Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Maunu Toiviainen

Near-Infrared Spectroscopy of Solids: Chemometrics and Signal Processing

Master's thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Technology

Kuopio, June 1st 2009

Supervisor: Prof. Olli Simula
Instructors: D.Sc. (Tech.) Pekka Teppola, D.Sc. (Eng.) Francesco Corona

HELSINKI UNIVERSITY OF TECHNOLOGY Faculty of Information and Natural Sciences Degree Programme of Engineering Physics and Mathematics		ABSTRACT OF MASTER'S THESIS
Author	Maunu Toiviainen	Date: June 1st 2009 Pages: x + 62
Title of thesis	Near-Infrared Spectroscopy of Solids: Chemometrics and Signal Processing	
Professorship	Computer and Information Science	Code: T-115
Supervisor	Prof. Olli Simula	
Instructors	D.Sc. (Tech.) Pekka Teppola, D.Sc. (Eng.) Francesco Corona	

Near-inrared (NIR) spectroscopy permits the remote analysis of solid samples in the diffuse reflectance (DR) measurement mode. However, uncontrolled physical variations between solid samples, such as changes in packing density and particle size distribution, have a complex non-linearizing effect on the NIR spectra which complicates the subsequent extraction of chemical information from the measured data.

The non-linearities may be removed from the NIR spectra by using model-based spectral preprocessing methods, such as extended multiplicative signal correction (EMSC) and optical path length estimation and correction (OPLEC). Such methods are designed to make the subsequently developed linear calibration model more robust to spectral artifacts of physical origin. In this work, an implementation for the optimized version of EMSC is proposed, in which the spectral model is chosen so that the error of cross validation is minimized for the subsequent calibration model. The performance of the method is tested on a laboratory data set comprising NIR DR spectra of ternary powder mixtures.

The physical interferences make the application of blind source separation (BSS) methods, which attempt to blindly factorize the measured mixture spectra into the pure analyte spectra and their concentration profiles, difficult on the NIR spectra of solids. The unique aspects of applying BSS in NIR DR spectra are discussed in detail, and a three-phase preprocessing procedure of the measured spectral signals, which is designed to improve the separation capability of BSS methods, is proposed in this work. The method is tested and the explanatory power of BSS is demonstrated using both the laboratory data set and process data measured during a pharmaceutical fluid bed granulation process.

Keywords: Near-infrared spectroscopy, Diffuse reflectance, Chemometrics, Multivariate calibration, Spectral preprocessing, Simulated annealing, Blind source separation, Independent component analysis

TEKNILLINEN KORKEAKOULU Informaatio- ja luonnontieteiden tiedekunta Teknillisen fysiikan ja matematiikan koulutusohjelma	DIPLOMITYÖN TIIVISTELMÄ

Tekijä	Maunu Toiviainen	Päiväys	1.6.2009
		Sivumäärä	x + 62
Työn nimi	Kiinteän aineen lähi-infrapunaspektroskopia: kemometria ja signaalinkäsittely		
Professuuri	Informaatiotekniikka	Koodi	T-115
Työn valvoja	Prof. Olli Simula		
Työn ohjaajat	TkT Pekka Teppola, D.Sc. (Eng.) Francesco Corona		

Kiinteitä aineita voidaan analysoida kajoamattomasti diffuusissa heijastusmittauksessa lähi-infrapunaspektroskopiaa (NIR) käyttäen. Näytteiden väliset fyysiset erot esimerkiksi pakkaustiheydessä ja partikkelikokojakaumassa voivat kuitenkin aiheuttaa mitattuihin spektrisignaaleihin epälineaarisuuksia, jotka vaikeuttavat kemiallisen tiedon louhintaa mittausdatasta.

Lineaaristen kalibraatiomallien herkkyyttä fysikaalisperäisille häiriöille voidaan vähentää esikäsittelemällä mitatut spektrit häiriöt mallintavilla menetelmillä kuten EMSC ja OPLEC. Tässä työssä toteutetaan joustava algoritmi EMSC-menetelmässä käytettävän spektrimallin optimoimiseksi siten, että esikäsittelyn jälkeen rakennetun lineaarisen kalibraatiomallin ristikkäisvalidointivirhe minimoituu. Menetelmän suorituskykyä verrataan muihin esikäsittelymenetelmiin käyttäen laboratorio-olosuhteissa mitattujen kolmikomponenttijauheseosten spektridataa.

