floc'h 1994, Micklander *et al.* 2002, Paper V, Paper VI], but also in other areas like the petrochemical industry [Nordon *et al.* 2002] and material sciences [Sharma *et al.* 2003].

This thesis focuses on the decomposition of fluorescence data and data from LF-NMR instruments. Recording one sample by fluorescence spectroscopy naturally produces two-way data for each sample; several emission spectra are recorded at several excitation wavelengths. Data from LF-NMR is normally one-way data; intensity of the signal is recorded at different times. There is, however, a method to rearrange this data into two-way data called DECRA/SLICING [Windig and Antalek 1997, Pedersen *et al.* 2002]. Stacking several two-way samples will form a three mode data array. This rearranged data can be handled well by parallel factor analysis (PARAFAC) [Carrol and Chang 1970, Harshman 1970, Bro 1997]. This work has sought to solve some practical aspects related to problems in the use of PARAFAC on data from fluorescence spectroscopy. These challenges include e.g. right number of components, light scattering effects (unwanted effects overlapping the signal of the analytes) [Lakowicz 1999, p. 39] and quenching (causing non-linearity in the signal with respect to concentrations) [Ingle and Crouch 1988]. The emphasis of the work with LF-NMR data has been on the application of PARAFAC in analysis of different food samples. In addition to these studies of fluorescence and LF-NMR data, the attributes of so-called second order prediction [Booksh and Kowalski 1994] have been investigated.

2 Fluorescence spectroscopy

The theory in this chapter does not intend to give the reader detailed knowledge in the field of fluorescence spectroscopy, but rather an introduction to the subject. A more detailed description can be found in the textbooks by Lakowicz (1999) and Ingle and Crouch (1988).

Fluorescence is one of the three luminescence methods used in spectroscopy [Skoog and Leary 1992, p. 174]. The other methods are phosphorescence and chemiluminescence. The difference between fluorescence and phosphorescence is found at the electronic level of the molecule. A paired electron in the ground state has two possible excitation levels (see Figure 2.1). In fluorescence the excited electron is paired to the second electron in the ground state, and thus the return to the ground state is spin-allowed and occurs rapidly (typically in 10^{-8}s). In phosphorescence the excited electron has the same spin as the electron in the ground state. In this configuration the transition to the ground state is said to be forbidden, and it thus occurs slowly (typically in the range of 1ms-1s).

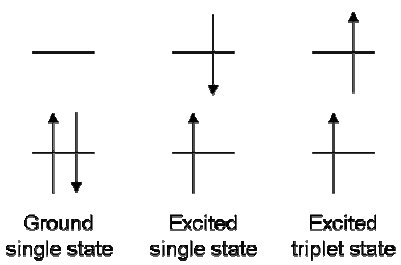


Figure 2.1: The ground state, and the two possible excitation states. The only one giving fluorescence is the excited single state.

In fluorescence and phosphorescence an electron is excited by absorbing energy in the form of light in the UV-VIS range. On the other hand, in chemiluminescence, the electron is excited through a chemical reaction. In all luminescence methods, after the excitation, the electron is relaxed through internal conversion and

vibrational relaxation to the lowest relaxational level of the excited state (S_1). The molecule will then emit energy and the electron will return to the non-excited state (S_0). Subsequently it will return to its ground single state through a series of vibrational relaxations. The only energy transfer that is of high enough intensity to be detected in fluorescence spectroscopy is the transfer between the excited and the non-excited state. The whole process is shown in a so called Jabłoński diagram in Figure 2.2.

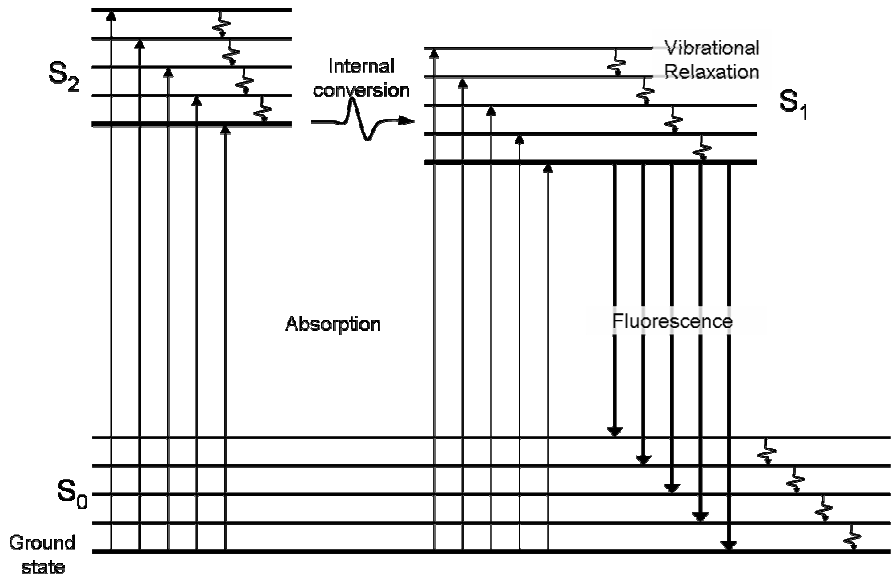


Figure 2.2: Jabłoński diagram of fluorescence.

Not all chemical components are fluorophores (molecules that are fluorescent). To be fluorescent a molecule needs to have double bonds or nucleophilic parts. The most intense fluorophores are aromatic rings and highly conjugated systems. Fluorophores are present in sources as different as food products (e.g. vitamins and proteins [Karoui *et al.* 2003]), pollution (e.g. polyaromatic hydrocarbons - PAH's [Fabbri *et al.* 2003]), and mining industry yields (e.g. uranium [Rathore and Kumar 2004]). Fluorophores behave differently by absorbing and emitting light at characteristic wavelengths. It is therefore possible to separate fluorophores based

on their spectral landscapes. However, the exact maximum for both absorption and emission is dependent on the temperature, pH and the solute of the sample.

2.1 Detection range

Fluorescence spectroscopy has low detection ranges compared to many other spectroscopic techniques. This makes it a valuable tool for measuring trace concentrations in products. In low concentrations, there is a linear relationship between the measured signal and the concentration of the fluorophore [Skoog and Leary 1992, p. 183]. Deviations from this linear relationship may be caused by too high concentrations of the fluorophore itself, causing inner filter effect, and/or by quenching [Ingle and Crouch 1988, p. 456 and 343-344], see Chapter 2.1.1. It may therefore be necessary to dilute the sample, measure in a smaller cuvette, or only measure the surface of the sample. If only the surface is to be measured, the representability of the information in the spectra is highly dependent on the homogeneity of the sample. Concentration determinations as low as the ppb-level are feasible for fluorescence spectroscopy, and measurement of concentrations even down to ppt have been reported [Li *et al.* 2003].

2.1.1 Quenching and inner filter effect

Quenching is the common term for the decrease of intensity caused by the sample itself. There are different types of quenching [Ingle and Crouch 1988, p. 343-344], the most common being mentioned below, together with the inner filter effect [Ingle and Crouch 1988, p456]:

1. *Dynamic quenching*: the excited-state fluorophore is deactivated on contact with other molecules in the sample. Its contribution dependents on the temperature of the sample. The higher temperature, the stronger is the effect of dynamic quenching is.
2. *Static quenching*: caused by the fluorescent forming non-fluorescent complexes with the quencher molecule.
3. *Inner filter effect*: the fluorophore itself or another molecule is absorbing some of the emitted light.

These different ways of quenching and the inner filter effect are visualized in Figure 2.3.

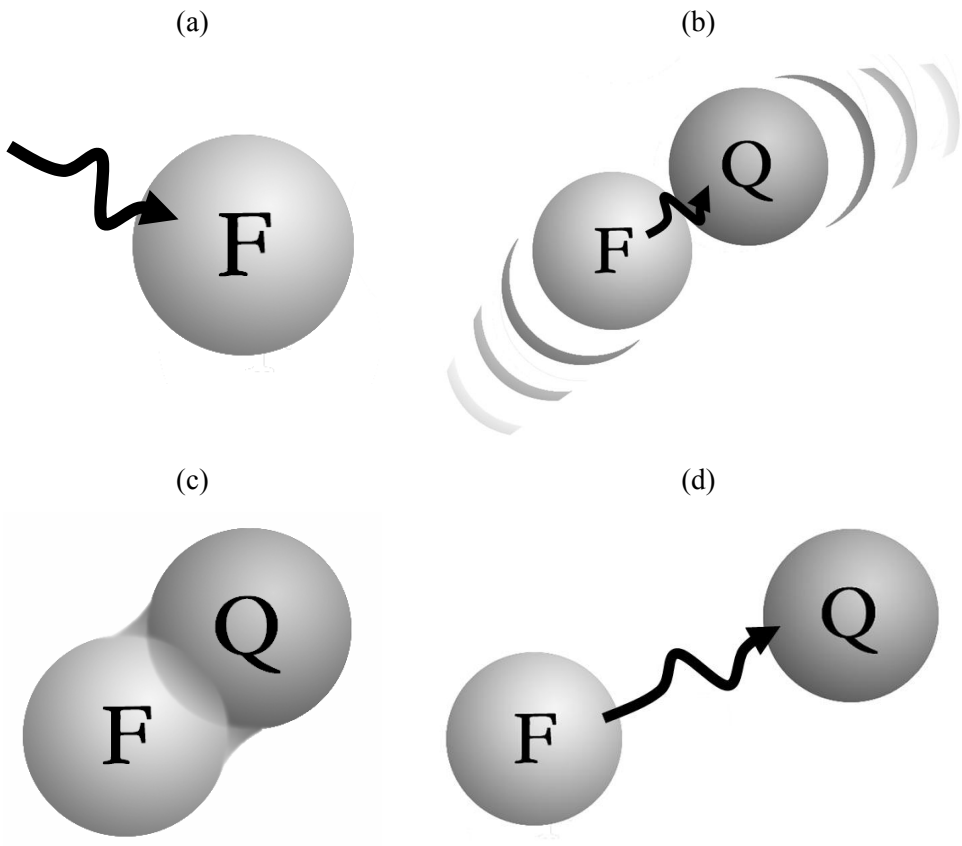


Figure 2.3: The fluorophore is first excited (a). Three ways of a decrease in the fluorescence intensity follow next: (b) Dynamic quenching, (c) static quenching, where the excitation energy is used in the formation, and (d) inner filter effect. 'F' denotes the fluorophore, and the 'Q' denotes the quencher.

2.2 Excitation-emission matrix

A common way to measure a fluorescence spectrum is exciting a sample at a certain wavelength, and detecting the emitted light in a range of wavelengths. It is also possible to excite the sample at different wavelengths, and then only measure the emission at one single wavelength [Karoui *et al.* 2003, Lakowicz 1999, p. 25].

This way of measurement makes it easier to visually inspect a group of samples, but by doing so all the available information is not recorded. Another way of collecting data from a fluorescence instrument is to collect several emission spectra at different excitation wavelengths [Matthews *et al.* 1996]. From this procedure the emission spectra can be set side-by-side thus creating a fluorescence landscape, with the excitation wavelength along the x-axis, the emission along the y-axis and the intensity of the signal along the z-axis. This landscape is also known as the excitation-emission-matrix, abbreviated EEM, see Figure 2.4. All the work in this thesis is performed on this type of fluorescence data.

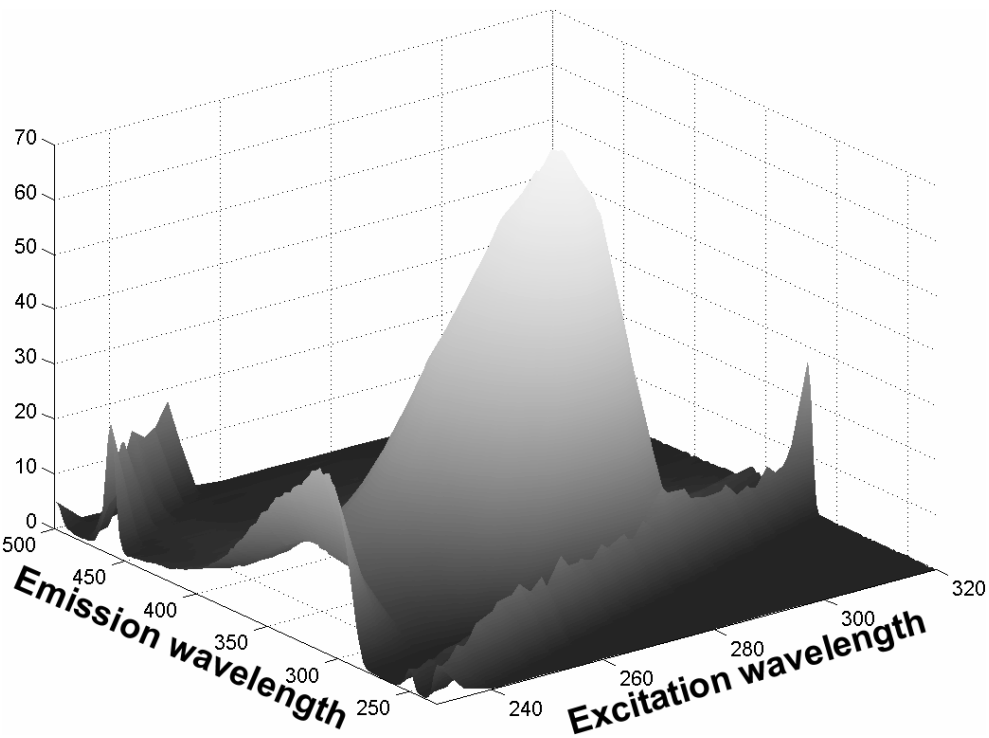


Figure 2.4: A typical EEM of a mixture of fluorophores in de-ionized water.

2.3 Scatter effects

An EEM will typically have areas with Rayleigh and Raman scatter [Lakowicz 199, p. 39-40, Ingle and Crouch, p. 462-463]. These two types of scatter can readily be seen in an EEM of a (non-fluorescing) water sample, as shown in Figure

2.5. In modeling the chemical information in fluorescence data, scatter effects are considered undesirable, because they do not hold any information about the fluorophores in the solution, and they can disturb the mathematical modeling of the fluorophores. Both of these scatter effects mainly originate from the solute.

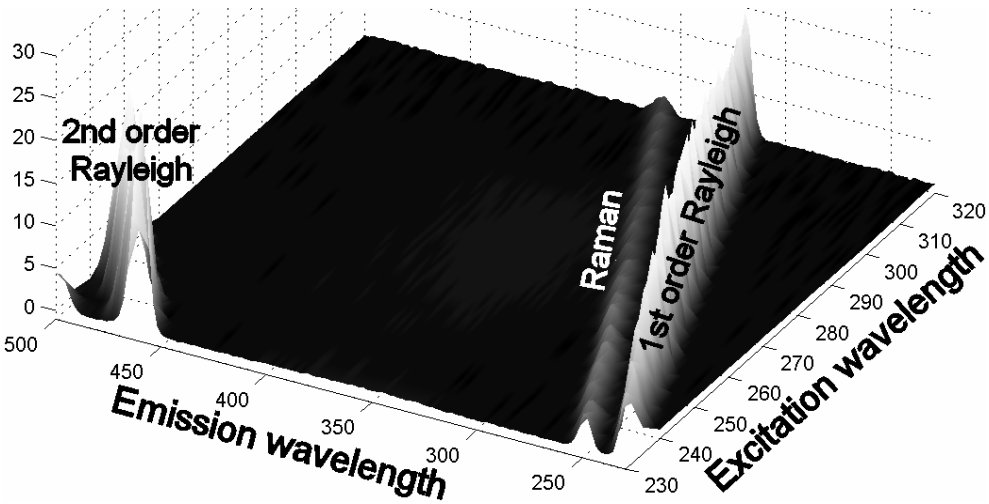


Figure 2.5: An EEM of de-ionized water showing three diagonal peaks: two Rayleigh peaks (1st and 2nd order) and one Raman peak.

2.3.1 Rayleigh scatter

Rayleigh scatter is predominantly caused by the solute, but may also originate from the fluorophores themselves [Ingle and Crouch 1988, p. 495-497]. Electrons in the molecules start to oscillate at the same frequency as the incident light, thus absorbing and emitting light at the same wavelength. Scatter lines at an integer-multiple of the absorbing wavelength will also occur. Therefore Rayleigh scatter is prenoted by a number, e.g. 1st order Rayleigh, 3rd order Rayleigh etc. Only the 1st and 2nd order Rayleigh scatters occurs in the EEM's in this thesis. Since there is no loss of energy in Rayleigh scattering it is a type of elastic scatter.

In fluorescence spectroscopy, one often only refers to Rayleigh scatter, and not to any other elastic scatter effects, although this is not always the correct term.

Rayleigh scatter requires that the dimension of the scatterer is much smaller than the wavelength of the incident light. Often this is the case in fluorescence EEM of solutions where water, methanol or other small solving agents are used. However, in the measurement of solids, or semi-solids, there may be other larger particles causing the scatter. If the size of the scatterer is close to the incident wavelength, the scattering is called Debye scatter. If the size of the scatterer is larger than the incident wavelength, the correct term to use is Mie scatter. However, in this work, the elastic scatter will only be referred to as Rayleigh scatter.

2.3.2 Raman scatter

While Rayleigh scatter is perfectly elastic, Raman is inelastic [Ingle and Crouch 1988, p. 497-499]. It is caused by the molecules of the solute absorbing some of the incident light, followed by the emission of a photon. However, the energy in this photon is less than the energy absorbed. Thus the molecule will be in an excited state (higher vibrational energy level). This is a semi stable state, from where the molecule finally will relax back to its ground state through small vibrational relaxations (not visible in fluorescence spectroscopy). This energy difference is constant and the Raman scatter line will be at a constant energy loss from the elastic Rayleigh scatter line. It is important to notice that a constant energy loss means a constant wavenumber shift – increasing the wavelength shift by increasing excitation wavelengths. The energy loss is dependent on the solute, e.g. for water it is 3600 cm^{-1}. An estimate of the specific energy loss of a solute can be calculated by recording an EEM of the solute, and then finding the mean energy difference between the 1st order Rayleigh and the Raman throughout the spectra. This information can be used under the analysis of the fluorescence spectra in order to separate the Raman scatter line from the signal of the fluorophores.

3 LF-NMR

Nuclear magnetic resonance (NMR) spectroscopy and relaxometry are widely used as analytical techniques in research and industry of e.g. food [Rutledge 2001, Zhen-Yi *et al.* 1996], pharmaceuticals [Fardella *et al.* 1995], biochemistry [Romão *et al.* 2000] and petrochemicals [Ahmad *et al.* 2002]. In the beginning, the focus was aimed towards obtaining higher resolution spectra. The only way to reach this goal was to build bigger and stronger superconductor magnetic fields. Strong magnetic fields require a lot of space, making it unsuitable for anything but laboratory work. However, it was realized that the high resolution these instruments give was not essential for all scientific and industrial applications. The development of smaller, lighter bench-top NMR instruments was a result. A bench-top NMR typically has a magnetic field-strength of 0.23 to 0.70 Tesla equal to 10 to 30 MHz for protons [Rutledge 1992]. Whereas high-field NMR has magnetic fields from 4.7 Tesla up to 21 Tesla equal to 900 MHz. Bench-top NMR instruments are typically used for obtaining relaxation curves, which is sample magnetization as a function of time. This is in contrast to high-field NMR where a spectrum is in the frequency domain is obtained (Figure 3.1).

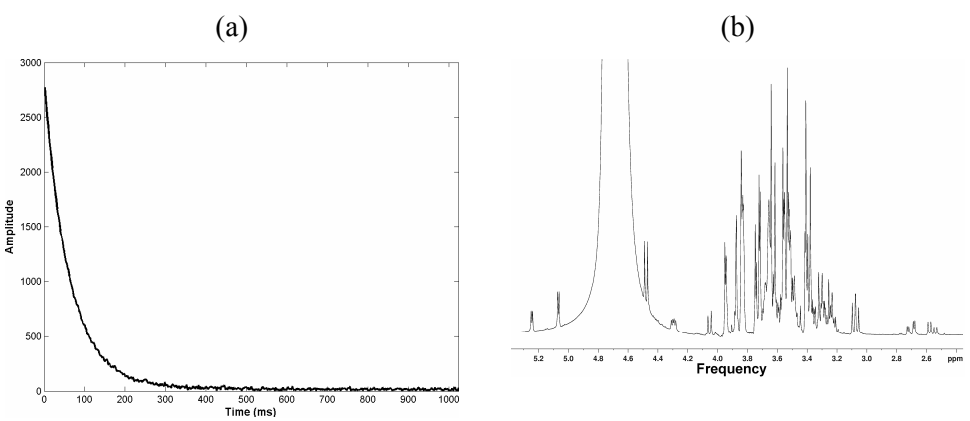


Figure 3.1: (a) LF-NMR spectra of fish meat, (b) a HF-NMR of apple-juice.

Relaxation curves primarily give information about water or fat content in the sample. It can further reveal whether the water is in a tightly bound, compartmentalized or in free state [Paper V and Paper VI].

This chapter deals with the basic theory of LF-NMR. If the reader is interested in a more detailed theoretical description, I would like to refer to special textbooks on the topic [Callaghan 1995, Williams and Fleming 1995, p. 63-169].

3.1 Basic NMR Theory

Some atomic nuclei have a nuclear spin (D), and the presence of the spin makes these nuclei behave like bar magnets [Callaghan 1995, Skoog and Leary 1992, Williams and Fleming 1995]. In the presence of an external magnetic field the nuclear magnets can orient themselves in $2D+1$ ways. Those nuclei with an odd mass number have nuclear spins of 1/2, 3/2 etc. Some of the most common nuclei of this type are: ^{1}H, ^{13}C, ^{19}F and ^{31}P. Of these hydrogen is the most frequently used in NMR, because it is the most abundant nucleic species and the ^{1}H isotope is the most common isotope of hydrogen (99.984%). ^{1}H has a nuclear spin of 1/2 and can therefore orient itself in two ways in the magnetic field, either being parallel or anti-parallel. The energy difference between these two states depends on the applied field strength.

In NMR experiments, the sample is exposed to an external magnetic field (B_0). This will force the majority of the nuclei in the sample to rotate in a manner such that their magnetic field is in accordance to the externally set field. If one applies a radio frequency signal orthogonal to the external magnetic field, the magnetic field of the nuclei begins to rotate thus changing the net magnetic field in a direction orthogonal to both the external magnetic field and to the radio frequency signal. This net magnetic field is recorded by a detector. Assigning axes to the system containing the nuclei, the external magnetic field, the radio frequency signal and the detector, may help in understanding how a signal is induced (Figure 3.2). The

external magnetic field (B_0) is along the z-axis. The radio frequency signal is sent through the sample along the x-axis, so that the magnetic field of the nuclei, and thus the nuclei itself is rotated around the x-axis. The net magnetic field along the y-axis is then recorded as the signal.

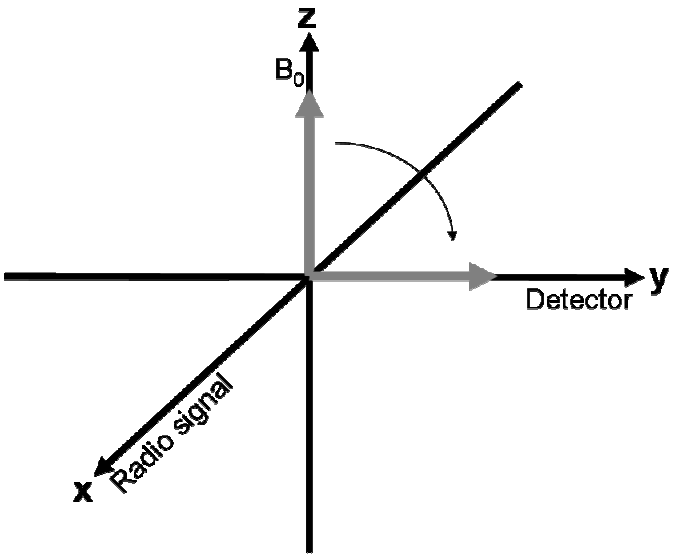


Figure 3.2: Diagram of the coordinates ascribed to the NMR instrument including the magnetic field (B_0), radio signal and the detector. The drawing also shows a 90° flipping of the net magnetic signal at equilibrium.

If the radio pulse is a 90°-pulse, the magnetic field of the nuclei will rotate from the z-axis in the direction of the y-axis. After this initial pulse, the nuclei will slowly rotate back into its equilibrium position aligned with the z-axis. This increase in the net-signal along the z-axis is described by a time constant – T_1 – known as the spin-lattice relaxation time. However, another mechanism that is faster than the above described phenomena is the dispersion of the signal in the xy-plane, which causes the net signal along the y-axis to decrease. After the frequency pulse is given along the x-axis, the magnetic field of the nuclei will be along the y-axis. Then, before it starts going back to equilibrium along the z-axis, it will completely disperse in the

xy-plane. This means that the net signal in y (and x) will eventually be 0. The decrease in this net signal is described by another time constant – T_2 – known as the spin-spin or transverse relaxation time. These T_2-values are of interest in this work.

3.2 Carr-Purcell-Meiboom-Gill (CPMG)

The time before the net-signal along the y-axis reaches zero is short and there is also some unwanted dispersion in this signal. It is of interest to correct for these effects to get more reliable measurements. Hahn introduced such an idea in 1950. Hahn realized that an additional pulse of 180° would refocus the net-signal and eliminate the dispersion, thereby also extending the time for which the relaxation can be measured. This idea was extended further by Carr and Purcell, and Meiboom and Gill in 1958, suggesting that a series of these 180° pulses should be added, all with a specific time, 2τ, between them. The time between the original 90° pulse and the first 180° is τ. The summit is defined as the maximum net-signal between each 180° pulse. Only every second spin-echo summit is recorded to correct for a small imprecision in the 180°-pulse. This can best be explained by an example:

> If the 180°-refocusing pulse in reality is a 185°-pulse, the maximum net signal would be 5° away from the y-axis. This means that the signal measured along the y-axis would be cos(5°)×true net signal and not the exact net signal. However, on the next pulse, which would be exactly the opposite of the first, -185°, the signal would be refocused at 0° and the signal recorded would be the exact net-signal. If the net signal is recorded for every refocusing pulse, an error will be present in the relaxation curve.

Thus the signal is measured at the times $4n*\tau$ (n = 1, 2, … N), as shown in Figure 3.3.

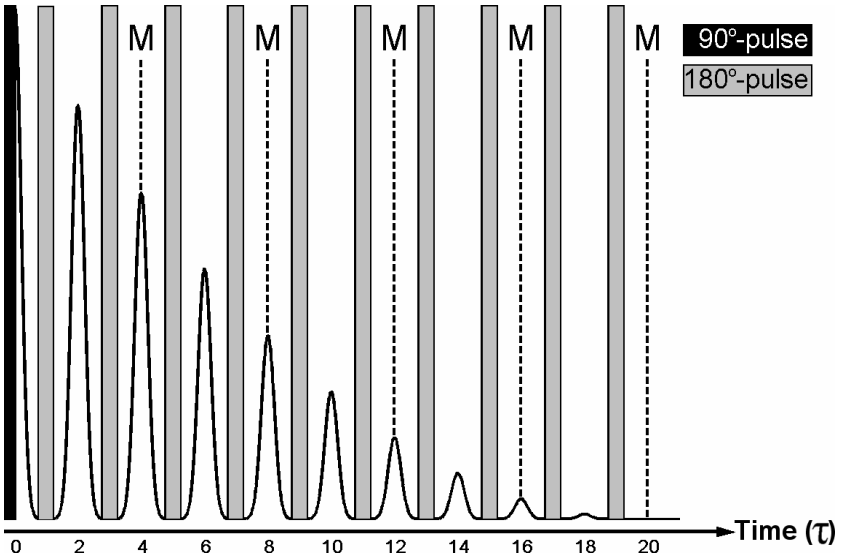


Figure 3.3: Signal from a Carr-Purcell-Meiboom-Gill measurement. 'M' indicates a measurement point. First a 90° pulse is given, then several subsequent 180° pulses.

Plotting these measurement points yields the relaxation curve (Figure 3.4).

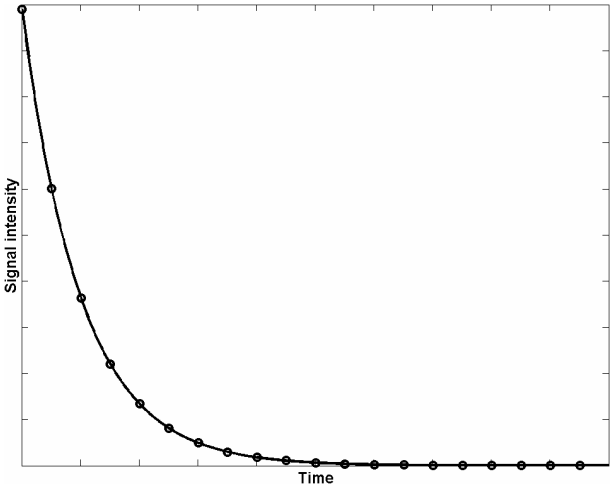


Figure 3.4: A typical CPMG relaxation curve without noise. 'o' are the measurement points in Figure 3.3.

The relaxation curve can mathematically be described as:

Equation 3.1

$$m(t) = \sum_{n=1}^{N} M_{0,n} \cdot \exp\left(-\frac{t}{T_{2,n}}\right) + E$$

where *m(t)* is the total relaxation signal, *N* is the number of underlying pure mono-exponential relaxation curves present in the raw data, M_0 is the magnitude of the relaxation curve, *t* is the time, T_2 is the characteristic transverse relaxation time, and *E* is the unmodeled part of the data (the noise). This equation is trivial to solve when $N = 1$. However, when $N > 1$, there are several ways of treating the data, some of which are specific to each sample, and others that take into account that similar samples should have equal T_2-values, see Chapter 4.3.

4 Chemometrics theory and major results

An essential part of chemometrics is, as the name indicates, chemistry. Chemometrics is the use of mathematics, statistics and computers to solve chemical problems. It is a field that expands as the speed of computers increases, the cost of spectroscopic instruments decreases and the general awareness of chemometrics increases.

This thesis mainly focuses on the analysis of three-way data from fluorescence spectroscopy and low-field NMR. Basic chemometric tools like principal component analysis (PCA), principal component regression (PCR) and partial least squares (PLS) regression are tools which are relevant for the understanding of the thesis, but will not be explained here. If the reader is unfamiliar with these tools, he/she can refer to Martens and Næs (1989) for a thorough explanation of the methods.

4.1 PARAFAC

Parallel factor analysis, or short PARAFAC [Carrol and Chang 1970, Harshman 1970, Bro 1997], is the main chemometric tool used throughout this thesis. In chemometrics, dimension is used for two related things describing the data. In order to prevent any confusion when dimension is used in this thesis, dimension will only refer to the size of the data, while modes will refer to the number of directions in the data: e.g. a matrix of the size $I \times J \times K$ has three modes, with dimension $I \times J \times K$. PARAFAC can be seen as an extension of the two mode PCA to the multi-mode system. While PCA only has a score and a loading matrix, PARAFAC has as many loading matrices as there are modes in the raw data. Often the first of these loading matrices is named scores and holds information about samples. In this thesis the highest number of modes of data analyzed was three, and therefore the explanation of PARAFAC will be in the three mode case, but the expansion to higher mode systems is straight forward.

A schematic explanation of the expansion from PCA to PARAFAC is shown in Figure 4.1.

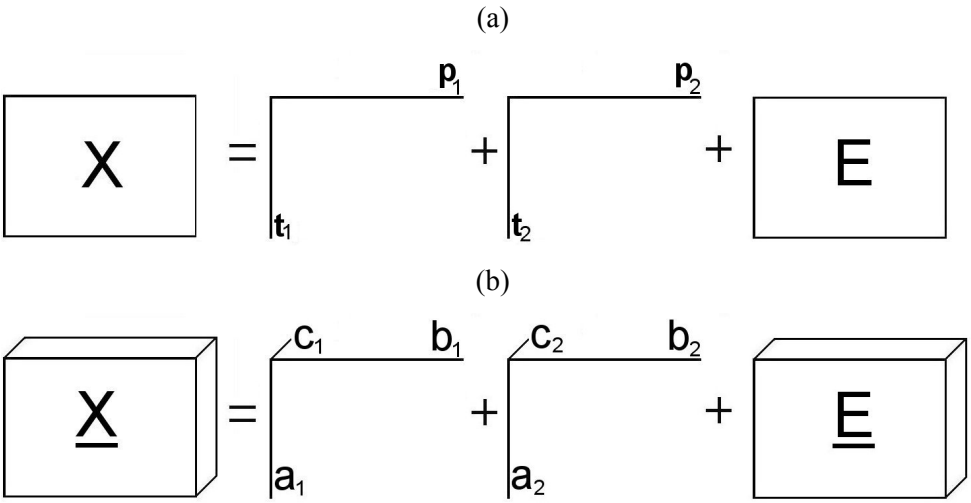


Figure 4.1: The expansion from (a) PCA to (b) PARAFAC.

This can also be shown with equations. The PCA in equation form can be written:

Equation 4.1
$$x_{ij} = \sum_{f=1}^{F} t_{if} p_{jf} + e_{ij} \qquad (\boldsymbol{i = 1,\ldots,I; j = 1,\ldots,J})$$

where x_{ij} is the original value in the position given by indices i and j; t and p denote the scores and loadings respectively, while e_{ij} is the unmodeled part of the data. PARAFAC can be written as follows

Equation 4.2
$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \quad (\boldsymbol{i = 1,..,I; j = 1,..,J; k = 1,..,K})$$

where x_{ijk} is the original value in the position given by indices i, j and k, while a, b and c are the scores and loadings and e_{ijk} is the unmodeled part of the data. For Equation 4.1 and Equation 4.2 f denote one factor and F is the total number of factors in the model. Both for PCA and PARAFAC the unmodeled part should optimally only contain noise. As can be seen from Figure 4.1, Equation 4.1 and

Equation 4.2, PARAFAC adds one more set of loadings to the decomposition, where this "extra" loading matrix explains the variance in the third or "extra" mode. Applying PARAFAC for parameter estimation (e.g. curve resolution), requires the data to be low-rank tri-linear, just as the data in the PCA case should be low-rank bi-linear. Low-rank bi-linear data means that the individual basic phenomena in the data can be explained by a small set of vectors in each of the modes of the original data. I.e. a diagonal variation in the data can not be explained by one vector in each of the directions in the original data (Figure 4.2). The extension from this example to low-rank tri-linearity is straight forward, just adding one more mode to the data and one vector in this new mode.

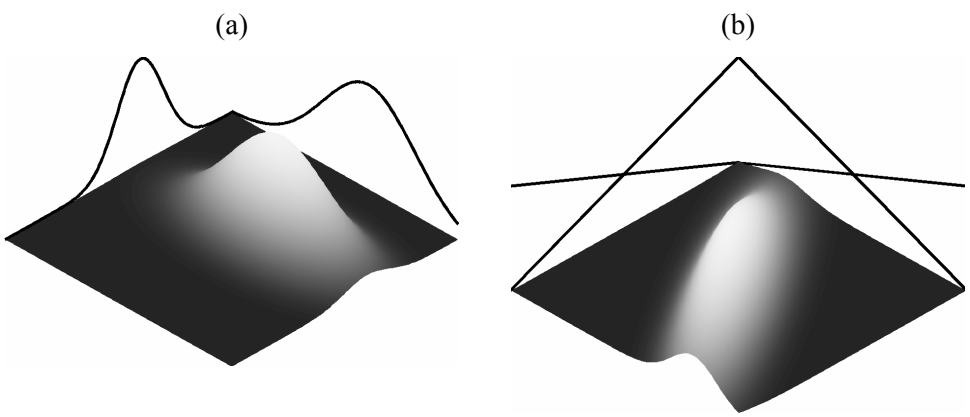


Figure 4.2: (a) Rank-one bi-linear data: The landscape can be described by one vector in each of the modes. (b) High-rank bi-linear data: The landscape can not be explained by one vector in each mode.

Not only does PARAFAC include another mode, but it also differs from PCA in how the decomposition is performed. While PCA can be calculated one factor at a time, PARAFAC necessarily computes them all simultaneously [Bro 1997]. Furthermore, the bi-linear PCA model has rotational freedom, while the tri-linear PARAFAC model, without any further constraints, is unique [Bro 1997]. This is a powerful feature of PARAFAC, making it an excellent tool for curve-resolution.

The advantage of PARAFAC compared to PCA in curve resolution can further be visualized by an example:

A data set of nine samples containing two fluorophores (indole and tryptophane) is decomposed by PARAFAC and PCA, Figure 4.3. In the case of PCA, the three-way array is unfolded into a two-mode matrix. The PCA loadings (Figure 4.3a-b) have been reshaped into the fluorescent landscapes for easier comparison. The PCA gives one component which closely resembles a fluorophore (Figure 4.3a), while the second component does not give any explicit physical or chemical meaning (Figure 4.3b). PARAFAC on the other hand, estimates two

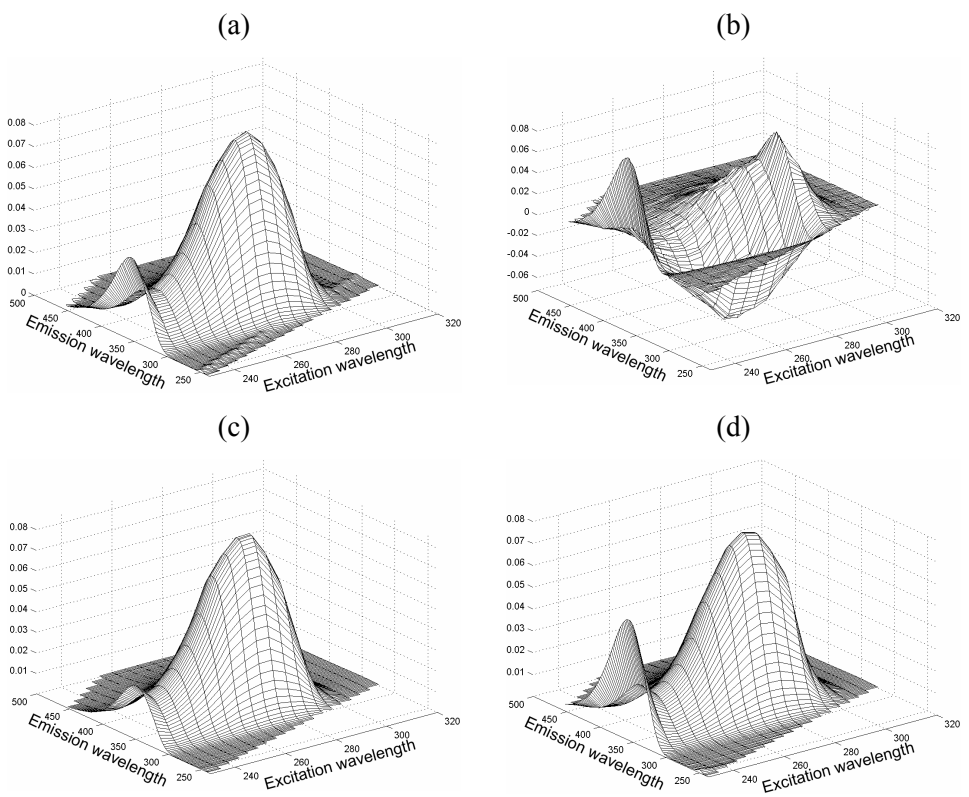


Figure 4.3: The decomposition of a system of two fluorophores – indole and tryptophane, by PCA (a and b), and PARAFAC (c and d). (a) and (c) are the first component, while (b) and (d) are the second component.

EEM's that closely resemble the EEM of indole (Figure 4.3c) and tryptophane (Figure 4.3d). As can be seen from Figure 4.3, if the original data are of a tri-linear structure, modeling it as only bi-linear may cause a less informative decomposition.

Not only can PARAFAC be used as a decomposition tool for qualitative analysis, but the score extracted may further be used in prediction [Bro 1997]. This type of prediction is called second order prediction [Booksh and Kowalski 1994], since a landscape of data (two-mode) is used for predicting one single number.

4.2 Chemometrics and Fluorescence Spectroscopy

PARAFAC has proven to be a valuable tool on the analysis of fluorescence spectra as a curve resolution method (Figure 4.3) [Booksh *et al.* 1996, Bro 1997, Bro 1999, Jiji *et al.* 2000, Lee *et al.* 1991, McKnight *et al.* 2001, Moberg *et al.* 2001, Wentzell *et al.* 2001].

Throughout this work, the fluorescence data has been arranged in a specific manner, with samples along the first mode, emissions along the second mode, and excitations along the third mode. Therefore the scores (**A** in Figure 4.1b and Equation 4.2) from PARAFAC are (ideally) proportional to the concentrations of the fluorophores in the samples, the second mode loadings (**B**'s) are the estimated emission spectra, while the third mode loadings (**C**'s) are the corresponding excitation spectra.

4.2.1 Number of components

The most important step in using PARAFAC (and for any chemometric tool in general) is to estimate the appropriate number of components in the data set. Bro and Kiers (2003) introduced the analytical tool core consistency diagnostics (CORCONDIA) for deciding the number of components. CORCONDIA is a method, which gives good estimates of the number of components on both fluorescence and other data [Bro 1998, Trevisan and Poppi 2003]. Other methods

to estimate the number of components are to assess the stability of the estimated loadings by jack-knifing [Martens and Martens 2001], split-half analysis [Harshman and de Sarboe 1994] or similar methods to estimate parameter uncertainty. The above mentioned methods all require visual inspection from the analyst and sometimes also include a-priori knowledge about the system. There are cases in which each of these methods fail, and thus relying on only one method can be fatal.

It is therefore of interest to have an automatic way to get a reliable estimate of the number of fluorophores in fluorescence data. In Paper I an automatic method of finding a good estimate of the number of components is investigated by looking at more than 80 different diagnostics, e.g. CORCONDIA, number of iterations, results from split-half analysis and jack-knifing. So far, the results indicate that this method and the diagnostics investigated are capable of estimating the right number of factors for fluorescence data. For nine of the twelve data sets, which were investigated, the estimated number of factors was equal to a-priori knowledge of the data sets. However, for three out of the twelve data sets, the estimated number of factors was uncertain, indicating a possible unstable additional factor. These data sets all contained catechol, which had a small impurity, giving rise to another weak PARAFAC component. This component, however, was of low intensity, and it did not disturb the decomposition of the other fluorophores; i.e. in an F+1 factor model where the extra factor was the impurity, the F first factors were equal to the F factor model. More work is needed to conclude this project. Different ways of analyzing the diagnostic tools will be tested in order to optimize the method for estimating the number of factors.

4.2.2 Light scatter

Light scattering, as described in Chapter 2.3, is a phenomenon having an unwanted bias effect in the decomposition step. Scatter cannot be described by a few PARAFAC factors, because it does not conform to the tri-linear structure which is necessary for PARAFAC to be able to model them. The Rayleigh and Raman

scatter lines, as seen in Figure 2.5, are diagonal and thus cannot be explained by one vector in the emission mode, and one in the excitation mode. The Rayleigh scatter line is of most concern since the intensity of this is higher than for Raman scatter [Skoog and Leary 1992, p. 298-299], thus affecting the decomposition to a larger extent. Sometimes the signal of the fluorophores lies away from the 1st order Rayleigh peak, and cutting away the area containing the scatter is sufficient [Moberg *et al.* 2001, Beltrán *et al.* 1998]. However, in several natural samples, the 1st order Rayleigh peak is partly overlapping the signal from one or more of the fluorophores. By removing this area with the scatter, one would also remove some of the information of the fluorophore(s). Therefore it is of interest to find a method to either remove exclusively the Rayleigh scatter, or to not let the Rayleigh scatter influence the decomposition.

4.2.2.1 Ways of removing Rayleigh

There are several ways of removing the Rayleigh scatter or enforcing PARAFAC not to take the Rayleigh scatter into account. The easiest way is by replacing the Rayleigh scatter and the area below and above these scatters by missing values [Munck *et al.* 1998], possibly using additional constraints [Andersen and Bro 2003, Bro 1999]. As explained above, this may lead to the loss of information, and as such is not generally recommended. Another solution is to subtract the spectrum of a standard from all the samples [Ho *et al.* 1978], which at the same time will reduce or remove the Raman scatter. This requires that such a standard is available, which is not always possible, especially for food, process or environmental samples. Further, this method would normally only reduce – not eliminate - the Rayleigh, and may introduce negative values to the EEM. Introducing negative values is not a problem in the decomposition step, but it indicates that more than the Rayleigh scatter in the sample has been removed, i.e. some information may have been lost. Inserting missing values [Christensen *et al.* 2003, Rodriguez-Cuesta *et al.* 2003, Trevisan and Poppi 2003, Jiji *et al.* 1999, Bro 1998, p. 235], and subtraction of a standard [Ho *et al.* 1980, McKnight *et al.* 2001] are the most common techniques for handling Rayleigh found in literature. It does not,

however, mean that these are the best methods for handling the scatter effect. Two different methods taking into account the shape and the position of the Rayleigh scatter are the use of weights (which also can take into account the Raman scatter), or modelling the Rayleigh scatter line separately.

Weights

Weights are used during the decomposition to focus the modelling on the fluorophores and not on the areas with the Rayleigh scatter (and possibly Raman). The areas containing the scatter will be weighted down, either decreasing steadily towards the peak of the scatter or plainly setting the weight to a fixed value as long as the scatter is present [Bro *et al.* 2002, Jiji and Booksh 2000], see Figure 4.4.

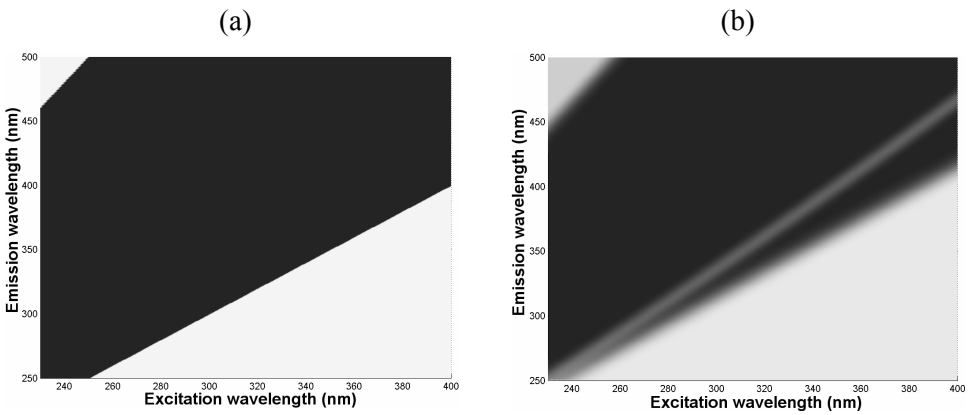


Figure 4.4: (a) Hard weights and (b) soft weights for the same data. Black means a high weight, while the lighter the area, the smaller weight. The lighter diagonal in (b) is the Raman scatter line.

Different ways of treating the 1st order Rayleigh through weighting and constraints are discussed in Paper II, where also the work from Paper IV (see below) is included. It is shown that using soft weights and introducing zeros to the area below the 1st order Rayleigh is the optimal existing method.

Modelling

A different approach to the problem is to model the light scatter as a separate part of the decomposition. In this work, the focus has been on modelling the 1st order Rayleigh scatter, as this is the most intense of the light scatters. The method proposed is shifting or rotating the original data, which makes the Rayleigh bi-linear in the new coordinate system, and then modelling this by PCA or PARAFAC. The modelled Rayleigh is then transformed back into the original coordinate of the EEM and subtracted from the original data. A PARAFAC of the fluorophores is then made on the residual. This model is subtracted from the raw data, and the Rayleigh scatter is modelled again and subsequently this model is subtracted from the raw data, etc. The complete model is then found through an iterative process swapping between modelling the Rayleigh scatter and the fluorophores until convergence.