Sokean lähteiden erottelun (BSS) menetelmien, jotka yrittävät ratkaista puhtaiden kemiallisten aineiden spektrit ja konsentraatioprofiilit mitatuista seosten spektreistä, soveltaminen kiinteiden aineiden NIR-signaaleihin on hankalaa fysikaalisperäisten häiriöiden takia. Tässä työssä käsitellään kyseisen sovelluskentän haasteita yksityiskohtaisesti, ja työssä esitellään kolmivaiheinen NIR-spektrien esikäsittelymenetelmä, jonka tarkoituksena on parantaa BSS-menetelmien erotuskykyä. Menetelmää testataan ja BSS-menetelmien havainnollistamiskykyä esitellään käyttäen sekä laboratoriossa, että leijupetirakeistusprosessin aikana mitattua dataa.

Avainsanat

Lähi-infrapunaspektroskopia, diffuusi heijastusmittaus, kemometria, monimuuttujakalibraatio, spektrisignaalien esikäsittely, simuloitu jäähdytys, sokea lähteiden erottelu, riippumattomien komponenttien analyysi

Acknowledgements

This thesis was made as a part of the PROMISENS project at Technical Research Centre of Finland (VTT) during fall 2008 and spring 2009. The project was funded by Tekes – Finnish Funding Agency for Technology and Innovation, as well as several Finnish and international industrial partners.

I would like to express my gratitude to my instructor D.Sc. (Tech.) Pekka Teppola for providing me with guidance and excellent working facilities at VTT. I would also like to thank Prof. Olli Simula for his supervision and D.Sc. (Eng.) Francesco Corona for his invaluable comments during the documentation of this work.

Furthermore, I would like to thank Ms. Maiju Järvinen, Mr. Sami Poutiainen and Ms. Anne Heikkilä for accompanying me in the fluid bed granulation process measurements. The staff of VTT Kuopio all deserve special thanks for creating a positive working atmosphere.

I want to thank my family for their support during my studies. Special thanks are reserved for Minna for standing by me all these years.

Maunu Toiviainen

Kuopio, June 1st 2009

Contents

List of Figures

List of Tables

Nomenclature

AD	Analog to digital
API	Active pharmaceutical ingredient
BSS	Blind source separation
DR	Diffuse reflectance
DSS	Denoising source separation
EMSC	Extended multiplicative signal correction
FBG	Fluid bed granulation
IC	Independent component
ICA	Independent component analysis
LBO	Leave-block-out
LOO	Leave-one-out
LS	Least-squares
LV	Latent variable
MCC	Microcrystalline cellulose
MCR	Multivariate curve resolution
MCT	Mercury-cadmium-telluride
MLR	Multiple linear regression
MSC	Multiplicative signal correction
NIR	Near-infrared
OPLEC	Optical path length estimation and correction
PC	Principal component
PCA	Principal component analysis

PCR	Principal component regression
PDF	Probability density function
PGP	Prism-grating-prism
PLSR	Partial least squares regression
PVP	Polyvinylpovidone
RMSECV	Root-mean-squared error of cross validation
RMSEP	Root-mean-squared error of prediction
SA	Simulated annealing
SMA	Subminiature A
SVD	Singular value decomposition

Chapter 1

Introduction

1.1 Motivations and objectives

Near-infrared (NIR) spectroscopy [1] utilizes the interaction of electromagnetic radiation with the vibrational energy states of molecular groups in providing indirect information on both chemical and physical properties of a broad variety of materials. The defining property of NIR spectroscopy is the relatively low absorption of the NIR active molecular groups which permits the possibility to perform non-contact measurements also on solid materials in the diffuse reflectance (DR) mode, where the sample is illuminated and the spectrum of the backscattered light is measured. Minimal sample preparation is thus required and fast inline analysis of even non-stationary samples is possible. Complemented by its robustness, flexibility and inexpensive instrumentation, NIR spectroscopy has thus gained a wide acceptance in both laboratory and industrial process analytical applications [2].