This method is described and discussed in detail in Paper III. Modeling the Rayleigh scatter by PCA or PARAFAC assumes that this scatter is equal in shape for all excitation wavelengths, and that it only differs in magnitude. This assumption is theoretically valid, and is also confirmed in the paper. The modeling of the Rayleigh scatter was so successful on one of the data sets that the Raman scatter, which did not influence the decomposition prior to the modeling of the Rayleigh scatter, influenced the PARAFAC decomposition on removal of the Rayleigh scatter (Figure 4.5). The next logical step would therefore be to model the Raman scatter as well, unless one has a standard and can hence subtract that from the samples. A natural extension of the 1st order Rayleigh results would also be to model the higher order Rayleigh scatter in the same way as the first order Rayleigh has been modeled.

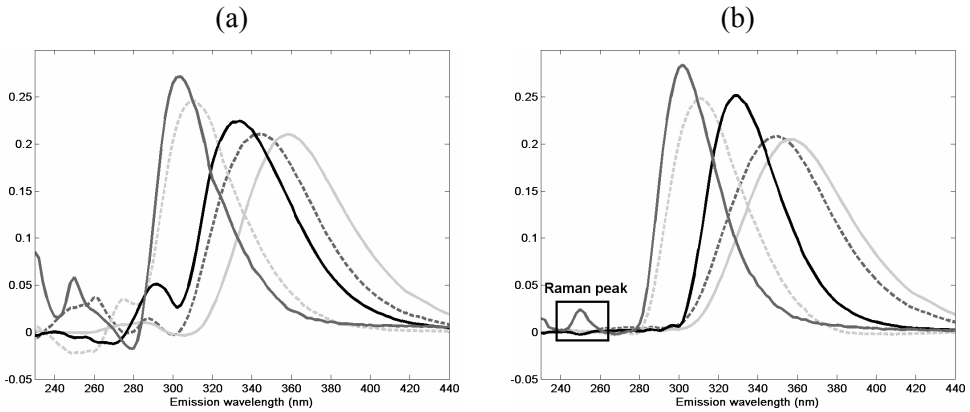


Figure 4.5: A five component system modeled without taking the Rayleigh into account (a), and with the Rayleigh as a separate component (b).

One of the advantages of modeling the Rayleigh scatter versus using weights is that the modeling is more automatic than the weighting scheme. On the other hand, weights can handle Raman scatter, which is not the case for the Rayleigh scatter model in its present form.

4.2.3 Stabilizing the PARAFAC decomposition

In some practical applications of fluorescence spectroscopy, the emission is only recorded at wavelengths higher than the incident excitation light. The missing part of the spectra is then filled with missing values [Christensen *et al.* 2003, Møller *et al.* 2003]. This, however, may cause the PARAFAC solution to include unwanted artefacts. Inserting zeros far below the 1st order Rayleigh scatter line (excitation wavelength larger than emission wavelength) to avoid these artefacts has been described earlier [Matthews *et al.* 1996, Stedmon *et al.* 2003]. It should be noted that on most datasets, inserting zeros all up to the Rayleigh scatter is not a good idea, since this can destroy the tri-linearity of the data, as explained by Andersen and Bro (2003). However, there has not been any thorough discussion on how inserting zeros affects the resolved spectra. In Paper IV several methods of inserting zeros and the effects on the resolved spectra are discussed. By inserting zeros the PARAFAC solution will be forced towards zero in these areas, and thus

removing large artefacts outside the data area (data area is between 1st and 2nd order Rayleigh scatter), see Figure 4.6. Further, it helps PARAFAC to converge faster (14650 iterations for the decomposition shown in Figure 4.6a compared to 398 iterations for the decomposition shown in Figure 4.6b). More work is still necessary in optimizing the amount of zeros to insert, and where to insert them.

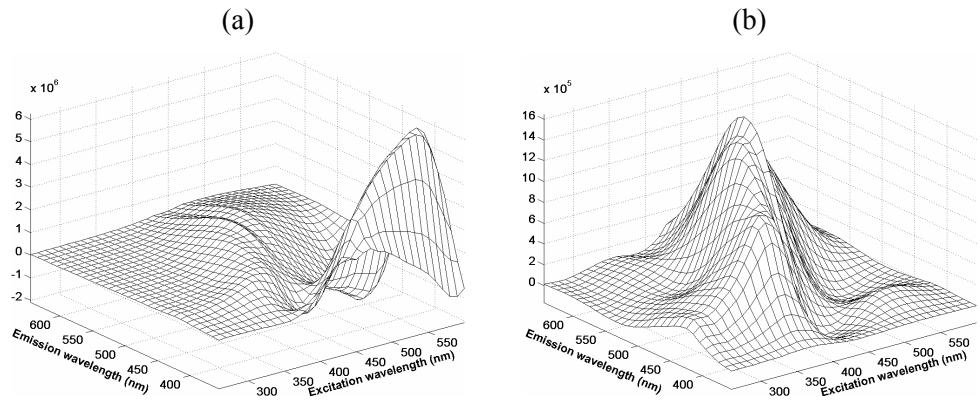


Figure 4.6: Decomposition by PARAFAC of a set of fluorescence EEM samples without (a) and with (b) the insertion of zeros some nm away from the 1st order Rayleigh scatter line. (Paper IV, Figure 5a and 5c)

4.3 Chemometrics and LF-NMR

The relaxation curves from LF-NMR are normally evaluated in one of four ways:

1. Exponential fitting
2. Distributional fitting
3. Matrix fitting
4. SLICING

The two first ones are both performed sample wise (one sample at a time), while the two last ones are performed at the whole data set at once. The three first methods will be explained shortly, while SLICING will be described more in detail.

4.3.1 Exponential fitting

In exponential fitting the summation of a low number of exponentials is fitted to the measurement profile in accordance with Equation 3.1 [Istratov and Vyvenko 1999]. The number of exponentials can e.g. be determined from the failure of a set of fewer exponentials to adequately model the curve. One of the main problems with this method is deciding the correct number of exponentials, because the fit of the curve will inherently improve with increasing number of exponentials, but using too many exponentials will lead to overfitting (starting to fit the noise in the data) [Pedersen *et al.* 2000]. As for matrix fit and SLICING, validation of the model is possible but not as straight forward as for these two methods [Pedersen *et al.* 2000]. Another problem is that it may not be a correct assumption that the relaxation curve only contains a few exponentials. A related method with the assumption that the number of exponentials should be large is distributed exponential fitting.

4.3.2 Distributed exponential fitting

In distributed exponential fitting, instead of trying to find a few specific T_2-values, the relaxation curve is described as a distribution of a set of (many) uni-exponentials [Butler *et al.* 1981, Provencher 1982a, Provencher 1982b]. This set of exponentials is defined prior to the analysis. The set of exponentials is typically large – e.g. 256 T_2-values - and distributed over a predetermined time window, which is based on a-priori knowledge on the system. A set of M_0's are then found, typically constrained to being smooth and positive. The major problem with this fitting procedure is that there are several distributions in the same time window that in practice give the same fit. Hence the problem is mathematically ill-defined. Further, the distribution curve is dependent on the number of uni-exponentials defined in the set and the range of T_2.

4.3.3 Matrix fit

Matrix fit is somewhat similar to exponential fitting. A set of sample relaxation curves are fitted with the same number of exponentials, and the T_2-values are set equal for all the samples [Pedersen *et al.* 2002]. The difference among the samples is the magnitude component (M_0-values in Equation 3.1). The advantage of this 2D method is that it is less sensitive to overfitting compared to the two previous methods. Moreover it is possible to evaluate the correctness of the model by techniques like split-half, jack-knifing or similar validation methods (explained in Chapter 4.5). The disadvantage is that it fits a discrete number of exponentials and all the exponentials are equal in all samples. First of all, natural samples may more correctly be described by a distribution of relaxation curves. Further, only similar samples can be expected to contain the same relaxation curves. The problem is to define when two samples are similar enough to be included in the same data set. In addition the T_2-value is highly dependent on the whole system under investigation, which might be a problem as for example the T_2-value will change with temperature [Micklander *et al.* 2002].

4.3.4 SLICING

The idea to use PARAFAC on LF-NMR data started from the direct exponential curve resolution algorithm (DECRA), a method introduced by Antalek and Windig (1997). This method was further developed into SLICING by Pedersen, Bro and Engelsen (2001). DECRA (and SLICING) takes into account the nature of the relaxation data. Since the data are exponentials, the different exponential contributions to the sum of exponentials (the different n's in Equation 3.1) have a different ratio along the relaxation curve. If one copies a part of the relaxation curve, and puts it behind the other, the same group of relaxation curves will be present in both slabs, but the ratio between the different T_2-times will differ in the two slabs. The faster relaxing components will have a larger part of the signal in the first slab, while the latter slab(s) will be dominated by the slower relaxing components. The terms used in SLICING for this reorganization of the data are

slabs for the number of copies in total, and *lag* for the difference in starting point of each slab from the original (Figure 4.7). In other words, if one makes three copies in total, starting from time-point 1, 2 and 5, there are three slabs, with lags 0, 1 and 4. The dimensionality of the data is thus decreased by four in the time-domain and increased by two in the new slab direction. This is because in the first slab, the four last time-points are removed, in the second, the first and the three last ones are removed, and in the third slab, the four first time-points are removed.

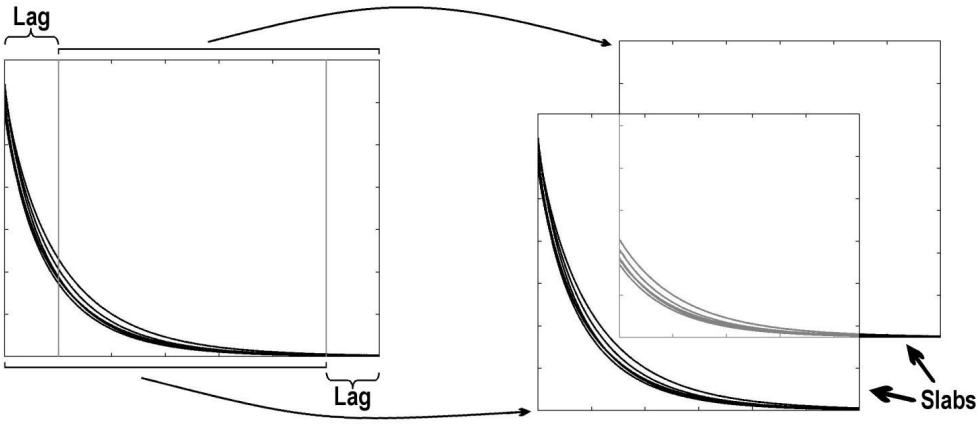


Figure 4.7: The idea behind SLICING. Only two slabs shown, but the number of slabs can be larger.

The dimensionality of the matrix will then increase from I (samples)×L (measurement points) to I×(L-maximum *lag*)×*number of slabs (K)*. Below this will simply be shortened to $I \times J \times K$.

The three mode array ($\underline{\mathbf{X}}$) has the size $I \times J \times K$ and it contains the elements x_{ijk}, where the first index (i) refers to the samples, the second (j) refers to the time and the third (k) refers to the slab number. The rearranged three mode data follow a PARAFAC model (Equation 4.2). In the case of LF-NMR, the scores (**A**'s) from the PARAFAC are proportional to the M_0-value in Equation 3.1, while the second mode loadings (**B**'s) are the estimated decay curves. The third mode loadings (**C**'s)

hold the same information as the second mode loadings, but with a smaller dimension and as such are not of any interest.

In the original work by Windig and Antalek (1997), the three mode array was treated by general rank annihilation method (GRAM) [Sanchez and Kowalski 1986], but because of that it was also limited to only include two slabs (or two samples). Pedersen, Bro and Engelsen (2001) extended this by using direct trilinear decomposition (DTLD) [Sanchez and Kowalski 1990], instead of GRAM, thus allowing the use of several slabs (and samples). The idea behind using several slabs instead of only two is that when using some slabs with small lags, the fast relaxing components contribute most to the total signal. The slower relaxing components are the major constituents in the slabs with larger lags. Including several slabs should then make it possible to extract the components more accurately. The problem with this, however, is that the number of possibilities for reorganizing the data is vast, and thus finding the optimal number of slabs and lags is, if not impossible, at least time consuming. This problem was realized by Engelsen and Bro (2003), who introduced power-slicing, a specific manner to define the lags to be used. However, by the time the work in this thesis was performed the work by Engelsen and Bro (2003) was not completed, and thus the more time consuming search for optimal numbers was used. Further, instead of using DTLD, the decomposition was optimized through the use of PARAFAC, with the DTLD solution as a starting point. Often the DTLD solution was the optimal solution, but in some cases refinement of the model was obtained through the use of PARAFAC. This is especially the case where one or more of the factors are small compared to the others [Paper V-VII].

If the model is adequate, each second mode loading (**B**) should be uni-exponential because the PARAFAC model can be shown to uniquely recover the underlying model when correctly specified [Windig and Antalek 1997]. If too many components are extracted, the curves will reflect this by one or more of them being

non-exponential. The number of factors can further be validated by the use of bootstrapping [Wehrens *et al.* 2000], jack-knifing [Martens and Martens 2001] and/or split-half analyses [Harshman and de Sarboe 1994] (see Chapter 4.5). In addition, the shape and distribution of the residuals of a model indicate whether there is information left in the data to model or not. The residuals should, in the perfect case, be randomly distributed with a zero mean. For an adequate model, the **B**'s, being uni-exponentials, can be fitted to Equation 3.1, and thus the T_2-values can be estimated. The M_0-value found in this fitting is dependent on the T_2-value of the curve, because in PARAFAC the **B**-loadings are normalized. The sample specific M_0-values are found by multiplying the M_0-values found in the exponential fitting with the corresponding **A**-scores from PARAFAC.

SLICING is in essence an alternative method to perform matrix-fitting, and thus also have the same advantages and disadvantages as mentioned under matrix fit, Chapter 4.3.3. The main difference between the two methods is that in SLICING the number of modes in the data is increased by one.

4.3.5 Applications of SLICING

In spite of the disadvantages mentioned in Chapter 4.3.3 SLICING was found to be useful in the study of two different types of food products. Paper V shows how SLICING successfully is used to separate five different potato tubers. In Paper VI good predictions of the water holding capacity in fish are shown. In both papers, the results from SLICING are compared with the above mentioned methods. They show that the use of distribution analysis does not contain the same information as SLICING, exponential fitting or matrix-fit. Furthermore, in Paper VI bootstrapping and split-half analysis was used to determine the number of components and investigating the effect of a baseline correction (Table 4.1). This investigation concluded that the baseline correction gave more stable results than modeling without any correction. In addition, the third factor in both cases is unstable, and thus two components were chosen.

Table 4.1: Mean relaxation times (in ms) and the standard error (in brackets) of second mode loadings obtained from two and three factor bootstrapping on 100 split-half models on raw and baseline corrected data. (Paper VI, Table 1)

	Uncorrected data		Corrected data	
	2 factors	3 factors	2 factors	3 factors
$T_{2,1}$	49.8 (0.6)	49.5 (0.6)	49.6 (0.4)	46.5 (1.8)
$T_{2,2}$	104.3 (2.8)	81.5 (1.7)	93.7 (1.2)	82.6 (5.7)
$T_{2,3}$		467.7 (100.8)		203.1 (73.6)

4.3.6 Regression analysis on LF-NMR data

In Paper VII there is a comparison of first-order prediction and semi-second order prediction (described in Chapter 4.4). The data inspected are LF-NMR spectra from three different food products, predicting from one to three different parameters through the use of PLS or OLS on the scores from SLICING compared to PLS on the raw data. There were no new interferents in any of these samples, and thus the second-order advantage (mentioned below) was not of any significance to the results. This study shows that both methods give similar prediction results. The decision on using one prior to the other depends on the focus of the analysis, PLS being faster while SLICING gives more qualitative information.

4.4 Second order prediction

The development and usage of PARAFAC has led to using this decomposition method also as a calibration tool. Regression based on the scores from PARAFAC is a so-called second order prediction. For better understanding about second order prediction, it may be of help to overview what zero to third order calibration is. Booksh and Kowalski (1994) gave such an overview. Multi-linear PLS (NPLS) was introduced by Bro (1996) as a regression method for multi-mode data.

However, there has not been any discussion in the literature to which type of calibration method it belongs. I have chosen to define NPLS as a 2nd order calibration method, focusing on the order of the data going into the calibration, thus 2nd order data lead to 2nd order calibration. Taking this definition into account, the overview by Booksh and Kowalski (1994) would be as follows:

- 0th order: calibration between two vectors – e.g. ordinary least squares (OLS).
- 1st order: calibration between a matrix and one (or more) vector(s) – e.g. PLS, PCR, or multi-linear regression (MLR).
- 2nd order: calibration between a set of matrices (forming a three mode array) and one (or more) vector(s) – e.g. OLS/MLR on PARAFAC scores and multi-linear PLS (NPLS).
- 3rd and higher order: calibration between sets of three or higher mode data and one (or more) vector(s) – e.g. OLS/MLR on PARAFAC scores and NPLS.

Booksh and Kowalski (1994) further discussed the second-order advantage, models capable of giving accurate predictions even in the presence of new and un-calibrated interferents in future samples. A method like PLS builds a model based on the calibration samples, and cannot handle any uncalibrated interferents. PLS can still give accurate predictions if the interferents do not overlap with the signal of the components in the calibration set, but if the overlap is extensive the predictions will be inaccurate.

In order for a method to posses the second-order advantage the sample data must be 2nd order (have three modes) and the decomposition should only be based on the **X**-data, and not on the **Y**-data. The reason is that the new samples with new interferents need more factors than the calibration model. There are no difficulties in adding more factors in a decomposition method like PARAFAC; one adds more components to the decomposition in Equation 4.2. However, a regression method

such as NPLS is built on a specific number of factors, and new unknown effects cannot be taken into account. Hence, a method such as NPLS, despite being a 2nd order calibration method, does not have the second-order advantage. PARAFAC on the other hand is a method which can be used in a way that maintains the second order advantage. Using PARAFAC as a calibration method, the scores from the decomposition would be used in an OLS/MLR setting with the concentration profiles. In the ideal case for spectroscopic data, an OLS without off-set should be used, as each factor in PARAFAC is an estimate of each of the components in the raw data. However, if there is any interaction between the components, or there is an offset in the raw data, other methods, such as OLS with offset, or even MLR would give better predictions. The second order advantage has been used with success on several occasions in the literature [Moberg *et al.* 2001, Saurina and Hernández-Cassou 2001, Nørgaard and Ridder 1994, Reis *et al.* 2000].

There has not been much discussion on how to perform second order prediction to optimize the regression model and the quality of prediction. It is not known how large the calibration set needs to be, or what degree of overlap can be handled. Other issues that have not been investigated are: whether the calibration set and the new sample(s) should be decomposed simultaneously or separately, and whether some factors from the decomposition step of the calibration set should be fixed or not when decomposing the set containing the new sample(s). Paper VIII discusses these and other questions concerning second order prediction. This work concludes that separating the new samples with new interferents from the calibration set is not a good idea. Instead, building a model on the whole dataset or fixing the loadings from the calibration data gives good predictions. During the investigation, the idea of fixing the scores from the calibration set also emerged, but there were problems in the computational step, and therefore these ideas had to be left out of the investigation. The cause of this problem should be investigated. Moreover, only simulated data was studied in this work, and therefore an extended study on real data should be performed in the future.

4.4.1 Uncertainty estimates in second-order prediction

In 0^{th} order prediction, a constant error variance is not accepted as good estimates for the uncertainty in the prediction models. On the contrary, the prediction error is smallest at the center of the data and increases towards the edges of the prediction line, as shown in Figure 4.8a. However, for some reason, uncertainty estimates in 1^{st} or 2^{nd} order prediction are normally given as constant values regardless of the actual sample (Figure 4.8b). These have been used and widely accepted. Faber and Bro (2002) established an easy and straightforward method to estimate the uncertainty sample-wise. This method is based on a-priori knowledge concerning the uncertainty in the reference measurements, and parameters from the regression. Paper IX shows an application of this method used on fluorescence data. The fluorescence data used in this work is a set of laboratory-made samples, and future work would be to apply the theory to even more complex data, e.g. food samples.

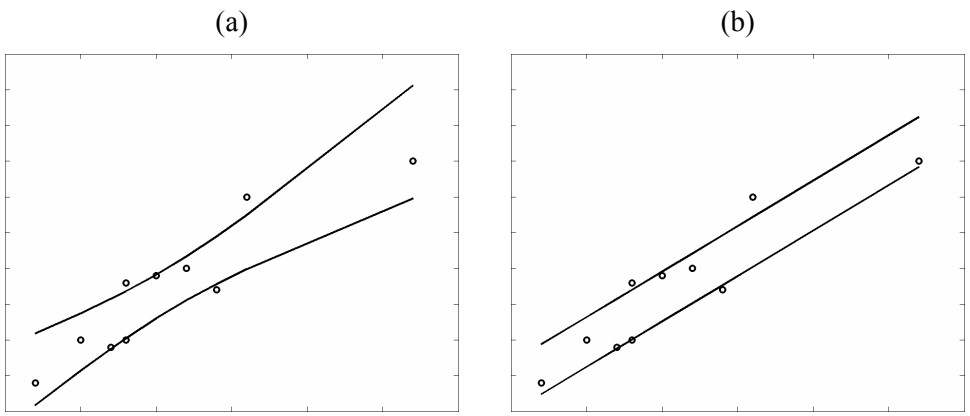


Figure 4.8: Confidence intervals in standard linear regression (a), and PLS (b).

4.5 Validation

Validation of a chemometric model is important in building the total model for reasons such as outlier detection, estimation of the right number of components and/or variable selection. There are several validation methods, each of which has specific properties and usages. A few of these methods are preferred over others because they are easy to understand, not cumbersome regarding computer-time,

and they give satisfying results. The focus is not to explain the mathematical background of the validation methods, but to give an introduction as to what the specific validation method does, and when or where it can be used. If a more mathematical explanation of these methods is sought I would like to refer to Martens and Næs (1989) or Martens and Martens (2001).

Detecting outliers is of vital importance when building a good model in order to prevent the model from focusing on the outliers. Even though the outlying samples might only be extreme cases of the other samples, this may cause the differences between the remaining samples to be drowned by this one (or possibly several) outlier(s). Note that extreme samples may stabilize a regression model, and thus it is important to know whether the outlier is an extreme sample of similar behavior as the others, or if it is different in behavior to the rest.

Estimating the right number of components can be done by using one or several validation methods. If they all point towards the same number of components, the rank of the system can be considered established. However, if there is some inconsistency in the results, the complexity of the model cannot be established, and caution should be maintained in the evaluation of the results. This inconsistency can both come from a vague definition of the problem, the diagnostic tools used to estimate the number of components and the data itself.

Variable selection may be appropriate in e.g. regression methods such as PCR, PLS and multi-linear PLS (NPLS), to achieve a better and more stable predictions for the future. It may also be necessary to use it prior to the final PARAFAC decomposition e.g. due to eliminating noisy measurement points.

There are several different ways of validation discussed in literature. Four of them will be discussed here: split-half analysis [Harshman and de Sarboe 1994], cross-validation [Martens and Martens 2001], jack-knifing [Martens and Martens 2001]

and bootstrapping [Wehrens *et al.* 2000]. All validations have been on decompositions by PARAFAC, and thus only this application of the validation schemes will be described and discussed.

4.5.1 Split-Half

The reason for performing split-half analysis is to investigate whether a model is adequate or not. The idea of the method is to split the dataset into two separate parts, and then compute a model with the same complexity on each of these halves. Split-half analysis is especially focused on estimating the number of factors in the system and for verifying that a postulated model is sensible. If the correct number of factors has been chosen and the data is homogenous (all containing the same underlying phenomena), the two models would ideally have the same loadings. Often this is assessed by visual inspection of the computed loadings, but may also be done by the use of qualitative criteria such as the regression coefficients between the similar loadings in the two models.

Small data sets may not give two identical sets of loadings due to the existence of a few samples having only one unique component. One should therefore show caution in the splitting, as this will affect the result, and thus the conclusion. Similar samples should be put in different splits. Thus for example in a process, the split should not be between early and late time-points, but as a random collection of half of the samples in one group, and the rest in another. However, if there are any replicates, these should be kept in the same group and not split up, because the interest lies in investigating the uncertainty of the model, and not the uncertainty between the replicates. If the dataset is large enough, splitting the data into three or more parts is a good idea.

4.5.2 Cross-validation

Cross-validation is a much used tool for estimating the right number of components in a regression model, as well as for determining the future uncertainty of the prediction of new samples. Cross-validation is a method that uses the variation

present in the data set available to the analyst. It is preferred instead of test-set validation when the data set is small and when an independent test set is infeasible.

In cross-validation a (small) number of samples is taken out from the data set and used as a test set. Then a model is built from the remaining samples in the dataset. The left-out samples are then predicted from this model. The left-out sample(s) is/ are then included in the data set again, and a new set of samples are taken out, and a model is built on the remaining samples. The left-out samples are predicted using the new model. This procedure is repeated until every sample has been left out once. The error of the prediction of each of the samples is calculated, and the average error term is found as:

Equation 4.3

$$RMSECV = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\left(\hat{y}_n - y_n\right)^2}$$

where $\hat{y}_n$ is the predicted value for the nth sample, y_n is the true value known from before and N is the total number of samples. The RMSECV is an approximation for the error of future (similar) samples. Systematically increasing the number of components in the model, the optimal number of components can be found at the (local) minimum of the RMSECV curve. For the RMSECV to be a good estimate for the uncertainty in future prediction, the data set should be built up of samples closely resembling future samples, and they should span out the variance and concentration levels (or other qualitative measurement) of future samples.

The number of samples taken out at a time can vary from one sample, so-called leave-one-out cross-validation, to a set of samples, so-called segmented cross-validation. The leave-one-out cross validation is a time-consuming approach, and may give a too optimistic estimate for future errors [Esbensen *et al.* 2000, p. 167], but is a good method if the number of samples is low. Segmented cross-validation, on the other hand, may be difficult if the number of samples is low, making the results from the sub-models too unstable.

Cross-validation has been used in Paper V and VI in order to estimate the model complexity for PLS models.

4.5.3 Jack-Knifing

Jack-knifing is closely related to cross-validation. However, while cross-validation only is used in the assessment of future prediction errors and estimating the right number of components, jack-knifing is used for investigating the uncertainty of estimated parameters, such as the regression vector, loadings, or scores. These uncertainties can further be used for variable selection.

As for cross-validation, a number of objects are taken out, and a model is built up on the remaining data. This procedure is repeated until all the objects have been left out once. Thus there will be as many regression vectors, loadings, and scores as there are models built. One of the parameter values is from when the object is predicted, while the rest of the parameter values are from when the object itself is included in the model building. By studying the uncertainty of the regression vector, unreliable variables and variables containing important information in the regression can be found. Uncertainty estimates of the loadings can be used to study how stable the model is concerning the shape of the loadings. For the model to be reliable, all the extracted loadings from all the models should be similar, or it is necessary to investigate why they are not similar. If the loadings vary considerably, the number of components may be wrong, and changing the number of components followed by another jack-knifing procedure can be helpful.

The uncertainty estimate of one parameter can mathematically be described by:

Equation 4.4:
$$s_{JK} = \sqrt{\frac{N-1}{N}\sum_{n=1}^{N}\left(o_n - \bar{o}\right)^2}$$

where o_n is the estimated value when fitting the model on the data excluding sample n, $\bar{o}$ is the mean value of the n estimated models, and N is the total number of samples.

It is also possible to take out more than one sample at a time from the data, and even do several resamplings, meaning that every sample is taken out more than once. This will, however, give a different correction factor in Equation 4.4. Instead of $(N\text{-}1)/N$ it is now needed to use $(N\text{-}m)/(rN)$ where m is the number of samples removed at a time and r is the number of times each sample has been removed. Further, the sum is now over all the estimations (rN/m estimations and not N). This is a resampling technique that is more time-consuming than the leave-one-out scheme, but may lead to better uncertainty estimates.

This last way of doing jack-knifing was used in Paper III to investigate the stability of modelling the Rayleigh scatter line as a separate component in the PARAFAC model.

4.5.4 Bootstrapping

Bootstrapping is another resampling approach used to estimate the uncertainty of model parameters. In this technique many new data sets, called bootstrap samples, are built up from the original dataset by sampling with replacement. As for jack-knifing a model is built on each of these bootstrap samples. This gives a number of estimated parameters from which statistical estimates such as mean and standard deviation can be obtained. If the number of bootstrap samples is selected large enough the bootstrap or so-called Monte Carlo error will be lower than the jack-knife error. In general the uncertainty estimate from bootstrapping is better than jack-knifing, but it requires more computer-time.

The estimate for the standard deviation is given by:

Equation 4.5:
$$s_{BS} = \sqrt{\frac{1}{N_{BS}-1}\sum_{n=1}^{N_{BS}}\left(o_n - \bar{o}\right)^2}$$

where o_n is the estimated value when fitting the model on bootstrap sample n, $\bar{o}$ is the mean value of the estimated o_n values, and N_{BS} is the total number of bootstrap samples.

Bootstrapping was used in Paper V to estimate the number of components in a SLICING model.

5 Practical perspectives

Fluorescence spectroscopy and LF-NMR have received considerable attention in recent years, also upon applying PARAFAC on the data. The PARAFAC model has provided insight into these spectroscopic methods, which is not gained with traditional analyses. Most notably this is due to the uniqueness properties of the PARAFAC model. However, since PARAFAC is a recently developed technique, there are still problems to solve and ties to untangle before widespread use of the technique is possible. For example, although PARAFAC is a powerful tool in analytical chemistry, traditional chemists do not yet support using it. Some reasons for the limited usage can be skepticism to the new technique, lack of a user friendly software, and requirement of knowledge of advanced chemometrics.

Fluorescence spectroscopy is a sensitive method, but it lacks the selectivity of other spectroscopic methods. However, fluorescence spectroscopy can accomplish its sought selectivity through the application of the PARAFAC model to the fluorescence data, and thereby splitting the signal into the different contributors. The combination of fluorescence spectroscopy and PARAFAC is thus a powerful tool in analytical chemistry. Work in this thesis showed that unwanted light scatter effects in fluorescence spectroscopy do not pose a problem to the decomposition of fluorescence data. It is possible, through the use of the right tools, to focus the decomposition of the fluorescence data on the fluorophores and not on the unwanted light scatter effects. The PARAFAC model further contains the powerful second-order advantage which makes accurate predictions possible even in the presence of uncalibrated new interferents in new samples.

Applying PARAFAC to LF-NMR data is a novel way of evaluating the relaxation signal, which gives easy access to information about the system at hand. It is an excellent tool for predicting the water and the oil content in a set of samples. This work also demonstrates that it can be used in classification of food products.

The thesis forms the basis of further development of multi-way chemometrics. Future research should aim at extending the estimation of the number of components, which was accomplished in this thesis with regard to fluorescence data, into a general automatic model applicable to any data set. Another remaining challenge is modeling all light scatter in fluorescence data as a separate set of components. Finally, the optimal method for second-order prediction still merits in-depth study.

6 Reference list

Ahmad, S., Ashraf, S. M., Sharmin, E., Zafar, F., and Hasnat, A.: Studies on ambient cured polyurethane modified epoxy coatings synthesized from a sustainable resource, *Progress in Crystal Growth and Characterization of Materials*, **45 (1-2)**, 2002, 83-88

Andersen, C. M. and Bro, R.: Practical aspects of PARAFAC modeling of fluorescence excitation-emission data, *Journal of Chemometrics*, **17 (4)**, 2003, 200-215

Beltrán, J. L., Ferrer, R., and Guiteras, J.: Multivariate calibration of polycyclic aromatic hydrocarbon mixtures from excitation–emission fluorescence spectra, *Analytica Chimica Acta*, **373 (2-3)**, 1998, 311-319

Booksh, K. S. and Kowalski, B. R.: Theory of analytical chemistry, *Analytical Chemistry*, **66 (15)**, 1994, 782A-791A

Booksh, K. S., Muroski, A. R., and Myrick, M. L.: Single measurement excitation/emission matrix spectrofluorometer for determination of hydrocarbons in ocean water .2. Calibration and quantitation of naphthalene and styrene, *Analytical Chemistry*, **68 (20)**, 1996, 3539-3544

Bro, R.: Multiway calibration. Multilinear PLS, *Journal of Chemometrics*, **10 (1)**, 1996, 47-61

Bro, R.: PARAFAC. Tutorial and applications, *Chemometrics and Intelligent Laboratory Systems*, **38 (2)**, 1997, 149-171

Bro, R.: Multi-Way Analysis in the Food Industry - Models, Algorithms and Applications, Universiteit van Amsterdam, The Netherlands, 1998, PhD thesis

Bro, R.: Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis, *Chemometrics and Intelligent Laboratory Systems*, **46 (2)**, 1999, 133-147

Bro, R., Sidiropoulos, N. D., and Smilde, A. K.: Maximum likelihood fitting using ordinary least squares algorithms, *Journal of Chemometrics*, **16 (8-10)**, 2002, 387-400

Bro, R. and Kiers, H. A. L.: A new efficient method for determining the number of components in PARAFAC models, *Journal of Chemometrics*, **17 (5)**, 2003, 274-286

Butler, J. P., Reeds, J. A., and Dawson, S. V.: Estimatin solutions of first kind integral equations with nonnegative constraints and optimal smoothing, *SIAM Journal of Numerical Analysis*, **18 (3)**, 1981, 381-397

Callaghan, P. T.: Principles of nuclear magnetic resonance microscopy, 1 ed., Oxford University Press, Oxford, 1995, 492 pages

Carr, H. Y. and Purcell, E. M.: Effects of diffusion on free precession in nuclear magnetic relaxation times, *Physical Review*, **29**, 1958, 630-638

Carrol, J. D. and Chang, J.-J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition, *Psychometrika*, **35 (3)**, 1970, 283-319

Christensen, J., Povlsen, V. T., and Sørensen, J.: Application of fluorescence spectroscopy and chemometrics in the evaluation of processed cheese during storage, *Journal of Dairy Science*, **86 (4)**, 2003, 1101-1107

Claret, F., Schäfer, T., Bauer, A., and Buckau, G.: Generation of humic and fulvic acid from Callovo-Oxfordian clay under high alkaline conditions, *The Science of the Total Environment*, **317 (1-3)**, 2003, 189-200

Dmitrieva, N., Rodríguez-Malaver, A. J., Pérez, J., and Hernández, L.: Differential release of neurotransmitters from superficial and deep layers of the dorsal horn in response to acute noxious stimulation and inflammation of the rat paw, *European Journal of Pain*, 2004, In press

Engelsen, S. B. and Bro, R.: PowerSlicing, *Journal of Magnetic Resonance*, **163 (1)**, 2003, 192-197

Esbensen, K., Guyot, D., and Westad, F.: Multivariate Data Analysis - in practice, 4 ed., Camo, Oslo, 2000, 600 pages

Fabbri, D., Vassura, I., Sun, C.-G., Snape, C. E., McRae, C., and Fallick, A. E.: Source apportionment of polycyclic aromatic hydrocarbons in a coastal lagoon by molecular and isotopic characterisation, *Marine Chemistry*, **84 (1-2)**, 2003, 123-135

Faber, N. M. and Bro, R.: Standard error of prediction for multiway PLS 1. Background and a simulation study, *Chemometrics and Intelligent Laboratory Systems*, **61 (1-2)**, 2002, 133-149

Fardella, G., Barbetti, P., Chiappini, I., and Grandolini, C.: Quantitative analysis of fluoroquinolones by 1H- and 19F-NMR spectroscopy, *International Journal of Pharmaceuticals*, **121 (1)**, 1995, 123-127

Greetham, G. M. and Ellis, A. M.: Laser-induced fluorescence spectroscopy of the gallium dimer: evidence for a 3u electronic ground state, *Journal of Molecular Spectroscopy*, **222 (2)**, 2003, 273-275

Grung, B. and Kvalheim, O. M.: Resolution of multicomponent profiles with partial selectivity - A comparison of direct-methods, *Chemometrics and Intelligent Laboratory Systems*, **29 (1)**, 1995, 75-87

Hahn, E. L.: Spin echoes, *Physical Review*, **80**, 1950, 580-594

Harshman, R. A.: Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis, *UCLA working papers in phonetics*, **16**, 1970, 1-84

Harshman, R. A. and de Sarboe, W. S.: An application of PARAFAC to a small sample problem, demonstrating preprocessing, orthogonality constraints and split-half diagnostic techniques, *Research methods for multimode data analysis*, Ed.: Law, H. G., Snyder, C. W., Hattie, J. A., and McDonald, R. P., Praeger, New York, 1994, 602-642

He, L. M., Kear-Padilla, L. L., Lieberman, S. H., and Andrews, J. M.: Rapid in situ determination of total oil concentration in water using ultraviolet fluorescence and light scattering coupled with artificial neural networks, *Analytica Chimica Acta*, **478**, 2003, 245-258

Hills, B. P. and Floc'h, G. L.: NMR studies of non-freezing water in cellular plant tissue, *Food Chemistry*, **51**, 1994, 331-336

Ho, C. N., Christian, G. D., and Davidson, E. R.: Application of the Method of Rank Annihilation to Quantitative Analyses of Multicomponent Fluorescence Data from the Video Fluorometer, *Analytical Chemistry*, **50 (8)**, 1978, 1108-1113

Ho, C. N., Christian, G. D., and Davidson, E. R.: Application of the Method of Rank Annihilation to Fluorescent Multicomponent Mixtures of Polynuclear Aromatic Hydrocarbons, *Analytical Chemistry*, **52**, 1980, 1071-1079

Ingle, J. D. Jr. and Crouch, S. R.: Spectrochemical analysis, Prentice Hall, New Jersey, 1988, 590 pages

Istratov, A. A. and Vyvenko, O. F.: Exponential analysis in physical phenomena, *Review of Scientific Instruments*, **70 (2)**, 1999, 1233-1257

Jiji, R. D., Andersson, G. G., and Booksh, K. S.: Application of PARAFAC for calibration with excitation-emission matrix fluorescence spectra of three classes of environmental pollutants, *Journal of Chemometrics*, **14 (3)**, 2000, 171-185

Jiji, R. D. and Booksh, K. S.: Mitigation of Rayleigh and Raman spectral interferences in multiway calibration of excitation-emission matrix fluorescence spectra, *Analytical Chemistry*, **72**, 2000, 718-725

Jiji, R. D., Cooper, G. A., and Booksh, K. S.: Excitation-emission matrix fluorescence based determination of carbamate pesticides and polycyclic aromatic hydrocarbons, *Analytica Chimica Acta*, **397 (1-3)**, 1999, 61-72

Karoui, R., Mazerolles, G., and Dufour, E.: Spectroscopic techniques coupled with chemometric tools for structure and texture determinations in dairy products, *International Dairy Journal*, **13 (8)**, 2003, 607-620

Kauppinen, R. A., Williams, S. R., Busza, A. L., and van Bruggen, N.: Applications of magnetic resonance spectroscopy and diffusion-weighted imaging to the study of brain biochemistry and pathology, *Trends in Neuroscience*, **16 (3)**, 1993, 88-95

Lakowicz, J. R.: Principles of Fluorescence Spectroscopy, 2 ed., Kluwer Academic/Plenum Publishers, New York, 1999, 698 pages

Lee, C. H., Kim, K., and Ross, R. T.: Trilinear Analysis for the Resolution of Overlapping Fluorescence Spectra, *Korean Biochemistry*, **24 (4)**, 1991, 374-379

Li, J.-S., Wang, H., Zhang, X., and Zhang, H.-S.: Spectrofluorometric determination of total amount of nitrite and nitrate in biological sample with a new fluorescenct probe 1,3,5,7-tetramethyl-8-(3',4'-diaminophenyl)-difluoroboradiaza-s-indacene, *Talanta*, **61 (6)**, 2003, 797-802

Lyons, W. H., Glascock, M. D., and Mehringer, P. J. Jr.: Silica from sources to site: ultraviolet fluorescence and trace elements identify cherts from Lost Dune, southeastern Oregon, USA, *Journal of Archaeological Science*, **30 (9)**, 2003, 1139-1159

Martens, H. and Næs, T.: Multivariate Calibration, Wiley, New York, USA, 1989,

Martens, H. and Martens, M.: Multivariate analysis of quality - An introduction, John Wiley, Chichester, 2001, 445 pages

Matthews, B. J. H., Jones, A. C., Theodorou, N. K., and Tudhope, A. W.: Excitation-emission-matrix fluorescence spectroscopy applied to humic acid bands in coral reefs, *Marine Chemistry*, **55 (3-4)**, 1996, 317-332

McKnight, D. M., Boyer, E. W., Westerhoff, P. K., Doran, P. T., Kulbe, T., and Andersen, D. T.: Spectrofluorometric characterization of dissolved organic matter for indication of precursor organic material and aromaticity, *Limnology and Oceanography*, **46 (1)**, 2001, 38-48

Meiboom, S. and Gill, D.: Modified spin-echo method for measuring nuclear relaxation times, *The review of scientific instruments*, **29**, 1958, 688-691

Micklander, E., Peshlov, B., Purslow, P. P., and Engelsen, S. B.: NMR-cooking: monitoring the changes in meat during cooking by low-field 1H-NMR, *Trends in Food Science & Technology*, **13**, 2002, 341-346

Moberg, L., Robertsson, G., and Karlberg, B.: Spectrofluorimetric determination of chlorophylls and pheopigments using parallel factor analysis, *Talanta*, **54**, 2001, 161-170

Møller, J. K. S., Parolari, G., Gabba, L., Christensen, J., and Skibsted, L. H.: Monitoring chemical changes of dry-cured parma ham during processing by surface autofluorescence spectroscopy, *Journal of Agricultural and Food Chemistry*, **51 (5)**, 2003, 1224-1230

Moshou, D., Strasser, R., Wahlen, S., Schenk, A., and Ramon, H.: Apple mealiness detection using fluorescence and self-organising maps, *Computers and Electronics in Agriculture*, **40 (1-3)**, 2003, 103-114

Munck, L., Nørgaard, L., Engelsen, S. B., Bro, R., and Andersson, C. A.: Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance, *Chemometrics and Intelligent Laboratory Systems*, **44 (1-2)**, 1998, 31-60

Nordon, A., Meunier, C., Carr, R. H., Gemperline, P. J., and Littlejohn, D.: Determination of the ethylene oxide content of polyether polyols by low-field 1H nuclear magnetic resonance spectrometry, *Analytica Chimica Acta*, **472 (1-2)**, 2002, 133-140

Nørgaard, L. and Ridder, C.: Rank annihilation factor analysis applied to flow injection analysis with photodiode-array detection, *Chemometrics and Intelligent Laboratory Systems*, **23 (1)**, 1994, 107-114

Pedersen, H. T., Bro, R., and Engelsen, S. B.: SLICING - A novel approach for unique deconvolution of NMR relaxation decays, *Magnetic Resonance in Food Science: A view to the Future*, Ed.: Webb, G. A., Belton, P. S., Gill, A. M., and Delgadillo, I., Royal Society of Chemistry, Cambridge, 2001, 202-209

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

Pedersen, H. T., Munck, L., and Engelsen, S. B.: Low-field H-1 nuclear magnetic resonance and chemometrics combined for simultaneous determination of water, oil, and protein contents in oilseeds, *Journal of the American Oil Chemists Society*, **77 (10)**, 2000, 1069-1076

Plugge, W. and van der Vlies, C.: Near-infrared spectroscopy as an alternative to assess compliance of ampicillin trihydrate with compendial specifications, *Journal of Pharmaceutical and Biomedical Analysis*, **11 (6)**, 1993, 435-442

Provencher, S. W.: A constrained regularization method for inverting data represented by linear algebraic or integral equations, *Computer Physics Communications*, **27**, 1982a, 213-227

Provencher, S. W.: CONTIN: A general purpose constrained regularization program fro inverting noisy linear algebraic and integral equations, *Computer Physics Communications*, **27**, 1982b, 229-242

Rathore, D. P. S. and Kumar, M.: Analytical applications of a differential technique in laser-induced fluorimetry: accurate and precise determination of uranium in concentrates and for designing microchemielectronic devices for on-line determination in processing industries, *Talanta*, **62 (2)**, 2004, 343-349

Reis, M. M., Gurden, S. P., Smilde, A. K., and Ferreira, M. M. C.: Calibration and detailed analysis of second-order flow injection analysis data with rank overlap, *Analytica Chimica Acta*, **422 (1)**, 2000, 21-36

Rodriguez-Cuesta, M. J., Boque, R., Rius, F. X., Picon Zamora, D., Martinez Galera, M., and Garrido Frenich, A.: Determination of carbendazim, fuberidazole and thiabendazole by three-dimensional excitation–emission matrix fluorescence and parallel factor analysis, *Analytica Chimica Acta*, **491 (1)**, 2003, 47-56

Romão, C. V., Luoro, R., Timkovich, R., Lübben, M., Liu, M.-Y., LeGall, J., Xavier, A. V., and Texeira, M.: Iron-coproporphyrin III is a natural cofactor in bacterioferritin from the anaerobic bacterium Desulfovibrio desulfuricans, *FEBS Letters*, **480 (2-3)**, 2000, 213-216

Rutledge, D. N.: Low resolution pulse nuclear magnetic resonance (LRP-NMR), *Analysis Magazine*, **20 (3)**, 1992, 58-62

Rutledge, D. N.: Characterisation of water in agro-food products by time domain-NMR, *Food Control*, **12**, 2001, 437-445

Sanchez, E. and Kowalski, B. R.: Generalized Rank Annihilation Factor-Analysis, *Analytical Chemistry*, **58 (2)**, 1986, 496-499

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

Saurina, J. and Hernández-Cassou, S.: Quantitative determinations in conventional flow injection analysis based on different chemometric calibration statregies: a review, *Analytica Chimica Acta*, **438 (1-2)**, 2001, 335-352

Sharma, S., Casanova, F., Wache, W., Segre, A., and Blümich, B.: Analysis of historical porous building materials by the NMR-MOUSE, *Magnetic Resonance Imaging*, **21 (3-4)**, 2003, 249-255

Silverman, D. C.: Corrosion prediction in complex environments using electrochemical impedance spectroscopy, *Electrochimica Acta*, **38 (14)**, 1993, 2075-2078

Skoog, D. A. and Leary, J. J.: Principles of instrumental analysis, 4 ed., Saunders College Publishing, Orlando, 1992, 700 pages

Stedmon, C. A., Markager, S., and Bro, R.: Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy, *Marine Chemistry*, **82 (3-4)**, 2003, 239-254

Trevisan, M. G. and Poppi, R. J.: Determination of doxorubicin in human plasma by excitation-emission matrix fluorescence and multi-way analysis, *Analytica Chimica Acta*, **493 (1)**, 2003, 69-81

Wehrens, R., Putter, H., and Buydens, L. M. C.: The bootstrap: a tutorial, *Chemometrics and Intelligent Laboratory Systems*, **54 (1)**, 2000, 35-52

Wentzell, P. D., Nair, S. S., and Guy, R. D.: Three-way analysis of fluorescence spectra of polycyclic aromatic hydrocarbons with quenching by nitromethane, *Analytical Chemistry*, **73 (7)**, 2001, 1408-1415

Williams, D. H. and Fleming, I.: Spectroscopic methods in organic chemistry, 5 ed., McGraw Hill, Cambridge, 1995, 329 pages

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

Zhen-Yi, Y., McCarthy, M. J., Klemann, L., Otterburn, M. S., and Finley, J.: NMR applications in complex food systems, *Magnetic Resonance Imaging*, **14 (7-8)**, 1996, 979-981

Paper I

Rinnan, Å., Bro, R.

Determining the number of components in a PARAFAC decomposition of fluorescence landscapes, In preparation

Determining the number of components in a PARAFAC decomposition of fluorescence landscapes

Åsmund Rinnan, Rasmus Bro

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

1 Abstract

2 Introduction

Since the beginning of chemometrics there has always been a discussion of how to decide the number of components to use in an analysis. Tools like explained variance plots, visual inspection of loadings, cross validation, etc, are all tools used to a large extent today. These tools are also used in the estimation of the number of components for a set of samples from fluorescence spectroscopy. There are, however, no clear rules of when the correct number of factors is reached, and each analyst interprets the results differently, and the conclusion may vary for the same data set. Also the diagnostic tools used to estimate the number of factors is to a degree a subjective choice, dependent upon what tool(s) the analyst trust the most. It is therefore of interest to find an automatic way of estimating the right number of components in chemometrics. This work focuses on the estimating of number of components in a PARAFAC model [Carrol and Chang 1970, Harshman 1970, Bro 1997] applied on a set of fluorescence landscapes.

In this work a novel automatic method to estimate the correct number of components in a PARAFAC model of a set of fluorescence landscapes is introduced. Several diagnostic tools are used in a combination to establish a robust method. All diagnostic tools fail to estimate the right number of components to a certain degree, but a combination may lead to a good estimate. Some diagnostic tools are better than others in determining the number of components, and therefore

these should be weighted as more important than the diagnostic tools that are uncertain in the estimation. Further, some diagnostic tools may be of such bad quality that they can be omitted altogether. In building an automatic model, the diagnostic tools used should all be digitized, meaning that each diagnostic tool should give a one number output. Thus upon finding the correct number of components, each PARAFAC model (of increasing complexity) gives a set of numbers corresponding to the diagnostic tools used. By analyzing these data, a good estimate should be found. Methods applied to these data can be PLS, fuzzy logic and setting limit values for each diagnostic tool.

3 Method

3.1 Data

Twelve data sets containing ten samples each, with three to five fluorophores in each data set was used as the basic data set. These 120 samples are taken from a bigger data set of 405 samples with the fluorophores: catechol, hydroquinone, indole, resorcinol, tryptophane and tyrosine. The six fluorophores are highly overlapping as can be seen from Figure 1.

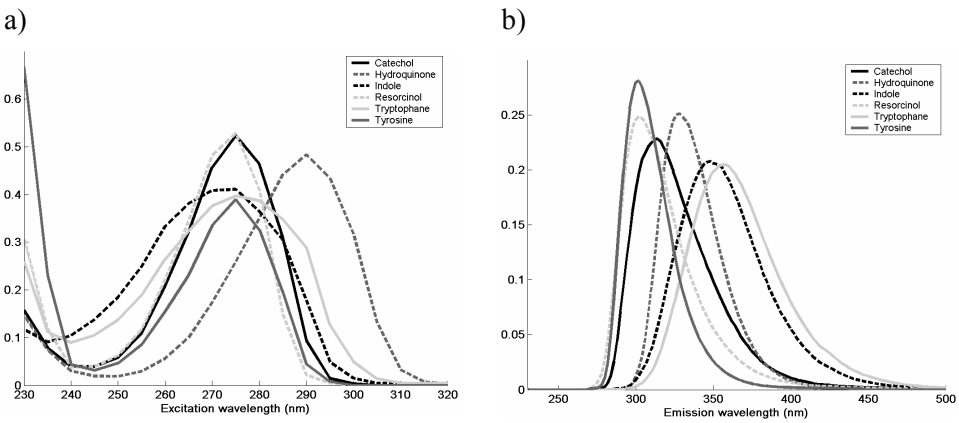


Figure 1: a) Excitation and b) emission spectra of the six fluorophores.

All the fluorophores were dissolved in de-ionized water, which was also used to dilute the final samples to the wanted concentration. The prepared samples were measured by a Varian Eclipse Fluorescence Spectrometer. The settings for the instrument were: Slit widths 5nm (for both excitation and emission), Emission wavelengths 230-500nm (recorded every 2nm), Excitation wavelengths 230-320 (recorded every 5nm) and scan rate 1920 nm/min. A PMT Detector voltage of 600V was used. The samples were excited with lowest energy (highest excitation wavelength) first, and then up to the highest energy excitation. Every sample was left in the instrument for a total of five replicate scans. The total recording time for one sample was approximately 15 min. In this paper, only the first replicate was used. In Table 1 there is an overview over the fluorophores present in the 12 data sets.

Table 1: Fluorophores present in each of the data sets.

Data set	Cat.	Hyd.	Ind.	Res.	Try.	Tyr.
1		×	×	×		×
2		×	×	×		
3			×	×	×	×
4		×	×	×		
5	×	×	×			×
6		×		×		×
7		×	×	×	×	
8	×	×	×		×	
9		×	×		×	×
10	×	×	×			
11		×		×		×
12		×	×	×	×	×

3.2 Diagnostic tools

81 diagnostic tools were used to estimate the number of components. These diagnostic tools each give a number to explain e.g. smoothness of loadings, degeneracy, uniqueness, and explained variance. They are calculated from PARAFAC-models, Jack-Knifed results [Martens and Martens 2001], split-half analysis [Harshman and de Sarboe 1994], PCA [Wold *et al.* 1987], and Tucker3 [Kroonenberg 1989] to mention some of the methods used. For a total list of all the parameters see Appendix I.

All the diagnostic tools were calculated from 1 factor and up to three more than the number of fluorophores the specific data set is built up of. I.e. in data set 1 there are 4 fluorophores, therefore the diagnostic tools are calculated from 1 and up to 7 factors. The diagnostic tools were studied in order to find possible limits (either upper or lower limits, depending on how it behaved) which could be used to define the number of factors that diagnostic tool estimates. All the diagnostic tools were investigated separately, and the "good" diagnostic tools – able to estimate correct of at least half of the data set – were kept and used in a final estimate.

4 Results

The first investigations indicated that of the 81 diagnostic tools, only 24 of them could give good estimates of the number of components. Of these 24, three are directly connected to PCA, two on the ratio between PCA and PARAFAC, and the remaining 19 from PARAFAC. The three PCA diagnostic tools are: relative eigenvalue, structure of P and maximum decrease in sample residual from jack-knifed PCA. The two ratios between PCA and PARAFAC are: ratio of residual from PCA with the residual from PARAFAC and the minimum ratio of eigenvalues from PCA with the norms from the A-scores in PARAFAC. The 19 diagnostic tools from PARAFAC is as follows: summed change in residual from jack-knifed PARAFAC, structure of B and C, minimum congruence between the modeled landscapes, median and minimum core consistency from 10 models, the

relative decrease in error, maximum percent negative values in B, C and the whole modeled cube, amount of modeled signal below 1st order Rayleigh scatter, number of unique error terms from 10 models, structure of B and C from a one factor PARAFAC model on the residual from PARAFAC, amount of modeled signal below 1st order Rayleigh scatter from a one factor PARAFAC model on the residual from PARAFAC, the structure in leverage in the B and C and the minimum equality in B and C with the known preset loadings from a PARAFAC model where two pre-known samples are added. (The structure used is the S_3 in Appendix I.)

Twelve tables containing only zeros (one for each data set) were made for these results, with the 24 diagnostic tools in the columns and the number of factors down the rows of the table. The diagnostic tools were then studied one by one, using the limits found in the first investigation. The number of factors a diagnostic tool estimated was set to 1 in the table. Subsequently each row was summed, and the maximum sum was found, see Table 2. If more than 50% of the parameters agreed the estimate was defined as a good estimate for the number of factors.

From Table 2 it can be seen that 9 out of the 12 data sets give the right estimate for the number of factors. There are, however, three that indicate two possible estimates. By investigating data set 5, 8 and 12, it becomes clear that these are the only three data sets containing catechol. This was investigated further by applying PARAFAC to a data set containing two samples: one pure catechol sample and one pure water sample. It was run with both one and two factors. Both the one factor and the two factor model converged fast, indicating that the two factor model was not overfitting the data. One of the spectra (Figure 2) equals the spectra of catechol, but the other is an unknown spectra. The score values for both of these factors were very low for the sample containing only water, while higher for the pure catechol sample. The factor most similar to catechol had the highest score value, while the

Table 2: Summing the results for the 24 diagnostic tools along the rows for each of the 12 data sets. Bold numbers indicate 50% or more of the diagnostic tools agree. Bold italic numbers indicate the data sets where no clear answer can be given.

		Number of factors							
Data set	True	1	2	3	4	5	6	7	8
1	4	1	0	7	**12**	3	0	0	-
2	3	0	1	**20**	0	2	0	-	-
3	4	1	0	2	**16**	3	1	0	-
4	3	0	2	**20**	0	2	0	-	-
5	4	3	1	3	***8***	***5***	2	0	-
6	3	1	2	**17**	2	1	0	-	-
7	4	0	3	2	**16**	2	1	0	-
8	4	1	3	2	***6***	***10***	0	1	-
9	4	0	1	1	**17**	2	2	0	-
10	3	0	2	***9***	***8***	4	0	-	-
11	3	2	1	**15**	4	2	0	-	-
12	5	1	1	0	3	**17**	2	0	0

other was in the magnitude of 10% of this. This explains the methods inability to establish the right number of factors for these three data sets, it really is half a factor higher than the given true value.

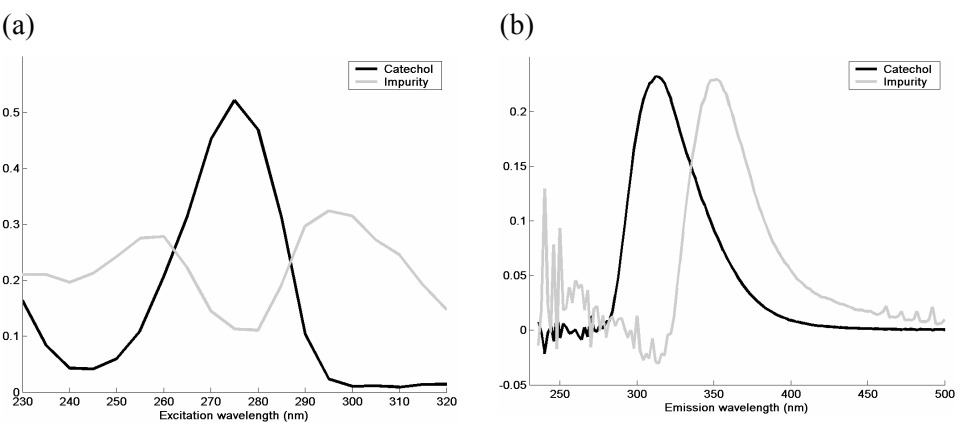


Figure 2: Two factor PARAFAC solution on Catechol and water.

5 Conclusion

By using 24 diagnostic tools it is possible to estimate the correct number of components in a set of fluorescence landscapes. The model used so far is a simple method, and more work should be performed to build a better, more refined and more robust model.

6 Acknowledgement

Rinnan wish to thank the STVF (Danish Research Council) for financial support through project 1179.

7 References

Bro, R.: PARAFAC. Tutorial and applications, *Chemometrics and Intelligent Laboratory Systems*, **38 (2)**, 1997, 149-171

Bro, R. and Kiers, H. A. L.: A new efficient method for determining the number of components in PARAFAC models, *Journal of Chemometrics*, **17 (5)**, 2003, 274-286

Carrol, J. D. and Chang, J.-J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition, *Psychometrika*, **35 (3)**, 1970, 283-319

Harshman, R. A.: Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis, *UCLA working papers in phonetics*, **16**, 1970, 1-84

Harshman, R. A. and de Sarboe, W. S.: An application of PARAFAC to a small sample problem, demonstrating preprocessing, orthogonality constraints and split-half diagnostic techniques, *Research methods for multimode data analysis*, Ed.: Law, H. G., Snyder, C. W., Hattie, J. A., and McDonald, R. P., Praeger, New York, 1994, 602-642

Kroonenberg, P. M.: Three-mode principal component analysis, DSWO Press, Leiden, Holland, 1989, 398 pages

Martens, H. and Martens, M.: Multivariate analysis of quality - An introduction, John Wiley, Chichester, 2001, 445 pages

Shen, H., Stordrange, L., Manne, R., Kvalheim, O. M., and Liang, Y. Z.: The morphological score and its application to chemical rank determination, *Chemometrics and Intelligent Laboratory Systems*, **51**, 2000, 37-47

Wang, J.-H., Liang, Y.-Z., Jiang, J.-H., and Yu, R.-Q.: Local chemical rank estimation of two-way data in the presence of heteroscedastic noise: A morphological approach, *Chemometrics and Intelligent Laboratory Systems*, **32 (2)**, 1996, 265-272

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

8 Appendix

8.1.1 Structure in loadings

The structure in the loadings is given by 5 different numbers. All numbers are based on the first derivative of the original loadings. The derivative method used is very crude, and is simply given by the following:

$$d\mathrm{x}_i = \mathrm{x}_{i+1} - \mathrm{x}_i$$

where dx_i is the estimate of the first derivative of the loading at point *i*, and *x* is the original data.. For the three first parameters, they are also based on a smoothing of this derivation spectra, by:

$$sd\mathrm{x}_i = \frac{\sum_{n=i-1}^{i+1} d\mathrm{x}_n}{3}$$

sdx_i is the smoothed first derivative at point *i*.

The first number is based on a jack-knifed congruence between d**x** and sd**x**. The jack-knifing is performed by removing one and one variable in the loading and calculating the congruence between the remaining normalized variables. The first structural parameter is given as the standard deviation of these jack-knifed congruences divided by the mean congruence. Congruence between two vectors is defined as:

$$cong_{\mathbf{a},\mathbf{b}} = \mathbf{a}_N{}^T \mathbf{b}_N$$

where the subscript *N* denotes that the vectors are normalized. When the structure in the loading decreases, the standard deviation of the congruence increases while the mean value decreases.

The second parameter is given by:

$$S_2 = \text{stdev}(d\mathbf{x} - sd\mathbf{x})$$

The third parameter is a normalization of S_2:

$$S_3 = \frac{\text{stdev}(d\mathbf{x} - sd\mathbf{x})}{\text{stdev}(d\mathbf{x})}$$

As for the first parameters, these numbers will also increase with decreasing structure.

The two last numbers are the so-called morphological factor (MF) [Wang *et al.* 1996] and the morphological score (MS) [Shen *et al.* 2000], and are discussed in these references and will not be discussed further here. Both of them decrease as the structure in the loadings decreases.

These five numbers are calculated on the loadings from PCA, PARAFAC, leverage of the PARAFAC loadings, one factor PARAFAC model on the residuals and the new loading from a *f*+1 factor PARAFAC model fixing the *f* first factors. For a *f* factor PARAFAC model, there are as many sets of structure parameters as there are factors in the model. The diagnostic tools recorded are then the maximum value of each of these five parameters. On all the PARAFAC models both B and C loadings are investigated separately. This adds up to a total of 45 diagnostic tools.

8.1.2 Relative eigenvalue

$$\lambda_{r,i} = \frac{\lambda_i}{\lambda_1}$$

where λ_i is the *i*th eigenvalue from a PCA. When there is no information left to explain the relative eigenvalue should drop. Only used on PCA; 1 diagnostic tool.

8.1.3 Jack-Knifing

The data has both been jack-knifed [Martens and Martens 2001] by PCA and by PARAFAC. The result from one jack-knifed PCA is given as the maximum decrease in the residual per factor, while for PARAFAC this is given as the summed change in the residual per factor.
Used on PCA and PARAFAC; 2 diagnostic tools.

8.1.4 Split-half

Data set is split into two halves, and a PARAFAC with equal complexity is computed on both of the two halves. The mean congruence of the similar loadings is used.
Used on PARAFAC, B and C loadings; 2 diagnostic tools.

8.1.5 Congruence between landscapes

This is only defined if the number of components is larger than 1. The congruence between two landscapes is found by multiplying the B and C loadings pair-wise, and subsequently vectorizing the landscapes. These landscapes are then normalized, and the congruence between the different vectorized landscapes is calculated.
Both the minimum and the maximum congruence between landscapes are used; 2 diagnostic tools.

8.1.6 Ten different PARAFAC models

Ten PARAFAC models from different starting points are calculated. The core consistency, number of iterations and the residual from these ten models are found. Core consistency is defined by Bro and Kiers [Bro and Kiers 2003]. The median and the minimum of these ten values are used, 2 parameters. When the congruence drops too many factors are extracted. The minimum number of iterations is found. The relative increase is then found as:

$$\Delta it_{r,i} = \frac{it_{i+1} - it_i}{it_{i+1}}$$

When PARAFAC starts having convergence problems, the number of iterations increases, indicating too many factors.

The amount of unique residuals is found as the number of residuals varying by more than 0.1%. A model with only one unique residual indicated that the result found is the grand minimum. This gives a total of 4 diagnostic tools.

8.1.7 Decrease in relative error

$$\Delta SSQ_{r,i} = \frac{SSQ_i - SSQ_{i+1}}{SSQ_i}$$

where SSQ_i is the sum of the squared error for the ith model. The error decreases with increasing components. It is only calculated from the PARAFAC models; 1 diagnostic tool.

8.1.8 PCA on loadings

A PCA is performed on the loadings from a PARAFAC model. The condition number is then found. Used on both B and C loadings; 2 diagnostic tools.

8.1.9 Percentage of negative values in loadings

$$\%neg = \frac{\sum_{i=1}^{N} d_i}{N}, \quad \begin{matrix} x_i < 0 \Rightarrow d_i = 1 \\ x_i \geq 0 \Rightarrow d_i = 0 \end{matrix}$$

where x_i is the loading value at point i, d_i is defined as shown above, and N is the total length of the loading. This is also calculated from the total modeled cube for each of the factors in the PARAFAC model, thus $d_i = 1$ if $a_{ij}b_{if}c_{if} < 0$ and $d_i = 0$ otherwise. N is then equal to IJK. This number should in general be small as fluorescence data per definition should be positive. For the PARAFAC model only the maximum value is recorded, both for the loadings, and for the modeled cube.

This descriptor is also calculated for the one factor PARAFAC model on the residual. Total number of diagnostic tools is 6.

8.1.10 Explained variance

$$\exp lained\ \mathrm{var}iance = \frac{SSQ_i}{SSQ_0}\%$$

where SSQ_i is the sum of squared model and SSQ_0 is the sum of squared raw data. Should flatten out when the right number of factors is reached. Computed on the PARAFAC models; 1 diagnostic tool.

8.1.11 PARAFAC factor based

$$cube\Sigma_r = \frac{\min_f \left(\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} a_{if} b_{jf} c_{kf} \right)}{\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} a_{i,1} b_{j,1} c_{k,1}}$$

where a_{if} is the ith score value, b_{jf} the jth **B**-loading value, c_{kf} the kth **C**-loading vector, all of factor f. The ",1" in the denominator refers to the 1 factor PARAFAC model on the raw data. This sum should always be positive, because of the nature of the data.

$$bad : good = \frac{\max_f \left(\sum bad_f \Big/ \sum good_f \right)}{\sum below_{raw} \Big/ \sum data_{raw}}$$

where bad_f is the sum of the f factor of the model lying below the 1st order Rayleigh scatter, $good_f$ the sum of the f factor of model above the 1st order Rayleigh scatter, $below_{raw}$ is the number of points below the 1st order Rayleigh scatter, and $data_{raw}$ is the amount of data points above the 1st order Rayleigh scatter. Should be low; no artifacts wanted.

These diagnostic tools are also computed for the one extra factor PARAFAC model on the residual, only then no minimum or maximum is computed (the number of factors is 1). For the cubeΣ_r for the one factor PARAFAC on the residual the denominator is set to 1.

This adds up to a total of 4 diagnostic tools.

8.1.12 Based on the residual

$$norm: Frobenius = \frac{\|\mathbf{a}\|}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} r_{ijk}^{\ 2}}$$

where ||a|| is the 2nd norm of the A-score from the one extra factor PARAFAC model on the residual, and r_{ijk} is the residual at point (i,j,k). Number of factors decided upon leveling of this number.

$$explained\ residual\ variance = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(a_i b_j c_k\right)^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} r_{ijk}^{\ 2}}$$

where notations are as shown above. Should be positive and less than one.

A [2 2 2] factor Tucker3 model is computed on the residual. The range of the core is computed. The relative decrease in the range is reported as:

$$\Delta range(G)_r = \frac{range(G_i) - range(G_{i+1})}{range(G_i)}, range(G_i) = \max(G_i) - \min(G_i)$$

where G_i is the core-array from the Tucker3 model. Should be high.

A t-test is performed on the vectorized difference in the residual from a f and f+1 factor model. If the null-hypothesis of the residual being equal is true, no more factors should be computed.

$$\Sigma r_r = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} r_{ijk}}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk}}$$

where r_{ijk} and x_{ijk} is the residual and value of raw data in the point (i,j,k). This gives the relative bias left to model.

$$\Sigma|r|_r = \frac{\sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} r_{ijk}^{\ 2}}}{\sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk}^{\ 2}}}$$

where the notation is as above. The residual left to model.

This gives a total of 6 diagnostic tools.

8.1.13 Ratios between PCA and PARAFAC

$$res(JK-PCA) : res(JK-PARAFAC) = \frac{\sum_{n=1}^{N} r_{JK-PCA,n}}{\sum_{n=1}^{N} r_{JK-PARAFAC,n}}$$

$r_{JK\text{-}PCA,n}$ is the residual of the nth sample from PCA, and $r_{JK\text{-}PARAFACni}$ is the residual of the nth sample from PARAFAC, both jack-knifed.

$$res(PCA) : res(PARAFAC) = \frac{SSQ_{PCA}}{SSQ_{PARAFAC}}$$

SSQ is the squared sum of squares, the subscript indicating the model used.

$$eig(PCA) : norm(PARAFAC) = \min_f \left(\frac{\|\mathbf{t}_f\|}{\|\mathbf{a}_f\|} \right)$$

||x|| denotes the norm of the number, $\boldsymbol{t}_f$ is the fth score from PCA and $\boldsymbol{a}_f$ is the fth score from PARAFAC. When these ratios flatten out the number of factors is reached. This gives a total of three diagnostic tools.

8.1.14 Preset sample

Two samples of two different known spectra are added to the rest of the dataset. A PARAFAC model with two extra factors is modeled and the known spectra are searched for. The congruence between the calculated and the known spectra is computed. The minimum congruence of the two known spectra is reported.

This is done both on the B and C loadings; 2 diagnostic tools.

Paper II

Rinnan, Å., Andersen, C. M.

Handling of first order Rayleigh scatter in PARAFAC modeling of fluorescence excitation-emission data, Submitted

Handling of First Order Rayleigh Scatter in PARAFAC Modeling of Fluorescence Excitation-Emission Data

Åsmund Rinnan and Charlotte M. Andersen*

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

*Corresponding author: cma@kvl.dk, +45 35 28 32 45

1 Abstract

The use of fluorescence spectroscopy has increased in use in recent years, mainly due to high sensitivity towards organic compounds typically present in e.g. food products and due to the instrumental development. The extension from measuring single emission or excitation spectra to so-called emission-excitation-matrices (EEM) has proven useful; increasing the selectivity between the components in the sample. However, EEMs can be rather complex and the analysis can become complicated due to interferences, scatter, overlapping signals, etc. This paper gives a comparison of several methods to handle the 1st order Rayleigh scatter when PARAFAC modeling is used to decompose the three-way fluorescence data. Furthermore, the paper provides suggestion for how to handle scatter in the modeling phase.

Keywords: Fluorescence, PARAFAC, Rayleigh scatter, weights, constraints, inserting zeros, missing values

2 Introduction

Fluorescence is a type of spectroscopy used in fields such as food science, analytical chemistry, biochemistry, environmental science and others. Fluorescence detects so-called fluorophores, molecules with a structure that allows emission of light when relaxing to the ground state from an excited singlet state. Such molecules include mainly aromatic compounds but also some carbonyls and molecules with highly conjugated double bonds. An advantage of the fluorescence technique is that it is capable of measuring concentrations down to one thousandth of what can be measured by normal absorption spectroscopy.

In fluorescence spectroscopy, the emission spectra are typically studied. A potentially more informative way to analyze the data is to study several emission spectra, taken at different excitation wavelengths. The data obtained this way can be seen as an excitation-emission matrix (EEM). Adding several matrices of equal excitation and emission wavelengths (EEMs of several samples) on top of each other makes up a cube of data (see Figure 1). Multi-way models, such as PARAFAC [1] can be used to study such data providing estimates of the spectra and concentration profiles of the underlying chemical analytes if the data are approximately tri-linear [2].

There are some parts of EEMs that may be problematic in the PARAFAC modeling because they may bias the estimated model parameters. The two most important are two types of scatter: Rayleigh and Raman, which show up as diagonal lines in the EEMs (see Figure 2). They both originate from the interaction between molecules in the solution and the incident light and do not contain information on the chemical properties of the sample. Rayleigh scatter is caused by molecules of the solute oscillating at the same frequency as the incident light, and thus emitting at the same wavelength as the incident light (1st order Rayleigh). It will also emit light at twice the wavelength of the incident light (2nd order Rayleigh), and also at higher multiples of the wavelength. While Rayleigh is

perfectly elastic, Raman is inelastic; the emitted light has less energy than the absorbed light. However, the energy loss for Raman is constant throughout the EEM for a given solute, e.g. for water the distance between the 1st order Rayleigh and the Raman scatter line is 3600 cm^{-1} [3].

The Rayleigh/Raman scatter signals may complicate the analysis of fluorescence data by PARAFAC since they do not conform to a low-rank trilinear model. Therefore, it is of interest to remove the influence of these scattering effects. Several methods to do this have been developed such as subtraction of a standard as discussed by Wentzell et al. (2001) [4], weighting [5-8], application of non-negativity and unimodality constraints [9, 10] and insertion of missing [11, 12] or zero values [13]. However, few comparisons between the different methods have been done, and in many publications only a few of the methods have been taken into account.

Some of the above mentioned methods can easily be combined without the loss of the strength of either method. This paper gives a comparison of many different methods to handle the scatter effects in PARAFAC modeling of three-way fluorescence data. The methods under investigations are combinations of the different weighting schemes or constraints, and insertion of zeros or missing values. Only the 1st order Rayleigh scatter is treated in the paper for simplicity. Further, the 1st order Rayleigh scatter has a larger influence on the PARAFAC modeling than the other scattering effects, mainly due to magnitude. The ability of each method to remove Rayleigh scatter and retrieve the pure components is evaluated. Three data sets are used: two consisting of designed data containing fluorescence measurements of solutions with known fluorophores and one consisting of fluorescence measurements of sugar samples taken from a production process.

3 Materials and methods

3.1 Data

3.1.1 Data set 1

Fifteen samples of mixtures of five fluorophores were recorded, with three or four fluorophores present in each sample. The spectra of these fluorophores are highly overlapping. The fluorophores are: catechol, hydroquinone, indole, tryptophane and tyrosine. The concentrations are shown in Table 1. All the fluorophores were dissolved in de-ionized water, which was also used to dilute the final samples to the wanted concentration.

The prepared samples were measured by a Varian Eclipse Fluorescence Spectrometer. The settings for the instrument were: Slit widths 5nm (for both excitation and emission), Emission wavelengths 230-500nm (recorded every 2nm), Excitation wavelengths 230-320 (recorded every 5nm) and scan rate 1920 nm/min. A PMT Detector voltage of 600V was used. The samples were excited with lowest energy (highest excitation wavelength) first, and then up to the highest energy excitation. Every sample was left in the instrument for a total of five replicate scans. The total recording time for one sample was approximately 15 min.

3.1.2 Data set 2

This data set contains EEMs of 16 samples containing different concentrations of four fluorophores with rather similar spectral properties [14]. The four compounds are: phenylalanine, 3,4-dihydroxyphenylalanine (DOPA), 1,4-dihydroxybenzene and tryptophan. The concentrations of the fluorophores are shown in Table 2.

The measurements were performed on a Perkin Elmer LS50 B fluorescence spectrofluorometer with excitation wavelengths ranging between 200 and 315nm (recorded every 5nm) and emission wavelengths ranging from 250 to 459nm

(recorded every 1nm). Both excitation and emission slit widths were set to 5nm and the scan speed was 1500nm/min.

3.1.3 Data set 3

Sugar samples taken from the final unit operation from a sugar plant in Scandinavia were obtained as described by Bro (1999) [15]. In total 268 samples were obtained. They were dissolved in un-buffered water (2.25g/15mL).

Fluorescence was measured on a Perkin Elmer LS50 B fluorescence spectrofluorometer with emission ranging from 275 to 560nm (recorded every 0.5nm). Seven excitation wavelengths were used. These were 230, 240, 255, 290, 305, 325 and 340nm.

3.2 Data analysis

The data were arranged in an $I \times J \times K$ three-way array where the first index (I) refers to the samples, the second (J) to the emission wavelengths, and the third (K) to the excitation wavelengths. The PARAFAC model was used to model the data. It can be written as

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

where x_{ijk} is the intensity if the i-th sample at the j-th variable (emission mode) and at the k-th variable (excitation mode). a_{if}, b_{jf} and c_{kf} are parameters describing the importance of the samples/variables to each component. The residuals, e_{ijk}, contain the variation not captured by the model.

3.3 Data pretreatment

For data set 1 a standard containing only the solvent is measured. The measured data of this standard is subtracted from all the samples in order to remove or at least minimize the Raman scatter line, and possibly the Rayleigh scatter line. Rayleigh scatter is mainly caused by the solvent, but is also dependent upon the

light source, and the solutes in the solvent. Thus, the Rayleigh effect will be reduced by subtraction of a standard, but it might cause some values in the EEM to become negative.

For the two other data sets the Raman scatter line is assumed to be small compared to the signal, and thus not disturbing the decomposition significantly.

Emission around twice the excitation wavelength is influenced by 2^{nd} order Rayleigh scatter. Since this paper is only focussing on the removal of 1^{st} order Rayleigh scatter, all the emission wavelengths which are more than twice that of the first excitation wavelength are removed (i.e. above the 2^{nd} order Rayleigh scatter line). Some wavelengths slightly below this are also removed in order to ensure that all the 2^{nd} order Rayleigh is removed.

Data set 1 is reduced by removing all the emission wavelengths exceeding 440 nm; a total of 30 emission lines. The new dimension of the dataset is thus $15 \times 106 \times 19$.

For data set 2, excitations from 200 to 230 nm are removed in order to reduce the amount of noise and missing values. Emissions from 441 to 459 are removed to remove any 2^{nd} order Rayleigh scatter. Furthermore, only every second emission wavelength is used. This gives a three-way array of the size $16 \times 101 \times 18$.

Data set 3 is reduced by removing all the emission wavelengths above 440 nm. In addition only every second emission wavelength is used, thus reducing the dimensionality to 268 x 165 x 7.

3.3.1 Insertion of zeros/missing values

All the data in the EEM where the emission is below the excitation wavelength is normally set to missing values. This may amount to a considerable part of the EEM, which may slow convergence and may also lead to spurious results [16]. Furthermore, it is typical to insert missing values in the area covered by the scatter lines (especially the Rayleigh line, since this normally has a higher intensity than Raman) ([16] and see Figure 2). However, the scatter lines may be confounded with chemical information, and thus it is of interest to keep these areas. Furthermore, it might be difficult to accurately estimate the exact width of a Rayleigh peak.

There is no emission below the excitation wavelength because emission will always be of lower energy. Therefore, emission is theoretically zero below the excitation wavelength. However, most of these zero values do not conform to the trilinear model [16] and may give misleading results. This problem may be handled by inserting missing values for a band of emission wavelengths below excitation and zeros for all other emissions below excitation [13]. This is done in the present paper. The bandwidth of missing values is varied from 0 to twice the estimated width of the Rayleigh scatter line, which is found through visual inspection of the data. In this way, the PARAFAC model is constrained to aim for zero signal in the areas of very low emission wavelengths. In Figure 3 an extreme example shows how these zeros can effect the decomposition. When all lower emissions are set to missing values, the actual model of the signal of the sample is hidden in high model-values in the spurious part. This approach will also be beneficiary in the sense that the PARAFAC algorithm tends to converge faster.

3.4 Models

Insertion of missing values in the band of the EEM where one knows that the Rayleigh scatter is present is not always a good method since some of the data, which actually hold some information, are removed. In order to circumvent this problem weighted PARAFAC or a combination of missing values and constraints can be used.

A total of four methods are investigated:

- "MILES" weighting
- Hard weighting
- Soft weighting (only possible for data set 1)
- Insertion of missing values instead of the Rayleigh scatter and the use of constraints. The constraints used are non-negativity for the A-scores (sample mode) and the C-loadings (excitation mode), and uni-modality and non-negativity for the B-loadings (emission mode).

The four methods are explained more thoroughly in the next section.

3.4.1 Weighting

Here, three types of weighting are investigated. In two of the methods, the Rayleigh scatter area is down-weighted, while for the third weighting scheme an instrumentally based set of weights are used.

Hard weighting

JiJi and Booksh (2000) [7] compared different ways of assigning weights to PARAFAC in the decomposition of fluorescence data. Their conclusion was that what they termed hard weighting was the best weighting scheme of the four they compared. This kind of weighting essentially mimics the use of missing values by assigning binary 0 or 1 weights to each data element. In this paper a slightly different approach to hard weights are used, where the area containing Rayleigh scatter is assigned a weight of 0 while the rest of the EEM is assigned a weight of 1. I.e. even the area of very low emission wavelengths is assigned a weight of 1.

"MILES"

Bro et al. (2002) [5] gave an explanation of maximum likelihood fitting of PARAFAC models and at the same time gave an example of a new weighting scheme for fluorescence data, which in this article is called "MILES" weights.

Here the idea is that there is always some information in the whole dataset, but because of the Rayleigh scatter line, the relevant information is partly hidden, and therefore this area should have smaller weights than the area only containing signals from the analytes. However, it is known that the intensity of the Rayleigh scatter line decreases as you go from the exact diagonal. Thus the farther away from the diagonal one moves the higher the weight. This increase in weights follows a Gaussian curve. The weights are set to 1 in the triangle below the 1st order Rayleigh scatter line where zeros are inserted. This method is similar to the soft-weights JiJi and Booksh (2000) [7] used in their study. The difference is, while the method proposed by JiJi and Booksh uses the relative intensity of the signal as the basis of the weights, MILES uses the shape of a theoretical Rayleigh scatter line as the base for the weights. So, the focus in MILES is in down-weighting the areas with Rayleigh scatter, while the soft weights used by JiJi and Booksh focuses on up-weighting the areas containing large signals.

Soft weigthing

The third method uses "smoother" weights and is only applied on data set 1. Russel and Gouterman (1988) [8] showed that by using a theoretically based measuring uncertainty for the instrument, the decomposition was faster and more precise. The theoretical measuring uncertainty of the instruments used in this work is not known, and thus it is necessary to use a different measurement of the uncertainty. An approximation to the measurement uncertainty can be found by measuring several instrumental replicates. However, this will lead to the use of different weights, than if the theoretical measurement uncertainty is known. In this paper the samples of dataset 1 were left in the instrument for a total of five replicate scans. The standard deviation of these replicates is then used as an estimate of the uncertainty. These standard deviations will reflect the uncertainty of the instrument. The detector in a fluorescence instrument typically has heteroscedastic noise, and by down-weighting the areas with a high noise level the decomposition of the spectra may converge faster and give better estimates. The inverse of these

estimated uncertainties is used as weights in a weighted PARAFAC. A typical weighting scheme for this method is shown in Figure 4. The use of soft weights does not in any way try to weigh down the Rayleigh scatter. However, it weighs down those areas in the EEM where the uncertainty is high.

3.4.2 Non-negativity and uni-modality constraints

All spectroscopic data should be positive in the ideal case. Furthermore, it is common for the emission spectra of a single fluorophore to only contain one single peak, thus allowing for the use of uni-modality as a constraint. Often this is not the case for the excitation mode, where it is more common for the fluorophore to have several peaks. However, if only a limited amount of excitation wavelengths are recorded, as to not enter the dual peak area, uni-modality may also be used with success. This is not the case for the data sets investigated here. Thus uni-modality is only applied on the emission mode, while non-negativity is used for the excitation mode. Non-negativity constraint is also applied to the concentration profiles, since clearly no concentrations can be negative.

It was initially tested if this approach alone would lead to a better modeling of the data. The results from this initial analysis (see Figure 5) show that the use of constraints only works if the Rayleigh scatter line is removed in another way, and will thus only be used together with the insertion of missing values where the Rayleigh scatter peak is present. However, insertion of missing values may remove some of the chemical information close to the Rayleigh scatter peak (if present).

From Figure 5 it can be seen that two of the resolved components have their emission maxima at a lower wavelength than their excitation maxima. This is physically impossible, and thus the estimates obtained when applying constraints and no missing values are not good.

3.5 Varying the parameters

In all the models there are two parameters that need to be set before the model is calculated: The width of the Rayleigh peak and the width of missing values to be inserted below the 1st order Rayleigh line. In order to investigate the importance of choosing the right parameters, both of these are changed independently of each other. Both the width of the band of missing data below the 1st order Rayleigh scatter line, and the estimated width of the 1st order Rayleigh peak is varied from 0 to twice the estimated width of the Rayleigh peak (done visually), if possible. The Rayleigh peak is estimated to 10 nm for data set 1, 20 nm for data set 2 and 40 nm for data set 3. The parameters are increased by 2.5 nm, 5 nm and 10 nm, for the three data sets, respectively. For example the bands of missing values for data set 1 used in the investigations have the size 0, 2.5, 5, 7.5, etc, up to 20nm. In addition to this, models are made with missing values for all emissions below excitation. This sums up to a total of 90 different combinations of parameters for every model. However, for data set 3, the maximum width of missing values is 60 nm, and thus only 72 combinations are possible.

3.6 Quality of a model

The quality of a model is evaluated by the Q^2 between the model and the reference. This is calculated as follows:

1. First the best model of all the models for one data set is found and used as the reference model. This reference is found by visual inspection of the models. Peak position, smoothness, absence of scatter-influence and amount of negative values in the loading modes were used as criteria for choosing the reference. An example of a good and a not-so-good model is given in Figure 6. All other models are then compared to this model.
2. The mean of the Q^2 between the factors in the model and the reference is calculated – both B and C-loadings (emission and excitation modes) are taken into account. The loadings of the tested model and the reference are compared in order to find the best match. Q^2 is defined as follows:

$$Q_{m,f}{}^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

where y is one of the loadings of one factor under investigation, $\hat{y}$ is the reference loading for this factor, while $\bar{y}$ is the average of the loading under investigation. The m and f indexes on the Q^2 denote the mode and factor number investigated (m=2 indicates the B-loadings, m=3 the C-loadings). If $Q_{m,f}^2$ equals one, the two loadings are equal, if it is zero or less, then there is no similarity between the loading under investigation and the loadings of the reference model. If $Q_{m,f}^2$ is below zero, it is set to zero.

3. The average Q^2-value for all the factors and loadings (B and C) is calculated, giving a number between zero and one. E.g. for a four factor PARAFAC model, the average Q^2 would be the average of 4×2 + 8 different $Q_{m,f}^2$-values.

The criterion for a model being good changes from data set to data set, where the easiest data set requires better similarity between the model and the reference than the most difficult one. Further, for data set 3, no reference model is found, but rather two that both looked reasonable. Thus the Q^2-value is not only averaged over the factors and loadings, but also across the two references. Each model is evaluated based on its average Q^2-value, time before convergence and number of iterations.

3.7 Software

MATLAB (The Mathworks Inc, Natick, MA) version 6.5 was used during the calculations. The algorithms in use were from the N-way toolbox version 2.10 [17], and some in-house algorithms.

4 Results and Discussion

The results from one data set comprises of three (or four for data set 1) matrices with 90 (or 72 for data set 3) numbers in each (one matrix for each method). The 10×9 matrix is made up of all the possible parameter combinations. The rows are

given by increasing band of missing values, while the columns are the estimation of the width of the Rayleigh scatter peak. One of these matrices given as the quality (Q^2) is visualized in Figure 7.

The three data sets are of different origin, so different values of Q^2 are used to determine whether a model is good or not. These values are set to 0.995, 0.95 and 0.85 for the three data sets, respectively. Q^2-values lower than 0.85 refer to models that are very different from the reference, and thus a lower limit for data set 3 was not chosen. Table 3 gives an overview of the results upon the investigation of how well the different methods perform in the ranges of the parameters. Relative time is the average of time needed before convergence of all the good models, divided by the method with the lowest average time. This version of relative time is chosen instead of averaging or summing all the times before convergence, because the interest lies in the good models, and not in the bad ones. The relative number of iterations is calculated in the same manner.

Generally, using hard weights is the method that needs the least amount of time to converge. However, its decomposition is not as good as those made by MILES and constraints. The amount of time needed for one computation is less for MILES compared to constraints when the data are easier to model, while for the process data, constraints uses less time than MILES. The average number of iterations is lowest when applying constraints or using hard weights, but MILES only needs a few more iterations. The difference in the number of iterations needed before convergence is largest for the process data. The results for soft weights will be discussed under Data set 1.

4.1 Data set 1

For this data set it was also investigated how the models performed if only the Raman part of the standard was subtracted, instead of the whole standard. In general this gave worse results than the ones presented here, with 86%, 87%, 0%

and 73% good models (with a limit on Q^2 of 0.99) for MILES, hard, soft and constraints, respectively. This is a lot lower than for the results obtained when the whole standard is subtracted and, therefore, these results will not be discussed further.

Almost all the models are good for these data, most probably because the signal is nicely separated from the 1st order Rayleigh scatter line and the signal from the analytes is of a considerable higher intensity than the Rayleigh peak. The main difference between the different approaches is that by the use of missing values and constraints, some of the results are not optimal in that the loadings have an inferior appearance. However, the greatest difference is between the calculation times. The fastest method is for using the hard weights, while MILES is almost as fast. Soft-weights generally need almost four times the amount of time to converge. The same tendencies are seen for the number of iterations where application of soft weights uses approximately 5 times more iterations before the model converges. The use of constraints needs approximately 60% more time to converge than hard weights. Even though the calculation time and the number of iterations for soft weights disfavor this method, it compensates somewhat for this by the need of only estimating one of the meta-parameters; width of missing values. So while the three other methods have a total of 90 parameter combinations under investigation, soft weights only have 10. It is difficult with the results from this dataset to tell which of the methods are better or worse, other than the use of constraints can result in inaccurate estimates, and that MILES and hard weighting are the two fastest. The setting of the band of missing and the size of the Rayleigh scatter does not seem to influence the decomposition considerably.

4.2 Data set 2

For this data set the number of good models per method vary slightly, with MILES performing the best, giving good estimates for as many as 86% of all the parameter combinations. For the other two methods, this number is 70% for hard weighting and 74% for constraints. However, the average time for those parameter

combinations that give good estimates is shortest for hard weighting, followed by MILES needing 35% more time, while constraints in general needs 60% more time than hard weighting. The number of iterations is lowest for MILES with constraints using 4% more and hard weights using 15% more. In general, setting the Rayleigh width to 0 nm, gives bad results. Also setting the Rayleigh width to 5 nm and using a big bandwidth of missing values cause results of lower quality. MILES is the method that is least sensitive to the correct estimate of the Rayleigh width, while hard weights is the most sensitive. Further, in the extreme parameter settings, with both parameters close to the maximum, both hard weights and constraints fail to give good estimates, while MILES does not suffer the same problems.

4.3 Data set 3

This is the only real data set, and in that perspective also the most interesting data set in the investigation. For this data set the results vary more than in the previous ones. In general MILES is the best giving good estimates in 28% of all the parameter combinations. For constraints, this percentage is 24% and for hard weights as low as 11%. Thus, the number of good parameter estimates is lower for data set 3 than for the other data sets. An interesting factor here is that the time needed before convergence, is the exact opposite to the amount of good decompositions, thus hard weights is fastest, followed by constraints, and MILES being the slowest. However, the time before convergence is only 50% more for MILES than for hard weights. The number of iterations follows the same trend, with the lowest upon applying hard weights. Constraints need 29% more iterations, while MILES requires 56% as many iterations.

In order for MILES to work it is essential to at least use an estimate of the Rayleigh peak width that is equal to the one found during visual inspection. However, overestimating the Rayleigh peak does not seem to have any influence on the decomposition. This means that although the visual inspection of the data suggests a Rayleigh peak width of 40 nm, setting this as high as 80 nm, still give good

estimates of the loadings using MILES. Constraints perform similarly, but the bandwidth of missing values is more limited. While for MILES the band of missing values could be set as high as 30 nm, and sometimes even up to 50 nm, constraints only behave well, when this number is between 0 and 20 nm. Constraints on the other hand are not as sensitive to estimating the Rayleigh peak correctly, and give good predictions for Rayleigh peak estimates as low as 20 nm. In general the best decompositions is found when the band of missing values is set very low meaning that a large number of zeros are inserted into the EEM.

Visual inspection of the data shows that a good estimate of the two parameters would be about 40 nm for the width of the Rayleigh peak and since it seems like the signal from the fluorophores is far from the Rayleigh scatter line, a band width of missing values of 0 or 10 nm should be appropriate. However, since the Rayleigh scatter peak is higher than the actual signal from the fluorophores, a width of as much as 60 nm may be an appropriate estimate for MILES. The Q^2-values for these parameter settings are given in Table 4.

From Table 4 it can be seen that for the parameter combinations shown MILES and constraints both give five models with good estimates of the loadings, while hard weighting gives three good models. However, since the visual inspection of the data indicated a width of the Rayleigh peak of 40nm, using 60 nm in the hard weights (and possibly also for the constraints) is not appropriate, and thus hard weights performs well on three out of four settings, almost as good as the other methods.

5 Conclusion

Of the four methods under investigation it seems that using MILES gives the best results. In order to use MILES it is of great importance to estimate the Rayleigh peak correctly, especially in cases where the intensity of the Rayleigh peak is higher than for the analytes themselves. It is less sensitive to changes in the

parameters compared to hard weights and constraints. However, MILES requires more work on estimating the Rayleigh scatter peak and the exact weights to apply than for example constraints. Of the four methods under investigation, MILES and soft weights are the only methods that use all the information in the EEM. However, based upon the analyses performed here it is impossible to conclude anything substantial regarding the use of soft weights. More data sets of this type are required. Further, it would be interesting to combine the soft weights with one of the other weighting methods to see if that could both speed up the process (focusing only on areas in the EEM with low uncertainty), and handle the influence of the scatter. Constraints, which is really hard weights with constraints, gives better results than by using hard weights by itself. From this perspective it would be interesting to apply constraints to the other weighting schemes as well.

6 Acknowledgement

Åsmund Rinnan would like to thank SJVT for the financial support through project 1179. The work by Charlotte M. Andersen is part of the FØTEK programme supported by the Danish Dairy Research Foundation (Danish Dairy Board) and the Danish Government. The authors want to thank Rasmus Bro for good discussions and help during the work with this article.

7 References

1. Harshman, R. A. UCLA Working Pap. Phonetics, 1970, 16, 1-84.
2. Leurgans, S. E.; Ross, R. T. Stat. Sci., 1992, 7 (3), 289-319.
3. Lackowicz J.R. *Principles of fluorescence spectroscopy 2nd edition*; Kluwer Academic/Plenum Publishers: New York, 1999.
4. Wentzell, P.D.; Nair, S.S.; Guy, R.D. Anal. Chem., 2001, 73 (7), 1408-1415.
5. Bro, R.; Sidiropoulos, N. D.; Smilde, A. K. J. Chemom., 2002, 16 (8-10), 387-400.
6. JiJi, R. D.; Andersson, G. G.; Booksh, K. S. J. Chemom., 2000, 14 (3), 171-185.
7. JiJi, R. D.; Booksh, K. S. Anal. Chem., 2000, 72 (4), 718-725.
8. Russel, M. D.; Gouterman, M. J. Chemom., 1988, 44 (9), 857-861.
9. Bro, R.; de Jong, S. J. Chemom., 1997, 11, 393-401.
10. Bro, R.; Sidiropoulos, N. D. J. Chemom., 1998, 12, 223-247.
11. Bro, R. Chemom. Intell. Lab. Syst., 1997, 38 (2), 149-171.
12. Jiji, R. D.; Cooper, G. A.; Booksh, K. S. Anal. Chim. Acta, 1999, 397 (1-3), 61-72.
13. Thygesen L.G., Rinnan, Å., Barsberg, S. and Møller, J.K.S. Submitted to Chemom. Intell. Lab. Syst., 2003.
14. Baunsgaard D. The Royal Veterinary and Agricultural University, Internal report, 1999, http://www.models.kvl.dk/research/data/dorrit/dorrit.pdf
15. Bro, R. Chemom. Intell. Lab. Syst., 1999, 46 (2), 133-147.
16. Andersen, C.M.; Bro R. J. Chemom., 2003, 17, 200-215.
17. Andersson, C. A.;Bro R. Chemom. Intell. Lab. Syst., 2000, 52 (1), 1-4, http://www.models.kvl.dk/source

Table 1: Concentrations of the five fluorophores in the 15 samples of data set 1.