In order to efficiently extract the desired information from the measured spectra, the implementation of a successful NIR application requires expertise from both instrumentation and data analysis. Chemometrics is a scientific discipline which involves the use of mathematical and statistical methods for efficient acquisition and analysis of chemical data [3]. The analysis of spectroscopic data involves the extraction of quantitative and qualitative information of chemical and physical nature from the measured spectra. In a typical occasion involving quantitative analysis, a mathematical transfer function, i.e., a multivariate calibration model, is developed between the response variable measured with a primary laboratory method, such as the concentration of a chemical species, and the secondary measurements such as the multidimensional NIR spectra. The model may then be used to predict the value of the explained variable from the subsequently measured NIR spectra, thus reducing expensive and time consuming laboratory analysis.

NIR DR spectra are usually transformed into apparent absorbance units by taking the negative logarithm of the measured reflectance. Under the simplifying assumption on the validity of the Beer-Lambert's law, the spectra are now assumed to follow the linear mixture model so that the measured absorbance values are linearly proportional to the concentrations of the pure analytes. Linear calibration models, such as principal

component regression (PCR) and partial least squares regression (PLSR) are thus widely applied in NIR spectroscopy [4].

Spectrophotometric analysis of solid materials often involves sample-to-sample variability caused by the scattering of light. Physical properties, such as packing density of the sample and the particle size and shape, etc., have a complicated nonlinear effect on the spectrum of backscattered light in DR measurements. These physical variations may be so prevailing that the desired chemical information is masked behind them, and the direct applicability of standard linear calibration models is not straightforward and may result in a deteriorated prediction ability. The linearity may be attempted to be regained with model-based spectral preprocessing methods, such as extended multiplicative signal correction (EMSC) [5, 6] and optical path length estimation and correction (OPLEC) [7]. Both methods utilize a modified version of Beer-Lambert's law in which the light scattering effects in an apparent absorbance spectrum are approximated as additive smooth wavelength-dependent terms and a multiplicative coefficient. Preprocessing involves the estimation and removal of these errors from the modeled spectra so that only the linear part of the spectral signal containing relevant chemical information is left. EMSC involves the assumption that chemically relevant spectral signals, which effectively span the vector space containing chemical information, be known *a priori*. OPLEC in turn requires the prior knowledge on the mass fraction of the target analyte in the calibration spectra already in the preprocessing phase.

The presence of nonlinearities evoked by physical light scattering effects and the subsequent deviations from the linear mixture model makes the application of blind source separation (BSS) methods in NIR DR spectra difficult. When applied in optical spectroscopy, BSS involves the estimation of the pure analyte spectra and their concentration profiles given only the measured mixture spectra and little or no additional prior information. BSS methods may prove useful, e.g., in providing chemically meaningful qualitative information for so called black systems for which no reference data or knowledge on the pure analyte spectra are available [8]. One approach to BSS is independent component analysis (ICA) [9] which linearly factorizes the measured mixture spectra into source signals whose statistical dependencies have been minimized. Most BSS methods assume the validity of the linear mixture model, and the estimation of the underlying signals from NIR DR spectra is thus a challenge which might be solved with spectral preprocessing.

1.2 Overview and contributions

In this work, the problematics of physical light scattering effects are dealed with in the context of NIR DR measurements of pharmaceutical powders. The manufacturing of solid dosage forms involves several subprocesses, such as powder blending and granulation, which need to be monitored inline, and in which NIR DR spectroscopy is often applied [10]. Two sets of NIR DR measurement data were created: A laboratory data set of ternary powder mixtures with different particle sizes and hence varying

light scattering properties was prepared, and an inline measurement on fluid bed granulation (FBG) process simulating a black system was conducted.

A flexible algorithm for optimized EMSC is proposed, in which the chemically relevant spectral signals in the EMSC model are optimized as linear combinations of some given base vectors so that the error of cross validation is minimized for the constructed calibration model. The performance of the method is compared to that of OPLEC and regular EMSC using the laboratory data set. Similar approach to the modeling of NIR DR spectra as in EMSC and OPLEC was applied also in the context of BSS. A combination of preprocessing methods designed for ICA with the application on NIR DR spectra, involving the removal of baseline offset by zero-meaning, smoothing by rank reduction and enhancement of the statistical propreties of the signals by differentiation, is proposed and justified. The combination of preprocessing and ICA is applied in the qualitative analysis of both the FBG process data and the laboratory data.