Sample no.	Catechol (10^{-6} M)	Hydroquinone (10^{-6} M)	Indole (10^{-6} M)	Tryptophane (10^{-6} M)	Tyrosine (10^{-6} M)
1	22	0	0.64	0	0.91
2	10.9	0	1.28	0	0.91
3	6.5	5.6	0.38	0	3.0
4	6.5	2.8	0	0.93	0.91
5	6.5	0	0.38	1.86	1.51
6	22	0	1.28	0	0
7	0	0	0.38	0.56	0
8	10.9	2.8	0	0	0
9	0	0	0.64	1.86	1.51
10	22	2.8	0	0.93	0
11	22	0	0.64	0	0.91
12	10.9	0	1.28	0	0.91
13	6.5	5.6	0.38	0	3.0
14	6.5	2.8	0	0.93	0.91
15	6.5	0	0.38	1.86	1.51

Table 2: Concentrations of the four fluorophores in the 16 samples of data set 2.

Sample no.	Hydroquinone (10^{-6} M)	Tryptophan (10^{-6} M)	Phenylalanine (10^{-6} M)	Dopa (10^{-6} M)
1	46	4	2800	18
2	17	2	4700	28
3	20	1	3200	8
4	10	4	3200	16
5	6	2	2800	28
6	3.5	1	350	20
7	3.5	0.5	175	20
8	3.5	0.25	700	10
9	1.75	4	1400	5
10	0.875	2	700	2.5
11	28	8	700	40
12	28	8	350	20
13	14	8	175	20
14	0.875	8	1400	2.5
15	1.75	8	700	5
16	3.5	2	700	80

Table 3: Amount of good estimates, relative time and relative number of iterations required per method.

	Method	1	2	3
% good	MILES weights	98.9	86.7	27.8
	Hard weights	100	70.0	11.1
	Soft weights	100	-	-
	Constraints	92.2	74.4	23.6
Rel. time	MILES weights	1.05	1.36	1.50
	Hard weights	1	1	1
	Soft weights	4.4	-	-
	Constraints	1.58	1.73	1.08
Rel. number of interations	MILES weights	1.25	1	1.56
	Hard weights	1.18	1.15	1
	Soft weights	5.2	-	-
	Constraints	1	1.04	1.29

Table 4: The Q^2-values for good estimates of the Rayleigh peak width and the band-width of missing values for data set 3.

		Method	Width of Rayleigh peak		
			40 nm	50 nm	60 nm
Width of band of missing values	0 nm	MILES	0.96	0.96	0.95
		Hard	0.65	0.96	0.46
		Constraints	0.95	0.95	0.94
	10 nm	MILES	0.19	0.95	0.94
		Hard	0.95	0.95	0.46
		Constraints	0.56	0.95	0.94

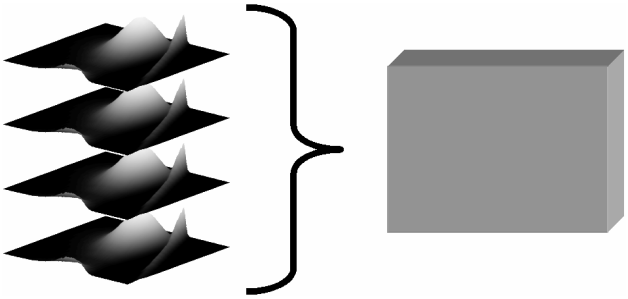


Figure 1: Going from a set of EEM's to a three-dimensional cube, ready to be handled by e.g. PARAFAC.

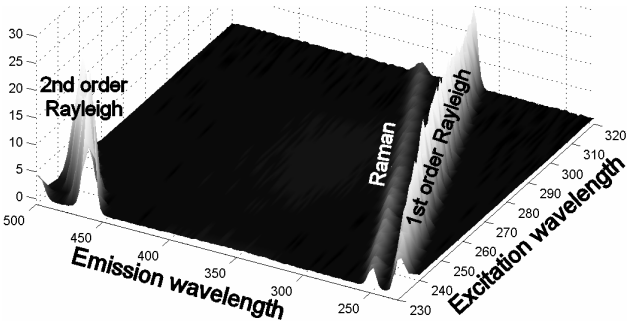


Figure 2: An EEM of a water sample. The scatter lines are clearly seen.

a) b)

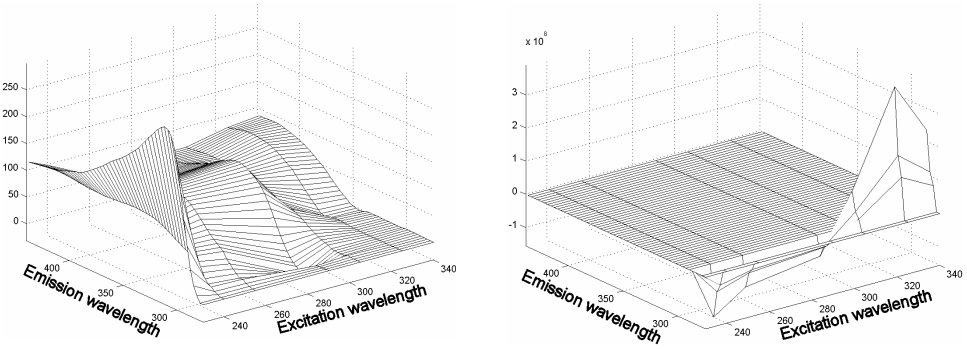


Figure 3: The predicted EEM of sample 1, data set 3, using hard weights. Inserting (a) only zeros – 454 iterations, or (b) only missing values, below the Rayleigh scatter line – 18118 iterations.

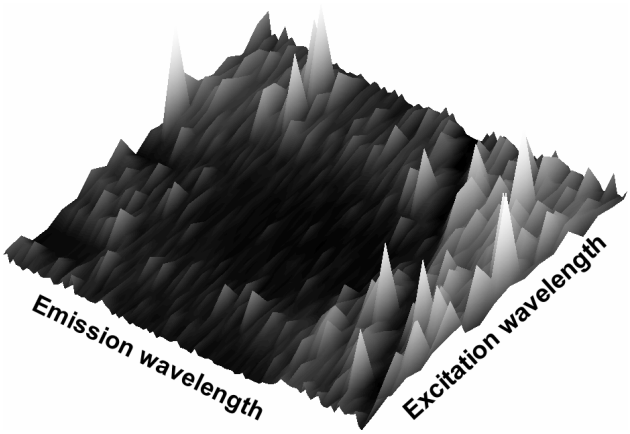


Figure 4: A typical set of soft weights calculated based on the inverse standard deviation of the instrumental replicates of one sample.

(a) (b)

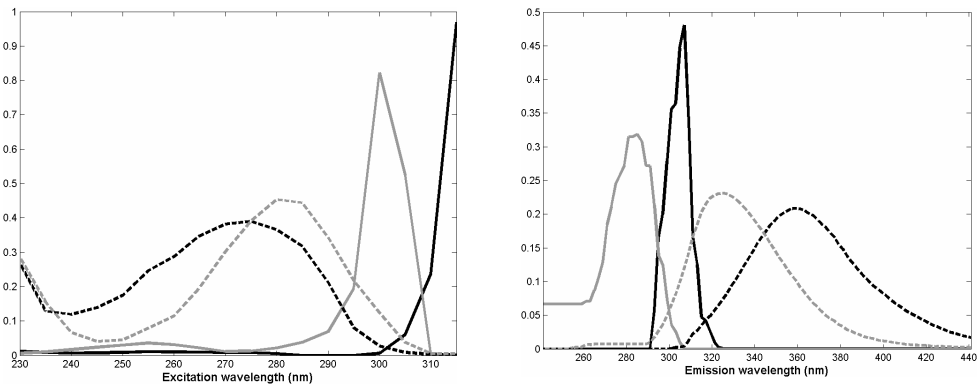


Figure 5: The excitation (a) and emission (b) loading when applying a non-negativity constraint on the A-scores and the C-loadings, while uni-modality on the B-loadings. No missing values are inserted. Data set 2.

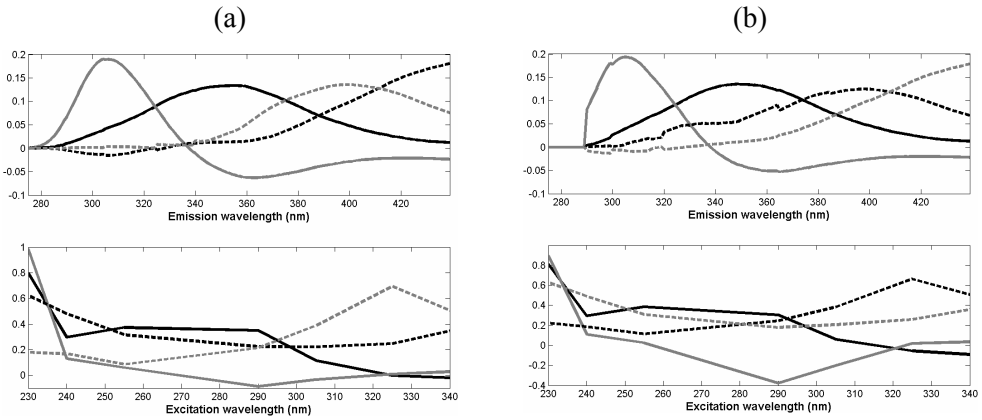


Figure 6: The loadings for a good (a), and a bad (b) model for dataset 3.

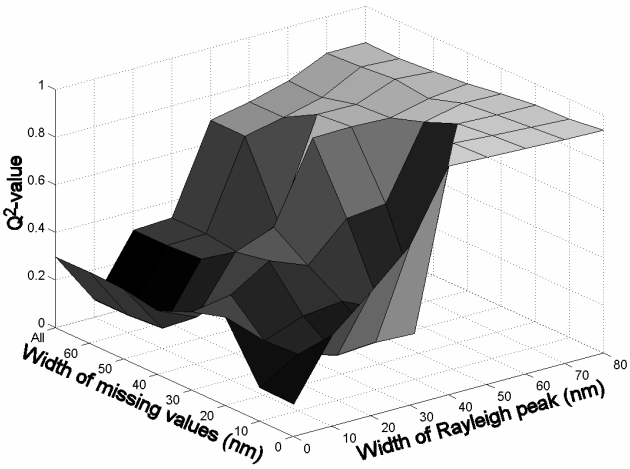


Figure 7: Visualization of the results from "MILES" on data set 3.

Paper III

Rinnan, Å., Booksh, K. S., Bro, R.

1st Order Rayleigh scatter as a Separate Component in PARAFAC Decomposition of Fluorescence Landscapes, In preparation

1st Order Rayleigh as a Separate Component in PARAFAC Decomposition of Fluorescence Landscapes

Åsmund Rinnan, Karl S. Booksh and Rasmus Bro

1 Abstract

2 Introduction

During the recent years, the use of fluorescence spectroscopy has increasingly focused on total luminescence spectra, or excitation-emission-matrices (EEM). The use of chemometrics, and more specifically parallel factor analysis (PARAFAC) [Carrol and Chang 1970, Harshman 1970, Bro 1997] has become recognized as a good and reliable tool for extracting chemical information from EEM spectra. By using PARAFAC to extract more information from the collected data, it is of vital importance that the data is low-rank tri-linear. This is the case for the behavior of all fluorophores in a sample, as long as no quenching or inner-filter effects are present. However, there are some light scatter effects in fluorescence that do not conform to the tri-linearity assumption. Instead these spectral features lie on a diagonal of the EMM landscape. The light scattering effects are called Rayleigh (1st and 2nd order are most common) and Raman scatter. The 1st order Rayleigh scatter line is centered at the emission equals excitation line, 2nd order Rayleigh scatter at the emission equals twice the excitation, and the Raman is at a certain energy difference from the 1st order Rayleigh scatter line. This energy difference is dependent upon the solute of the sample. 1st order Rayleigh scatter is the most intense of the light scattering, and therefore has a higher influence on the PARAFAC model.

If the signal from the fluorophores themselves lie away from these scatter lines, they can simply be omitted. However, in many food or environmental samples the signal from the fluorophores of interest lies close to, or on top of, one or more of the scatter effects. If these scatter effects are not accounted for in the modeling process they may cause errors in the decomposition of these spectra. Several ways of handling the scatter effects have been proposed in literature: subtraction of a standard [Ho *et al.* 1978, Ho *et al.* 1980, McKnight *et al.* 2001], inserting missing values [Christensen *et al.* 2003, Munck *et al.* 1998, Rodriguez-Cuesta *et al.* 2003, Trevisan and Poppi 2003], constraints in the PARAFAC decomposition [Andersen and Bro 2003, Bro 1999] and weighted PARAFAC [Bro *et al.* 2002, Jiji and Booksh 2000, Rinnan and Andersen 2004]. In this paper a novel way of treating the 1st order Rayleigh scatter is proposed, namely modeling it as a separate component in the decomposition step. This is done through reorganization of the data so that the 1st order Rayleigh scatter line is low-rank bi-linear. The following bi-linear Rayleigh scatter is either modeled by PCA (one model per sample) or PARAFAC (one model for all samples).

3 Materials and Methods

3.1 Data sets

For analyzing the ability of the new method, three different fluorescent data sets are used.

3.1.1 Data set 1

Fifteen mixtures of five fluorophores were recorded, with three or four fluorophores present in each sample. The spectra of these fluorophores are highly overlapping. The fluorophores are: catechol, hydroquinone, indole, tryptophane and tyrosine. The concentrations are shown in Table 1. All the fluorophores were dissolved in de-ionized water, which was also used to dilute the final samples to the wanted concentration.

Table 1: Concentrations of the five fluorophores in the 15 samples of data set 1.

Sample no.	Catechol (10^{-6} M)	Hydroquinone (10^{-6} M)	Indole (10^{-6} M)	Tryptophane (10^{-6} M)	Tyrosine (10^{-6} M)
1	22	0	0.64	0	0.91
2	10.9	0	1.28	0	0.91
3	6.5	5.6	0.38	0	3.0
4	6.5	2.8	0	0.93	0.91
5	6.5	0	0.38	1.86	1.51
6	22	0	1.28	0	0
7	0	0	0.38	0.56	0
8	10.9	2.8	0	0	0
9	0	0	0.64	1.86	1.51
10	22	2.8	0	0.93	0
11	22	0	0.64	0	0.91
12	10.9	0	1.28	0	0.91
13	6.5	5.6	0.38	0	3.0
14	6.5	2.8	0	0.93	0.91
15	6.5	0	0.38	1.86	1.51

The prepared samples were measured by a Varian Eclipse Fluorescence Spectrometer. The settings for the instrument were: Slit widths 5nm (for both excitation and emission), Emission wavelengths 230-500nm (recorded every 2nm), Excitation wavelengths 230-320 (recorded every 5nm) and scan rate 1920 nm/min. A PMT Detector voltage of 600V was used. The samples were excited with lowest energy (highest excitation wavelength) first, and then up to the highest energy excitation. Every sample was left in the instrument for a total of five replicate scans. The total recording time for one sample was approximately 15 min.

3.1.2 Data set 2

This data set contains EEMs of 16 samples containing different concentrations of four fluorophores with rather similar spectral properties [14]. The four compounds are: phenylalanine, 3,4-dihydroxyphenylalanine (DOPA), 1,4-dihydroxybenzene and tryptophan. The concentrations of the fluorophores are shown in Table 2.

Table 2: Concentrations of the four fluorophores in the 16 samples of data set 2.

Sample no.	Hydroquinone (10^{-6} M)	Tryptophan (10^{-6} M)	Phenylalanine (10^{-6} M)	DOPA (10^{-6} M)
1	46	4	2800	18
2	17	2	4700	28
3	20	1	3200	8
4	10	4	3200	16
5	6	2	2800	28
6	3.5	1	350	20
7	3.5	0.5	175	20
8	3.5	0.25	700	10
9	1.75	4	1400	5
10	0.875	2	700	2.5
11	28	8	700	40
12	28	8	350	20
13	14	8	175	20
14	0.875	8	1400	2.5
15	1.75	8	700	5
16	3.5	2	700	80

The measurements were performed on a Perkin Elmer LS50 B fluorescence spectrofluorometer with excitation wavelengths ranging between 200 and 315nm (recorded every 5 nm) and emission wavelengths ranging from 241 to 481nm

(recorded every 2 nm). Both excitation and emission slit widths were set to 5nm and the scan speed was 1500nm/min.

3.1.3 Data set 3

Sugar samples taken from the final unit operation from a sugar plant in Scandinavia were obtained as described by Bro (1999). In total 268 samples were obtained. They were dissolved in un-buffered water (2.25g/15mL).

Fluorescence was measured on a Perkin Elmer LS50 B fluorescence spectrofluorometer with emission ranging from 275 to 560nm (recorded every 0.5nm). Seven excitation wavelengths were used. These were 230, 240, 255, 290, 305, 325 and 340nm.

3.2 Reduction of original data

The goal of this work was to model the 1st order Rayleigh scatter. Therefore, the parts of the fluorescence landscape that was affected by 2nd order Rayleigh scatter, or other artifacts were removed prior to any analysis. Little fluorescence information was lost because no significant overlap with the second order Rayleigh scattering occurred. Data set 1 was reduced from 15×136×19 down to 15×106×19, by removing the 30 last emission wavelengths due the presence of 2nd order Rayleigh scatter. Data set 2 was reduced from 16×121×24 down to 16×101×18, the six first excitation wavelengths due to high noise-level, and the 20 last emission wavelengths due to the presence of 2nd order Rayleigh scatter. Data set 3 was reduced from 268×571×7 down to 194×165×7. The last 242 points was removed due to the presence of 2nd order Rayleigh. The dataset was reduced even further by removing every second measured emission wavelength. 71 samples were removed because of intensities reaching max, and three samples were removed as outliers.

3.3 Constraints

For all data sets, both models without constraints and models with nonnegativity constraints were investigated. Fluorescence spectra can readily be decomposed by

PARAFAC, and an unconstrained solution is often also in accordance with the physical/chemical premises of the data. Sometimes it may, however, be valuable to enforce some constraints on the PARAFAC solution. In the data used in this investigation, there were no missing values, and thus no danger of artifacts in the areas with missing values. Instead constraints, like non-negativity [Bro and de Jong 1997, Bro and Sidiropoulos 1998] can be used in order to smooth the estimated spectra. Non-negativity in all modes is a valid constraint in fluorescence, because all concentrations and spectra should be strictly positive. In decomposition, however, there may be small negative numbers, and by using non-negativity constraint these are removed. The effect of non-negativity constraints in all modes in the PARAFAC model was tested together with the parameters mentioned above.

3.4 Methods to model Rayleigh

In this paper, two different techniques of modeling the 1st order Rayleigh scatter have been investigated:

1. Rotating the spectra into a new coordinate system such that the Rayleigh scatter makes a line in the landscape and not a diagonal
2. Shifting the different emission spectra according to the excitation wavelength, cutting off the part away from the Rayleigh scatter line, making the Rayleigh scatter low-rank bi-linear

The Rayleigh scatter is sequentially modeled by either PCA or PARAFAC. By using PCA, the Rayleigh scatter of the different samples are allowed to have different shapes, while by the use of PARAFAC all the samples should have similar shapes of the Rayleigh scatter. This Rayleigh model is then reshaped into the original data matrix and subtracted from the original data. A PARAFAC is run on the corrected spectra. The PARAFAC modeling of the fluorophores has both been computed with and without non-negativity in all modes.

3.4.1 Rotating the EEM

The rotation of the EEM to make Rayleigh a line in the new coordinate system, would ideally be a rotation by 45º. However, since there are uncertainties in the

lamp and the detector, there may be small deviations from this. Thus the optimal rotation from 43º-47º was searched for, by finding the smallest model error in the Rayleigh scatter model. The step length in the new coordinate system was defined as the step length of the old system, with the step length in the emission mode being the new step length in the width of the Rayleigh, while the step length in the excitation mode is along the length of the Rayleigh scatter peak. The rotation was performed according to the following steps:

1. The coordinates of the landscape is found, and formed into a long matrix containing two columns, and as many rows as the size of the landscape. I.e. a landscape of the size $J \times K$ would have a coordinate axis of $JK \times 2$.
2. This coordinate matrix is multiplied by a rotational 2×2 matrix.
3. If the rotated coordinates do not fit to the new coordinate system, then it is necessary to find the closest coordinates in the new system.
4. This corrected rotation matrix is rotated back into the old coordinate system in order to find where the corrected raw coordinates lie in the original coordinate system.
5. Since these corrected raw coordinates are not in correspondence with the original raw coordinates it is necessary to perform interpolation in the old system in order to find the amplitude values of the corrected rotation matrix.
6. The corrected rotated matrix is reduced in order to not take into account fluorescence far away from the Rayleigh scatter line.
7. The signal intensity at desired excitation-emission wavelength pairs is interpolated from the nearest 4 measurements.
8. The reduced corrected rotated matrix is modeled by PCA or PARAFAC.
9. This model is rotated back into the old coordinate system by a similar procedure as the one described in 1-7 (not including point 6).

3.4.2 Shifting the EEM

For the shifting to work it is essential that the uncertainty in the excitation and the emission wavelengths are negligible. This is because this method is purely based

on the coordinate system recorded from the instrument itself. The idea here is simple: the Rayleigh scatter peak should behave similarly at different excitation wavelengths. The shifting is done as follows:

1. For each row in the data, the position of the Rayleigh peak is located where the emission wavelength equals the excitation wavelength. Each row in the spectra are shifted such that the Rayleigh scattering forms a vertical structure in the matrix. The smaller wavelength spectra are shifted to the right (higher wavelengths), while the larger wavelength spectra are shifted to the left (shorter wavelengths).
2. This procedure is repeated for all the different excitation wavelengths of the landscape.
3. The shifted landscape is reduced by removing points away from the Rayleigh scatter line.
4. A PCA or a PARAFAC is used to model the Rayleigh scatter line.
5. The modeled Rayleigh line is shifted back to the original matrix.

3.4.3 Number of factors to model the Rayleigh scatter

It was investigated whether the modeling of the Rayleigh scatter line should be performed with one or more components. This was done through re-computing the same total model with a different amount of factors for the Rayleigh scatter model, keeping the numbers of factors explaining the fluorophores constant. The criterion to estimate the right number of components for the modeling of the Rayleigh scatter was set by maximizing the modeled landscape. If the sum for the current model was higher than for the previous model with less complexity, a new model with increased complexity was computed. This was repeated until the sum decreased or was equal, and the estimated model complexity was set as the previous less complex model.

3.4.4 The complete model

The complete model was calculated in an iterative fashion. First the Rayleigh was modeled. Then the Rayleigh model was subtracted from the original data, and this

Rayleigh corrected data was then modeled for the fluorophores. This fluorophore model was again subtracted from the original data, and the fluorophore corrected data was then put through the Rayleigh modeling step again. This procedure was repeated until convergence. The convergence criterion was set to 10^{-6} relative change in the residual between two consecutive models – one model being both the Rayleigh and the fluorophore parts.

3.5 Stability of a model

The stability of each of the models was investigated by the use of bootstrapping [Wehrens *et al.* 2000] for the two first datasets and by jack-knifing [Martens and Martens 2001] for the last dataset. The reason for these choices is that the two first data sets are rather small, and thus a Jack-knifing could cause biased results. For both of these methods, 20 models were built on the data. The excitation- and emission-loadings for each set of factors and each model was multiplied making a landscape. This means that for each stability tests a number of three-way arrays equal to the numbers of factors for the model were made. The dimension for one of these arrays is $20\times$*(length of excitation)*$\times$*(length of emission)*. The standard deviation landscape for each of these arrays was calculated, and the mean standard deviation of these standard deviation landscapes is used as a parameter describing the stability of a model. The lower number the more stable model.

3.6 Quality of a model

In addition to the stability of a model it is of interest to determine how good the model estimates the pure spectra, and hence a quality parameter. The quality of a model was quantified by comparing the result from one model with the pure spectra for data sets 1 and 2. There were no known spectra for data set 3, and instead the best model with hard-weights, constraints and a band of missing values from [Rinnan and Andersen 2004], which is a refined decomposition of what Bro (1999) shows in, was used as the reference. It should, however, be noted that because there are no known spectra for this data set, the standard deviation of the

Q^2-values are of more importance than the mean Q^2-value in order to define a stable, good model.

As in [Rinnan and Andersen 2004] the quality is defined using the Q^2-statistics:

1. The mean of the Q^2 between the factors in the model and the reference is calculated – both emission and excitation modes are taken into account. The loadings of the tested model and the reference are compared in order to find the best match. Q^2 is defined as follows:

$$Q_{m,f}{}^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

   where y is one of the loadings of one factor under investigation, $\hat{y}$ is the reference loading for this factor, while $\bar{y}$ is the average of the loading under investigation. The m and f indexes on the Q^2 denote the mode and factor number investigated (m = 2 indicates the excitation-loadings, m = 3 the emission-loadings). If $Q_{m,f}^2$ equals one, the two loadings are equal, if it is zero or less, then there is no similarity between the loading under investigation and the loadings of the reference model. If $Q_{m,f}^2$ is below zero, it is set to zero.
2. The average Q^2-value for all the factors and loadings (excitation and emission) is calculated, giving a number between zero and one. E.g. for a four factor PARAFAC model, the average Q^2 would be the average of 4×2 = 8 different $Q_{m,f}^2$-values.

Each of the total models is computed 20 times according to bootstrapping/ Jack-knifing as explained above. The mean and the standard deviation of the Q^2-values can therefore be calculated.

4 Results

The first step was to decide the number of factors to use in order to model the Rayleigh scatter line as a separate component. For both of the two first data sets,

one factor was optimal using either PCA or PARAFAC. For the last, more complex system, two factors were needed both for the PCA and for the PARAFAC case.

As a reference method, a "MILES" PARAFAC [Bro *et al.* 2002, Rinnan and Andersen 2004] with insertion of zeros below the 1st order Rayleigh scatter was used, as explained in Rinnan and Andersen (2004). The PARAFAC model was both computed without any constraints and with non-negativity in all modes.

Table 3: The stability and quality of the 10 different PARAFAC models on the three data sets. The stability is given as the average standard deviation in 10^{-5}. The quality is given as the mean and the standard deviation of the Q^2-values.

		Data set 1			Data set 2			Data set 3		
Model	Cons	JK	Q^2	Std-Q^2	JK	Q^2	Std-Q^2	JK	Q^2	Std-Q^2
Reference										
MILES	-	34	0.992	0.0004	717	0.989	0.002	125	0.907	0.006
MILES	NN	30	0.993	0.0004	723	0.992	0.001	73	0.929	0.013
Rotation										
PARAFAC	-	364	0.797	0.033	250	0.734	0.012	104	0.359	0.003
PARAFAC	NN	72	0.916	0.018	356	0.775	0.082	87	0.149	0.0003
PCA	-	376	0.791	0.042	394	0.676	0.080	77	0.382	0.020
PCA	NN	72	0.916	0.018	345	0.792	0.115	68	0.159	0.0004
Shifting										
PARAFAC	-	16	0.992	0.0004	40	0.988	0.002	63	0.933	0.005
PARAFAC	NN	13	0.993	0.0004	27	0.992	0.001	116	0.960	0.018
PCA	-	23	0.991	0.002	39	0.986	0.003	86	0.918	0.004
PCA	NN	12	0.992	0.0005	28	0.989	0.001	128	0.913	0.034

From Table 3 it is clear that by modeling the 1st order Rayleigh scatter as a separate factor in the modeling step gives a more stable model. The average quality of the model is similar for both data set 1 and 2, while for data set 3 (the "real" data set) the quality of the model is higher; the model is better. Rotating the landscape, making the Rayleigh scatter line low-rank bilinear, seems not to be a good idea.

For data set 3, the models are the most stable, with a standard deviation of the Q^2 values below those of both the reference model and the shifting of the Rayleigh scatter line. However, upon looking at the loadings from the model (see Figure 1) it becomes clear that the model is not adequate, and thus even though the model is stable, the solution reached is not the correct one, which also can be seen from the very low Q^2-values.

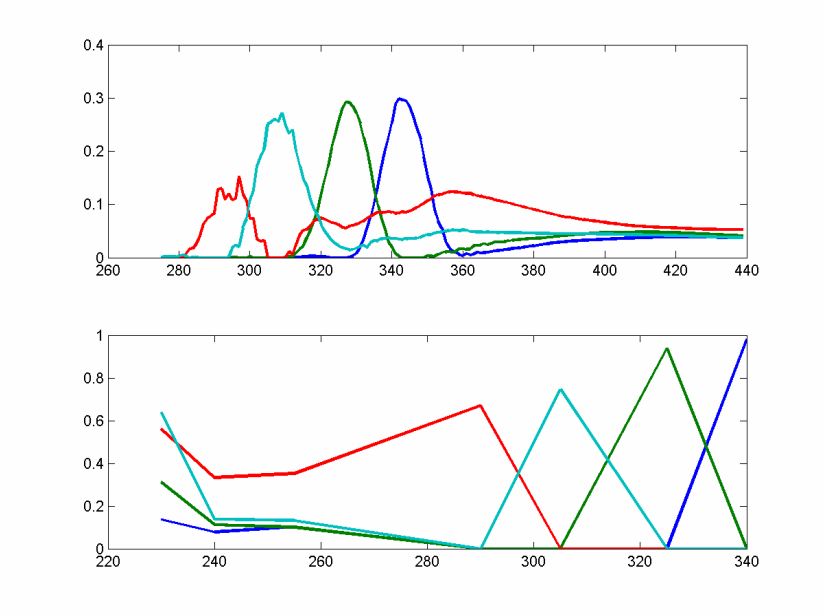


Figure 1: The fluorophore loadings, modeling the Rayleigh scatter by rotation. (a) emission loadings, (b) excitation loadings.

Figure 1 show that all the four fluorophore factors have excitation and emission maxima at the same wavelength, meaning that what this model mainly models the Rayleigh scatter line itself. The tails of the emission loadings seems to resemble that of the real fluorophores. This means that the modeling of the Rayleigh scatter line fail, it is not modeled by the PCA/ PARAFAC that should model the Rayleigh. There are some assumptions in the rotation method: 1) the modeling should start with modeling the Rayleigh and not the fluorophores, 2) the rotational angle can be optimized by explained variance, 3) keeping the steps in the old coordinate system

gives a good estimation of the rotated coordinate system, 4) the area between points should readily be modeled by a line and 5) the chosen model complexity for modeling the Rayleigh scatter is correct. It does not seem like one or some of these assumptions hold. The rotation model was therefore re-analyzed several times on data set 2 in order to check if these assumptions where correct. The alternatives for the five points are: 1) using hard weights prior to modeling Rayleigh for the first time, 2) keeping the rotational angle constant, 3) the steps in the new coordinate system equal to the average step length in the old coordinate system, 4) using second or third order fitting in the interpolation step and 5) modeling the Rayleigh model with one or two components. All these combinations were run to investigate if there was an optimal setup for data set 2. The best setup gave a decomposition as good as the shifting method. However, when the six best setups for data set 2 were tried on data set 1, the results did not give any good decomposition. It therefore seems like the rotation methods needs to be setup specifically for each data set. In the optimal case it can give as good results as the shifting method, but in order to find the optimal setup for the rotation method, the time used compared to the shifting is multiplied by at least 50.

The shifting method on the other side, being straightforward, fast and stable seems to be a reliable and good method for modeling the Rayleigh scatter line. The results are slightly better than for the reference method. The time needed before convergence is between 4 and 10 times longer for the shifting methods, but on the other hand, there are less parameters to estimate. While for the MILES algorithm to work optimally, the Rayleigh width and the bandwidth of missing values are both important to estimate correctly in order to get good decompositions, as was shown by Rinnan and Andersen (2004). For the shifting method, the only parameter to estimate is the width of the Rayleigh scatter. However, it is not as important to estimate this correctly as for the MILES method. There are less parameters to optimize and less important to estimate the parameters correctly for the shifting method rather than the MILES method. Therefore the total time consumption may

not be as large as for the optimal case. It thus seems like a better method than MILES if a more automatic decomposition is wanted.

5 Conclusion

Rotating the landscape does not work, while shifting the emission spectra according to the excitation spectra gives good estimates of the Rayleigh scatter peak. Generally, imposing non-negativity constraints on the modeling of the fluorophores gives a better model, but may cause some unstability. For modeling the Rayleigh scatter line, PARAFAC is a better method than by using PCA, although the difference is not large.

This work has shown that modeling the Rayleigh scatter line as a separate component in the decomposition with the shifting method gives more reliable and better results than MILES – the best method found in the literature. The shifting method is, model to model, slower than MILES, but taking into account the optimization steps necessary for MILES, equal or even less time is consumed for the shifting method. The shifting method is therefore a better and more automatic model than the best method found in the literature.

6 Acknowledgement

Rinnan wish to thank the STVF (Danish Research Council) for financial support through project 1179.

7 References

Andersen, C. M. and Bro, R.: Practical aspects of PARAFAC modeling of fluorescence excitation-emission data, *Journal of Chemometrics*, **17 (4)**, 2003, 200-215

Bro, R.: PARAFAC. Tutorial and applications, *Chemometrics and Intelligent Laboratory Systems*, **38 (2)**, 1997, 149-171

Bro, R.: Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis, *Chemometrics and Intelligent Laboratory Systems*, **46 (2)**, 1999, 133-147

Bro, R. and de Jong, S.: A fast non-negativity-constrained least squares algorithm, *Journal of Chemometrics*, **11**, 1997, 393-401

Bro, R. and Sidiropoulos, N. D.: Least squares algorithms under unimodality and non-negativity constraints, *Journal of Chemometrics*, **12**, 1998, 223-247

Bro, R., Sidiropoulos, N. D., and Smilde, A. K.: Maximum likelihood fitting using ordinary least squares algorithms, *Journal of Chemometrics*, **16 (8-10)**, 2002, 387-400

Carrol, J. D. and Chang, J.-J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition, *Psychometrika*, **35 (3)**, 1970, 283-319

Christensen, J., Povlsen, V. T., and Sørensen, J.: Application of fluorescence spectroscopy and chemometrics in the evaluation of processed cheese during storage, *Journal of Dairy Science*, **86 (4)**, 2003, 1101-1107

Harshman, R. A.: Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis, *UCLA working papers in phonetics*, **16**, 1970, 1-84

Ho, C. N., Christian, G. D., and Davidson, E. R.: Application of the Method of Rank Annihilation to Quantitative Analyses of Multicomponent Fluorescence Data from the Video Fluorometer, *Analytical Chemistry*, **50 (8)**, 1978, 1108-1113

Ho, C. N., Christian, G. D., and Davidson, E. R.: Application of the Method of Rank Annihilation to Fluorescent Multicomponent Mixtures of Polynuclear Aromatic Hydrocarbons, *Analytical Chemistry*, **52**, 1980, 1071-1079

Jiji, R. D. and Booksh, K. S.: Mitigation of Rayleigh and Raman spectral interferences in multiway calibration of excitation-emission matrix fluorescence spectra, *Analytical Chemistry*, **72**, 2000, 718-725

Martens, H. and Martens, M.: Multivariate analysis of quality - An introduction, John Wiley, Chichester, 2001, 445 pages

McKnight, D. M., Boyer, E. W., Westerhoff, P. K., Doran, P. T., Kulbe, T., and Andersen, D. T.: Spectrofluorometric characterization of dissolved organic

matter for indication of precursor organic material and aromaticity, *Limnology and Oceanography*, **46 (1)**, 2001, 38-48

Munck, L., Nørgaard, L., Engelsen, S. B., Bro, R., and Andersson, C. A.: Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance, *Chemometrics and Intelligent Laboratory Systems*, **44 (1-2)**, 1998, 31-60

Rinnan, Å. and Andersen, C. M.: Handling 1st order Rayleigh scatter effect in PARAFAC, *Chemometrics and Intelligent Laboratory Systems*, 2004, Submitted

Rodriguez-Cuesta, M. J., Boque, R., Rius, F. X., Picon Zamora, D., Martinez Galera, M., and Garrido Frenich, A.: Determination of carbendazim, fuberidazole and thiabendazole by three-dimensional excitation–emission matrix fluorescence and parallel factor analysis, *Analytica Chimica Acta*, **491 (1)**, 2003, 47-56

Trevisan, M. G. and Poppi, R. J.: Determination of doxorubicin in human plasma by excitation-emission matrix fluorescence and multi-way analysis, *Analytica Chimica Acta*, **493 (1)**, 2003, 69-81

Wehrens, R., Putter, H., and Buydens, L. M. C.: The bootstrap: a tutorial, *Chemometrics and Intelligent Laboratory Systems*, **54 (1)**, 2000, 35-52

Paper IV

Thygesen, L.G., Rinnan, Å.,Barsberg, S., Møller, J.K.S.

Stabilizing the PARAFAC decomposition of Fluorescence Spectra by insertion of zeros outside the data area, *Chemometrics and Intelligent Laboratory Systems*, 2004, In press

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Chemometrics and Intelligent Laboratory Systems 71 (2004) 97–106
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Stabilizing the PARAFAC decomposition of fluorescence spectra by insertion of zeros outside the data area

Lisbeth Garbrecht Thygesen[a,1], Åsmund Rinnan[a,*], Søren Barsberg[b], Jens K.S. Møller[c]

[a] Royal Veterinary and Agricultural University, Department of Dairy and Food Science, Food Technology, Rolighedsvej 30, DK-1958 Frederiksberg, Denmark
[b] Royal Veterinary and Agricultural University, Plant Fibre Laboratory, Agrovej 10, DK-2630 Taastrup, Denmark
[c] Royal Veterinary and Agricultural University, Department of Dairy and Food Science, Food Chemistry, Rolighedsvej 30, DK-1958 Frederiksberg, Denmark

Received 15 June 2003; received in revised form 10 November 2003; accepted 7 December 2003

Available online 5 March 2004

Abstract

The use of fluorescence spectroscopy for recording multiple excitation and corresponding emission wavelengths and the subsequent technique of analyzing the resulting fluorescence landscapes is a rather new method as opposed to the use of just a single excitation wavelength. In a fluorescence landscape, several light-scatter effects are usually present, and often the part of the landscape containing information on the chemical and/or physical characteristics of the sample is surrounded by two Rayleigh scatter lines. When such landscapes are decomposed using parallel factor analysis (PARAFAC), the scatter effects may have detrimental effects on the resolved spectra, especially if the peaks from the analytes lie close to or on the Rayleigh scatter lines. Normally, all values close to and outside the Rayleigh scatter lines are set to missing values before decomposing the fluorescence landscapes by PARAFAC. In this paper, we introduce a novel pretreatment method applicable for two-dimensional fluorescence landscapes, where instead of inserting only missing values a mixture of zeros and missing values are inserted close to and outside the Rayleigh scatter lines. It is shown that, by the use of this technique, a physically and chemically meaningful decomposition is obtained, and furthermore the modeling converges faster. Constraining the PARAFAC solution to positive values in all modes gave results similar to those obtained for the unconstrained model, except that the loadings where less smooth and the number of iterations before convergence was smaller.

Keywords: PARAFAC; Fluorescence; Rayleigh scatter; Missing values; Inserting zeros into the EEM

1. Introduction

The role of fluorescence spectroscopy in the analysis of organic products has increased. Handling these data with parallel factor analysis (PARAFAC) [1] is a very powerful way of extracting information from an excitation–emission matrix (EEM), i.e. several samples where the emission intensity is depicted as a function of both excitation and emission wavelengths. Often these data also include light scattering effects, such as Rayleigh scatter. Since PARAFAC only decomposes trilinear structures and the scatter is on the diagonal (excitation = emission), this causes some mathematical difficulties in the decomposition. It is therefore of interest to remove this effect, or at least to reduce it as much as possible. Several ways of handling these scattering effects have been presented previously: weighting the scatter areas down (or areas containing information up) [2,3], inserting missing values [4] or plainly avoiding the part of the matrix that includes the scatter. The last method, however, can only be used in cases where the removed wavelengths contain little or no information. For some analysis, this is mostly the case, and thus there need not be any concern about the scattering effect. The method of inserting missing values, on the other hand, can lead to unacceptable decomposition of the spectra. Some of these models show photophysical impossible features, like a component emitting light at a higher energy level than the absorbed light.

In this paper, we describe a new way of handling these problems. The method can be seen as a pretreatment, where

* Corresponding author.
E-mail address: aar@kvl.dk (Å. Rinnan).
[1] Present affiliation: Department of Civil Engineering, The Danish Technical University, Building no. 118, DK-2800 Kgs. Lyngby, Denmark.

0169-7439/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2003.12.012

zeros are inserted into the matrix. PARAFAC will then converge faster and give better estimates of the spectra. This method can be used on any type of fluorescence landscapes, but it does not have a significant effect on the resolved spectra unless one or more of the peaks lie close to or on the Rayleigh scatter lines, as illustrated in the emission spectra in Fig. 1.

Andersen and Bro [5] explained that inserting zeros into the dataset instead of using missing values does not conform to the trilinear structure of the EEM. While this is true in general, this paper will show that in some practical situations, even though mathematically being untrue, inserting zeros in part of the spectra can be very handful for the decomposition of the underlying emission and excitation spectra.

1.1. Theory

1.1.1. Terms used

The area between the Rayleigh scatter lines will in this paper be called the ''data area''. The part of the landscape where the emission wavelength is smaller than the excitation wavelength will be termed below the first order Rayleigh scatter. Above the second order Rayleigh scatter is where the emission wavelength is larger than twice the excitation wavelength.

1.1.2. Information in the data

There is rarely any additional chemical information outside the data area. More specifically, the area below the first order Rayleigh scatter does not give any physical meaning, since a molecule cannot emit light of higher energy than what it absorbed, and can thus be deleted. However, removing this part is not a simple matter since it makes up a triangular area in the matrix. The easiest way of circumventing this problem would be to enlarge this area into the smallest possible rectangle and remove this part. By this, you might, however, remove some interesting information and, therefore, it is not a good method for solving the problem. A different, and much used method, is to insert missing values below the first order Rayleigh scatter. To ensure that all the Rayleigh scatter is removed, some extra data points around the first order Rayleigh scatter are also replaced with missing values. This often gives satisfying results, but the amount of missing values can affect the convergence of PARAFAC and the quality of the results. A third way of handling this type of data is to either weight down the part outside the data area, or weight up the data area. This is a powerful method, but it is computationally cumbersome, increasing the computational time before convergence by a factor of 10, or even as high as 100 compared to non-weighted PARAFAC.

The area above the second order Rayleigh scatter, on the other hand, is meaningful, but rarely holds any new chemical information. The information in this part is mostly an echo of the information in the data area. These values can be treated in the same way as for the values below the first order Rayleigh scatter, with the limitations described.

1.1.3. Constraints

PARAFAC modeling gives the least squares solution, and this solution is often also in accordance with the physical/chemical premises of the data. Sometimes, it may, however, be valuable to enforce some constraints on the PARAFAC solution, especially in situations where the number of missing values is relatively high. In these situations, small model-errors may strongly bias the estimated spectra. Thus, the estimated spectra may give no or little chemical meaning. By constraining the PARAFAC model with sound constraints in agreement with a priori knowledge about the system, such as non-negativity and maybe unimodality, the resolved spectrum will more clearly

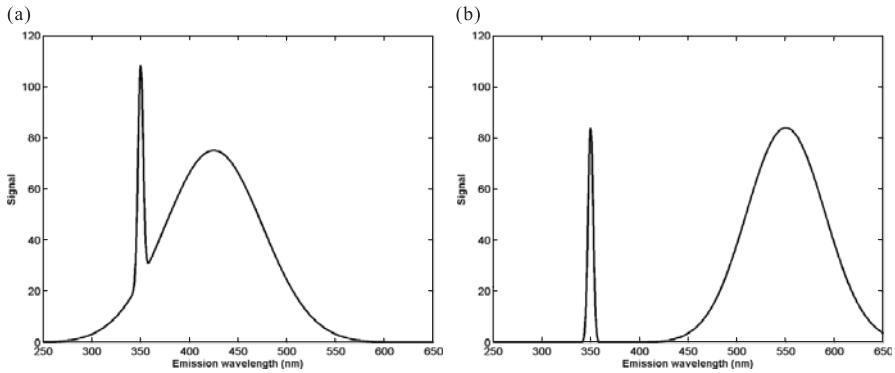


Fig. 1. Emission spectrum with a Rayleigh scatter line and a peak from a chemical component when excited at 350 nm (simulated data). (a) The problematic case, (b) the insertion of zeros does not affect the solution very much.

indicate the chemical attributes of the sample. Spectroscopic data should be strictly positive, and hence the use of non-negativity may be enforced upon the data to ensure this [6–8]. Fluorescence spectra of a single fluorophore without vibrational fine structure will only have one emission peak, making it adequate to use unimodality constraints. Normally, this constraint is not valid in the excitation mode, which—even in the absence of vibrational fine structure—may contain several excitation peaks (electronic transitions). In the case where only a limited range of excitation wavelengths is recorded, it may, however, be valid.

2. Experimental

2.1. Methods

A novel way of pretreating fluorescence spectra is to insert zero-values in parts or in the whole of the landscape outside the data area.

Four different pretreatment methods of the data are tested in this paper (see Fig. 2):

1. Only missing values—conventional method
2. Zeros below first order and missing values above second order—mixed method
3. Only zeros—all zeros method
4. Mostly zeros, but with a ribbon of missing values around the Rayleigh scatter lines—ribbon method

As Andersen and Bro clearly stated, methods 2–4 above do not conform to the trilinear structure of the EEM. However, a large amount of missing values may cause problems in the decomposition. Especially if the number of missing values at one emission/excitation wavelength is very large (typically 80% or more), small artifacts close to the missing values may lead to large artifacts in the extracted spectra (see Fig. 4). So, even though inserting zeros is not mathematically true, it can help PARAFAC to decompose into a solution that is more

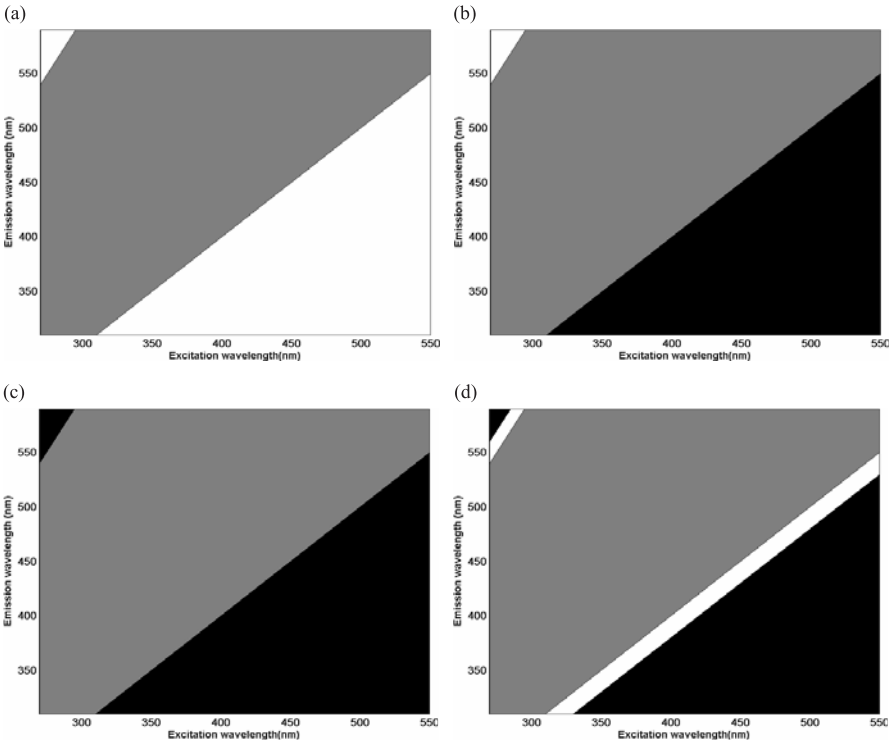


Fig. 2. Different strategies for inserting missing (white) and zeros (black) outside the data area of an excitation–emission matrix (data area: grey). (a) Only missing values, (b) zeros outside the first order, missing outside the second order, (c) only zeros and (d) mostly zeros with a ribbon of missing values around the Rayleigh scatter lines.

liable than upon keeping all the missing values (method 1). Inserting zeros may be seen as a weak form of non-negativity constraint, since the spectra are guided towards zero, but not set to strictly be non-negative. However, since one fluorophore's excitation and emission spectra might have some overlap (which is the case of, e.g., tyrosine), inserting zeros will destroy this possibility of overlap. Inserting zeros below the first order Rayleigh indicates that there should not be any overlap in the emission and excitation spectra.

2.1.1. Conventional method

This method has already been discussed and will therefore not be explained further.

2.1.2. Mixed method

Sometimes the areas with missing values especially below the first order Rayleigh scatter can be quite large. This can cause PARAFAC to convergence slower, as it has to estimate a large number of missing values. It can also give unstable or non-meaningful results. Therefore, inserting zeros outside the scatter line can stabilize the decomposition and make it converge faster. From a photophysical point of view this method correctly reflects the fact that no emission is expected below the first order Rayleigh scatter, whereas emission tails may extend into the part above the second order Rayleigh scatter, as is seen for some samples analyzed in the present work.

2.1.3. All zeros method

Instead of missing values above the second order Rayleigh scatter, zeros are inserted. This may cause PARAFAC to converge even faster and give better estimates of the spectra.

2.1.4. Ribbon method

In some of the above-mentioned methods, where zeros are inserted adjacent to the data area, PARAFAC may produce chemically unsound solutions due to this introduction of discontinuity in the data. One way to circumvent this problem is to insert a ribbon of missing values around the Rayleigh scatter lines, which also permits some overlap between one fluorophore's emission and excitation spectra. In this way, PARAFAC will be "free" to estimate a continuous shape of the peaks. Three different widths of the ribbon are chosen: 1, 5 and 10 units. A unit equals the width of a single step in the excitation or emission (wavelength) dimension, which is defined by the experimental conditions. It should be noted that, for the data analyzed in this paper, there is no difference in a width of 5 and 10 units for dataset II above the second order Rayleigh scatter, since the dimensions of the excitation–emission matrix are transgressed with the increase from 5 to 10 units. The ribbon was applied in two ways: only around the first order Rayleigh scatter (A) or both around the first and second order Rayleigh scatter (B). Ribbon method "1A" thus denotes a ribbon width of 1 unit in conjunction with using the ribbon only around the first order Rayleigh line.

2.2. Quality of a model

The pretreated dataset is decomposed into its pure chemical components by the use of PARAFAC (with or without constraints). This results in a number of factors for every model, and it is of vital importance to evaluate which models are adequate. In order to evaluate this, five different criteria were set (they are set in the order of importance):

- No emission spectrum of a component can have its peak placed at lower wavelength than the corresponding excitation spectrum peak
- Only small negative values in the resolved spectra are allowed
- The explained variance of the model should be high (above 97%)
- All factors should be smooth and not only describe noise
- Finally, the spectra were evaluated by visual inspection to make sure they were reasonable

All analyses were performed in Matlab 6.1 and 6.5 for Windows (Mathworks, Natick, MA, USA), with algorithms taken from the N-way toolbox version 2.10 [9], available at www.models.kvl.dk. The algorithms for the insertion of zeros were written in-house.

2.3. The data

2.3.1. Dataset I (wood fibers)

Thermo mechanical pulp (wood fibers) of spruce (*P. abies*) was supplied by Sunds Defibrator, Sundsvall, Sweden. The pulp was dried at ambient temperature and humidity before use. The pulp is autofluorescent due to its content of the fluorescent plant polymer lignin. Samples of different emission intensities were produced by the adsorption of (non-emissive) *p*-benzoquinone into the fiber cell walls. The control sample had no *p*-benzoquinone adsorbed, whereas three less emissive samples had three different quantities of the quinone adsorbed, the larger the adsorbed amount the smaller the emission. The experimental setup is described in more details elsewhere [10].

The samples were pressed into disks of 1 mm thickness by a Perkin-Elmer hydraulic press (at 5 bar). The disks were used for fluorescence measurements by a SPEX 1680 0.22-m double monochromator fluorescence spectrometer in front-face setup. The samples were measured at 35 excitation wavelengths (259–599 nm), and 30 emission wavelengths (355–645 nm), both with a step of 10 nm. For the control sample, as well as for the quinone-containing

samples, five independent excitation–emission intensity matrices were produced. However, for the quinone-containing sample with the highest content of quinone, only 3 of the 5 matrices have been included because the remaining 2 matrices were highly deviating from the other 18 matrices. The cause of this deviation was not further investigated and, because the aim of this study is to show the effects of inserting zeros into the EEM and not how to detect outliers, no further explanation as to why these were outliers will be given. Fluorescence data for sample number 1 of this dataset is shown in Fig. 3a and b.

2.3.2. Dataset II (dry-cured hams)

Lean raw pork from fresh hams was obtained from the local market, whereas Parma hams were from a processing plant in Parma, Italy. Parma ham ages ranges from salted (3 months) to matured (11 and 12 months) and further to aged (15 and 18 months). Samples were thermostated in a water bath at 25 °C before measuring. A total of 67 meat samples were submitted to duplicate measurements of surface autofluorescence spectroscopy.

The measurements were done using a BioView instrument (Delta Light and Optics, Lyngby, Denmark) equipped with a fiber optics measuring probe giving an open-end 180° excitation–emission geometry. The instrument used a pulsed xenon lamp for excitation, and a surface area of ~ 6 mm in diameter was sampled in each measurement. The samples were measured at 15 excitation wavelengths (270–550 nm), and 15 emission wavelengths (310–590 nm), both with a step of 20 nm. The emission wavelengths were shifted by 40 nm from each excitation wavelength applied. Before analysis of the data, the excitation wavelengths above 470 nm, and the emission wavelengths below 350 nm were removed. Thus, the dimension of the dataset was reduced from $67 \times 15 \times 15$ to $67 \times 11 \times 13$ (samples × excitation × emission). The areas deleted only contain very few measurement values and were thus removed to stabilize the model. The data and their chemical interpretation in relation to process control is described in detail elsewhere [11]. Fluorescence data for sample number 50 of this dataset is shown in Fig. 3c and d.

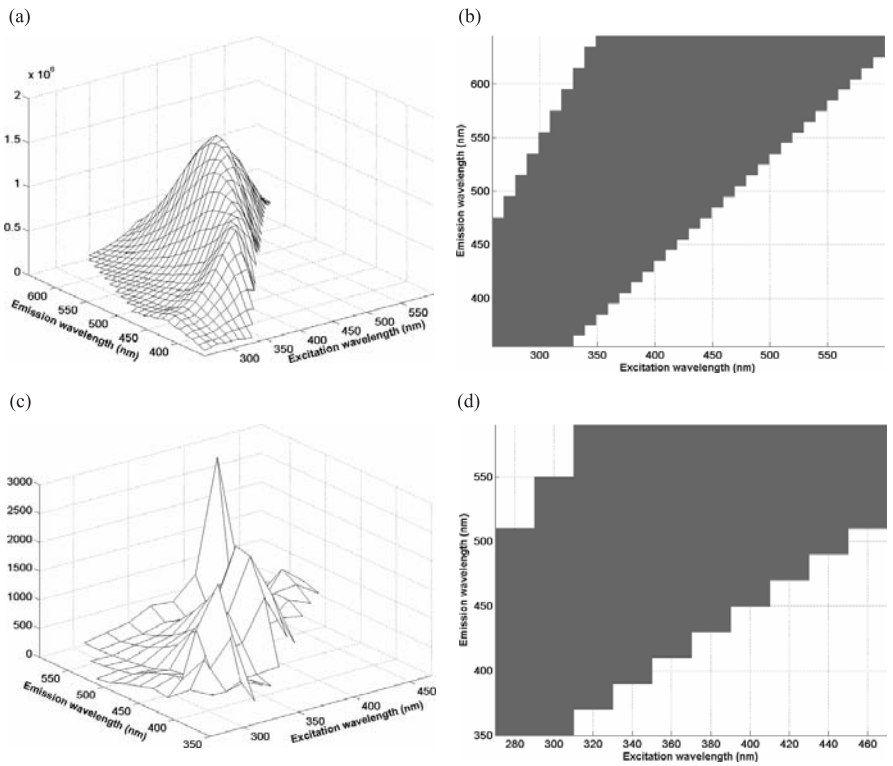


Fig. 3. Fluorescence landscape and data area for sample no. 1 from dataset I (a and b) and from sample no. 50 from dataset II (c and d).

2.4. Data analysis

Thirty-six PARAFAC models were calculated for each dataset, i.e. all combinations of the nine methods for insertion of zeros into the data matrices (of which six are the different ribbon methods), and four different ways of constraining the PARAFAC model: no constraints, non-negativity constraint on the two spectroscopic modes, non-negativity constraint on all three modes and unimodality constraint on the two spectroscopic modes. Apart for the constraints, PARAFAC was carried out using default input parameters. For dataset I, the number of components in the models was set to three based on an evaluation of model stability and the shape of the loadings. For dataset II, five component models were calculated in accordance with a previous publication on the dataset [11]. Normally, there is little sense in assessing model quality without discussing the number of components to include in the model. However, in the present study, focus is on the effect of inserting zeros into the data matrices, and thus a fixed number of components for each dataset have been chosen.

3. Results and discussion

Fig. 4 shows the scores and loadings for five different PARAFAC models obtained for dataset I. All models were calculated without constraints. The sample scores are very similar, both within each model and within the models, indicating that the fluorophores in the wood fiber samples co-vary. Furthermore, the sample score follow the intensity of the treatment (quinone-adsorption), i.e. in reality the dataset only contains four uniquely different sample types. These two characteristics are bound to make the PARAFAC model less stable for dataset I, and a Tucker3 model with dimensions [1 3 3] might have been more appropriate. However, a PARAFAC split-half analysis of dataset I gave the same sets of loadings for the two subsets, and thus it seems reliable to use PARAFAC with three components. In Fig. 4, the excitation and emission loadings are very different between models, even though the explained variance does not differ with more than 1–2%. From a physical point of view, the model for the conventional method is not acceptable, as the emission peak appears at a shorter wavelength than the excitation peak for at least one of the

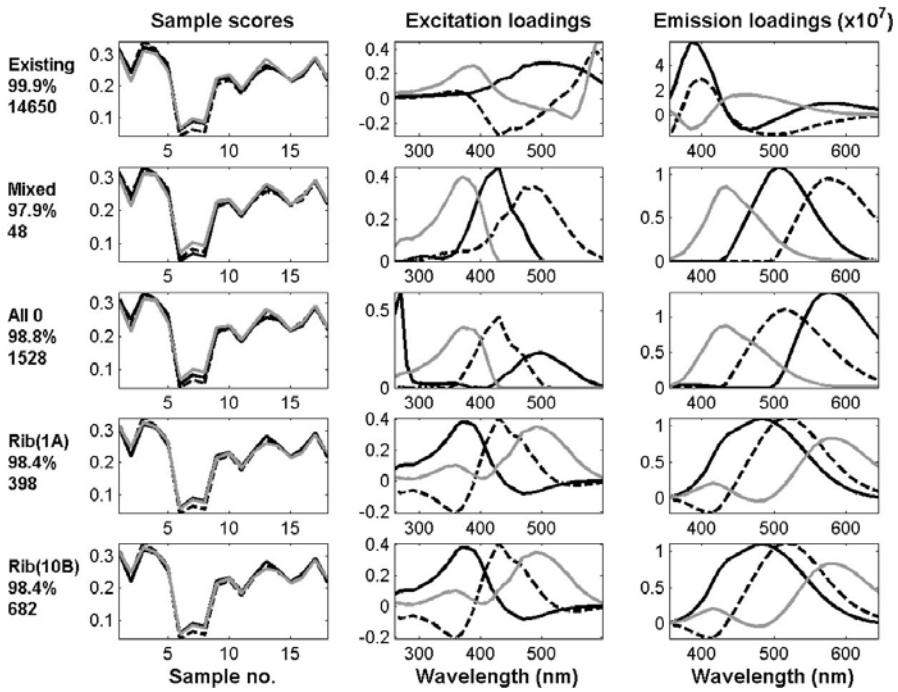


Fig. 4. Scores and loadings for five different unconstrained PARAFAC models for dataset I. The text to the left gives for each set of loadings the method used for insertion of zero values (line 1), the percent explained variance (line 2) and the number of iterations necessary before convergence was reached (line 3).

components, and more than half of at least one loading is negative. Furthermore, the emission loadings are not unimodal, as they should be if each component represented a single fluorophore. The model for the all zeros method, on the other hand, has three perfectly unimodal loadings with a reasonable spacing between each excitation peak and its corresponding emission peak. Another nice feature of the model for the all zeros method is that it converged in only 48 iterations, while the model for the conventional method took 14,650 iterations. The model for the mixed method gave emission loadings more or less identical to the model for the all zeros method, except that one of the excitation loadings has an extra peak below 300 nm. The models for the ribbon method were all rather similar (only some results shown), independent of whether a narrow or a broad ribbon of missing values was inserted and independent of whether ribbons of missing values were inserted both above and below the data area or only below. In Fig. 4, only two of the models for the ribbon method are included, namely methods 1A and 10B. For both models, the excitation and emission loadings are reasonably unimodal, but all have small negative values at the base of the main peaks. Using the criteria for an acceptable model set up in Section 1, the model for the conventional method is unacceptable. The other four models, however, are acceptable, with the model for the all zero method being the best from a physical point of view, albeit it explains slightly less of the variance compared to the other models.

Fig. 5 shows the model of the fluorescence landscape for sample no. 1 of dataset I for four different combinations of PARAFAC modeling constraints and zero insertion methods. Fig. 5a confirms what is already indicated in Fig. 4, namely that the unconstrained model for the conventional method has a spurious peak outside the data area. Fig. 5a shows that the spurious peak is an order of magnitude larger than the part modeling the data area. The non-negativity constrained model for the all zeros method (Fig. 5b) also shows a spurious peak outside the data area, but in contrast to what is the case for the model in Fig. 5a, the modeling of the data area is acceptable according to the criteria set up in the introduction. Fig. 5c and d illustrates the difference between "constraining" the data, i.e. the input of PAR-

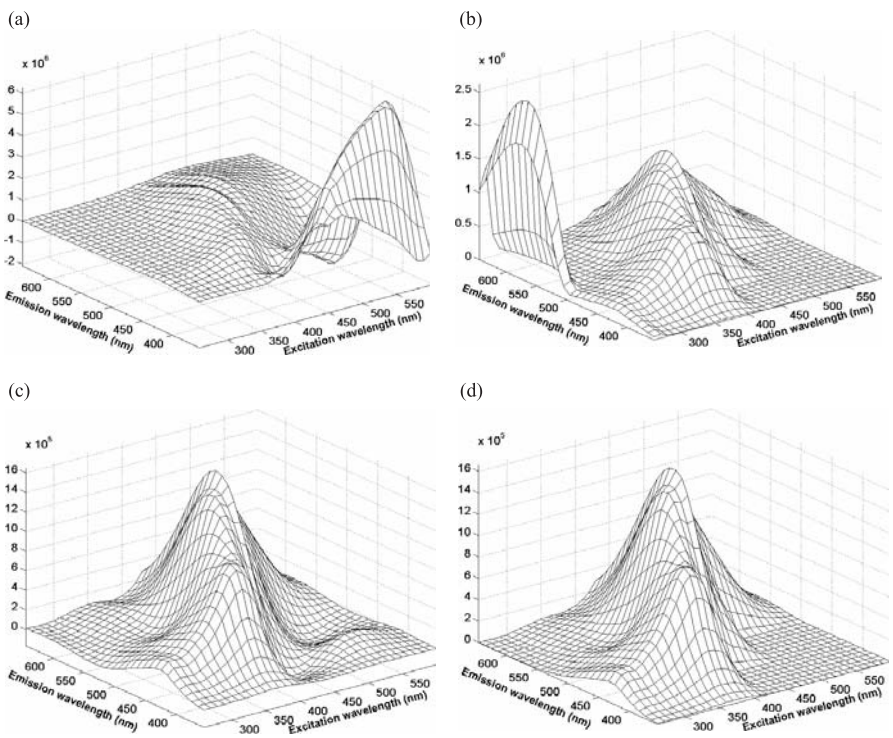


Fig. 5. Selected PARAFAC models of the fluorescence landscape of sample no. 1 from dataset I: (a) conventional method, unconstrained, (b) all zeros method, non-negativity constraint on all three modes, (c) ribbon (1A) method, unconstrained and (d) ribbon (10B) method, non-negativity constraint on all three modes.

AFAC, and constraining PARAFAC directly, for data subjected to pretreatment according to the ribbon method. In Fig. 5c, the missing values were "constrained" to zero, except for one row below the first order Rayleigh scatter, while PARAFAC was unconstrained. In Fig. 5d, zero values were only inserted far from the actual data area (10 measurement points from it at both sides), while PARAFAC was constrained to non-negativity in all three modes. The modeling of the core of the data area is similar. The difference is that the peaks of the model in Fig. 5c continue a bit below zero at the base, thus providing continuous, smooth loadings, while the peaks of the model in Fig. 5d are abruptly cut off at zero because of the non-negativity constraint. In turn, this may lead to spurious peaks outside the data area, as a form of compensation. Replacing missing values outside the data area with zeros, while keeping PARAFAC unconstrained, thus appears to be a more gentle way of guiding the model towards non-negativity.

Figs. 6 and 7 are similar to Figs. 4 and 5, but here dataset II is under investigation. A big difference between the two datasets is that the number of wavelengths in the excitation and emission modes for dataset II is only half the number of wavelengths included in dataset I. The resolved spectra for the components of dataset II therefore appear less smooth than those of dataset I. From Fig. 6, it can clearly be seen that the conventional method gives excitation and emission spectra that are unacceptable from a photochemical point of view, as more than one emission peak appears at a shorter wavelength than its corresponding excitation peak. The other four results are all very similar, giving one component with negative score values, but all the emission spectra are close to unimodal, and the excitation spectra have only small negative values. Both excitation and emission spectra are smooth. The three last methods result in identical decompositions, while the second from the top is very similar to these. From an analytical point of view, it can be argued that only four factors should be included in these models, since two of the spectra in both the excitation and the emission mode behave similarly. However, since five is the estimated number of components [11] for this dataset, it has also been used here. None of the models in Fig. 6 are good since one of the score factors is negative, but for all the four last models, the two other modes seem reasonable, and thus these models will be evaluated as acceptable. It should

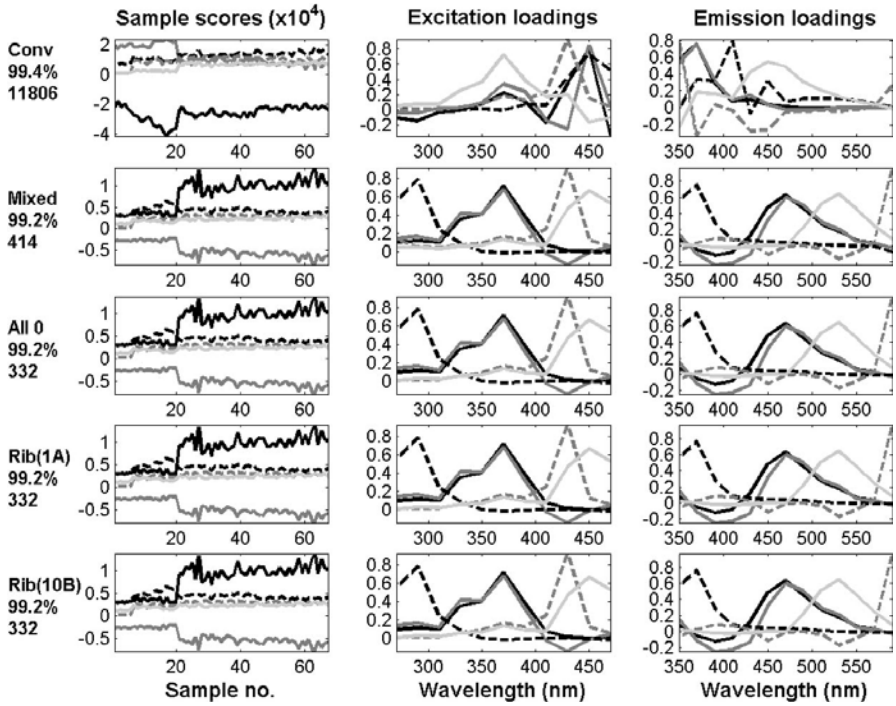


Fig. 6. Scores and loadings for five different unconstrained PARAFAC models for dataset II. The text to the left gives for each set of loadings the method used for insertion of zero values (line 1), the percent explained variance (line 2) and the number of iterations necessary before convergence was reached (line 3).

also be noted that modeling using the conventional method (missing values outside the data area) with three or four components also resulted in models where two components showed ''complementary'' or ''mirrored'' sample scores, i.e. each sample was modeled to ''contain'' either one or the other (i.e. to have a high score value for one and a low score value for the other). This is also seen in all five-factor models in Fig. 6, with one of the two ''complementary'' scores vectors being negative for all models, and thus this pattern appears to be unaffected by the replacement of missing values with zeros. We have also observed this pattern for PARAFAC modeling of fluorescence data from beet sugar processing juices (not published), and speculate that it may occur whenever a ''discrete'' PARAFAC model implying a few well-defined fluorophores is forced upon data of natural/biological origin, which in reality contains a continuum of only slightly different fluorophores, maybe because the same fluorescent groups are part of many different macromolecules. In other words, the negative scores vector is maybe a sign that PARAFAC modeling of dataset II is a simplification of a complex reality. Nevertheless, it may lead to something useful.

Fig. 7 shows that, for dataset II, the conventional method gives spurious peaks below the first order Rayleigh scatter line, where there should not be any peaks. The three other methods are all similar, showing all the same peaks. There are only minor differences in these predicted spectra, with the Ribbon (1A) method giving some small negative values in the spectra, and a small peak in the area below the first order Rayleigh scatter line.

Table 1 gives a condensed overview of the quality of the calculated models, evaluated from plots like those in Figs. 4–7. Each model was categorized as either ''acceptable'' or ''unacceptable'' based on the criteria mentioned in the introduction. No acceptable models where achieved at all for the datasets with missing values at all positions outside the Rayleigh scatter lines (the conventional method). For the pretreatments of the datasets with zeros inserted the PAR-

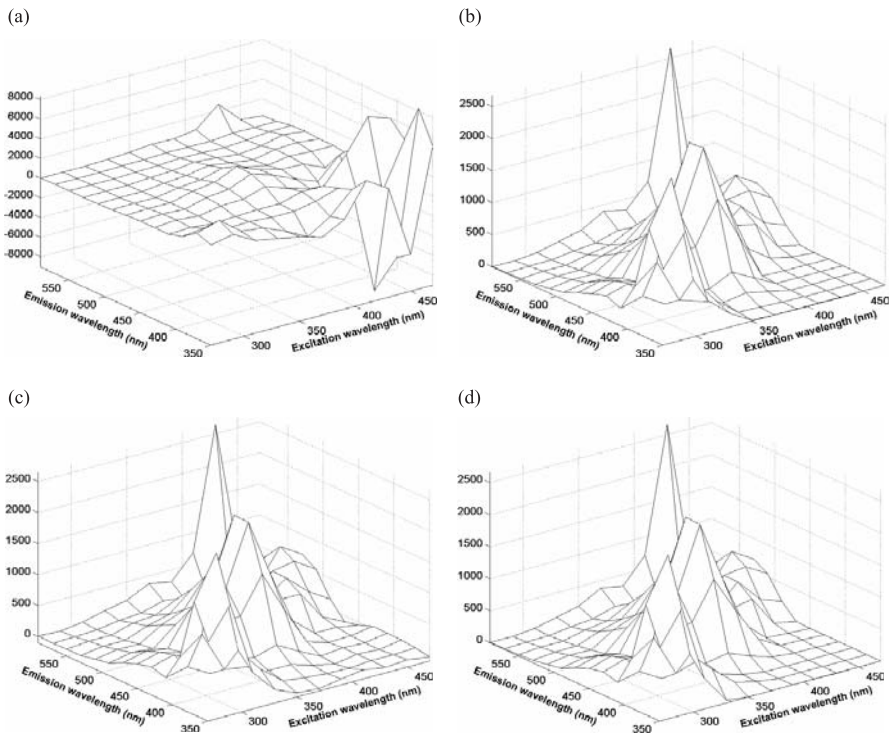


Fig. 7. Selected PARAFAC models of the fluorescence landscape of sample no. 50 from dataset II: (a) conventional method, unconstrained, (b) all zeros method, non-negativity constraint on all three modes, (c) ribbon (1A) method, unconstrained and (d) ribbon (10B) method, non-negativity constraint on all three modes.

Table 1
The results on the two datasets studied (I and II)

PARAFAC constraints	Conv		Mix		Zeros		Ribbon													
							1				5				10					
							A		B		A		B		A		B			
	I	II	I	II	I	II	I	II	I	II	I	II	I	II	I	II	I	II	I	II
None			x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x
Non-negativity, spectroscopic modes																				
Non-negativity, all modes			x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Unimodality, spectroscopic modes																				

'x' means that the model was acceptable according to the criteria described in Section 1.1. An 'A' in the ribbon method means that the ribbon was inserted outside the first order Rayleigh scatter only and 'B' the ribbon was inserted on both sides of the data area.

AFAC constraints rather than the amount/placement of the zeros where decisive for the quality of the models. Unconstrained models and models with a non-negativity constraint on all three modes stand out as acceptable in contrast to models with a non-negativity or unimodality constraint on the spectroscopic modes only. The unconstrained models for all except the conventional method generally gave smoother loadings than the non-negativity constrained models for these methods, and fewer of them resulted in spurious peaks. The number of iterations before the model converged was only about 1/10 for the non-negativity constrained model relative to the corresponding unconstrained model—but the ratio in modeling time was more equal, as each iteration took more time for a non-negativity constrained modeling.

4. Conclusion

Inserting zeros instead of missing values in parts of the data area helps PARAFAC to converge faster, and leads to solutions that are physically and chemically meaningful. Insertion of zeros into the data matrices was also combined with constraining the PARAFAC model. Constraining only the spectroscopic modes had a detrimental effect on the model quality, while a non-negativity constraint on all three modes resulted in models that were essentially identical to the corresponding unconstrained models, but with less smooth loadings.

Future work lies in optimization of the method of inserting zeros. In this respect, several topics would be worth examining, such as evaluation of what kind of bandwidth of missing values to use for the ribbon method, and other, new and better methods of selecting the positions to insert zero values at. An investigation of different methods to initialize the PARAFAC model may also be appropriate.

Acknowledgements

Rinnan and Thygesen wish to thank the STVF (Danish Research Council) for financial support through project 1179. Barsberg wishes to thank the STVF (Danish Research Council) for financial support, and W. Kessler and R. Kessler, FH Reutlingen, Germany, for obtaining parts of the fluorescence data and for helpful discussions. The work on dry-cured hams was partly sponsored by an EU innovation program (OPUS, IN30905I), and J. Møller wishes to thank Dr. R. Bro for valuable advice during the preliminary data analysis.

References

[1] R. Bro, Chemometrics and Intelligent Laboratory Systems 38 (1997) 149–171.
[2] R. Bro, N.D. Sidiropoulos, A.K. Smilde, Journal of Chemometrics 16 (2002) 387–400.
[3] R.D. JiJi, K.S. Booksh, Analytical Chemistry 72 (2000) 718–725.
[4] R. Bro, Chemometrics and Intelligent Laboratory Systems 46 (1999) 133–147.
[5] C.M. Andersen, R. Bro, Journal of Chemometrics 17 (2003) 200–215.
[6] R. Bro, S. de Jong, Journal of Chemometrics 11 (1997) 393–401.
[7] R. Bro, N.D. Sidiropoulos, Journal of Chemometrics 12 (1998) 223–247.
[8] R. Bro, Multi-Way Analysis in the Food Industry-Models, Algorithms and Applications, PhD Thesis, Universiteit van Amsterdam, 1998, 135.
[9] C.A. Andersson, R. Bro, Chemometrics and Intelligent Laboratory Systems 52 (2000) 1–4.
[10] S. Barsberg, T. Elder, C. Felby, Chemistry of Materials 15 (2003) 649–655.
[11] J.K.S. Møller, G. Parolari, L. Gabba, J. Christensen, L.H. Skibsted, Journal of Agricultural and Food Chemistry 51 (2003) 1224–1230.

Paper V

Andersen, C.M., Rinnan, Å.

Distribution of water in fresh cod, *Lebensmittel Wissenschaft und Technologie*, **35**, 2002, 687-696

Lebensm.-Wiss. u.-Technol., 35, 687–696 (2002)

Distribution of Water in Fresh Cod

C. M. Andersen and Å. Rinnan*

C. M. Andersen, Å. Rinnan: The Royal Veterinary and Agricultural University, Department of Food Science, Food Technology, Rolighedsvej 30, DK-1958 Frederiksberg C (Denmark)
C. M. Andersen: Danish Institute of Fisheries Research, Department of Seafood Research, DTU Building 221, Søltofts Plads, DK-2800 Lyngby (Denmark)
(Received February 4, 2002; accepted July 11, 2002)

Low-field ^1H nuclear magnetic resonance (NMR) transverse relaxation was used to measure water mobility and distribution of water in fresh cod fillets. The NMR relaxations were analysed with the so-called SLICING method giving uni-exponential profiles from which the transverse relaxation time (T_2-values) and the relative sizes of the water populations were calculated. Two water populations with the T_2-values of 50 and 94 ms were obtained. The shortest relaxation time was primarily found near the head, and water with the longest relaxation time was primarily found near the tail. This variation can be explained by the smaller muscle cells and muscle fibers in the tail, which may influence the distributions of water into the different pools. The amount of one of the water populations was correlated to the overall water content with a correlation coefficient of −0.94.

Keywords: water distribution; water populations; NMR; SLICING

Introduction

Water is one of the most important components for the quality of food matrices including fish muscle. Water influences quality attributes such as appearance, texture and storage stability. The water content of fish muscles can be separated into different populations according to the mobility and how tight the water molecules are bound to the muscle structure. It is not only the total amount of water that is important for the overall quality of the product, but also its state and mobility (Ruan and Chen, 1998).

The heterogeneity within biological materials has implications on the results of most types of analyses made on that material. Therefore, it is important to know about this variation before analysing muscle-based food. Only very few studies have tried to characterize the heterogeneous distribution of the water content in lean gadoid fish species. Dambergs (1963) showed an increase in the overall water content of cod from the head-end to the tail-end. There was an inverse relationship between water and protein content within the fish muscle basically because these are the only significant components in the muscle.

This paper describes the distribution of water within cod fillets. The variation in water content and the variation in the populations of water within cod fillets will be investigated by measuring the water content physically and by measuring NMR relaxations. Low-field ^1H nuclear magnetic resonance (NMR) measures the mobility of protons and is therefore a direct technique for investigating both the total quantity of water and the state of water within the fish muscle. Low-field ^1H NMR has been used to measure and describe changes in properties of fish muscle occurring during frozen storage and processing (Lambelet *et al.*, 1995; Steen and Lambelet, 1997) and for investigating quality attributes of pork meat (Larsson and Tornberg, 1988; Brøndum *et al.*, 2000; Bertram *et al.*, 2001).

The use of a fast bench-top NMR hardware makes it possible and easy to acquire entire relaxation curves. In this paper, these will be analysed with a multivariate technique—SLICING—enabling interpretation of the underlying phenomena of the measurements. SLICING is a chemometric tool that resolves multivariate data, where the signals measured are an additive sum of exponentials. SLICING has been used for describing NMR relaxations of frozen and chill stored cod fillets (Jensen *et al.*, 2002), and minced and processed meat (Pedersen *et al.*, 2001).

*To whom correspondence should be addressed.
E-mail: aar@kvl.dk

0023-6438/02/$35.00
doi:10.1006/fstl.2002.0924

Materials and Methods

Materials

Five cod were caught in Øresund, Denmark, in August 2000. They were immediately brought to the laboratory where they were stored on ice for 5–7 days such that all fish were in a post-rigor state when they were measured. The five fish were coded A–E and were weighed before filleting. The weights were 1.0, 0.7, 0.6, 1.3 and 1.5 kg for the fish, A, B, C, D and E, respectively. Only one fillet from each cod was used for the analyses, since it was supposed that there was no difference between the two sides of the fish.

NMR measurements

The cod fillets were divided in squares of 1.5 cm^2 retaining information on the anatomical position of the square. The samples were taken from each of the squares such that one sample is measured for every 1.5 cm both in the horizontal and the vertical direction of the fillet. The number of samples measured was 59, 37, 38, 58 and 62 for the five fish, respectively, giving 254 samples in total. Each of the samples were weighted and prepared for the measurements. Blood, bone and red muscle, though, were manually excluded from the samples.

The measurements were performed on a Maran Benchtop Pulsed NMR analyser (Resonance Instruments, Witney, U.K.) operating at 23.2 MHz and equipped with an 18 mm variable temperature probe head. The receiver was adjusted to 3% and the receiver delay was set to 6 s. Previous experiments showed that a receiver delay of 6 s was enough for recovering of the magnet. Transverse relaxations were measured using the Carr–Purcell–Meiboom–Gill (CPMG) sequence (Carr and Purcell, 1954; Meiboom and Gill, 1958). For each measurement eight scans were performed with 1024 echoes and Tau at 500 μs. Tau is the 90–180° interpulse spacing in the CPMG sequence. Only even echoes were recorded, which gives 512 echoes measured for each sample.

All measurements were performed at 4 °C. Before measuring, the samples were equilibrated for 30 min at the chosen temperature. The fish were introduced into the NMR probe by placing samples of 2–4 g into glass tubes that matched the inner diameter of the 18 mm NMR sample tubes. The sample preparation was performed as carefully as possible in order not to alter the muscle structure by the manual treatment.

Water content

The water content was determined on exactly the same samples as were used for the NMR measurements, after the NMR relaxations were measured. The fish samples were kept in the small glass tubes and dried overnight at 110 °C. They were weighed before and after drying.

Data analysis

The measurements of six of the squares taken from the fillet were considered as outliers. Three of these were removed because of extreme water content, due to measurement errors. The three remaining outliers were removed due to high modeling residuals in X and Y, caused by errors in the NMR-measurements. The decay of mobile protons is measured by ^1H low-field NMR relaxations. In cod, the protons that can be measured with low-field NMR will almost exclusively be found in the water molecules. Thus, the amplitude of the signal will depend on the amount of water in the sample and thus the weight of the sample since all the samples contain approximately the same concentration of water. In order to properly handle the fact that the measured samples are of different weight, the measurements are normalized using maximum normalization. By this method, the

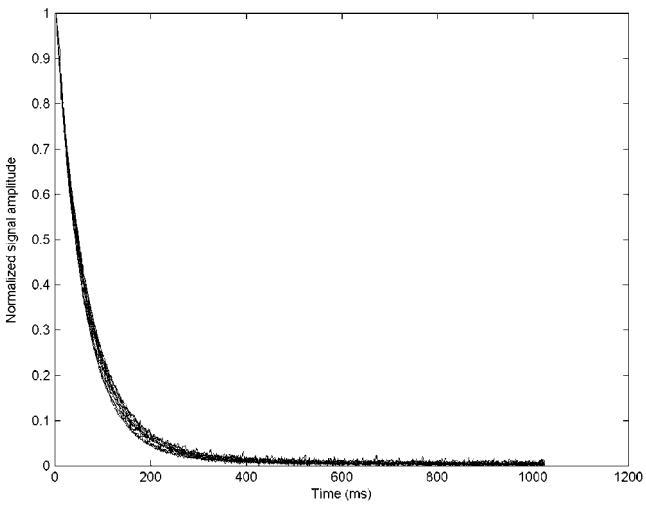


Fig. 1 Normalized CMPG curves. Each line represents the relaxation curve of one sample. Only some of the samples are shown

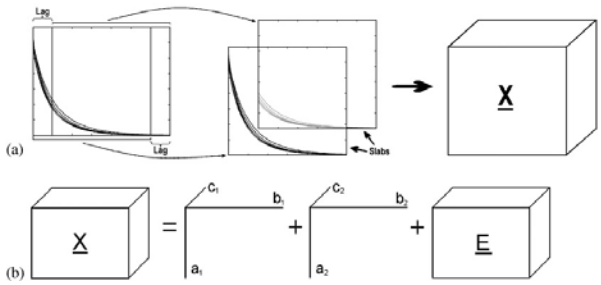


Fig. 2 (a) The SLICING method yielding three-way data from two-way data. (b) A visualization of the PARAFAC model. (b) is made on the three-way data from (a)

maximum of each relaxation curve is set to one and all other elements in the profile scaled accordingly.

Theory of SLICING. The normalized CMPG relaxation curves, illustrated in **Fig. 1**, are sums of exponentially decaying curves.

$$A(t) = \sum_{n=1}^{N} M_n e^{-t/T_{2,n}} \qquad \text{Eqn [1]}$$

Where $A(t)$ is the NMR signal, N is the number of exponential components, and M_n is the amplitude of the nth exponential and thus a measure of the relative concentration. $T_{2,n}$ is the corresponding spin–spin relaxation time constant and t is the acquisition time. Due to the normalization, the sum of the amplitudes (M_n) is one. Equation [1] represents the ideal case of the NMR signal. In real samples there will always be some noise, and a noise factor should then be included in the equation.

SLICING is a method based on the principles of direct exponential curve resolution algorithm (DECRA) (Windig and Antalek, 1997). The idea is to split the spectrum into two (or more) overlapping parts (slabs), where the size of the overlap is determined by the lag term generating a three-dimensional array. Most of the relaxation curve is present in both slabs. This operation is illustrated in **Fig. 2a**. The dimensionality of the matrix will then increase from I (samles) $\times$ L (measurement points) to $I \times$ (L-maximum *lag*) $\times$ number of slabs (K). Below, this will simply be shortened to $I \times J \times K$.

The three-dimensional array (**X**) has the size $I \times J \times K$ and contains the elements x_{ijk}, where the first index (i) refers to the samples, the second (j) refers to the time and the third (k) refers to the slab number. It can be shown that the rearranged three-way data follow a so-called PARAFAC model (Eqn [2]) (Harshmann, 1970; Bro, 1997) when the original data is of the form described by Eqn [1]

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

$$(i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K) \qquad \text{Eqn [2]}$$

The element x_{ijk} is the original value in the position (i, j, k) of the data cube **X**. $a_{\bullet f}$ is the object score (magnitude) for factor f (first mode), $b_{\bullet f}$ is the estimated decay curve for the pure component f (second mode), and loading $c_{\bullet f}$ gives the ratio between the different slabs (third mode). The term e_{ijk} contains residual variation not captured by the model. Another way of presenting the PARAFAC model is shown in **Fig. 2b**. The **X**-cube represents the sliced data array. In the figure, the array is decomposed into two sets of factors, where each of these sets is called a triad. The factors (triads) are found simultaneously via an alternating least-squares algorithm (Bro, 1997) and are represented by a set of a–c. The **E**-cube to the right represents the noise that is left unmodeled. Noise is here defined as the variation in the data that contains no chemical information. Ideally, the estimated decay curves (b) are uni-exponential and by fitting one exponential to these, the corresponding T_2-value can be found.

If the residuals show random behavior and no systematic trend, it can be presumed that only noise is left unexplained and hence the N estimated profiles explain the variation in the data up to the noise. Furthermore, if the model is adequate each second mode loading should be exponential because the PARAFAC model can be shown to uniquely recover the underlying model when correctly specified (Windig and Antalek, 1997). If too many components are extracted, the curves will reflect this, one or more being nonexponential. The appearance of the spectral loadings, bootstrapping (Wehrens *et al.*, 2000) by the use of split-halves (Harshmann and de Sarbo, 1994) and the distribution of the residuals will be used to estimate the right number of components.

Results and Discussion

According to Love (1970) there are no systematic differences between the chemical composition of the left and the right fillets of cod muscles. Thus, analyses performed on one of the fillets (left or right side) are used in this study to determine the general distribution of water within a fish.

Distribution of the water content

Figure 3 visualizes the distribution of the water content over the entire fillet. The figure shows the water content

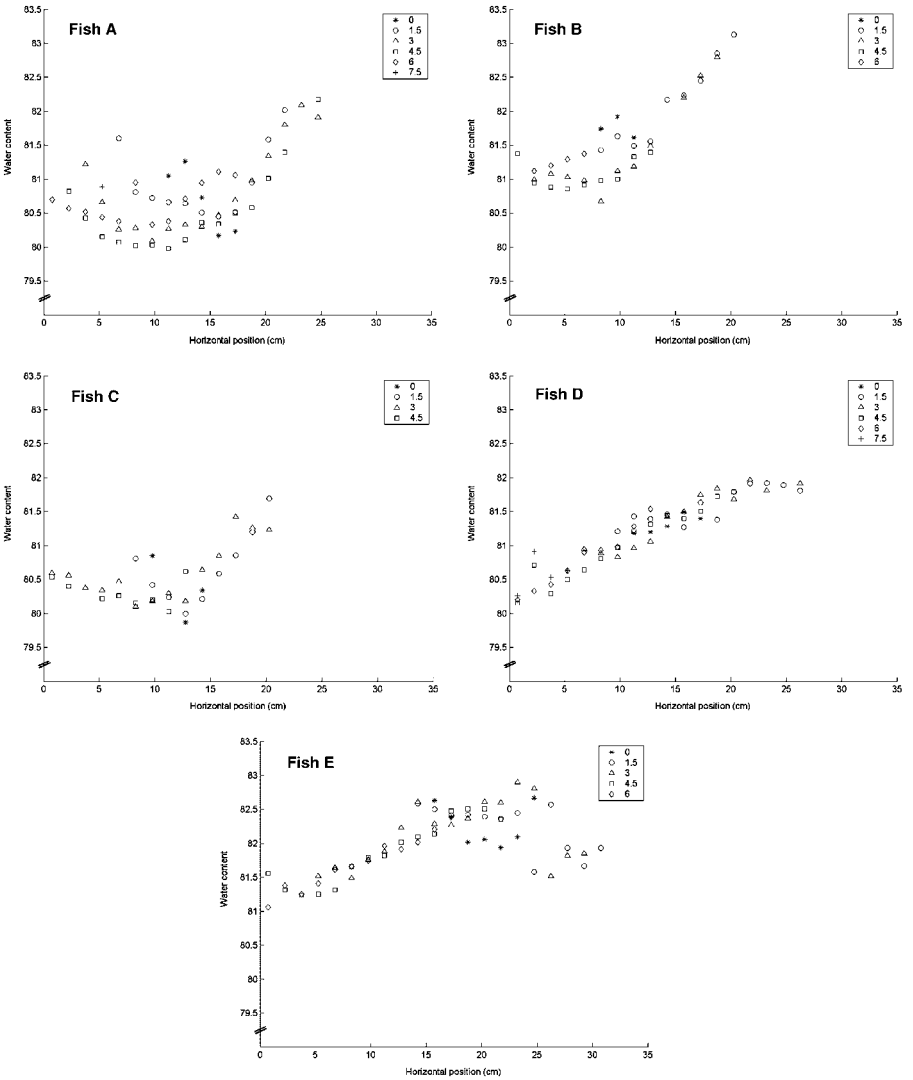


Fig. 3 Distribution of the water content within the fillets of the five fish. The water content is denoted in g water per g fish muscle and is shown as a function of the distance from the head (position 0). Differences across the fillet are shown with different marks: (*) is the sample closest to the belly, (○) is 1.5 cm from the belly, (△) is 3 cm from belly, (□) is 4.5 cm from belly, (◇) is 6 cm from belly and (+) is 7.5 cm from belly

as a function of the distance from the head. Generally, there is an increase in the water content going from the head to the tail. It seems that the smaller fish A–C have a similar water content distribution. Going from the head to the tail, the water content is getting a little lower towards the middle of the fish. Beyond that the water content increases with the highest content found at the tail-end of the fish. Fish D and E do not have the same distribution. The water content increases constantly from the head towards the tail. However, at some point the water content stops increasing, and in the last points it even decreases, especially for fish E. It is probable that this is reflecting a fundamental difference between smaller and bigger fish, but this cannot be investigated further with the present data.

The results correspond to the results obtained by Dambergs (1963). There, cod fillets were divided in three parts (head, middle/belly and tail) and showed an increase in water content going from the head to the tail. The results obtained by Dambergs (1963) were found as

an average over the whole year and it is likely that the same variation would have been obtained in this study if more experiments were performed at other seasons. The opposite variation in water content was found in catfish and herring (Brandes and Dietrich, 1953; Jafri, 1973), which both have a higher fat content than cod.

Cell sizes and muscle fibers are smaller towards the tail compared with the rest of the fillet. This may have an influence on the distribution of the water content within the fillet corresponding to the results shown in the figure. Furthermore, the lower water content around 10 cm from the head seen for fish A–C can be explained by the fact that the largest muscle cells and muscle fibers are found around myotome number 12 (Love, 1988). An effect of sampling may explain some of the unexpected variations such as the decrease in water content in the outermost end of the tail. In some cases especially in the tail where very small samples are obtained, it is possible that a little part of red muscle is included in the samples because of the difficulty in separating red and white muscle in those small pieces of muscle flesh. Since red muscle has lower water content (Mannan *et al.*, 1961), it can explain why the water content decreases at the tail end.

SLICING modeling

The two-dimensional data were sliced using a lag of 1, and with two slabs, increasing the dimensionality of the matrix from 248×512 to $248 \times 511 \times 2$.