The work is organized as follows. Chapt. 2 starts with an overview on the physical principles of NIR spectroscopy, the unique characteristics and the interpretation of NIR spectra and the applications of NIR spectroscopy in the analysis of pharmaceutical solids. The utilized data analysis methods, viz., multivariate calibration, spectral preprocessing methods and BSS are discussed thereafter. The algorithmic contribution of the Author is given in Sect. 2.4.3, where the optimized version of EMSC is proposed. Sect. 2.5.2 and 2.5.3 describe the preprocessing methods which are needed when BSS methods are applied on NIR DR spectra. Chapt. 3 begins with a description on the phases and monitoring needs of an FBG process. The utilized NIR instrument, the measurements conducted both inline and in laboratory as well as the origins of the ulitized algorithms are detailed thereafter. The results achieved by applying both BSS and spectral preprocessing methods on the data sets are reported in Chapt. 4.

Chapter 2

Theory

2.1 Physical principles of near-infrared spectroscopy

Fundamental transitions in the vibrational energy states of molecules are observed generally in mid-infrared (MIR) region of the electromagnetic spectrum, corresponding to the wavenumber range of 4000–400 cm^{-1}, or the wavelength range of 2.5–25 μm. In a typical spectroscopic measurement a polychromatic beam of MIR radiation is incident on a sample containing MIR active molecular groups, and the transmitted radiation exhibits significant attenuation at the distinct wavenumbers that correspond to the transition energies between the fundamental and the first excited vibrational states of the molecules. In the process, radiation energy is transferred into mechanical energy associated with the vibrational motion of MIR-active molecules with permanent dipole moment. If near-infrared (NIR) radiation, generally accepted as the wavelength range of 740–2500 nm (or 13 500–4000 cm^{-1}), is used instead, the absorption peaks are seen to be less intense, broader and heavily overlapping. Characteristics of MIR absorption bands, such as location on the wavelength axis and intensity, can be understood using the quantum mechanical harmonic oscillator to model the potential energy in molecular vibrations, as is described in Sect. 2.1.1. The origin of NIR absorption bands are combination and overtone bands of the fundamental transitions, and their origin can be understood with the use of the anharmonic oscillator model (cf. Sect. 2.1.2).

The physical principles of vibrational spectroscopy are now described at sufficient depth so that the reader is able to understand the unique characteristics of NIR spectra. The overview is based on the Ref. [1, 11, 12], where references to more comprehensive treatments of the subject can be found.

2.1.1 The harmonic oscillator model

At small displacements from the equilibrium r_e, the potential energy function of the vibrational oscillations of the diatomic molecule in free space illustrated in Fig. 2.1a can be approximated with the ideal harmonic oscillator [1]

$$V(r) = \frac{1}{2}k\,(r_e - r)^2 \tag{2.1}$$

depicted in Fig. 2.1b. Here, k is the force constant, which describes the strength of the molecular bond, and r is the distance between the two atoms. The classical vibrational frequency of the model (in s^{-1}) is

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}}, \tag{2.2}$$

where $\mu = m_1\, m_2/(m_1 + m_2)$ is the reduced mass of the system such that m_1 and m_2 are the masses of the two atoms.

A quantum mechanical treatment of the diatomic system reveals that the molecular vibrational energy can now take only discrete levels given by

$$E_{\mathrm{vib}} = h\nu\left(\mathrm{v} + \frac{1}{2}\right), \tag{2.3}$$

where h is the Planck's constant and v is the vibrational quantum number which takes only nonnegative integer values $\mathrm{v} = 0, 1, 2\ldots$. In optical spectroscopy, Eq. (2.3) is usually written in wavenumber units as follows

$$G(\mathrm{v}) = E_{\mathrm{vib}}/hc = \bar{\nu}\left(\mathrm{v} + \frac{1}{2}\right), \tag{2.4}$$

where c is the speed of light and $\bar{\nu}$ is the wavenumber (in cm^{-1}) corresponding to the classical vibrational frequency. The probability of transition from the vibrational state $\mathrm{v_i}$ to $\mathrm{v_f}$ is given by the transition dipole moment [1]

$$P_{\mathrm{v_i}\to\mathrm{v_f}} = \int \psi^*_{\mathrm{v_i}}\epsilon\psi_{\mathrm{v_f}}\,\mathrm{d}^3r, \tag{2.5}$$

where ϵ is the dipole moment of the molecule, ψ_i denotes the wave function of the molecule at the vibrational state i and * indicates complex conjugation. Integration is performed over all space. The numerical value of the transition dipole moment is proportional to the intensity of the absorption band. The transition probability is nonzero only if the initial and final vibrational states both involve a change in the dipole moment of the diatomic molecule. Hence, a transition is possible only for a heteronuclear diatomic oscillator with permanent dipole moment. If the dipole moment of the system can be represented as a linear function of the displacement, i.e.,

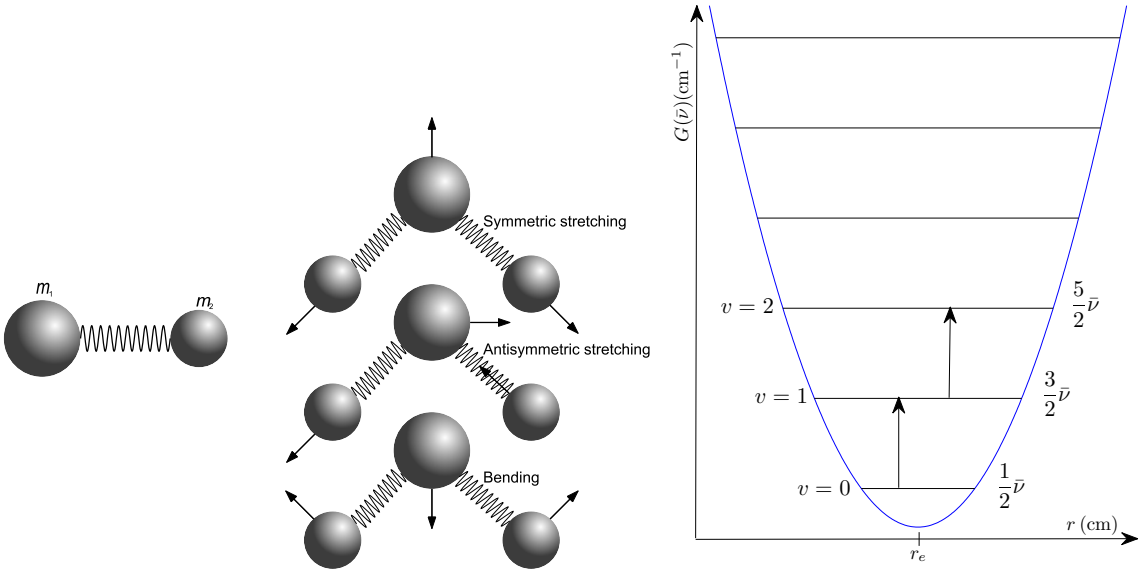
$$\epsilon = \epsilon_{\mathrm{e}} + \left(\frac{\mathrm{d}\epsilon}{\mathrm{d}r}\right)_{\mathrm{e}} r, \tag{2.6}$$

which is the case for the classical definition of the electric dipole moment, it can be shown that $P_{\mathrm{v_i}\to\mathrm{v_f}} \neq 0$ only if $\mathrm{v_i} - \mathrm{v_f} = \pm 1$ [11]. Due to the equally spaced energy levels (cf. Eq. (2.4)), the selection rule thus states that an increase in the vibrational energy state of the diatomic harmonic oscillator can occur only through the absorption of a photon of the single wavenumber

$$\bar{\nu} = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}}. \tag{2.7}$$

Strong chemical bonds and light atoms thus generally exhibit absorption peaks at shorter wavelengths.

The majority of the absorption phenomena can be attributed to the fundamental transition ($\mathrm{v_i} = 0$) $\rightarrow$ ($\mathrm{v_f} = 1$) as most molecules are in their ground vibrational state in room temperature. According to the Boltzmann distribution of vibrational energies, the probability of the "hot bands", i.e., the transitions between the higher quantum numbers such as ($\mathrm{v_i} = 1$) $\rightarrow$ ($\mathrm{v_f} = 2$), is increased at higher temperatures [1].