Choosing the optimal number of factors. **Figure 4** illustrates the results obtained from slicing models with two or three factors. From the loading plot (**Fig. 4a**) it is clear that the two first factors are both exponential, indicating that at least two populations of water are present. It is imperative that the right number of components are chosen in the model, because all components change with the total number of components. For example, the faster relaxing factor has a T_2-value of 49.8 and 49.5 ms for the two- and three-factor models, respectively. The second T_2-values are 104.3 and 81.5 ms. The third factor for the three-factor model is probably not exponential, since it does not seem to decrease towards zero, but rather to a limit of approximately 0.03. There might be two reasons for this behavior: either the model is overfitted, or there is some kind of offset in the data. The offset may be due to the way data are collected in that all data were constrained to have positive values (H.T. Pedersen,

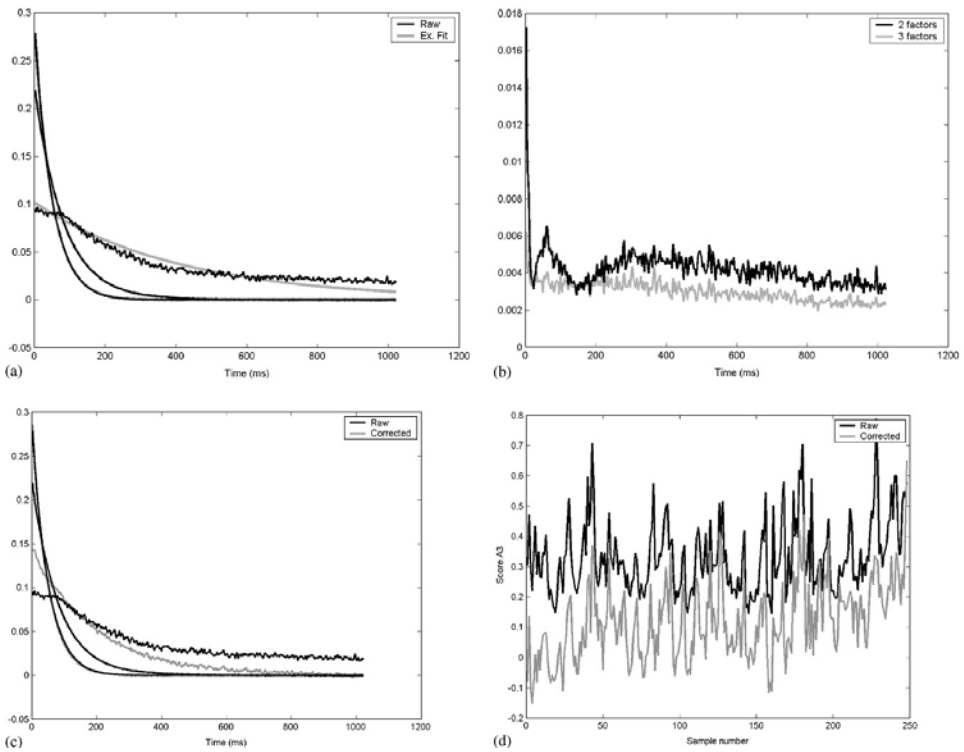


Fig. 4 (a) Second mode loadings (black) and exponential fit (gray) after three components. (b) Squared residuals summed over relaxation time from PARAFAC models with two (black) and three (gray) factors. (c) The B-loadings from three-factor models on the raw data (black) and the corrected data (gray). (d) The A-scores for the third factor. Black is the model on the raw data and the gray is the model on the corrected data

pers. commun.). **Figure 4b** shows the residuals along the time axis for the two- and three-factor models. From this figure, it is clear that some information is gained by going from two to three factors. Thus, application of a three-factor model could be an advantage. Furthermore, when increasing the model complexity from two to three factors, the residuals become less structural as can be seen in **Fig. 4b**.

Corrections of the data. The possible presence of a minor third factor could be caused by an offset effect (**Fig. 4a**), as noted above. In order to investigate if the model could be improved by correcting for this artifact, the mean of the last 50 measuring points are subtracted from the whole spectra. The correction is done sample wise and is based on the assumption that the decays are measured sufficiently long so that the signal should effectively be zero unless there are offsets. As can be seen from **Fig. 4c**, the second mode loadings for the third factor has become exponential by the correction indicating the presence of a third minor factor. However, upon studying the scores (**Fig. 4d**) of the third factor, it becomes clear that some of these are negative, which gives no chemical meaning. Thus, only the two main factors give strictly positive score values.

Bootstrapping. To support the conclusion of a two-factor model, and further investigate the effect of correcting the data, bootstrapping (Wehrens *et al.*, 2000) is performed. Two- and three-factor SLICING models are made on 100 split-halves (Harshmann and de Sarbo, 1994) of the dataset with raw data and of the corrected data, giving rise to 4×200 models in total.

For comparing the results from the bootstrapping, uni-exponential fits to the second mode loadings are made. Mean and standard errors of the uni-exponential fits are calculated and are given in **Table 1**. Since the sample sets contain the same water populations, the standard error of the T_2-values should be as small as possible. Large standard errors indicate an unstable model, hence too many factors. By investigating the results in **Table 1**, it becomes clear that for both datasets the standard error of the third factor is very large. Furthermore, it was found that only 45% of the three-factor models made with corrected data have three exponential second mode loadings. The correction seems appropriate for the two-factor model since the standard error decreases going from the noncorrected to the corrected data. The bootstrapping thus supports the conclusion that the two-factor model is the optimal, but that the corrected data should be used.

For the two-factor model made on all data corrected for the offset, the fast relaxing factor has a T_2-value of approximately 50 ms and the slower relaxing factor has a T_2-value of 94 ms. Thus, one of the water populations is relatively tightly bound to the muscle structure and the other is less tightly bound.

Lambelet *et al.* (1995) obtained T_2-values of 1 and 65 ms by exponential fitting. The low T_2-value was only obtained when tau was as low as 12 μs. In the present experiment a tau of 500 μs was used and it therefore seems reasonable that a component with a T_2-value of 1 ms could not be detected. Modeling only one component gave a T_2-value of approximately 66 ms, which agrees with the slower relaxing component obtained by Lambelet *et al.* (1995). It seems that SLICING gives a possibility for identifying two water populations with great similarities, which could not be identified with the bi-exponential fitting as was used by Lambelet *et al.* (1995). Another study made on frozen–thawed cod gave three exponential components (Jepsen *et al.*, 1999). These had T_2-values of 34, 62 and 526 ms. Four populations of water with T_2-values of 37, 56, 126 and 361 ms were obtained by the SLICING technique applied on NMR-relaxations of minced frozen–thawed cod (Jensen *et al.*, 2002). The two components with the T_2-values of 56 and 126 could correspond to the water populations obtained in the present study. Thus, it seems as two more water populations are developed during processing such as mincing or freezing. Furthermore, the variation in T_2-values between this study and other studies may be due to variations in the sample preparation, season, places of catch, etc. Differences in instrumentation, parameters used and the ways of analysing the data may also cause a variation in the experimentally determined T_2-values.

For whole, minced and homogenized pork three exponential components were obtained by distributed exponential fitting (Fjelkner-Modig and Tornberg, 1986; Larsson *et al.*, 1988; Bertram *et al.*, 2001). As mentioned, only two components were obtained in the present study. It might be that fresh fish muscle in contrast to pig muscle contains only two populations of water. Furthermore, the variation and/or the content of the third water population may be so small that it cannot be described by the method used here.

Table 1 Mean relaxation times (in ms) and the standard error of second mode loadings obtained from two and three factor bootstrapping on 100 split-half models performed on both noncorrected and corrected data.

	Uncorrected data				Corrected data			
	Two factors		Three factors		Two factors		Three factors	
	Mean	s.e.	Mean	s.e.	Mean	s.e.	Mean	s.e.
$T_{2,1}$	49.8	0.6	49.5	0.6	49.6	0.4	46.5	1.8
$T_{2,2}$	104.3	2.8	81.5	1.7	93.7	1.2	82.6	5.7
$T_{2,3}$			467.7	100.8			203.1	73.6

The number of samples for these tests is 248.

Interpretation of the SLICING model

Distribution of the different types of water between the cod fillets. The score values (first mode) describe how much each water population contributes to the NMR relaxation phenomena and may thus be used as a measure of the relative concentrations of the populations of water. **Figure 5** shows a score plot of factor one (relating to $T_{2,1}$) *vs*. factor two (relating to $T_{2,2}$) of all the fish. The plot is coded according to each of the fish (A–E). The ellipses are made for helping in the interpretation of the plot. The ellipses are centered in the mean of each fish along each of the two axes. The range in each direction is the standard deviation for that fish's scores in the specific direction of variation.

By comparing **Fig. 5** with **Fig. 3**, it might seem odd that fish B and D look so similar in **Fig. 5**. However, by investigating **Fig. 3** further, and taking into account that the NMR-signal is mostly due to the water in the fish, it comes as no surprise that fish B and D overlap in **Fig. 5**. The range of the water content for fish B is higher than for fish D, but the range of the water content of fish D is inside the range of fish B. Therefore, it is expected that the ellipse for fish D should be inside the ellipse for fish B, which also is the case in **Fig. 5**.

From **Fig. 5,** it is seen that the samples from the five fish are grouped together illustrating some variation among the fish. The grouping shown in **Fig. 5** can be related to the average water content of the five fish. Fish E has the highest water content of 82.02/100 g, fish B and D have the next highest of 81.53/100 g and 81.20/100 g, respectively, whereas fish A and C have the lowest water content of 80.74/100 g and 80.53/100 g, respectively. There is an inverse relation between these two types of water. The reason for this is that by normalizing the spectra to one, the scores will approximately add up to a constant value. Thus, when there are only two factors in the model and the noise level of the raw data is low, the score values will be highly correlated.

Comparing heterogeneity within the fish. **Figure 6** illustrates the scores of the first factor *vs*. the relative position of each sample. The scores are obtained from the two-factor SLICING model made on all fish with correction for offset. The results will be more or less the same if the figures are made using the results obtained from PARAFAC models made on each fish (figure not shown). Further, if the scores from the second factor were used instead of the scores from the first, the same results would be obtained. This, however, is of no surprise since the two factors are highly correlated.

There is a clear relation between the scores obtained from the PARAFAC model and the position of the sample. The black line in each figure is the 'best line' fitted to minimize the distance from every sample to the line (corresponding to the first component in a PCA). The correlation coefficients illustrate that fish A, C and E are fairly homogenous, whereas the largest correlation between the score values and the horizontal position is found for fish B (**Table 2** and **Fig. 6**). However, the same results are not obtained by looking at the range of the values, where the greatest variation is found for fish E. The low correlation and large range obtained for fish E seems to be due to a curvature in the score-values, and is consistent with what is seen in **Fig. 3**. Fish A has two samples close to the head, which have fairly low score-values compared to the others. By removing these two samples, the correlation coefficient increases to −0.66. It

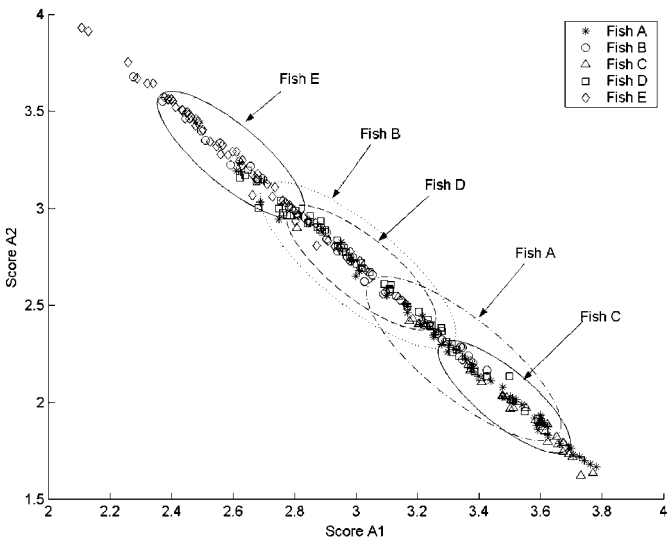


Fig. 5 Scatter plot of the scores of factor 1 *vs*. the scores of factor 2 obtained from a two-factor PARAFAC model on the raw data. Samples are coded according to fish: fish A (*), fish B (○), fish C (△), fish D (□) and fish E (◇)

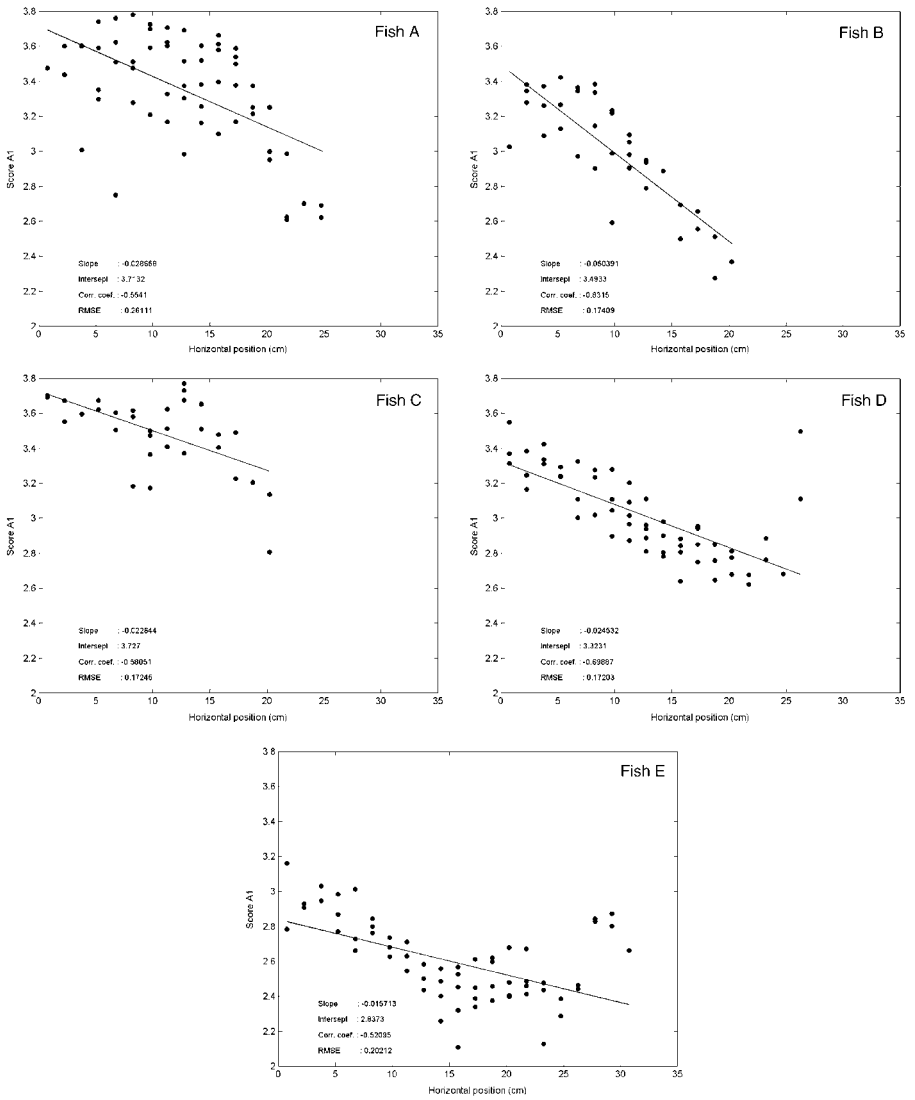


Fig. 6 Plot of the horizontal position of each sample *vs*. the A-scores of the first factor obtained from a two-factor PARAFAC model on the raw data

Table 2 Correlation between the first score and the horizontal position (r), and the range in the first score. Both are done for each fish separately

	A	B	C	D	E
No. of samples	59	36	34	57	62
r	−0.57	−0.85	−0.69	−0.74	−0.75
Range	1.20	1.05	0.88	1.06	1.09

becomes clear that fish A is homogenous up to the tail-end where the score values start to drop. Fish C seems to be homogenous in the middle of the fish, while to both sides of the middle there is a steady decrease of the score values. The reason why fish A and C in some part of the fillet are more homogenous than the other fish is not known.

There are some samples near the tail of fish D, which clearly do not follow the trend of the other samples. This might be due to some red muscle being included in the samples, and as mentioned earlier, red muscle contains less water than white muscle and will thus give a different signal. If these few samples at the tail-end of the fish are removed, the results for fish D improve,

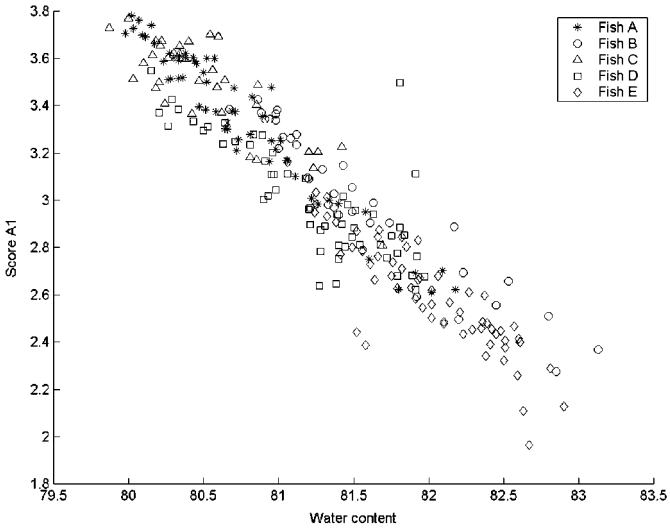


Fig. 7 Relation between the water content and the score values of the first factor

giving a correlation coefficient of −0.88. Thereby, the relation between the score values and the sample position increases resulting in a larger correlation coefficient than is obtained for fish B.

The negative correlation between the score values of the two factors means that samples with a high content of the water population described by factor one will have a low content of the water population described by factor two. Thus, samples from the head part contain more of the fast relaxing component, whereas samples from the tail part contain more of the slow relaxing component. This is consistent with theory. The muscle fibers are smaller towards the tail where the cell sizes also get smaller (Love, 1970). It is suggested that a separation of the water into two populations is due to an anatomical compartmentalization. The water with the shortest relaxation time is likely intracellular water and the water with the longest relaxation time is extracellular water (Cole *et al.*, 1993). In that case the intracellular water that is found in a higher amount near the head is bound more firmly to the muscle fibers than the extracellular water. The more loosely bound extracellular water is found in a larger amount near the tail. The relative size of intra- and extracellular water is in agreement with the variation in cell and muscle fiber sizes. Furthermore, the distribution of the score values is in accordance with the theory of the separation of water into the intracellular and the extracellular parts.

Above, it was shown that the score values of the first factor decrease when samples are taken from a position closer to the tail. This was also the case for the overall water content (**Fig. 3**). Therefore, the relative amount of each water population may correlate to the total amount of water. This is illustrated in **Fig. 7** where the water content is plotted against the score values of the first factor. It is illustrated that there is a linear relation between the two parameters. Furthermore, a correlation coefficient of −0.94 is found. This variation in water content and content of the different water populations illustrates the importance of how sampling is performed when instrumental measurements are performed on a small part of the fish.

As mentioned, other studies have already shown the multi-exponential distribution of water within muscle tissue. However, low-field NMR combined with SLICING is a new way of analysing the states of water in various products. Compared with exponential fitting, SLICING is a fast, easy and stable method giving the underlying exponential decays of the NMR relaxations. When analysing the NMR relaxations of fresh cod fillets, it has been shown that two populations of water can be identified. The concentration of the populations varies within the length of the cod fillet and seems to be related to the overall water content. Thus, these results illustrate the possibility for studying the distribution of the populations of water between and within other fish species.

Acknowledgements

Andersen thanks for financial support through Advanced Quality Monitoring (AQM), while Rinnan wish to thank the STVF (Danish Research Council) for financial support through project 1179. Rasmus Bro, Frans van den Berg and Bo Jørgensen are gratefully acknowledged for valuable advice during the data analysis and preparation of the manuscript.

References

Bertram, H. C., Karlsson, A. H., Rasmussen, M., Pedersen, O. D. Dønstrup, S. and Andersen, H. J. Origin of

multiexponential T_2 relaxation in muscle myowater. *Journal of Agricultural and Food Chemistry*, **49**, 3092–3100 (2001)

Brandes, C. H. and Dietrich, R. Die Fettverteilung im Koerper des Herring. *Veroffentliches Institut für Meeresforshung in Bremerhaven*, **2**, 109–121 (1953)

Bro, R. Parafac: tutorials and applications. *Chemometrics and Intelligent Laboratory Systems*, **38**, 149–171 (1997)

Brøndum, J., Munck, L., Henckel, P., Karlsson, A., Tornberg, E. and Engelsen, S. B. Prediction of water-holding capacity and composition of porcine meat by comparative spectroscopy. *Meat Science*, **55**, 177–185 (2000)

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

Cole, W. C., Leblanc, A. D. and Jhingran, S. G. The origin of biexponential T2 relaxation in muscle water. *Magnetic Resonance in Medicine*, **29**(1), 19–24 (1993)

Dambergs, N. Extractives of fish muscle. 3. Amounts, sectional distribution, and variations of fat, water-solubles, proteins and moisture in cod (*Gadus morhua* L.) fillets. *Journal of Fisheries Research Board, Canada*, **20**, 909–918 (1963)

Fjelkner-Modig, S. and Tornberg, E. Water distribution in porcine *M. longissimus dorsi* in relation to sensory properties. *Meat Science*, **17**, 213–231 (1986)

Harshmann, R. A. Foundations of the PARAFAC procedure: model and conditions for an "explanatory" multi-mode factor analysis. *UCLA Working Papers in Phonetics*, **16**, 1–84 (1970)

Harshmann, R. A. and De Sarbo, W. S. An application of PARAFAC to a small sample problem, demonstrating preprocessing, orthogonality constraints and split-half diagnostic techniques. In: Law, H. G., Snyder, C. W., Hattie, J. A. and McDonald, R. P. (Eds), *Research Methods for Multimode Data analysis*. New York: Praeger (1994)

Jafri, A. K. Fat and water distribution patterns in the flesh of the common cat-fish *Wallago attu* (bl. & schn.). *Fishery Technology*, **10**, 138–141 (1973)

Jensen, K. N., Guldager, H. S. and Jørgensen, B. M. Three-way modelling of NMR relaxation profiles from thawed cod muscle. *Journal of Aquatic Food Product Technology* (2002) in press

Jepsen, S. M., Pedersen, H. T. and Engelsen, S. B. Application of chemometrics to low-field ^{1}H NMR relaxation data of intact fish flesh. *Journal of the Science of Food and Agriculture*, **79**, 1793–1802 (1999)

Lambelet, P., Renevey, F., Kaabi, C. and Raemy, A. Low-field nuclear magnetic resonance relaxation study of stored or processed cod. *Journal of Agricultural and Food Chemistry*, **43**, 1462–1466 (1995)

Larsson, G. and Tornberg, E. An attempt to relate meat quality og pork (*M. longissimus*) to meat structure. *34th International Congress of Meat Science and Technology. Congress Proceedings Part B*, pp. 588–591 (1988)

Love, R. M. *The Chemical Biology of Fishes*. New York: Academic Press Inc. (1970)

Love, R. M. *The food fishes, their intrinsic variation and practical implications*. London: Farrand Press, (1988)

Mannan, A., Fraser, D. I. and Dyer, W. J. Proximate composition of Canadian Atlantic fish 1. Variation in composition of different section of the flesh of atlantic halibut (*Hippoglossus hipoglossus*). *Journal of Fisheries Research Board, Canada*, **18**(4), 483–493 (1961)

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

Pedersen, H. T., Bro, R. and Engelsen, S. B. A novel approach for unique deconvolution of NMR relaxation decays. In: Webb, G. A., Belton, P. S., Gill, A. M. and Delgadillo, I. (Eds), *Magnetic Resonance in Food Science: A View to the Future*. Cambridge, UK. The Royal Society of Chemistry, pp. 202–209 (2001)

Ruan, R. R. and Chen, P. L. *Water in Foods and Biological Materials. A Nuclear Magnetic Resonance Approach*. Lancaster: Technomic Publishing Co., Inc (1998)

Steen, C. and Lambelet, P. Texture changes in frozen cod mince measured by low-field nuclear magnetic resonance spectroscopy. *Journal of the Science of Food and Agriculture*, **75**, 268–272 (1997)

Wehrens, R., Putter, H. and Buydens, L. M. C. The bootstrap: a tutorial. *Chemometrics and Intelligent Laboratory Systems*, **54**, 35–52 (2000)

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

Paper VI

Povlsen, V.T., Rinnan, Å., van den Berg, F., Andersen, H.J., Thybo, A.K.

Direct decomposition of NMR relaxation profiles and prediction of sensory attributes of potato samples, *Lebensmittel Wissenschaft und Technologie*, **36**, 2003, 423-432

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

lwt

www.elsevier.com/locate/lwt

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

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

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

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

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

Received 13 March 2002; accepted 12 January 2003

Abstract

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

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

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

1. Introduction

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

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

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

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

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

2. Materials and methods

2.1. Potatoes

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

2.2. Sensory analysis

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

2.3. NMR measurements

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

Table 1
Tuber samples used in the experiments

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

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

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

2.4. Data treatment

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

3. Data analysis and modelling

3.1. Description of the NMR curves

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

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

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

3.2. Regression by PLS on the raw CPMG curves

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

3.3. Bi-exponential fitting

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

3.4. Distribution analysis

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

4. SLICING

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

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

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

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

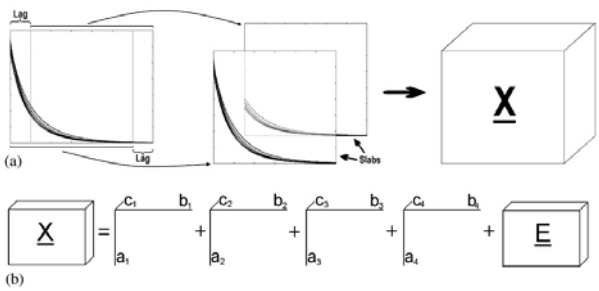


Fig. 1. Going from NMR signal to the data cube for PARAFAC modelling: (a) illustrating the principles of creating a three-way array from NMR relaxation curves; (b) the data cube $\boldsymbol{X}$ [23 × 1991 × 3] is decomposed into four triads with sample scores **a** (23 exponential loadings), **b** (1991)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

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

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

4.1. Validation

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

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

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

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

5. Results and discussion

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

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

5.1. Data analysis using SLICING

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

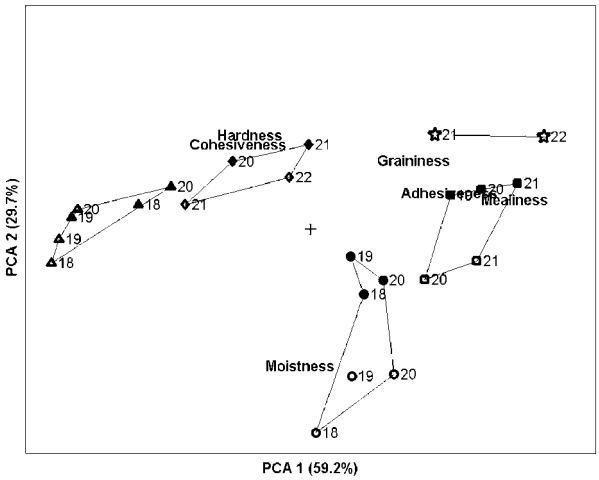


Fig. 2. Bi-plot from PCA on the sensory data. Ditta (◇), Sava (Δ), Bintje—low dry matter (□), Bintje—high dry matter (★), and Berber (○) for storage 1999 (open) and 2000 (filled). The numbers from 18 to 22 represent the % dry matter bins (see Table 1) where the range of the % dry matter bin is 18: 18.0–18.9,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

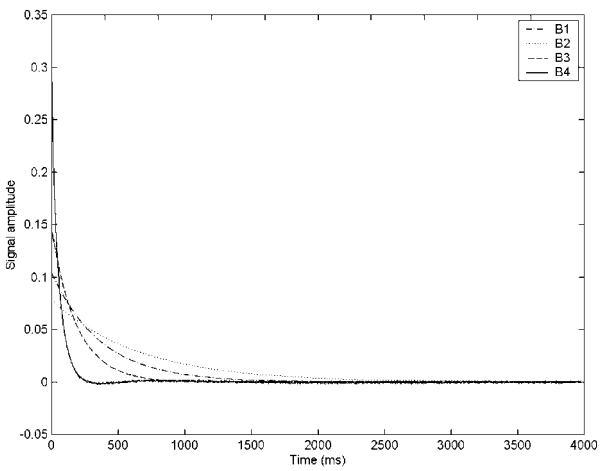


Fig. 3. Exponential loadings for components one to four from the PARAFAC model.

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

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

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

Table 2
Overview of the transversal relaxation time (T_2)

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

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

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

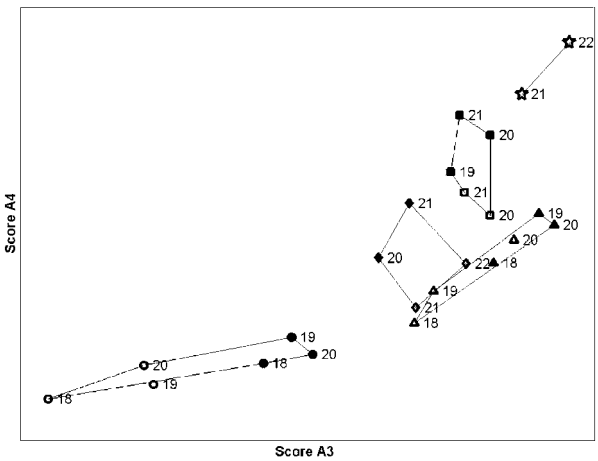


Fig. 4. Score plot of sample score 3 and 4 for the PARAFAC . Ditta (◇), Sava (Δ), Bintje—low dry matter (□), Bintje—high dry matter (★), and Berber (○) for storage 1999 (open) and 2000 (filled). The numbers from 18 to 22 represent the % dry matter bins (see Table 1) where the range of the % dry matter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

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

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

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

5.2. Regression models

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

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

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

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

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

6. Conclusion

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

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

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

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Paper VII

Rinnan, Å.

Alternative regression method to PLS on NMR-relaxation curves, *Rivista di Statistica Applicata – Italian Journal of Applied Statistics*, **15 (3)**, 2003, 393-402

ALTERNATIVE REGRESSION METHOD TO PLS ON NMR-RELAXATION CURVES

Åsmund Rinnan

Food Technology, Dept. Dairy and Food Science, The Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark, aar@kvl.dk

ABSTRACT:

Predictions with NMR-relaxation curves as descriptor variables are commonly performed by PLS regression (PLSR). An alternative method for regression modelling is based on the scores from 'slicing'. Slicing is a method based on the principles of DECRA and serves as an alternative tool to decompose NMR relaxation curves. The loadings from slicing have the advantage, compared to PLSR, of being directly interpretable in terms of underlying exponentials. The scores from slicing provide information about the samples, and may in addition be used for regression. In this work it will be shown that this procedure gives as good predictions as PLSR and also has exploratory advantages.

Key words: PLSR, slicing, PARAFAC, prediction, NMR relaxation

INTRODUCTION:

In the food industry, fast prediction methods are called for e.g. for quality control. Predictions based on NMR-relaxation measurements have become more popular in recent years. The use of PLS [5] on such data for regression analysis gives good results [9]. In this work an alternative method will be demonstrated, where the dimension of the data-table is rearranged into a data-cuboid, by the use of 'slicing' [11, 12]. A PARAFAC (PARAllel FACtor analysis) model [3, 6] is fitted to the three-way array, and the scores are used as descriptor variables. There is little or no gain in the prediction, but in the interpretation of the results, there is an advantage in using slicing. This method is capable of estimating the underlying exponential decays, and thus gives more information about the system than the loadings from PLSR do. In this paper the rearrangement into a data-cuboid and the following decomposition by the use of PARAFAC will be referred to as slicing.

The use of PLSR in the prediction of food quality is not new. Both Thybo et al. (2000) [16] and Thygesen et al. (2001) [17] successfully used PLSR in the

prediction of sensory attributes of potatoes. Povlsen et al. (2002) [14] has showed that the use of slicing gives equally good predictions, but with a gain in the understanding of the potatoes. Slicing has further been used in the decomposition of NMR relaxation curves of frozen and chill stored cod [8], on processed meat [11], and fresh cod [2]. In this paper more data sets have been investigated in order to verify these previous results.

In the next section of the paper, there will be a description of the slicing method, followed by the comparison of PLSR and slicing on four different datasets. The comparison will mainly be on the quantitative perspective, but there will be some comments as to the advantages of slicing to PLSR when it comes to understanding and interpretation of the data.

THEORY AND METHODS

NUCLEAR MAGNETIC RESONANCE (NMR) SPECTROSCOPY

Nuclear Magnetic Resonance (NMR) spectroscopy can provide useful information about molecular structure within the sample and has become a powerful non-destructive analytical tool in chemistry. In the present study the relation between low field proton NMR spectroscopy (^{1}H-NMR) from Carr-Purcel-Meiboom-Gill (CMPG) pulse relaxation curves and different predictor variables in various food products is investigated. The relaxation signals used in this paper are transverse relaxation curves, and can ideally be written as a sum of exponentials:

$$I(t) = \sum_{n=1}^{N} M_{0,n} \cdot \exp\left(-\frac{t}{T_{2,n}}\right) + E \qquad \text{(Eq. 1)}$$

Where N is the number of exponentials, $T_{2,n}$ is the transversal relaxation time and $M_{0,n}$ is the magnitude value. The last term E is an error term, and should be as small as possible.

CHEMOMETRICS

PARAFAC is a model that can be fitted on cuboids or data-arrays of higher dimension. It is closely related to and an extension of Principal Component Analysis – PCA [5, 19]. There are, however, two major differences between PARAFAC and PCA, except the dimension of the data set to be analysed. In PARAFAC, the factors (triads, see Figure 1) are found simultaneously via Alternating Least Squares algorithms. This differs from PCA where the factors often are computed one after the other, gradually increasing the model complexity. Another important difference is the freedom of rotation that is present in the bilinear model. To handle this, the results from PCA are restricted such that the first component explains the maximum

variance; the next is orthogonal and explains maximum residual variance etc. Due to the necessity for these abstract restrictions PCA cannot provide unique estimates of the underlying components, even if the true model is bilinear. For PARAFAC the factors do not have rotational freedom, and thus the solution is essentially unique [10, 15]. This means that the parameters from PARAFAC can be directly interpreted as estimates of 'pure' components when the data follow a low-rank trilinear model.

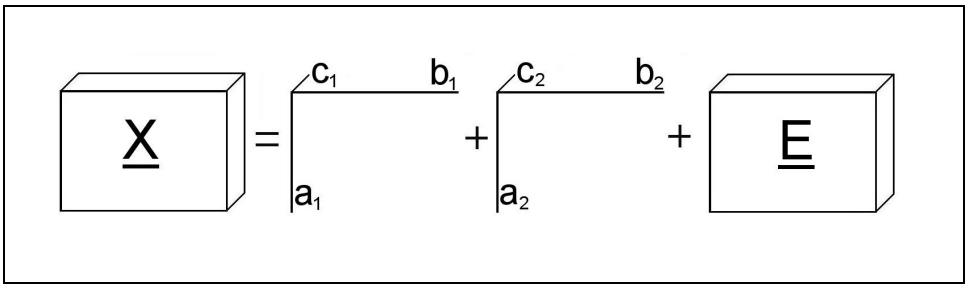


Fig. 1: The decomposition by PARAFAC of a data-cuboid X into triads and a residual-cuboid E.

Slicing is a method based on DECRA which again is intrinsically connected to PARAFAC [18], and can only meaningfully be performed on data consisting of exponential decaying signals. NMR relaxation curves are of such a nature, and thus this method can be used. A data-table containing NMR relaxation curves of the mathematical form shown in equation 1 above, follow a PARAFAC model [18], if it is properly rearranged, as is the case for slicing. The idea behind slicing is to split the spectrum into two (or more) overlapping parts (slabs), where the size of the overlap is determined by the lag term, thus generating the three dimensional array. Most of the relaxation curve is present in all the slabs. This operation is illustrated in Figure 2. The size of the data-matrix is thus increased from $I \times J$ to $I \times (J\text{-}lag \times (slabs\text{-}1)) \times slabs$. It is also

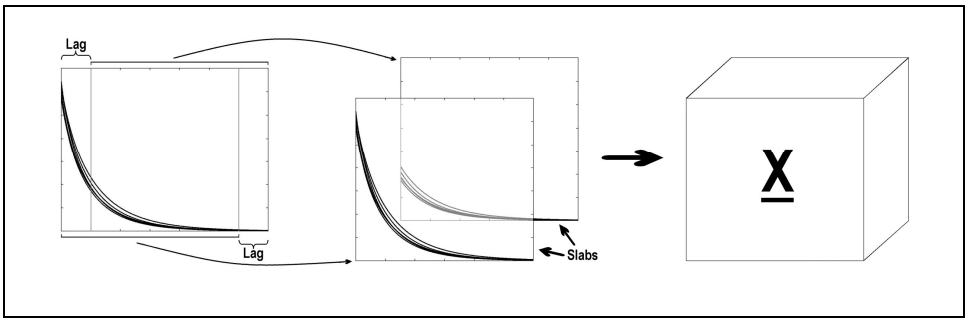


Fig. 2: Slicing of the raw CMPG relaxation curves into a three dimensional array.

possible to have different lags for the different slabs. The effect of this has not been investigated closely. The idea, however, is that a short lag extract the faster decaying components, while the slower decaying components is easier extracted by the use of a longer lag.

The PARAFAC model derives estimates of the true underlying exponential decays from the generated data cube. Each resulting relaxation loading can be fitted by a single exponential decay (N=1 in the above equation). Using slicing for the decomposition of the relaxation curves gives simple estimates of the underlying exponential decay curves, which directly can be related to proton properties. This is not the case for the loadings from PLSR models, where the NMR relaxation curves are used as descriptor variables. The PLSR models are mathematically forced to obey certain orthogonality constraints [7], and thus the loadings from PLSR do not hold the same information as the PARAFAC loadings (see Figure 3.) In the regression step for the sliced data, the scores from PARAFAC are used as the descriptor variables in a multivariate regression model.

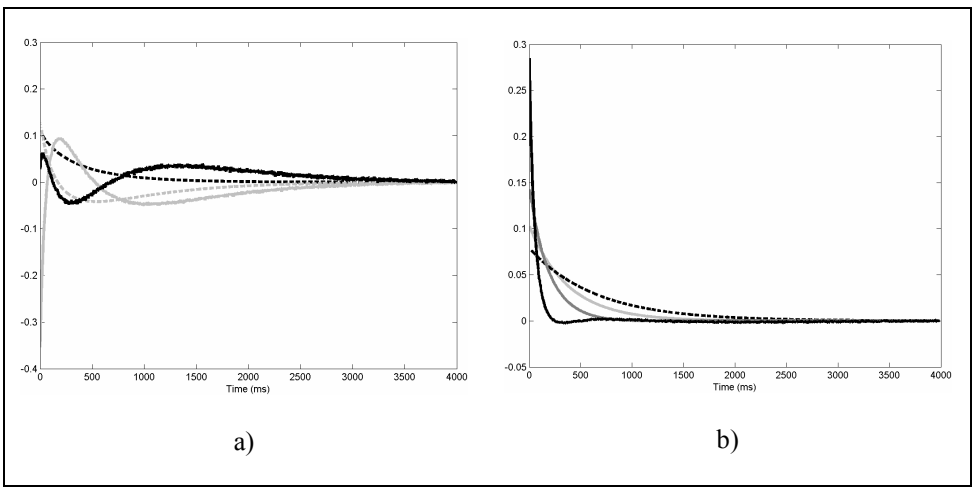


Fig. 3: Typical loadings from CMPG relaxation curves decomposed by a) PLSR, and b) PARAFAC.

In the present work, the validation of the PLSR and the slicing predictions is performed by segmented cross validation [4]. In the cross validation the datasets are divided into four subsets. The size of these subsets is different for each dataset, but for all datasets the choice of samples in each subset is random. The Root Mean Square Error of Cross Validation (RMSECV) and correlation coefficients are used as indicators of the predictive ability of a model. RMSECV are calculated using the following equation:

$$RMSECV = \sqrt{\frac{(y_m - y_p)^2}{n-1}} \qquad \text{(Eq. 2)}$$

y_m is the measured value, y_p is the predicted value, and n is the number of samples.

RMSECV and the correlation coefficients are sensitive to how the dataset is segmented in the cross-validation, and thus 100 different randomly segmented cross-validations were performed for each model. Random segmentation is appropriate for these data because all the samples are similar. The values reported in this paper are the median of the RMSECV, together with the standard error of the RMSECV and the median of the correlation coefficient based on the 100 cross-validations.

All data analysis and modelling is performed using Matlab 5.3 software (Mathworks) for Windows with algorithms taken from the Matlab PLS-Toolbox (www.eigenvector.com) and the N-way Toolbox [1] (www.models.kvl.dk/source).

DATA ANALYSIS AND MODELLING

The data used in this paper are taken from three different sources all from the field of food science. Data from potatoes are from Thybo et al. [16], fish data are from Andersen et al. [2], and the data on seeds are from Pedersen, et al. [13].

Due to noise in two of the datasets – potatoes and seeds – some of the first data points are removed. The noise occurs because the recording of the signal starts too early after the pulse is given. This gives rise to erroneous measurement points – non-exponential decay in the beginning of the curve. The size of the samples measured was not constant for the fish data. As the magnitude of the measured signal varies linearly with the amount of sample, it is important to remove the variation due to sample size before modelling the data. To minimize the effect of the sample size, the relaxation curves were maximum normalized, setting the maximum of each curve to one. Since the intensity of the signal is linearly dependent upon the amount of the sample, as noted above, dividing each sample by its maximum would minimize the influence of the sample size on the data. It can be noted that maximum normalization will not influence the calculated T_2-values from slicing. The pre-treatments are summarized in Table 1.

Tab. 1: The pre-treatment of the different datasets, and the slicing parameters.

Dataset	Points removed	Pre-treatment	Lag	Slab
Potatoes	5 first + 2000 last	None	1, 4	3
Fish	None	Normalized by first measurement point	1	2
Seeds	2 first	None	1, 4, 10	4

The pre-treatment reported in Table 1 is applied both to the PLSR and the slicing analysis.

The correct number of components in the PLSR model is determined primarily using the minimum RMSECV. In addition a visual assessment of the loadings and the residual (in the variable direction) is performed. The number of components in the slicing model is estimated by looking at the mono-exponential fitting of the loadings and the residuals in the second mode (the spectral mode). The number of factors in the different models is summarized in Table 2.

Tab. 2: The number of factors and the predictor variables for the different analysis.

Data	Number of factors		Predictor	Number of samples	Y centered?
	PLSR	Slicing			
Potato	4	4	3 sensory	23	Yes
Fish	2	2	1 chemical	248	Yes
Mustardseed	3	3	2 chemical	51	No
Rapeseed	3	4	2 chemical	60	No

RESULTS AND DISCUSSION

QUANTITATIVE RESULTS

The results from the quantitative analysis are given in Table 3-5. In each of these tables there is a column called range. This is the difference between the lowest and the highest value measured for the specific variable.

POTATO DATA

The NMR measurements were performed on pre-boiled potatoes. After cooking, a sensory panel evaluated the quality of the potatoes using six different sensory variables (only three included in this paper), all on a scale from 0-15. The results from this analysis can be seen in Table 3.

Tab. 3: Results from the regression analysis of the potato data.

Sensory variables		PLSR			Slicing		
Attributes	Range*	RMSECV		r	RMSECV		r
		Median	Std	Median	Median	Std	Median
Cohesiveness	5.7	1.13	0.09	0.77	1.19	0.09	0.74
Mealiness	7.4	1.36	0.14	0.79	1.48	0.15	0.75
Moistness	7.0	1.11	0.15	0.77	1.11	0.14	0.77

* Difference between the highest and the lowest value.

The results show that PLSR is slightly better than slicing in the regression analysis.

FISH DATA

Five different fish were measured, with samples taken from different parts of the fish. The water content both differs from fish to fish, and internally in each fish as is shown by Jepsen et al. [9]. Therefore the water content was measured on every sample.

Tab. 4: Results from the regression analysis of the fish data.

Method	Range*	RMSECV		r
		Median	Std	
PLSR	3.26	0.242	0.001	0.95
Slicing	3.26	0.243	0.001	0.95

* Difference between the highest and the lowest measured value.

From Table 4 it becomes clear that upon predicting the water content, the two methods give the same good predictions.

SEED DATA

Two types of seeds - mustardseed and rapeseed - were measured by CPMG together with three chemical properties (oil-, protein-, and water content). The seeds were measured as untreated, wet, and dry seeds. The contents of the different chemical properties were calculated as the relative content. In this paper, only the predictions for the oil and water content are shown.

As can be seen from Table 5, the results are quite similar, except from the prediction of water in the rapeseeds, where slicing performs better than PLSR. This might be due to the ability of slicing to find four important factors, while PLSR only finds three.

Tab. 5: Results from the regression analysis of the seed data.

Seed	Variable	Range*	PLSR			Slicing		
			RMSECV		r	RMSECV		r
			Median	Std	Median	Median	Std	Median
Mustardseed	Oil	8.9	0.35	0.01	0.99	0.36	0.01	0.99
	Water	16.4	2.42	0.10	0.92	2.38	0.06	0.92
Rapeseed	Oil	13.0	0.60	0.01	0.98	0.58	0.02	0.98
	Water	13.9	2.03	0.05	0.92	1.36	0.05	0.97

*Difference between the highest and the lowest measured value.

QUALITATIVE RESULTS

These above results show that the predictive abilities of PLSR and slicing-based regression are quite similar. However, upon looking at the loadings from

the two methods, it becomes clear, that slicing gives loadings that can be interpreted directly by the NMR spectroscopist whereas this is not at all the case for PLSR (Figure 3). From the slicing results it is possible to calculate the T_2-values for the different proton populations present in the sample. The NMR signal is often mainly caused by water, as is the case for these datasets. Very often in NMR-relaxation measurements, it is of interest to study the T_2-values, since these provide the qualitative information as to what the detected phenomena represent and also provide means for comparing the results to previous work. Further, the scores from slicing can be of more interest than the scores from PLSR, namely because of the effect explained above. Povlsen et al. [14] showed that the scores from slicing were able to separate the five potato tubers, while PLSR scores were not. Another interesting attribute from this method, is shown by Andersen et al. [2], which were able to detect an offset in the data in one of the factors from slicing. The third factor showed a non-exponential behaviour. Rather than decreasing towards zero, the third factor decreased to a positive constant value. A single exponential can thus not explain this factor, and hence there is an offset in the data. Since there is no a priori knowledge about the behaviour of the PLSR-loadings, this offset would be harder to detect. The reason for this offset was investigated further, and was most likely caused by magnitude correction of the raw data. Unfortunately the raw data was no longer available, so the offset needed to be corrected otherwise. The slicing method is based upon the loadings being exponential, and it may thus not perform well if it turns out not to be true. The correction caused all the loadings to become exponential, but the T_2-values calculated from these loadings, will only be the best values based on the data, and not necessarily the true values. Even though the fish data was of this type the predictions by the use of slicing was identical to the predictions by PLSR.

CONCLUSION

It is shown that the two methods give equally good predictions for the applications shown here. An important difference between the two methods can be seen upon looking at the spectral loadings. The loadings from slicing are clearly exponentials (as shown in Figure 3), and T_2-values can be calculated for each loading. The T_2-values can be related to the major proton containing components. For the PLSR loadings only the first loading is exponential, and sometimes not mono-exponential, thus making it impossible to find the T_2-values corresponding to the raw data. These T_2-values are essential in the understanding of the data, though not explicitly needed for building calibration models.

The slicing method is more time consuming than the PLSR method, and often the predictions are of highest importance. However, if the predictive ability of PLSR is unsatisfying, slicing may show a better predicting ability.

Further, as noted above, this approach has the advantage that the model directly provides estimates of the mono-exponential decays present in the raw data.

ACKNOWLEDGEMENT

The author wishes to thank the frame program AQM and STVF (Danish Technical Research Council) for financial support through project 1179. The author is grateful to A. Thybo, C.M. Andersen, and H.T. Pedersen for the kindness of using their data. Thanks to R. Bro are also due for valuable inspection of the results.

REFERENCES

[1] ANDERSSON C.A., BRO R. (2000): *The N-way Toolbox for MATLAB*. Chemometrics and Intelligent Laboratory Systems, 1, 52, 1-4.

[2] ANDERSEN, C. M., RINNAN, Å. (2002): *Distribution of water in fresh cod.* Food Science and Technology-Lebensmittel-Wissenschaft & Technologie, Accepted.

[3] BRO R. (1997): *PARAFAC. Tutorial and applications*. Chemometrics and Intelligent Laboratory Systems, 2, 38, 149-171.

[4] EASTMAN H.T., KRZANOWSKI W.J. (1982): *Cross-validatory choice of the number of components from principal component analysis*. Technometrics, 1, 24, 73-77.

[5] ESBENSEN K., SCHÖNKOPF S., and MIDTGAARD T. (1994): *Multivariate analysis in practice*. Wennbergs Trykkeri, Trondheim.

[6] HARSHMAN R.A. (1970): *Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis*. UCLA working papers in phonetics, 16, 1-84.

[7] HÖSKULDSSON, A. (1988): *PLS regression methods*. Journal of Chemometrics, 2, 211.