(a) Diatomic and triatomic molecules

(b) Harmonic potential

Figure 2.1: (a) Diatomic molecule consisting of atoms with masses m_1 and m_2. The three vibrational modes of a triatomic molecule are also shown citebrittainsops; (b) The harmonic potential function illustrated with the equidistant energy levels. The fundamental transition and a the first "hot band" are illustrated with arrows. The symbols are explained in the main text.

A nonlinear (linear) polyatomic molecule containing N atoms has $3N-6$ ($3N-5$) vibrational degrees of freedom, or normal modes of vibration, in which all atoms oscillate in phase with a unique, mode-specific frequency [11]. The three vibrational modes of a nonlinear triatomic molecule, such as H_2O, are shown in Fig. 2.1a [11]. Each of the three modes can be regarded as an independent harmonic oscillator with the fundamental wavenumber $\bar{\nu}_i$ and the quantum number v_i, $i = 1, 2, 3$. The wavefunction of the triatomic molecule can thus be written as a product of the wavefunctions describing the states of the independent vibrational modes, i.e., [11]

$$\psi_{\mathrm{v_1 v_2 v_3}} = \psi_{\mathrm{v_1}} \psi_{\mathrm{v_2}} \psi_{\mathrm{v_3}}. \tag{2.8}$$

The vibrational energies are now given by [11]

$$G(\mathrm{v_1},\, \mathrm{v_2},\, \mathrm{v_3}) = \sum_{i=1}^{3} \bar{\nu}_i \left(\mathrm{v}_i + \frac{1}{2} \right). \tag{2.9}$$

The selection rule for the system consisting of independent harmonic oscillators states that only one quantum number of the system is allowed to change at a time by only one step, i.e., $\sum_{i=1}^{3} |\Delta \mathrm{v}_i| = 1$. The molecule is thus allowed to absorb electromagnetic energy only at three discrete wavenumbers $\bar{\nu}_i$, $i = 1, 2, 3$. Again, the initial and final vibrational states must involve a change in the dipole moment of the molecule.

The harmonic oscillator model adequately explains the fundamental transitions $\mathrm{v_i} = 0 \rightarrow \mathrm{v_f} = 1$ in the vibrational energy state of both diatomic and polyatomic heteronuclear molecules which are observed as strong, discrete absorption peaks in their MIR absorption spectra. With most molecules, the fundamental transition energies are restricted to the MIR range on the wavelength scale. The harmonic oscillator model does not predict any transitions to occur in the NIR range.

2.1.2 The anharmonic oscillator model

In reality, the chemical bond between two atoms is expected to rupture if high energy is transferred to the molecule and the inter-atomic distance is sufficiently increased. Also, the potential energy should approach infinity as the inter-atomic distance approaches zero. The realistic potential energy function between two atoms can be approximated by supplying Eq. (2.1) with higher order polynomial terms or using the Morse function [1]

$$V(r) = D_e \left(1 - e^{-a(r-r_e)}\right)^2, \tag{2.10}$$

where D_e is the dissociation energy and a quantifies the rigidity of the chemical bond. The energy levels of the anharmonic oscillator can be solved with the perturbation method, and their second order approximation is

$$G(\mathrm{v}) = \bar{\nu}\left(\mathrm{v} + \frac{1}{2}\right) - \chi_e \bar{\nu}\left(\mathrm{v} + \frac{1}{2}\right)^2. \tag{2.11}$$

The nonnegative anharmonicity constant χ_e quantifies the amount of mechanical anharmonicity which is observed as unequal spacings between adjacent energy levels. The spacings decrease as the quantum number is increased. The anharmonic potential and the second order approximations for the energy levels are illustrated in Fig. 2.2a.

The electrical dipole moment of a diatomic molecule deviates from the classical linear expression of Eq. (2.6) and its real value can be approximated with a polynomial series. The higher order terms in its expression give arise to electrical anharmonicity which results in a nonzero transition probability in Eq. (2.5) for overtone transitions involving quantum number changes greater than one ($\Delta \mathrm{v} = \pm 2, \pm 3, \ldots$) [11]. The overtone transitions are observed in the NIR wavelength range as broader and fainter absorption peaks when compared to the fundamental transitions in the MIR range.