[8] Jensen, K.N.; Guldager H.S., Jørgensen, B.M. (2002): *Three-way modeling of NMR relaxation profiles from thawed cod muscle.* Journal of Aquatic Food Product Technology, In press.

[9] JEPSEN S.M., PEDERSEN H.T., and ENGELSEN S.B. (1999): *Application of chemometrics to low-field H-1 NMR relaxation data of intact fish flesh*. Journal of the Science of Food and Agriculture, 13, 79, 1793-1802.

[10] KRUSKAL, J. B. (1977): *Three-way arrays: Rank and uniqueness of trilinear decompositions with applications to arithmetic complexity and statistics.* Linear Algebra and its Applications, 18, 95-138.

[11] PEDERSEN, H. T., BRO, R., ENGELSEN, S. B. (2002): *Towards Rapid and Unique Curve Resolution of Low-Field NMR Relaxation Data: Trilinear SLICING versus Two-Dimensional Curve Fitting,* Journal of Magnetic Resonance, 157**,** 141–155.

[12] PEDERSEN, H. T., BRO, R., ENGELSEN, S. B. (2001): *A novel approach for unique deconvolution of NMR relaxation decays.* In G.A. Webb, P.S. Belton, A. M. Gill and I. Delgadillo (Eds): Magnetic Resonance in Food Science: A view to the future, The Royal Society of Chemistry, 202-209.

[13] PEDERSEN H.T., MUNCK L., and ENGELSEN S.B. (2000): *Low-field H-1 nuclear magnetic resonance and chemometrics combined for simultaneous determination of water, oil, and protein contents in oilseeds*. Journal of the American Oil Chemists Society, 10, 77, 1069-1076.

[14] POVLSEN V., RINNAN Å., VAN DEN BERG F., ANDERSEN H.J., THYBO A.K. (2002): *Direct Decomposition of NMR relaxation profiles and Prediction of Sensory Attributes of Potato samples*. Food Science and Technology-Lebensmittel-Wissenschaft & Technologie, Submitted.

[15] SIDIROPOULOS, N. D., BRO, R. (2000): *On the uniqueness of multilinear decomposition of N-way arrays.* Journal of Chemometrics, 14, 229-239.

[16] THYBO, A. K., BECHMANN, I. E., MARTENS, M., ENGELSEN, S. B. (2000): *Prediction of sensory texture of cooked potatoes using uniaxial compression, near infrared spectroscopy and low field H-1 NMR spectroscopy.* Food Science and Technology-Lebensmittel-Wissenschaft & Technologie, 33, 103-111.

[17] THYGESEN, L. G., THYBO, A. K., ENGELSEN, S. B. (2001): *Prediction of sensory texture quality of boiled potatoes from low field ^{1}H NMR of raw potatoes. The role of chemical constituents*. Food Science and Technology-Lebensmittel-Wissenschaft & Technologie, 34, 469-477.

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

[19] WOLD S., ESBENSEN K., and GELADI P. (1987): *Principal Component Analysis*. Chemometrics and Intelligent Laboratory Systems, 2, 37-52.

Paper VIII

Rinnan, Å., Riu, J., Bro, R.

Multi-way prediction in the presence of uncalibrated interferents, Submitted

Multi-way prediction in the presence of uncalibrated interferents

Åsmund Rinnan[a], Jordi Riu[b*] and Rasmus Bro[a]

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

[b] Universitat Rovira i Virgili
Department of Analytical and Organic Chemistry
Pl. Imperial Tarraco 1, 43005-Tarragona, Catalonia - Spain

21 Pages, 3 Tables, 7 Figures

*Please address correspondence to:

Jordi Riu
Departament of Analytical and Organic Chemistry
Universitat Rovira i Virgili
Pl. Imperial Tarraco 1
43005 – Tarragona
Catalonia – Spain

Tel.: +34 977 558 629
Fax.: +34 977 559 563
e-mail: rusell@correu.urv.es

1 Abstract

The second order advantage states that second-order calibration methods based on intrinsically unique decomposition should give good predictions for new samples even if they contain new interferents not taken into account in the calibration model. In this paper we test the second order advantage using regression models based on PARAFAC, focussing on practical issues such as the importance of the type of the PARAFAC model to use, the size of the calibration set, the number and degree of overlap of new interferents, and the type and magnitude of noise. In order to control all these factors, simulated data is used to get the scenarios in which regression models based on PARAFAC can take profit of the second order advantage.

Keywords: Second-order advantage, PARAFAC, prediction

2 Introduction

In the last years, data provided by analytical instrumentation has changed from univariate signals (a single number for a single sample, for instance a pH measurement; a vector for a set of samples) to two-way data (a vector for a single sample, for instance a NIR spectrum; a matrix for a set of samples), and further to multi-way data (a matrix for a single sample, for instance a fluorescence emission-excitation landscape; a three-way array for a set of samples in case of three-way data). Extension to higher orders is straightforward. During the last years multi-way analysis has become increasingly important because it has proved to be a valuable tool in interpreting such complex data. One of the most used multi-way models in chemometrics is PARAFAC (parallel factor analysis) [1-3]. PARAFAC can to some extent be seen as an extension of more classical multivariate models (i.e. models that handle two-way data) as PCA (principal components analysis) [4].

Since PARAFAC only deals with one array of data (i.e. PARAFAC only works with an $\underline{\mathbf{X}}$ array), it is not primarily seen as a prediction method, but rather as a

decomposition tool. However, the results (scores) given by PARAFAC can easily be related to concentration values by e.g. ordinary least squares (OLS), giving PARAFAC the attributes of a second order prediction method.

When making calibration models it is common practice to include samples with varying amounts of all future interferents in order to be able to provide essentially unbiased predictions. This is necessary in regression-based calibration because the regression vector is essentially determined by orthogonalizing the analyte signal to the signals from all interferents. In some situations, e.g. in process analysis or environmental analysis, the new samples to predict can contain unknown interferents that can not be taken into account in building the model. With standard regression-based methods it is not possible to handle interferents that are not present during calibration, and it is therefore of interest to build models that can handle this situation. Previous work [5] has described the advantages of using second order prediction for these scenarios instead of classical first order prediction – e.g. PLS (partial least squares), PCR (principal components regression) and MLR (multi linear regression) [6].

However, there has not been much work on how to perform second order prediction using PARAFAC in practice, and some practical doubts arise when dealing with this subject. The general principle of second-order calibration agreed on so far is that a model of the multi-way calibration data is established using PARAFAC. With this model, the concentration of the analyte of interest can be predicted using one or several scores. When the concentration in new samples is to be predicted, then either the scores in the new samples are found from the earlier PARAFAC model loadings or by including the new sample(s) in the calibration set and redo the PARAFAC model including the regression step.

In order to choose a reasonable strategy for calibration there are, however, many additional issues that must be decided: E.g. should the calibration set be included in

the new set to be predicted? Should the loadings from the calibration part be fixed during prediction? Should new components be allowed for in the new set? Should the scores of the calibration set be fixed? The current paper tries to address these and similar issues, looking at different strategies for performing second order prediction by the use of PARAFAC. Practical aspects such as the size of calibration set, degree of overlap of spectra, noise patterns and number of analytes in the calibration set, are investigated to verify which factors are important to consider when choosing prediction strategy. Apart from looking at practical aspects like the ones mentioned above, PARAFAC has shown to take profit of the second order advantage and has shown to make better predictions than other methods such as GRAM.

3 Theory

3.1 PARAFAC model

PARAFAC [1-3] is a decomposition method that can be considered as one possible generalization of PCA to higher order arrays. For three-way data, PARAFAC decomposes the original $I \times J \times K$ three-way array $\underline{\mathbf{X}}$ (extension to higher order arrays is straightforward) into F trilinear components (factors), each one of the F factors consisting of one score vector $\mathbf{a}_f$ (column of $\mathbf{A}$) and two loading vectors $\mathbf{b}_f$ and $\mathbf{c}_f$, columns of $\mathbf{B}$ and $\mathbf{C}$ (also sometimes called first, second and third mode loadings respectively), with respective elements a_{if} (i=1…I, f=1…F), b_{jf} (j=1…J) and c_{kf} (k=1…K). In this paper, $\mathbf{A}$ corresponds to the sample mode, while $\mathbf{B}$ and $\mathbf{C}$ are spectral modes. The factors in PARAFAC are found simultaneously and are not nested as for instance in the PCA case. The model is found minimizing the sum of squares of the residual e_{ijk} in the model:

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

Each factor from a PARAFAC model consists of a set of triads as illustrated in Fig 1.

Even with no additional restrictions on the parameters in Eq. (1), the PARAFAC solutions are unique (up to trivial scaling and permutation) in contrast with the rotational freedom of the bilinear model, where additional constraints are necessary e.g. to make the PCA model unique [7]. For example this implies that for correctly modelled fluorescence data, the **B** and **C** loadings are estimates of the pure emission and excitation profiles (arbitrarily scaled) of the *F* analytes in the data set.

Although PARAFAC is primarily a decomposition method, it can be used for regression relating the scores of each component (i.e. of each analyte) to its concentration. This is possible because, as the loadings are estimates of relative spectra of specific analytes, then the scores are estimates of the relative concentrations of these analytes. The prediction is in this paper performed using OLS (which relates the concentration of each analyte only with their scores) according to the following expression:

$$y_{i,f} = \beta_{0,f} + a_{i,f} \cdot \beta_{1,f} \tag{2}$$

where $y_{i,f}$ represents the concentration of analyte f in sample i, $a_{i,f}$ is the score corresponding to analyte f for the ith sample, $\beta_{1,f}$ is the multiplicative factor for the score and $\beta_{0,f}$ is the offset term. β_0 and β_1 are estimated using the scores (from the PARAFAC model using the calibration set) and the concentration values of the calibration set. Eq. (2) is then used to predict the new concentration(s) from the score(s) of the new sample(s). Other alternatives for Eq. 2 include using no offset or using all the scores in a MLR regression. For well-behaved data, one would expect the simplest regression method, OLS with no offset, to be the most appropriate. In this paper, however, OLS with offset was preferred because in some of the calibration sets with few samples (two or three) the scores tended to be slightly biased, and using an offset helped in the quality of the predictions. MLR on

the other hand is thought of as too complicated since there are ideally no interactions between the factors upon building the datasets used in these analyses.

One of the main difficulties in using Eq. 2 for predicting concentrations in new samples is to calculate the PARAFAC scores of the new samples. If the new samples contain only the same analytes as the samples in the calibration set (i.e. no interferents are present in the new samples), then the scores of the new samples, $\mathbf{A}_{new}$, are simply found using the loadings from the PARAFAC model on the calibration set:

$$\mathbf{A}_{\mathbf{new}} = \mathbf{X} \cdot \left(\left(\mathbf{C}_{\mathbf{cal}} \otimes \mathbf{B}_{\mathbf{cal}} \right)^{\mathbf{T}} \right)^{+} \tag{3}$$

where $\mathbf{X}$ corresponds to the vectorized experimental data for the new sample(s) and $\mathbf{B}_{cal}$ and $\mathbf{C}_{cal}$ are the loadings from the PARAFAC model using the calibration set. The superscript '+' refers to the Moore-Penrose inverse and the sign '$\otimes$' stands for the Khatri-Rao product [3]. Possible missing data in the experimental values ($\mathbf{X}$) can be handled by using suitable PARAFAC algorithms [8].

If, on the other hand, the new samples to be predicted contain interferents not taken into account in the calibration model, the scores of the new samples have to be calculated from a PARAFAC model with a higher number of components than the one on the calibration samples. The increased number of components is necessary to account for the interferents in the new sample(s), and is important in order to get good predictions of the analytes. There are several possible strategies for finding the analyte scores of the new samples. For instance, the loadings of the analytes of interest found from the PARAFAC on the calibration samples can be fixed upon calculating these scores in the new samples, or the scores can be found by calculating a new PARAFAC model mixing the calibration and the new samples. One of the goals of this paper is to find the best PARAFAC strategy to get the best estimates of the scores of the new samples and hence the best predictions of the concentrations of the analytes in the new samples.

An alternative to the use of PARAFAC is using the generalized rank annihilation method (GRAM) [9], which only uses one sample for calibration and one sample for prediction at a time. This can be an advantage, because GRAM only needs two samples, but also a drawback because working with only one calibration sample implies that predictions are only adequate for very low-noise well-modeled data. The properties of GRAM in relation to these simulations are described in the end of the paper.

3.2 The second-order advantage

Booksh and Kowalski [5] discussed the different calibration techniques, moving from 0th order calibration up to 3rd or higher order calibration.

- 0th order: calibration between two vectors – e.g. OLS.
- 1st order: calibration between a matrix and one (or more) vector(s) – e.g. PLS, PCR, MLR.
- 2nd order: calibration between a set of matrices (forming a cube) and one (or more) vector(s) – e.g. OLS/MLR on PARAFAC scores and multilinear PLS (NPLS) [10].
- 3rd and higher order: calibration between sets of three or higher dimension data and one (or more) vector(s) – e.g. OLS/MLR on PARAFAC scores.

They further stated that second-order calibration should make good predictions even in the presence of new interferents; the second order advantage. This, however, only holds for second-order methods based on intrinsically unique decompositions, hence not for NPLS. OLS/MLR on PARAFAC scores on the other hand, is such a method, as described above, and GRAM is another widely used method.

Methods that use the second order advantage have already been applied to several real problems. For instance, Moberg et al (2001) [11] investigated the use of PARAFAC on excitation/emission matrices (EEM) on six fluorophores followed

by regression. They further compared it with CLS (classical least squares), and concluded that PARAFAC gave better predictions. Further they checked the second-order advantage applied to the use of PARAFAC, where they were able to predict two analytes precisely, although only those two analytes were present in the calibration set, and there were six analytes (i.e. four new interferents) in the new samples.

Second-order calibration has also been used in flow injection analysis [12] where Nørgaard and Ridder were the first to show the benefits and drawbacks of three-way methods applied to such data [13,14]. Their results were later elaborated and compared with other techniques [15,16].

Wilson et al. [17] used rank annihilation on MS/MS spectra of samples containing warfarin, different hydroxywarfarins, and phenylbutazone. Rank annihilation was also compared to and outperformed different ordinary curve-resolution and regression techniques.

Li et al. [18] showed how GRAM can be used to estimate low concentrations of hydrocortisone in urine using LC-UV data. They noted that using the original rank annihilation algorithm of Lorber [19] inaccurate results were obtained.

Xie et al. [20] compared the merits of second-order calibration for quantifying binary mixtures of p-, o-, and m-amino benzoic acids and orciprenaline reacting with diazotized sulfanilamide in a kinetic UV/VIS study.

Ho et al. [21,22] exemplified second-order calibration with simple one- and two-component mixtures of perylene and anthracene and showed that they were able to determine the concentration of one analyte in presence of the other one using only one pure standard and using fluorescence excitation-emission landscapes. They

also used samples of a six-component polynuclear aromatic hydrocarbon to show the same principle in a more complex matrix.

Gui et al. [23] used direct trilinear decomposition [24] for quantifying initial concentrations of two components (glycine and glutamine) based on the kinetic development of fluorescence in a thin-layer chromatographic system.

Poe and Rutan [25] compared the use of GRAM, peak height, peak area, and adaptive Kalman filter for quantifying analytes in a separation of polycyclic aromatic hydrocarbons using reversed phase liquid chromatography coupled with fluorescence detection. GRAM outperformed the other methods, but was found to be more sensitive to retention time shifts.

4 Experimentation

4.1 Simulation of data

Simulated data was used during this study in order to control all possible aspects concerning calibration and prediction of new samples. Ninety-six different scenarios of simulated data sets were obtained according to the following steps and the factors in Table 1 (See Fig. 3 for an overview of the problem.):

1. The second (B) and third loadings (C) of the analytes and interferents were constructed as Gaussians curves.
2. A number of equally spaced Gaussian curves were made. The total number of curves was equal to the total number of analytes and interferents that were present in the calibration set and the new samples.
3. The loadings where orthogonalized and normalized to norm one.
4. The congruence matrix (see Table 1 and Fig. 2) that was to be used to assure the degree of overlap between the analytes and the interferents were Cholesky factorized [26].
5. The B and C loadings were then found by multiplying the orthogonalized and normalized loadings (from 3) and the Cholesky factorized congruence

(from 4). This provides a set of loadings with congruences as defined above.

6. The A loadings (concentrations) were made from numbers drawn from a uniform distribution.
7. The random numbers (with the number of columns equal to the total number of analytes and interferents) were then orthogonalized and normalized to norm 100 in order to for all analytes and interferents to have similar magnitude.
8. These orthogonal and randomized numbers were then multiplied with Cholesky factorized congruence (different from the loadings).
9. The concentration values for the prediction interferents (i.e. those that only occur in the prediction set) were all set to zero in the calibration set.
10. The total matrix of pure data was then calculated as the outer product of **A**, **B** and **C**.
11. Noise (hetero- and/or homoscedastic, see Table 1) was added to the dataset.
12. The noise addition was done 10 times to give 10 replicates of each of the 96 scenarios.
13. All the selected PARAFAC models were fitted (see 'Prediction models' section) on each one of the 10 replicates of each of the 96 scenarios.

In all cases, the prediction set was made up of three samples.

4.2 Prediction models

Nine initial different methods for performing second order prediction using PARAFAC were compared. These were reduced to five final methods for reasons explained below. A schematic overview of the setup is given in Fig. 3.

The five final different PARAFAC models were:

- Predicting using the whole dataset in one step. I.e. the calibration set and the new samples are put together making one cube. Using the scores and

known concentrations of the calibration samples, the calibration regression model is determined and then applied to the scores of the new samples. This calibration model will be coded as model 1.

- Instead of including all the new samples when computing the new model, only one new sample is modelled at a time (model 2), and hence a PARAFAC model is recalculated for every new sample.

Calculating a PARAFAC model on the calibration set and then:

- Computing a separate model on the new samples and searching for the common factors with the calibration model (model 3).
- Fixing the B and C-loadings from the calibration model when fitting the PARAFAC model to the new samples. The prediction model is either calculated only on the new samples (model 4), or on the calibration set and the new samples put together (model 5). In both cases, the PARAFAC model for the new samples is calculated with a higher number of components than the PARAFAC model for the calibration samples (due to the interferents).

There were furthermore four other PARAFAC models that initially were investigated:

- Fixing the A, B and C-loadings from the calibration model. Calculate a model on the calibration set with all or one of the new samples at a time.
- As previous, but setting the new interferents to zero for the scores in the calibration set. Calculating with all or one of the new samples at a time.

These four models, however, suffered from convergence problems and reached local minima. The initialization for these methods was the factors from model 1, as was also used for the other methods. The reason for the problem with local minima is not known, and needs further investigation.

4.3 Software

MATLAB (The Mathworks Inc, Natick, MA) version 6.5 was used during the calculations. The algorithms in use were from the PLS_Toolbox version 3 beta (Eigenvector Research Inc, Manson, WA), and some in-house algorithms.

4.4 Evaluating the prediction models

The five different PARAFAC methods/models were evaluated by the use of a goodness-of-fit (GOF) factor. If the prediction was perfect compared to the real simulated value, this factor was 1. The factor decreased linearly to 0, as the prediction error increases to 10, see Eq. 4.

$$\mathrm{GOF} = 1 - \frac{|y_p - y|}{10} \tag{4}$$

where y_p is the predicted value, while y is the corresponding real simulated value. If the GOF factor is less than 0 (prediction error larger than 10), the GOF was set to 0. We decided to use such a factor instead of using RMSEP (root-mean-square-error-of-prediction) or relative RMSEP, since this will put too much focus on extreme errors which are e.g. a factor 20 off. The value of 10 was chosen as a limit because the prediction errors should be small (all real simulated values were below 76).

Some of the PARAFAC models gave imaginary results in some scenarios, due to ill-posed problems. In the predictions based on those imaginary results the GOF-values were set to 0. After this procedure there was a total of 10560 different GOF-values per regression model (64 scenarios with 4 analytes plus 32 scenarios with 3 analytes, all with three new samples and 10 noise additions). Data analyses were carried out in order to investigate:

- which of the scenarios shown in above provide the best and the worst prediction results
- what PARAFAC regression model gave the overall best predictions, what gave the worst

- the why's for some of the results

This investigation was done by the use of analysis of variance (ANOVA) [27], and distribution plots of the GOF's.

4.5 Analysis of variance (ANOVA)

For the calibrations results obtained in this study, the ANOVA factors were the different congruences, set-up for calibration data, noise level, and type of PARAFAC prediction model (all the factors were coded as discontinued variables, from 1 up to the number of different ways each parameter was set), while the dependent variable was the GOF-value. Interaction terms were included in the ANOVA model primarily based on P-values.

In this analysis the prime objective was to study the effect of the different prediction models. For the study of the difference in the means of the five different methods, the Games-Howell Pairwise Comparison test [28] was used. This test checks every possible pair of prediction models to see if their means are different. The distribution used is the studentized range statistics.

5 Results and discussion

5.1 ANOVA

The number of samples in the data set analysed by ANOVA was 52800 (10560 GOF-values per prediction model). Taking this into account the significance limit was set as low as 0.001. The first ANOVA run showed that there were only a few interactions that were non-significant. However, upon the removal of these and running the analysis again, no more interactions showed up as non-significant. The results from the first ANOVA can be seen in Table 2.

Since the interest of the ANOVA was to investigate the effect of the choice of the model (X4), only these sources were investigated further: X4, X1X4, X3X4, X1X3X4 and X1X2X3X4.

First X4 was investigated to see if there was some significant difference in the means of the methods. The significance limit here was also set to 0.001 as for the ANOVA. The results from this investigation are given in Table 3.

From Table 3 it can be seen that model 1 is the best model (all its values in its column are positive), followed by model 2, 4 and 5. However, the only significant difference in the means is that model 3 predicts significantly poorer than the four other models.

Further analyses were performed on the four interactions mentioned above. Here the data was investigated in two different ways in order to explore the results. All the interactions were analyzed taking into account the distance between the PARAFAC prediction models. Either all the distances were summed directly, or they were multiplied by 1 if they were significant and 0 if they were insignificant, according to a significant level of 0.001. All the results from these analyses will not be shown due to the large amount of tables necessary to show the results. The results of the analyses indicated that model 1 is the best, closely followed by models 4 and 5, with model 2 also giving good results. Worst is model 3 by far. It is of no surprise that model 3 gives the worst predictions, since this method decomposes only a small data set containing many more components than samples (3 samples with 6 or 7 components). Model 2 is partly subject to a bit of the same problem, but this time there is only one new sample with the interferents. The problem now is to correctly find the interferents. Since there is only one sample included at a time, the decomposition suffers from partial rotational freedom (of the new interferents), and this will affect the predictions to a larger degree than is the case for model 1, where there are several samples with the same interferents, where all of them vary independently of each other. Models 4 and 5 are very similar in behavior, which was also expected, as these two models are very similar; both of them fix the loadings from the calibration set. The only difference is that model 5 includes the samples from the calibration set, and thus recalculating the

scores for the analytes, while model 4 keeps the scores from the calibration set. These two models do give slightly worse predictions than model 1. By close inspection of the results it seems that model 1 is better when the problem is the worst-case scenarios (two samples in the calibration with four analytes) while models 4 and 5 are slightly better in the more "normal" cases (i.e. when the size of the calibration set is higher than two samples). This is probably because the most important factor, alongside with the selection of the PARAFAC model, is the number of samples in building the model. In the worst-case scenarios, model 1 uses five samples (two from the calibration set and three from the prediction set) to estimate six or seven loadings, while model 4 and 5 estimates three or four (depending on the number of analytes) loadings using only two samples. By calculating the loadings based only on two samples, the estimated loadings are not as good as the ones used in model 1.

5.2 Checking the goodness-of-fit factor

A close examination of the 52800 goodness-of-fit factors can give the scenarios which correspond to the best and the worst prediction results. The GOF-values are binned into 100 different bins ranging from 0 and up to 1, with 0.01 between the bins. The relative cumulative sums of the bins are then investigated.

An important factor in building a calibration model is the size of the calibration set. Fig. 4 shows the relative cumulative sum for the six different calibration set types.

From Fig. 4 it can be seen that the smallest calibration set is the dataset causing the biggest problems in prediction analyses. The worst scenario should intuitively be the calibration set containing only two samples with four analytes. It however seems that the calibration set with two samples and three analytes gives the overall worst predictions. If we strictly look at the number of samples in the calibration set, it can clearly be seen that the higher number of samples in the calibration set, the better predictions one will get as expected. In this way, the calibration set

containing seven samples with four analytes is the best one since it is the one with lowest percentage of low GOF factors (only 17% of the GOF factors are below 0.9 while more than 42% of the GOF factors are below 0.9 for the calibration set containing two samples with three analytes). For low GOF factors, the second best scenario is three samples and three analytes. It might have been expected that the four samples and four analytes scenario to be the second best. This is because it seems that three/four samples is the borderline to have good predictions (two samples is very low, seven samples is a good size). Apparently, adding another sample (from three to four), does not pay back the effect of adding another analyte, and three samples and three analytes remains as the second best for low GOF factors.

After studying the effect of the size of the calibration set it is of interest to see if there is an effect of the overlapping pattern of the analytes and interferents, thus the congruence is studied closer.

From Fig. 5 it can be seen that the (a) congruence pattern is the one with better GOFs, while the congruence pattern with the highest amount of overlap, (d), between the analytes and the interferents gave the overall worst results. The two others with intermediate overlap perform equally, with the congruence pattern with generally low overlap, but each interferent highly overlapping one analyte each, (b), giving slightly better overall results. This fits nicely with what was expected.

The last factor we investigated is the influence of the noise on the goodness-of-fit factor. Fig. 6 shows the sorted GOF values of the four noise patterns. It shows that the datasets with most noise give the worst predictions as expected. Furthermore, the dataset with only homoscedastic noise gives overall quite good predictions, while there is only a small difference between the two other noise patterns, with the one with only heteroscedastic noise slightly better than the one with both.

5.3 Comparison with generalized rank annihilation (GRAM) results

The same simulated data used in the PARAFAC prediction models studied above were used to find the predicted values of the test samples using the GRAM method, because GRAM is one of the most widely used methods that takes profit of the second order advantage. In this study we did not study factors that could help in getting better predictions with GRAM, because the focus of this paper is prediction by the use of PARAFAC. To show the advantage of PARAFAC to GRAM, the overall results of PARAFAC model 1 is compared with the overall results from GRAM. The comparison of the results is shown in Fig. 7. Looking at this figure, it can be concluded that GRAM results generally are worse than the PARAFAC results; about 11% of the GRAM results have a GOF-factor of zero, more than twice of the PARAFAC results. Contrarily, more than 23% of the PARAFAC results have a GOF higher than 0.99, which virtually means a perfect prediction, while only 8% of the GRAM GOFs are. Fig. 7 shows that PARAFAC uses the second order advantage better than GRAM for these simulated data.

6 Conclusion

An overall conclusion based on the above results is that the models have the following order (from best to worst): 1, 4, 2, 5, and 3. However, the differences between the four best models are marginal, and thus the clearest conclusion that can be made is that it is not wise to make a new model on only the new samples and then try to match the factors found with the factors in the calibration set. It is important to be aware that this work is only done on simulated data, which are completely trilinear up to the noise. This may be the cause of model 1 (raw PARAFAC without fixing anything) giving the best results; it is the model which finds the loadings (including calibration parameters) using the highest number of samples. All the other models divide the samples in calibration and prediction samples, or find the loadings using only the calibration set. Real data may not be as perfectly modeled by PARAFAC or may not have as perfect linear relation with

the concentrations as the ones in this paper, and therefore the idea of fixing factors from the calibration set may play a more important role. Other special characteristics such as e.g. slightly varying spectra over time are also problem-specific and can hence not be tested in general even though they may have significance for which approach is most feasible in practice.

There are more that can be done in order to find the best way to do second order prediction: Why did the fixing of the A-scores fail to converge? What is the actual difference in the models 1, 2, 4 and 5, when do they differ significantly? Is there some scenario where one model is clearly better than another? How should the number of new interferents be determined in practice? Some of these issues should be tested by similar simulations as performed here, whereas others are better left for more problem-specific testing.

The influence of the tested congruence pattern and the type of noise follow the beforehand expected behavior, but these are difficult factors to control in acquiring real data. What appears to be very important in order to achieve good predictions is the size of the calibration set, and fortunately this factor is often easy to control by the analyst. As is apparent, increasing the calibration set size (with appropriate samples) leads to better predictions of future samples. A calibration set size of three-four samples gives fairly good results, and a calibration set size of seven samples gives very good results in most of the situations. Hence, a calibration set size of at least one sample more than the number of analytes, alongside with calculations using PARAFAC models 1 or 4 should be adequate to achieve good predictions of future samples in the presence of interferents. This conclusion also points to why the GRAM results seem to be worse than the PARAFAC ones. In the GRAM approach to second order calibration only one calibration sample (or one-pseudo sample in case of direct trilinear decomposition) is used and hence the predictions will only be adequate for very low-noise well-modelled data.

7 Acknowledgements

Åsmund Rinnan would like to thank SJVT for the financial support through project 1179. Jordi Riu would like to thank the Spanish Ministry of Science and Technology for the financial support through the 'Ramón y Cajal' project.

8 References

1. R. Bro, Chemom. Intell. Lab. Syst., 38 (1997) 149-171.
2. R.A. Harshman, UCLA Work. Pap. Phon., 16 (1970) 1-84.
3. R.A. Harshman, J. Chemom., 15 (2001) 689-714.
4. S. Wold, K. Esbensen, P. Geladi, Chemom. Intell. Lab. Syst., 2 (1987) 37-52.
5. K.S. Booksh, B.R. Kowalski, Anal. Chem., 66 (1994) 782A-791A.
6. H. Martens, T. Næs, Multivariate Calibration, Wiley, Chichester, 1989, pp. 97-116.
7. J.B. Kruskal, in: R. Coppi, S. Bolasco (Eds.), Multiway Data Analysis, Elsevier, Amsterdam, 1989, p. 7.
8. C.A. Andersson, R. Bro, Chemom. Intell. Lab. Syst., 52 (2000) 1-4.
9. E. Sánchez, R. Scott-Ramos, B.R. Kowalski, J. Chrom., 385 (1987) 151-164.
10. R. Bro, J. Chemom., 10 (1996) 47-61.
11. L. Moberg, G. Robertsson, B. Karlberg, Talanta, 54 (2001) 161-170.
12. J. Taurina, S. Hernández-Cassou, Anal. Chim. Acta, 438 (2001) 335-352.
13. L. Nørgaard, C. Ridder, Talanta, 41 (1994) 59-66.
14. L. Nørgaard, C. Ridder, Chemom. Intell. Lab. Syst., 23 (1994) 107-114.
15. M.M. Reis, S.P. Gurden, A.K. Smilde, M.M.C. Ferreira, Anal. Chim. Acta, 422 (2000) 21-36.
16. A.K. Smilde, R. Tauler, J. Saurina, R. Bro, Anal. Chim. Acta, 398 (1999) 237-251.
17. B.E.Wilson, W. Lindberg, B.R.Kowalski, J. Am. Chem. Soc., 111 (1989) 3797-3804.

18. S.S. Li, P.J. Gemperline, K. Briley, S. Kazmierczak, J. Chrom B-Biom. App., 655 (1994) 213-223.
19. A. Lorber, Anal. Chem., 57 (1985) 2395-2397.
20. Y.L. Xie, J.J. Baeza-Baeza, G. Ramis-Ramos, Chemom. Intell. Lab. Syst., 32 (1996) 215-232
21. C.N. Ho, G.D. Christian, E.R. Davidson, Anal. Chem., 50 (1978) 1108-1113.
22. C.N. Ho, G.D. Chistian, E.R. Davidson, Anal. Chem., 52 (1980) 1071-1079.
23. M. Gui, S.C. Rutan, A. Agbodjan, Anal. Chem., 67 (1995) 3293-3299.
24. E. Sanchez, B.R Kowalski, J. Chem., 4 (1990) 29-45.
25. R.B. Poe, S.C. Rutan, Anal. Chim. Acta, 283 (1993) 845-853.
26. D.C. Lay, Linear algebra and its applications, 2nd edition, Addison-Wesley, Reading (MA), 1997, p. 457
27. K. Esbensen, S. Schönkopf, T. Midtgaard, Multivariate Analysis in Practice, Wennbergs Trykkeri, Trondheim, 1994
28. http://www.uvm.edu/~dhowell/StatPages/More_Stuff/MultComp/lab1answer.html

Table 1. The factors that were varied in the simulations. *The first number indicates the amount of homoscedastic noise, the second the amount of heteroscedastic noise.

Factor	Variations
Congruence	4 patterns (shown in Fig. 2)
Noise addition*	[5 0]%, [0 2]%, [5 2]% or [10 2]%
Number of samples in the calibration set	7, 4, 3, or 2
Number of analytes	4 (for 7, 4, 3 or 2 samples in the calibration set) or 3 (for 3 or 2 samples in the calibration set)

Table 2. Results from the complete ANOVA for goodness-of-fit. Sources: X1 = congruence, X2 = size of calibration set, X3 = noise, and X4 = model. Significant factors in bold.

Source	d.f.	Sum of sq.	Mean sum of sq.	F	Prob>F
X1	**3**	**229.29**	**76.431**	**1970.34**	**0**
X2	**5**	**62.833**	**12.567**	**323.96**	**0**
X3	**3**	**123.02**	**41.005**	**1057.09**	**0**
X4	**4**	**78.293**	**19.573**	**504.59**	**0**
X1X2	**15**	**60.535**	**4.0357**	**104.04**	**0**
X1X3	**9**	**84.62**	**9.4023**	**242.38**	**0**
X1X4	**12**	**20.258**	**1.6882**	**43.52**	**0**
X2X3	**15**	**46.932**	**3.1288**	**80.66**	**0**
X2X4	20	0.89043	0.044521	1.15	0.29109
X3X4	**12**	**6.2227**	**0.51856**	**13.37**	**0**
X1X2X3	**45**	**174.83**	**3.8851**	**100.15**	**0**
X1X2X4	60	2.5934	0.043224	1.11	0.25351
X1X3X4	**36**	**2.6618**	**0.07394**	**1.91**	**0.00085**
X2X3X4	60	3.2528	0.054214	1.4	0.02278
X1X2X3X4	**180**	**10.385**	**0.057693**	**1.49**	**0.00002**
Error	52320	2029.5	0.038791	-	-
Total	52799	2936.2	-	-	-

Table 3. The difference in the mean value between the five methods. Bold numbers refers to significant differences in the mean value at a 0.001 level. * The numbers down the columns are the differences in the mean from the model on top to the model on the left. If the difference is positive, it means that the model on top has a higher GOF factor than the model to the left.

		Model number				
		1	2	3	4	5
Model number	1		-0.0075	**-0.1019**	-0.0076	-0.0089
	2	0.0075		**-0.0945**	-0.0001	-0.0014
	3	**0.1019**	**0.0945**		**0.0944**	**0.093**
	4	0.0076	0.0001	**-0.0944**		-0.0013
	5	0.0089	0.0014	**-0.093**	0.0013	

Figure captions

Figure 1: A two-factor decomposition by PARAFAC of a data-cube X into triads and a residual-cube E.

Figure 2: The four different congruence matrices used in the simulations. (d) is the congruence matrix that was used for the scores. The congruence matrices in the figure are for the 4-analytes case. The matrices for the 3-analytes case is identical but with one analyte less. The number of uncalibrated interferents was set to three in all cases.

Figure 3: The scores and loading of PARAFAC coded according to Analytes, Interferents, Calibration set and New Samples.

Figure 4: The cumulative relative sum of the binned GOF-factors for the six different calibration set types.

Figure 5: The cumulative relative sum of the binned GOF-factors for the four different congruence matrices. 'a'-'d' indicates the congruence patterns shown in Figure 2.

Figure 6: The cumulative relative sum of the binned GOF-factors for the four different noise patterns coded according to the legend.

Figure 7. Comparison between the GOFs obtained using PARAFAC model 1 and GRAM.

Figure 1

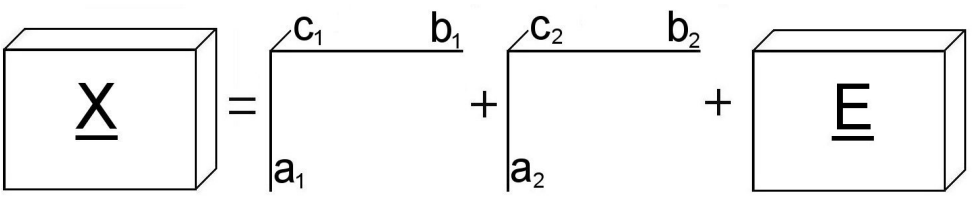


Figure 2

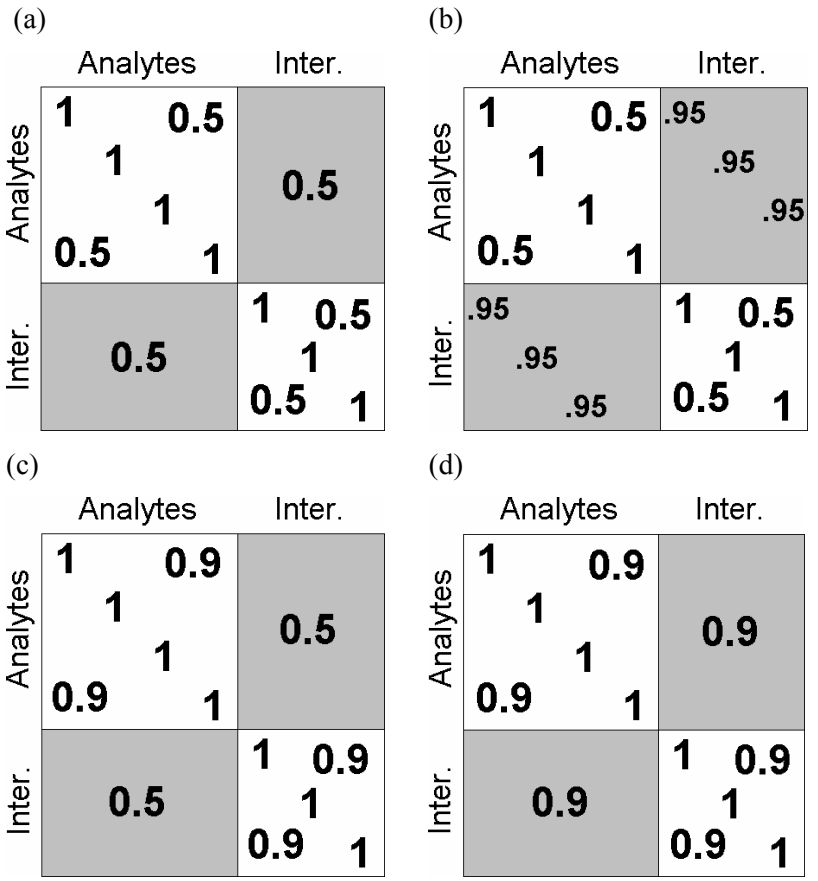

(a)
Analytes
Inter.
Analytes
Inter.
1
0.5
1
1
0.5
1
0.5
0.5
1
0.5
1
0.5
1
(b)
Analytes
Inter.
Analytes
Inter.
1
0.5
1
1
0.5
1
.95
.95
.95
.95
.95
.95
1
0.5
1
0.5
1
(c)
Analytes
Inter.
Analytes
Inter.
1
0.9
1
1
0.9
1
0.5
0.5
1
0.9
1
0.9
1
(d)
Analytes
Inter.
Analytes
Inter.
1
0.9
1
1
0.9
1
0.9
0.9
1
0.9
1
0.9
1

Figure 3

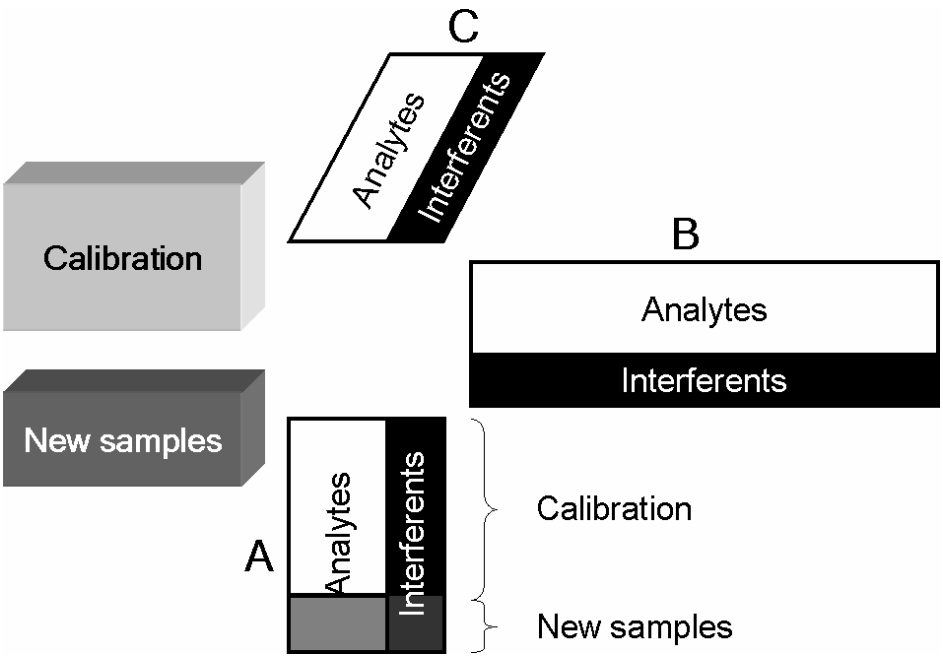

C
Analytes
Interferents
Calibration
B
Analytes
Interferents
New samples
A
Analytes
Interferents
Calibration
New samples

Figure 4

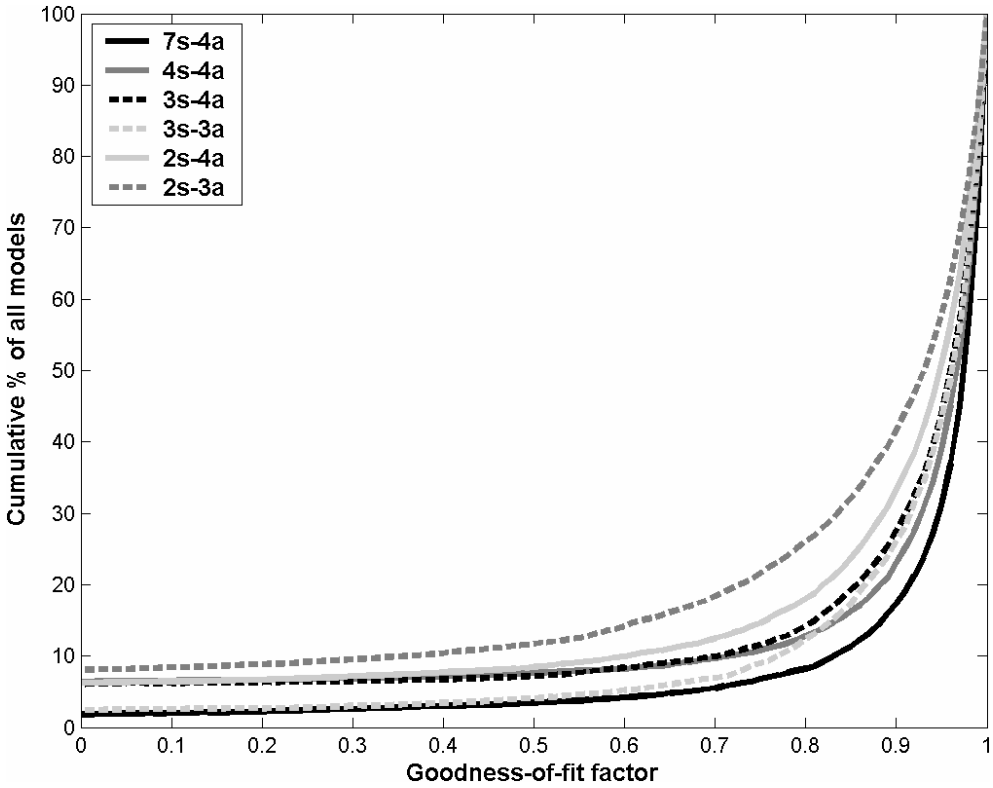

7s-4a
4s-4a
3s-4a
3s-3a
2s-4a
2s-3a
Cumulative % of all models
Goodness-of-fit factor
0
10
20
30
40
50
60
70
80
90
100
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 5

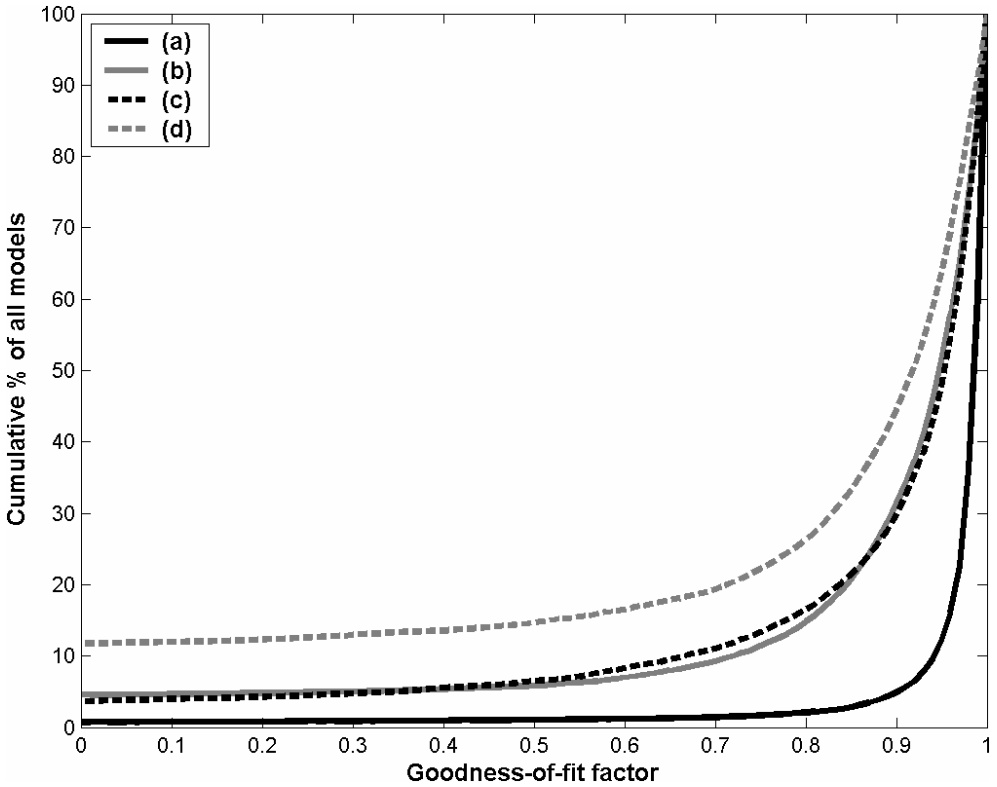

(a)
(b)
(c)
(d)
Cumulative % of all models
Goodness-of-fit factor
0
10
20
30
40
50
60
70
80
90
100
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 6

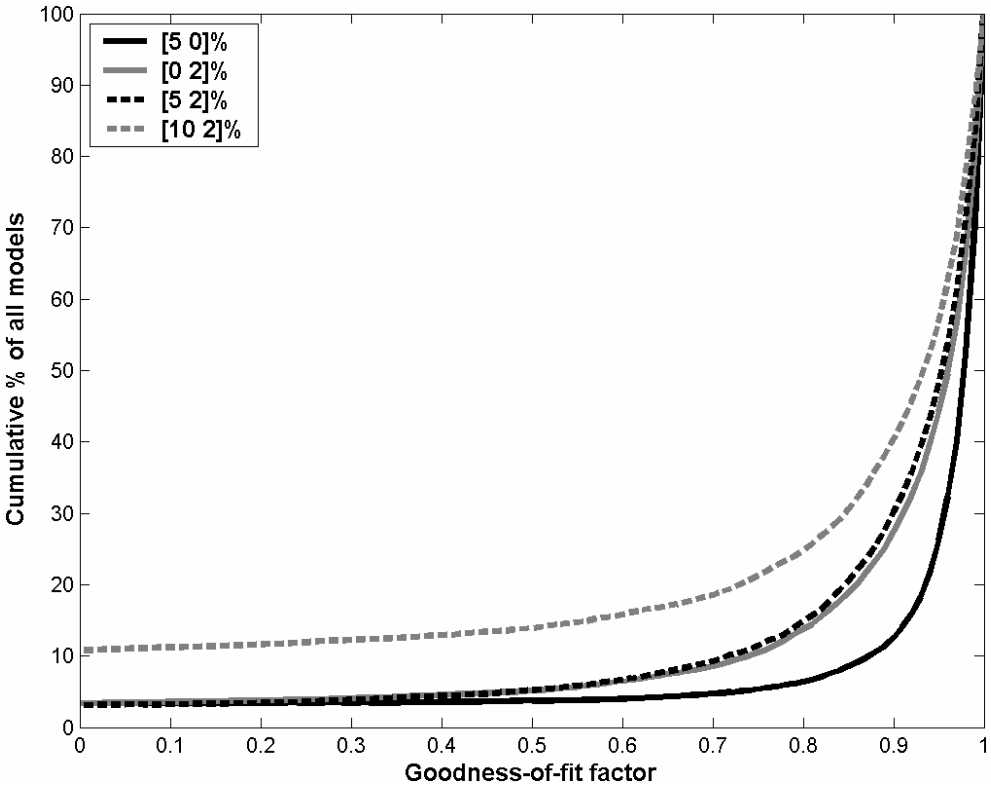
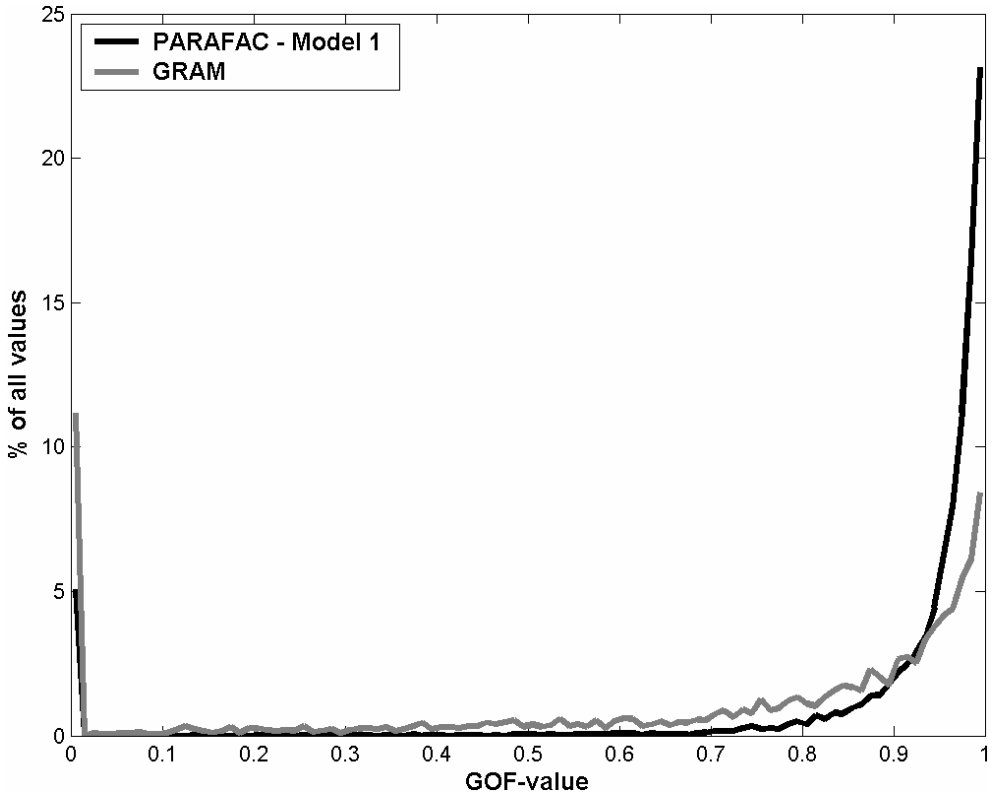

[5 0]%
[0 2]%
[5 2]%
[10 2]%
Cumulative % of all models
Goodness-of-fit factor
0
10
20
30
40
50
60
70
80
90
100
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 7


PARAFAC - Model 1
GRAM
% of all values
GOF-value
0
5
10
15
20
25
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Paper IX

Bro, R., Rinnan, Å., Faber, N.M.