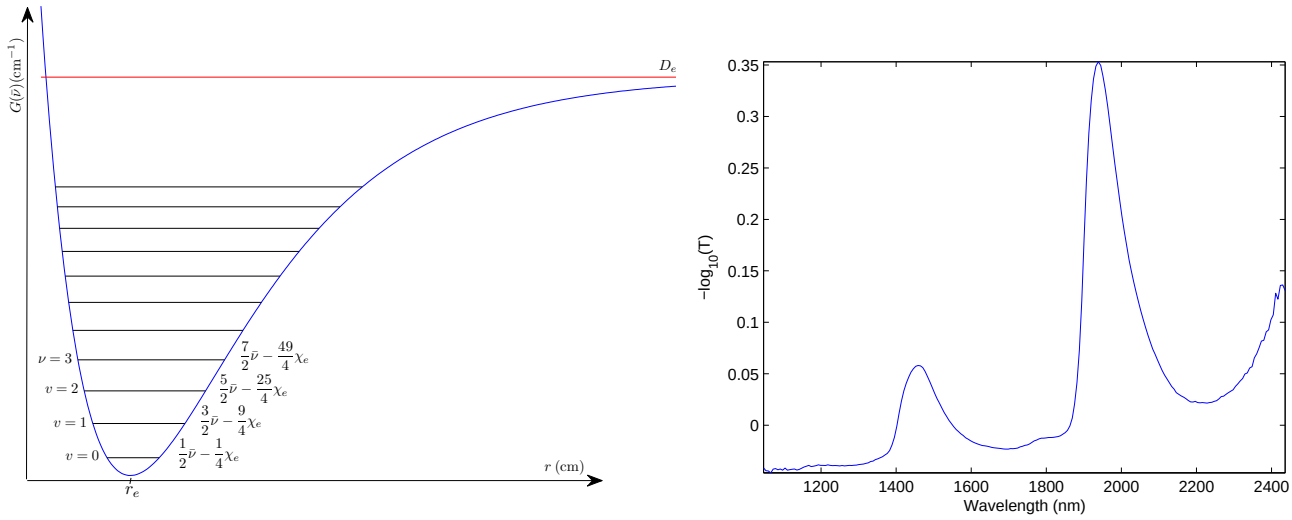
The electrical anharmonicity permits overtone transitions also for the normal modes of vibration in polyatomic molecules, so that only one quantum number changes at a time as $\Delta \mathrm{v}_i = \pm 2, \pm 3, \ldots$. Furthermore, the combination transitions, where several $\Delta \mathrm{v}_i \neq 0$ in a single absorption event, become possible. The intensity of a given

transition can again be evaluated by the transition moment of Eq. (2.5). At least one of the normal modes of vibrations must induce a change in the dipole moment of the molecule in both the initial and final vibrational states. Moreover, the integrand must be symmetric in order to produce nonzero transition probability. The feasibility of a given transition can be also examined with molecular symmetry-based group theory [11].

The most common absorption bands observed in the NIR region originate from polar molecular bonds between the light hydrogen atom and a heavier atom. The NIR functional bonds O–H, C–H, N–H and S–H have small reduced mass and a strong chemical bond which results in a relatively high fundamental vibrational energy in the region of 3000–4000 nm (cf. Eq. (2.7)). Their first overtones are thus observed in the NIR region. Many molecules have lower fundamental vibrational energies and they subsequently exhibit their first overtones already in the MIR region, leaving only the weak second overtones to occur the NIR range. The low-wavenumber part of NIR absorption spectra between 1600–2500 nm consists of combinations of fundamental transitions and the first overtones of the functional groups containing X–H bonds. The second overtones are generally observed in the range 1300–1600 nm and the third overtones in 750–1300 nm. NIR absorption spectra thus exhibit a sloping baseline as the probability of transition decrease by a factor of 10–100 with each increment in the order of overtone [1]. When moving towards the high-energy end of a representative NIR spectrum, the absorption peaks also tend to become broader, as the vibrational modes begin to decouple and move more independently at higher energy levels.