Standard Error of Prediction for Multiway PLS. 2. Practical Implementation in Fluorescence Spectroscopy, *S*ubmitted

Standard Error of Prediction for Multilinear PLS. 2. Practical Implementation in Fluorescence Spectroscopy

Rasmus Bro[a,*], Åsmund Rinnan[a], Nicolaas (Klaas) M. Faber[b]

[a] *Food Technology, Royal Veterinary and Agricultural University, Rolighedsvej 30, 1958 Frederiksberg C, Denmark*

[b] *Chemometry Consultancy, Rubensstraat 7, 6717 VD Ede, The Netherlands*

Abstract

In Part 1 of this series, a new simplified expression was derived for estimating sample-specific standard error of prediction in inverse multivariate regression. The focus was on the application of this expression in multilinear partial least squares (N-PLS) regression, but its scope is more general. In this paper, the expression is applied to a fluorescence spectroscopic calibration problem where N-PLS regression is appropriate. Guidelines are given for how to cope in practice with the main assumptions underlying the proposed methodology. The sample-specific uncertainty estimates yield coverage probabilities close to the stated nominal value. Similar results were obtained for standard (i.e., linear) PLS regression and principal component regression on data rearranged to ordinary two-way matrices. The two-way results highlight the generality of the proposed expression.

KEYWORDS: Multiway calibration; Partial least squares; Standard error of prediction; Fluorescence spectroscopy; Noise addition

[*] Corresponding author.

E-mail address: rb@kvl.dk (R. Bro)

Content

1 Introduction

It is considered good analytical practice to report a result together with an estimate of its uncertainty. For example, when fitting a line through scattering x-y points, it is desirable to construct the familiar confidence and prediction bands using well-known expressions from basic statistics. Interestingly, calibration methodology differs greatly when moving to more complex data structures, i.e., multivariate and multiway data. Beyond univariate calibration, the only generally accepted approach to prediction uncertainty is to use an overall measure such as root mean square error of prediction (RMSEP), hence for any prediction the uncertainty is set to a constant value [1]. Clearly, a negative aspect of this global uncertainty estimate is that it does not yield realistic prediction intervals. However, the required sample-specific prediction uncertainty estimates are available for latent variable (LV) methods in general, see [2] for a review of various proposals. Very recently, a simple and user-friendly expression for standard error of prediction (SEP) has been derived and tested on multiway models using Monte Carlo simulations [3]. It performed well for standard (i.e., linear) partial least squares (PLS) regression applied to near infrared (NIR) data sets [4]. The purpose of the current work is to:

- demonstrate the practical utility of our simple expression on fluorescence data, and
- identify the conditions that are critical for reliable use of the proposed methodology.

2 Theory

Since a detailed discussion of the proposed expression is given in Part 1 [3], we restrict ourselves to pointing out the relationship with variations of RMSEP, as well as another expression for sample-specific SEP. These relationships may lead to a better understanding of the properties of our proposed expression. To simplify the presentation, we have found it convenient to slightly adapt the earlier notation.

2.1 Prediction uncertainty on the global set level

Current practice is to characterise (multivariate or multiway) prediction uncertainty on the set level. An RMSEP–value is calculated as the root mean squared difference between predictions and reference values. It is important to stress that this procedure is only sound when the noise in the reference values is negligible compared with the true prediction uncertainty. The reason for this is that prediction errors are defined with respect to the true quantities, rather than noisy reference values. Consider the ideal situation where one has the perfect model and noisy reference values – a mental experiment. Of course, this limit is not practical, but adding noise to the reference values as described by DiFoggio [5] and Coates [6] can always approach it to some extent. Clearly, the predictions should be perfect and the only contribution to RMSEP would originate from the measurement error in the reference values. In this extreme case, RMSEP would just estimate the standard deviation (square root of the variance) of the measurement error of the reference value – it would not relate to the true prediction uncertainty at all! Thus, in general, the presence of this spurious error component leads to a so–called *apparent* RMSEP [5]:

$$\mathrm{RMSEP}_{\mathrm{app}} = \left[\frac{1}{I} \sum_{i=1}^{I} \left(\hat{y}_i - y_{\mathrm{ref},i} \right)^2 \right]^{½} \tag{1}$$

where I denotes the number of samples in the test set, $\hat{y}_i$ is the prediction of property y for sample i and $y_{\mathrm{ref},i}$ is the associated reference value. A simple but effective correction for the spurious error component leads to a so–called *corrected* RMSEP [5,7] :

$$\mathrm{RMSEP}_{\mathrm{cor}} = \left[\mathrm{MSEP}_{\mathrm{app}} - V_{\Delta y} \right]^{½} \tag{2}$$

where $V_{\Delta y}$ is an estimate for the measurement error variance associated with the reference method. This correction has been used successfully for NIR [8] and Raman [9] applications. If knowledge on $V_{\Delta y}$ is lacking, it can be set to zero and the more pessimistic apparent RMSEP is then obtained.

2.2 Prediction uncertainty on the individual sample level

Characterising prediction uncertainty on the set level is the best way to answer important questions like "how good is my calibration?" It is therefore logical, for example, to monitor changes in the (set level) RMSEP when optimising a calibration model (spectral pre–treatment, factor selection, etc.). However, as noted before, this procedure does not lead to sample–specific prediction intervals with good coverage probability. The American Society for Testing and Materials (ASTM) has recognised the need for a sample–specific SEP [10] and recommends using the expression originally proposed by Höskuldsson [11]:

$$\mathrm{SEP}_{\mathrm{app},i} = \left[(1 + h_i)\mathrm{MSEC}_{\mathrm{app}}\right]^{½} \tag{3}$$

where h_i symbolises the leverage for sample *i* and MSEC stands for the mean square error of calibration. This expression is implemented in certain commercial software [12].

The leverage is related to the distance of a sample to the mean of the calibration set data. The calculation of MSEC is similar to the calculation of the apparent (set level) RMSEP, i.e. Eq. (1), but now one has to account for the degrees of freedom of the calibration model. Because MSEC is explicitly based on reference values, Eq. (3) leads to an apparent sample–specific SEP when the reference method is imprecise. In other words, Eq. (3) is the sample–specific analogue of Eq. (1). Clearly, the correction in Eq. (2) can also be applied on the sample level, leading to our proposal [3]

$$\mathrm{SEP}_{\mathrm{cor},i} = \left[(1 + h_i)\mathrm{MSEC}_{\mathrm{app}} - V_{\Delta y}\right]^{½} \tag{4}$$

which was derived in Part 1 based on an approximation of a local linearization approach.

Notice that only the leverage reflects the individual differences. As leverages are simple to calculate this expression is highly operational. The quality of the

expression is crucially dependent on the quality of the estimate of $MSEC_{app}$. Ensuring that the calibration model is sound, robust and representative is therefore of significant importance to be able to rely on the use of Eq. 4. This is in fact similar to the situation in univariate regression. For example, a model where the MSEC changes significantly by the in- or exclusion of one sample cannot be considered to provide a robust hence reliable estimate of MSEC. Likewise, if the cross- or testset-validated mean square error differs markedly from MSEC, then most likely the estimate can not be considered to be appropriate. Some practical guidelines on how the adequateness can be assessed are provided in the experimental section.

2.3 Validation of proposed sample-specific uncertainty estimate

Comparing the coverage probabilities of the resulting prediction intervals with the nominal value would validate Eq. (4). Unfortunately, this requires error-free reference values. However, the direct relationship between Eqs. (3) and (4), ensures that an equivalent test follows from the studentised *apparent* prediction residuals,

$$t_i = \frac{\hat{y}_i - y_{\text{ref},i}}{\text{SEP}_{\text{app},i}} \qquad i = 1, \cdots, I \tag{5}$$

These should be approximately distributed as Students *t* with degrees of freedom (*f*) associated with the MSEC estimate [3]. In particular, the standard deviation should be close to $\sqrt{f/(f-2)}$. As SEP_{app} does not rely on the actual reference measurements in the validation set, the measurement errors in these are of no consequence for the evaluation.

2.4 Checking validity of the regression model

For a specific regression model it is possible to verify whether the approach for estimating the sample-specific SEP (Eq. 4) is appropriate. As described in Part 1, the formula is based on a local linearization and the validity of this first-order

approximation can be tested. In this paper a noise-addition approach will be used for this purpose. From a principal component analysis (PCA) model of the predictor (spectral) data, the noise level in **X** can be determined. Adding different multiples of this level of random Gaussian noise to **X** leads to different realizations of the regression vector, and associated predictions of **y**. If the first-order approximation is valid, the standard error in the regression vector and hence in the predictions should increase linearly with the noise added. The presence of linearity can then be verified formally or visually in simple manners as will be shown in the experimental part. Test set predictions are used for assessing linearity as fitted values can led to spurious results when performed in an unsupervised fashion (fixing the number of components). Noise is added to both the calibration and the test data.

3 Experimental

3.1 Generation of the data

The data set used in this paper is part of a larger data set prepared for the study of several topics in fluorescence spectroscopy. The part not used here is characterised by specific artificially induced problems (e.g. co-varying components). The selected analytes have very similar excitation and emission spectra (see Figure 1). Consequently, the calibration problem is fairly difficult.

Figure 1: The pure excitation and emission spectra for the five fluorophores used in the data set.

The fluorescence spectra of 131 samples were recorded. Five different analytes were used: catechol (Sigma, approx. 99%), hydroquinone (Riedel-deHaën, min. 99.5%), indole (Riedel-deHaën, min. 99%), L-tryptophan (Merck, min. 99%) and/or DL-tyrosine (Sigma, min. 98%). All samples were mixtures of 2 to 4 of

these fluorophores. The concentration ranges of the fluorophores in the samples are stated in Table 1.

Table 1: Concentration ranges for the analytes.

Analyte	Concentration in 10^{-6}M
Catechol	0-87.0
Hydroquinone	0-22.5
Indole	0-5.46
Tryptophan	0-7.44
Tyrosine	0-12.14

The samples were prepared through several dilution steps with deionised water. First, a small amount of each analyte was weighted and transferred to a container. It was further diluted into standard strength, before they were mixed and diluted to the desired concentrations. The prepared samples were then measured by a Varian Eclipse Fluorescence Spectrometer. The settings for the instrument were: slit widths 5nm (for both excitation and emission), emission wavelengths 230-500nm (recorded every 2nm) and excitation wavelengths 230-320 (recorded every 5nm), scan rate 1920 nm/min and a PMT (photo multiplier tube) detector voltage of 600V. The sample was excited with lowest energy (highest excitation wavelength) first and then up to the highest energy excitation. Every sample was left in the instrument for a total of five replicate scans. The total recording time for one sample was approximately 15 min.

Every day a standard was run before and after analysis in order to ensure that there was no drift in the instrument.

3.2 Outlier detection and wavelength selection during calibration

Eq. (4) only accounts for the effect of random noise in the data as well as the model [3]. Consequently, it is crucial for the analyses performed here, that the calibration

model is well behaving. The term well behaving is difficult to quantify, but it means that the model is primarily reflecting the systematic variation foreseen in new samples. This can be quantified by assessing how close the model fit is to the (cross-) validated fit or by influence analysis. In line with this requirement, it is of importance to remove abnormal and extreme samples as well as irrelevant variables. Such outliers and irrelevant variables deteriorate any sound statistical evaluation of the model and lead to misleading statistics. Outlying behaviour can come from: pollution in the sample, irregular behaviour of the instrument, incorrect sample preparation, extremely high or low concentration of an analyte, etc.

One method for outlier detection is based on initial PLS analyses, and the visual inspection of the T vs. U-score plots. These plots describe the relationship between X and Y. For a good prediction model with relevant variables, these plots should be approximately straight lines indicating a predictive relation between X and Y. If a sample diverges significantly from this line compared to the other samples, it is an outlier – its relationship between X and Y is different from the rest.

Prior to analysis, part of the recorded data was removed in order to avoid any scattering effects that are present in fluorescence spectroscopy [13]. The emission wavelength ranges 230-296 nm and 422-500 nm were removed, together with the excitation wavelength ranges 230-240nm and 300-320nm. This reduced the landscapes from 136×19 (emission × excitation) down to 62×10. Further, only the first replicate measurement of each sample was used in this analysis.

3.3 Calculation of the measurement error in the reference values

The samples were prepared in several dilution steps. Consequently, there are several uncertainties to be accounted for upon estimating the final uncertainty of the reference values. In order to calculate the uncertainty induced by the different dilution steps, the following error propagation formula has been used:

$$\sigma(y) = \sqrt{\sum_{q=1}^{Q}\left(\frac{\partial y}{\partial x_q}\right)^2 \cdot \sigma(x_q)^2} \tag{6}$$

where 'σ' denotes the standard deviation in the associated variable, '∂' symbolises a partial derivative, Q is the total number of parameters with uncertainties and x_q ($q = 1, \cdots, Q$) is a parameter of y.

The steps involved in making the solution were as follows: an amount of solid was weighed and transferred to a 250ml container, 10 ml of this was taken out using a pipette, and transferred to a 100 ml container and 0.5 to 2 ml of this was taken out using an adjustable pipette and transferred to a 10 ml container.

An example of the uncertainty for one of these steps is as follows:

$$\sigma(V_{new}) = \sqrt{\left(\frac{\partial V_{new}}{\partial V_{old}}\right)^2 \cdot \sigma(V_{old})^2 + \left(\frac{\partial V_{new}}{\partial vol_P}\right)^2 \cdot \sigma(vol_P)^2 + \left(\frac{\partial V_{new}}{\partial vol_C}\right)^2 \cdot \sigma(vol_C)^2} \Rightarrow$$

$$\sigma(V_{new}) = \sqrt{\left(\frac{vol_P \cdot \sigma(V_{old})}{vol_C}\right)^2 + \left(\frac{V_{old} \cdot \sigma(vol_P)}{vol_C}\right)^2 + \left(\frac{V_{old} \cdot vol_P \cdot \sigma(vol_C)}{(vol_C)^2}\right)^2} \tag{7}$$

where V_{new} and V_{old} denote the new and old concentration, respectively, vol_P is the volume of the pipette used to transfer the analyte, and vol_C is the volume of the new container. Estimates of the measurement uncertainties for the different steps are most often given on the measuring equipment itself (see Table 2).

Table 2: Measurement uncertainties in the equipment used in the preparation of the samples.

Parameter	Amount	Uncertainty
Weight	0-1g	0.0001 g
$Volume_{C,1}$	250 ml	0.23 ml
$Volume_{C,2}$	100 ml	0.1 ml
$Volume_{C,3}$	10 ml	0.038 ml
$Volume_{P,1}$	10 ml	0.03 ml
$Volume_{P,0.5ml}$	0.5 ml	0.0075 ml
$Volume_{P,1ml}$	1 ml	0.008 ml
$Volume_{P,2.5ml}$	2.5 ml	0.015 ml

Each analyte was present at four levels of concentration, and can thus be coded from 1 to 4. Since the measurement error of the weight was constant over its range, and the relative uncertainty of the last pipette ($V_{P,0.5ml}$, $V_{P,1ml}$, and $V_{P,2.5ml}$) increased with decreasing volume, the uncertainties varied among the analytes, as well as for the different concentrations (see Table 3). The relative uncertainties vary from 0.8% for the highest concentration of catechol, hydroquinone and tyrosine, to 2.1% for the lowest concentration of indole. In the remainder of this paper the lowest reference uncertainty – 0.8% – is used as a lower bound of the uncertainty. This will give the most pessimistic results for the confidence limits, see Eq. (4). In this way, the presentation of unrealistically good results is avoided.

Table 3: Uncertainty in the reference value for the different analytes with varying concentration.

Analyte	Uncertainty in % Relative concentration			
	1	2	3	4
Catechol	1.6	0.9	0.9	0.8
Hydroquinone	1.6	1.0	1.0	0.8
Indole	2.1	1.6	1.6	1.5
Tryptophane	1.7	1.1	1.1	0.9
Tyrosine	1.6	1.0	1.0	0.8

4 Results and discussion

Calibration models were constructed using both N-PLS as well as standard PLS and PCR. Only the results of N-PLS are shown in the following, but similar results were obtained for the two-way calibration models.

4.1 Constructing the model

The data set was divided into a calibration set of 35 samples and a validation set of 86 samples.

Initial PLS models were made for each analyte in order to check for gross outliers. Two T vs. U-score plots – showing the relationship between X and Y – showed one outlier each, see Figure 2, hence the total number of samples were reduced to 129. Other traditional outlier diagnostics were also investigated, but no additional *gross* outliers were identified. By close inspection of the experimental setup and the two spectra in question, it was clear that the two samples were prepared wrongly. For the first outlier, the amount of catechol in the sample was less than it should be according to the experimental setup. For the second outlier the hydroquinone concentration was higher than planned. So, not only do these outliers make sense

from a statistical point of view, but it is also possible to backtrack and investigate the actual reason for the outlying behaviour.

Figure 2: T vs. U plots from PLS of a) Catechol – PC5, and b) Hydroquinone – PC3, both showing one clear outlier.

Eight additional samples were excluded during specific model building as potential outliers. These samples were not extreme outliers as the above-mentioned. For most, reasonable explanations for the outlying behaviour was possible to find but not for all. However, in order for the distributional properties of the uncertainties to be meaningful, even moderate outliers are necessary to remove especially when validating the method. Thus, given the large sample size, even some debatable outliers were removed in order to be absolutely certain that these did not bias the calibration model or the evaluation of the prediction results. In a more realistic setting, the decision on which samples to remove could be different depending on purpose, but in this paper, the main issue is to show that the formulas work for data of good quality. Univariate regression statistics do not work well when influential samples are present, and the same holds in multivariate regression. The issue here is not finding the limit for when the formulas work but to show that they do provide meaningful results for absolutely meaningful data. Hence, we want to remove (a little too many) objects to make sure that the results are not due to an unfortunate choice of samples.

The choice of the calibration set size was based on having enough samples to adequately span the space of interferents and analyte. The remainder of the samples was then assigned as the test set. A relatively large test set is required to verify the properties of the distribution of studentised prediction residuals. Only tryptophan is discussed in the following as an example. Its concentration ranged from zero to 7.443 μmol/L, but the concentrations were scaled such that the maximum value was 1 (for convenience of plotting). The design was purposely chosen so that some

of the prediction samples were slightly outside the region of the calibration samples. This was done in order to test the approach in a demanding (extrapolation) situation (Figure 3).

Figure 3: Second (14% variance explained)) versus first (68%) principal component using calibration and prediction spectral data. The plot shows how the prediction samples extend beyond the space of the calibration samples.

The adequacy of Eq. (4) depends directly (and almost solely) on having an accurate estimate of the calibration error. Recall that Eq. (4) is obtained by propagating random errors through the model; any systematic error in the model itself is not accounted for. The presence of significant model error shows up in, for example, a proper test set or cross-validated mean square error. If the model error is insignificant, then it follows that MSEC and MSECV should be of similar size. In case there is a significant model error, larger deviations are expected. Thus, by monitoring the gap between MSEC and MSECV, an indication of the validity of the formula can be obtained.

Figure 4: Calibration errors as a function of the number of components used.

In this investigation, the number of components was chosen manually based on a comparison of the cross-validated and fitted predictions. In Figure 4 the calibration (RMSEC) and cross-validation (RMSECV) results are shown for the current data problem. The prediction results (RMSEP) are also shown for convenience as well as the concentration reference uncertainty (SDy) which is situated almost on top of the 0 axis. A six-component model was chosen due to the minimum value of RMSECV and also due to the small gap between RMSEC and RMSECV at this number of components.

The predictions obtained with the six-component model are shown in *Figure 5*.

Figure 5: Predictions for calibration (left) and test set (right). Target line superimposed.

To verify that the sample-specific standard errors of predictions can be trusted it is necessary to test that the local linearization of the error in the regression vector estimate is a valid approximation. By noise-addition as described in the theory section, predictions of the test set samples are obtained for different levels of random noise added. The noise level was determined from the residuals of a six-component PCA model of the spectral data. With the added noise, six-component PLS models were determined and used to predict the test set samples.

Figure 6: Prediction standard errors for test set predictions using N-PLS and unfold PLS with different levels of noise.

As can be seen in Figure 6, the standard error increases linearly with the noise level at least up to twice the intrinsic amount of noise. After this point, the added noise leads to a non-linear relation between amount of added noise and prediction standard error. Because of the linearity for low levels of added noise, the local linearization and the equations for sample-specific standard errors can be assumed to be valid from this point of view.

Figure 6 also shows that the predictions from unfold PLS are more sensitive to the added noise than is the case for N-PLS. This is in agreement with expectations for low-noise low-rank trilinear data where the excess free parameters in unfold PLS are not needed for describing the systematic part of the data and hence lead to more noisy regression coefficient estimates.

4.2 Validation of proposed sample-specific uncertainty estimate

Using Eq. (4), the predictions in Figure 5 (right) can be assigned individual uncertainty estimates. In Figure 7, this uncertainty is reported as twice the estimated standard error. Also shown is the prediction interval that is calculated as twice the RMSEP. In that case, the same interval is obtained for all samples. For low concentration samples this typically (though not automatically) leads to too large intervals while the converse holds for high concentrations.

Figure 7: Test set predictions with uncertainty. For each prediction, two 'confidence limits' are given (almost superimposed). The left-most is calculated from Eq. 4, while the right-most is twice the RMSEP. The uncertainty in the reference value is also shown but is of insignificant magnitude.

For the intervals in Figure 7 the sample specific intervals vary by 115% as opposed to the constant size obtained from RMSEP. It is possible to assess how the sample-specific error coverage compares with an overall RMSEP based coverage. The set level RMSEP assumes identical prediction error but as this is not a valid assumption (compare with the univariate case), a systematic bias is expected from the RMSEP based coverage. Indeed, the RMSEP-based coverage of all samples is too high for low levels and too low for high levels (Table 4). The coverage based on sample-specific prediction errors is in better agreement with the theoretically expected results.

Table 4. Percentage coverages.

Level	.70	.80	.90	.95	
Theoretical	60	69	77	82	All samples
Sample-specific	62	70	76	82	
RMSEP	67	72	79	80	
Theoretical	34	39	44	47	Zero-
Sample-specific	36	40	41	46	concentration
RMSEP	40	41	46	46	samples

In Figure 8 the studentized residuals are shown for the prediction residuals in all and for the zero-concentration sample residuals. The theoretically expected standard deviation is $\sqrt{f/(f-2)}$ where f = 35-7 which equals 1.04. Although not perfect, the empirically observed values are close. This even holds for the zero-concentration samples; a result that is important e.g. for determining limit-of-detection. Clearly a similar approach based on RMSEP would hold little promise (Table 4).

Figure 8. Studentized prediction residuals calculated using Eq. (5) for (a) entire test set and (b) zero-concentration samples.

5 Conclusions

In this paper the approach for estimation of sample-specific prediction errors developed in part 1 has been has been put to the test. The example shows that when the assumptions are carefully assessed to be valid, the sample-specific errors provide a more adequate and detailed view on the prediction errors than e.g. obtained from traditional RMSEP based overall errors. The approach is based on a local linearization, but in this paper, the direct connection to earlier approaches to sample-specific prediction errors has also been highlighted. The obtained results are stimulating and point to several useful developments e.g. for limit-of-detection estimation.

Acknowledgement

The authors gratefully acknowledge support from the Centre for Advanced Food Studies and the Danish Technical Research Council (project 1179) for financial support.

Reference List

1. Martens H, Næs T, *Multivariate calibration*. Wiley & Sons, Chichester, 1989.

2. Faber NM, Song XH, Hopke PK, Sample-specific standard error of prediction for partial least squares regression, *Trac-Trends in Analytical Chemistry,* 2003, **22**, 330-334.

3. Faber NM, Bro R, Standard error of prediction for multiway PLS 1. Background and a simulation study, *Chemom Intell Lab Syst,* 2002, **61**, 133-149.

4. Fernandez Pierna JA, Jin L, Wahl F, Faber NM, Massart DL, Estimation of partial least squares regression prediction uncertainty when the reference values carry a sizeable measurement error, *Chemom Intell Lab Syst,* 2003, **65**, 281-291.

5. DiFoggio R, Examination of some misconceptions about near-infrared analysis, *Appl Spectrosc,* 1995, **49**, 67-75.

6. Coates DB, *Spectroscopy Europe,* 2002, **14**, 24-26.

7. Faber NM, Kowalski BR, Improved prediction error estimates for multivariate calibration by correcting for the measurement error in the reference values, *Appl Spectrosc,* 1997, **51**, 660-665.

8. Sørensen LK, *Journal of Near-Infrared Spectroscopy,* 2002, **10**, 15-25.

9. Wolthuis R, van Aken M, Fountas K, Robinson JS, Bruining HA, Puppels GJ, Determination of water concentration in brain tissue by Raman spectroscopy, *Anal Chem,* 2001, **73**, 3915-3920.

10. The American Society for Testing and Materials (ASTM), Practice E1655-00. ASTM Annual Book of Standards, *Vol. 03.06*, *ASTM*, 2001, 573-600.

11. Höskuldsson A, PLS regression methods, *J Chemom,* 1988, **2**, 211-228.

12. PerkinElmer Inc., Quant+ software.2002,

13. Lakowicz JR, *Principles of Fluorescence Spectroscopy*. Kluwer Academic, New York, 1999.

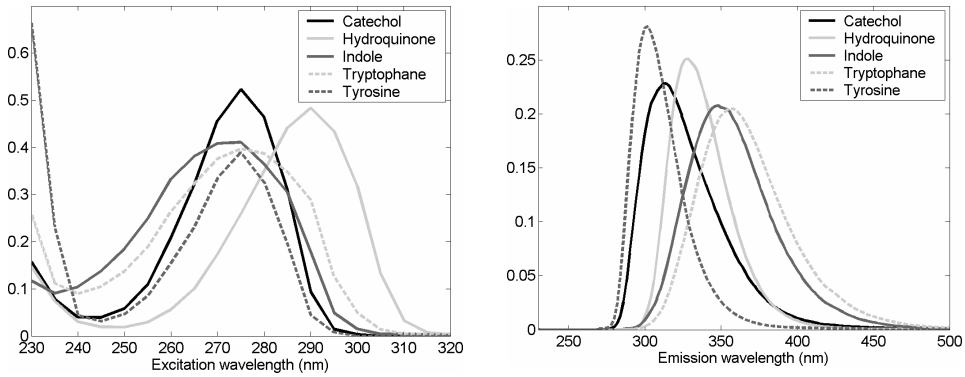
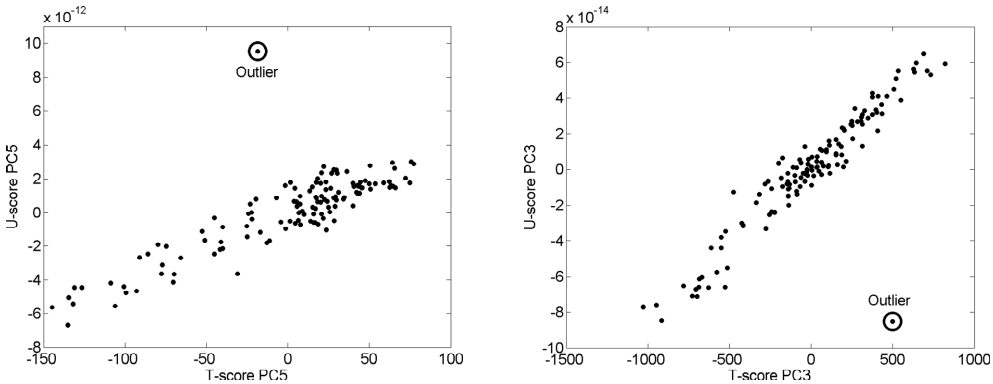


FIGURE 1



a) b)

FIGURE 2

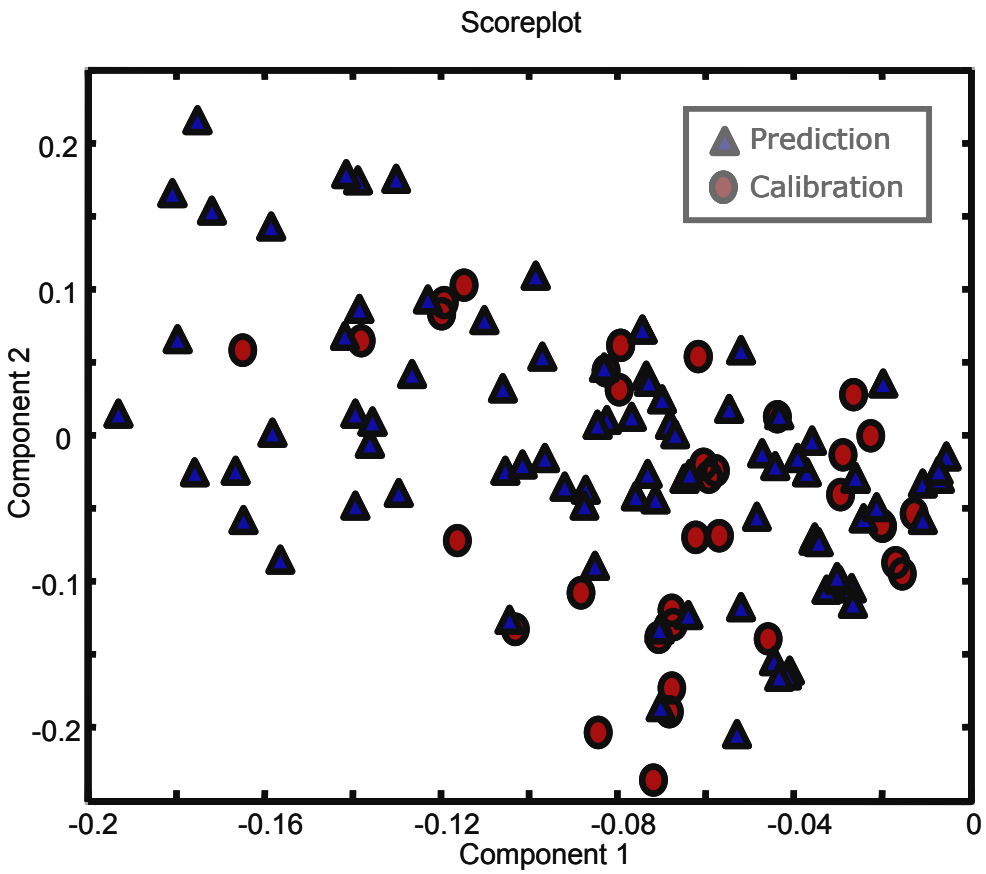

Scoreplot
Prediction
Calibration
Component 2
Component 1
0.2
0.1
0
-0.1
-0.2
-0.2
-0.16
-0.12
-0.08
-0.04
0

FIGURE 3

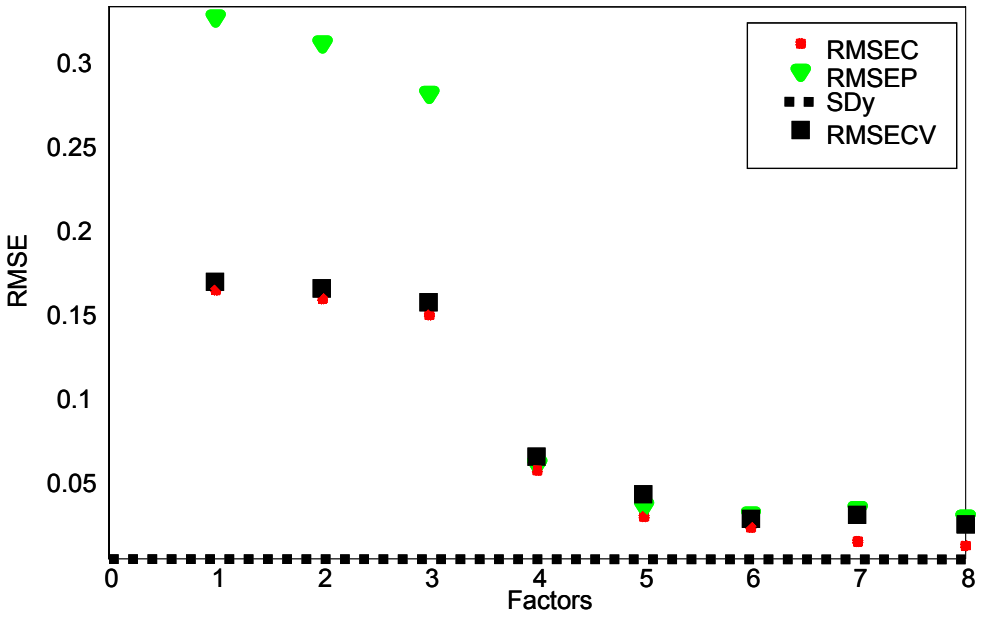


FIGURE 4

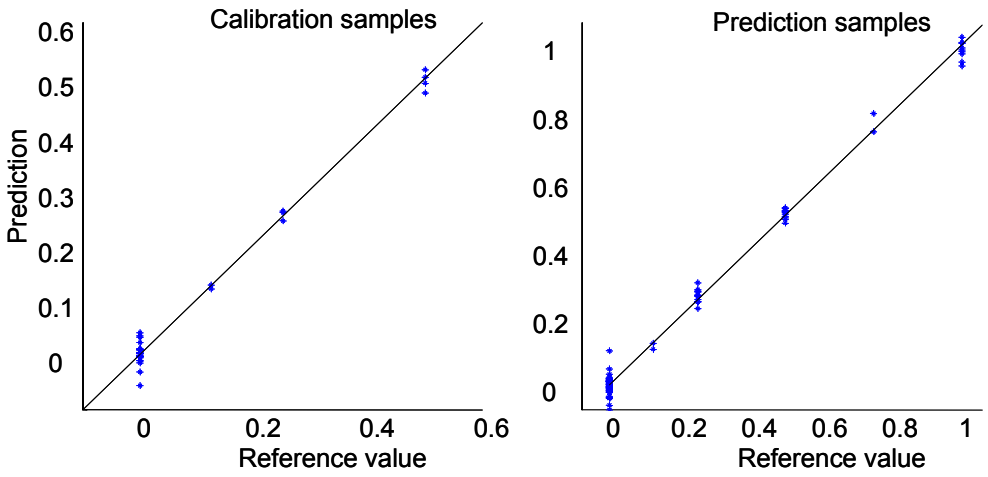


FIGURE 5

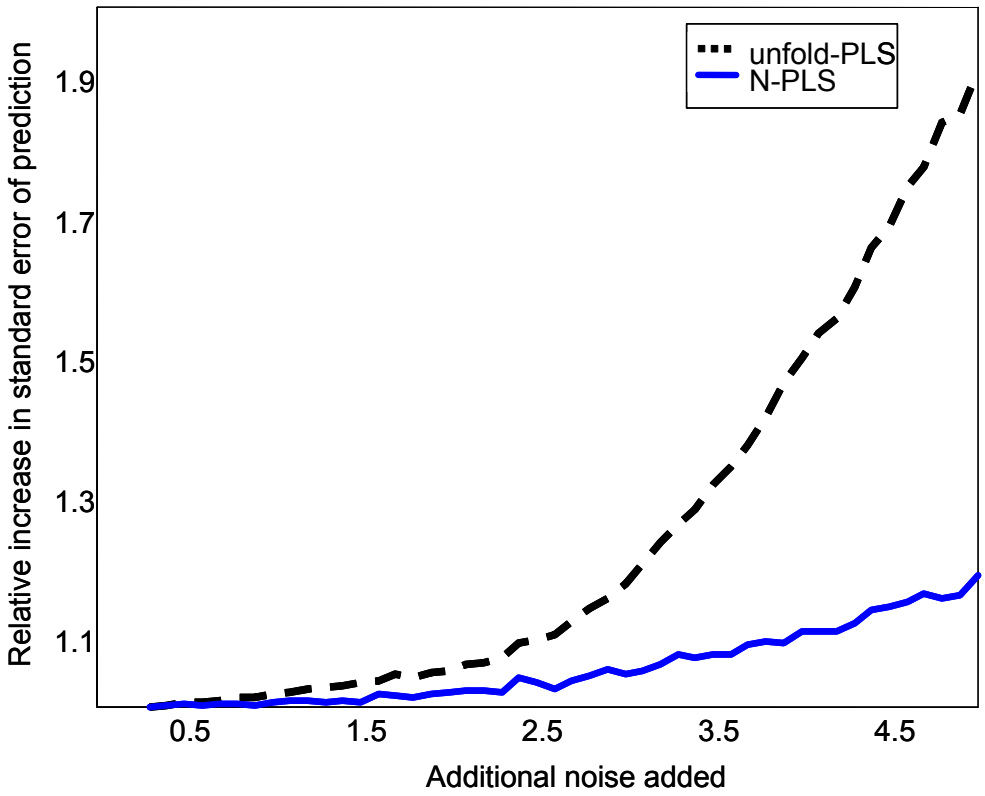

unfold-PLS
N-PLS
Relative increase in standard error of prediction
1.9
1.7
1.5
1.3
1.1
0.5
1.5
2.5
3.5
4.5
Additional noise added

FIGURE 6

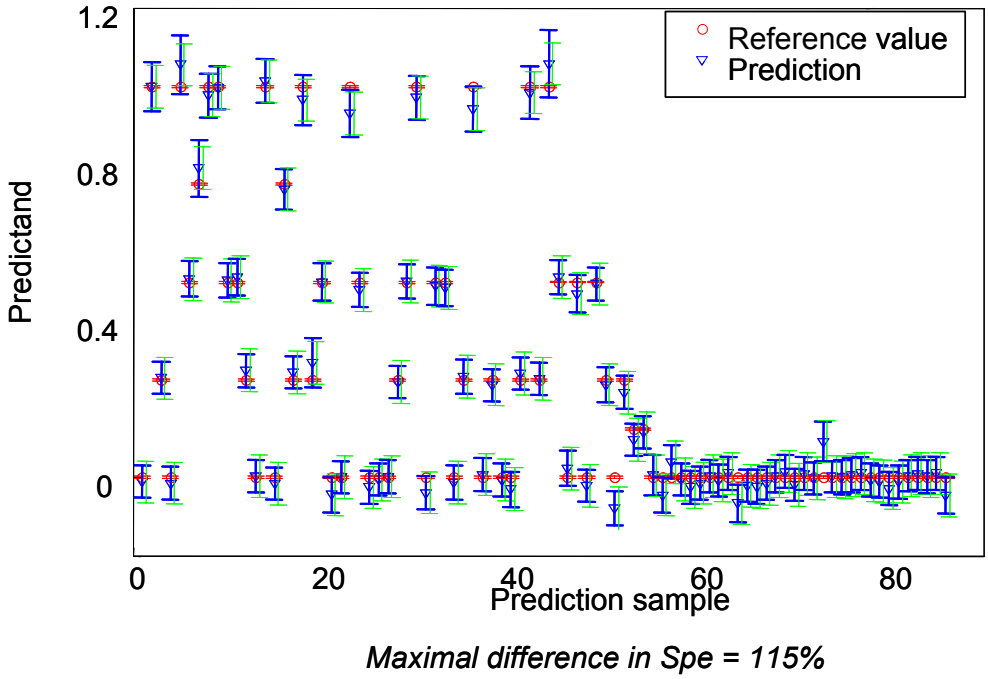

Reference value
Prediction
Predictand
1.2
0.8
0.4
0
0
20
40
60
80
Prediction sample
Maximal difference in Spe = 115%

FIGURE 7

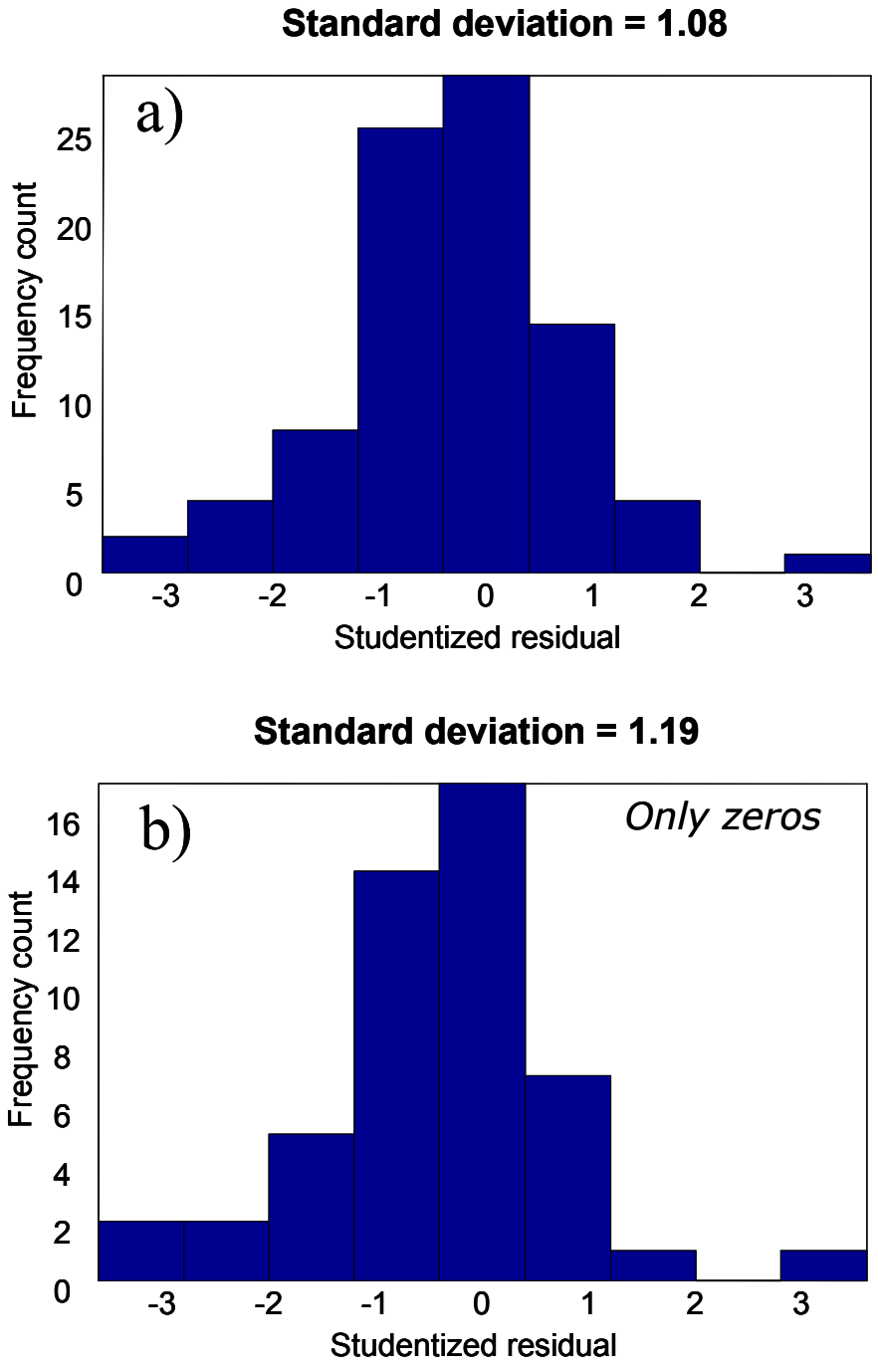

Standard deviation = 1.08
a)
Frequency count
25
20
15
10
5
0
-3
-2
-1
0
1
2
3
Studentized residual
Standard deviation = 1.19
b)
Only zeros
Frequency count
16
14
12
10
8
6
4
2
0
-3
-2
-1
0
1
2
3
Studentized residual

FIGURE 8