NIR spectra are further complicated by resonance effects, interactions between molecular groups and temperature changes. Fermi and Darling-Dennison resonance phenomena may occur under certain conditions involving small differences between fundamental and overtone or combination bands, symmetry of two vibrational modes and large degree of anharmonicity. They may be observed as two absorbance bands at a location where only one is expected [12]. Changes in the location, intensity and width of an absorbance band occur when the bond strength, dipole moment or anharmonicity of an involved vibrational state is altered [11]. These alterations occur, e.g., in the neighboring group effect, where the inspected NIR active group interacts with the neighboring molecular groups within the same molecule. In bulk matter, the NIR active group may interact with the neighboring molecules in the sample matrix through, e.g., hydrogen bonding which tends to make the vibrations more harmonic. The transition band positions and intensities of a NIR active group thus vary according to the structure of rest of the molecule and the presence and nature of the surrounding molecules. A sample has, e.g., slightly different NIR spectra in the gas and bulk phases. Increase in temperature tends to increase the anharmonicity of especially light molecules involving hydrogen bonding which again results in changes in band position, width and intensity. Water, which consists of light molecules with hydrogen bonding interactions, has a high propensity for temperature effects which often complicates NIR spectroscopic measurements in the presence of moisture and unstable temperature conditions.

The strong absorption bands of water near 1936 and 1455 nm, and a medium band near 1800 nm, are shown in Fig. 2.2b. The bands can be assigned to the combina-



(a) Anharmonic potential

(b) NIR spectrum of water

Figure 2.2: (a) The anharmonic potential approximated with the Morse function. The second order approximations of the energy levels are shown. The fundamental transition and the first and second overtones are illustrated with arrows. The symbols are explained in the main text; (b) The NIR absorption spectrum of water measured with the instrumentation detailed in Sect. 3.2.

tion transitions $\psi_{000} \rightarrow \psi_{011}$, $\psi_{000} \rightarrow \psi_{101}$ and $\psi_{000} \rightarrow \psi_{021}$, respectively, where the quantum numbers describe the symmetric stretching, antisymmetric stretching and bending vibrational states of the triatomic H_2O molecule (cf. Fig. 2.1a), respectively [11]. Numerous fainter absorption bands are superposed on the water spectrum, and their presence can be observed, e.g., by taking a second order difference with respect to the wavelength channels [1].

2.2 NIR spectroscopy of pharmaceutical solids

NIR spectroscopy can be used in the analysis of both clear and turbid materials. Transmittance measurements made on non-scattering samples can be transformed into absorbance units which, according to Beer-Lambert's law, results in a simple linear relationship between the absorbance and the concentrations of the chemical species present in the measured sample. Turbid materials deviate from the assumptions of Beer-Lambert's law since the chemically relevant absorption of light is now convoluted with light scattering effects in the measured spectra. The transformation of transmittance spectra into apparent absorbance units results then in signals with difficulty of interpretation. Propagation of light in turbid and strongly scattering media can be modeled using the equation of radiative transfer [13], from which the absorption and scattering coefficients of the sample can be solved using, e.g., the numerical Monte Carlo method [14] or the analytical diffusion approximation [15]. In this work, however, only the absorbance transformation is utilized, and it is later shown in Sect. 2.4

and 2.5 that the desired chemical information may then be extracted at sufficient accuracy even from DR measurements made on scattering samples.

Next, Sect. 2.2.1 overviews the transmittance measurement mode in NIR spectroscopy, the Beer-Lambert's law and the linear mixture model inherent to the absorbance spectra. The interaction of light with solid material is qualitatively described in Sect. 2.2.2, where the DR measurement mode and the challenges involved with it are overviewed. The effect of light scattering on apparent absorbance spectra are illustrated with examples considering changes in moisture content and particle size distributions of the measured samples. Applications of NIR spectroscopy in the analysis of pharmaceutical solids are given in both sections.

2.2.1 Beer-Lambert's law and transmittance measurements

According to Beer-Lambert's law [11], light intensity exhibits exponential decay as a function of distance in transparent, purely absorbing media. The absorbance, defined as the natural logarithm of inverted transmittance, is linearly proportional to both the optical path length and the concentrations of the chemical species present. For a mixture of J non-scattering analytes, the absorbance at the wavelength λ is given by

$$A(\lambda) = \log\left(\frac{1}{T(\lambda)}\right) = \log\left(\frac{I_0(\lambda)}{I(\lambda)}\right) = l\sum_{j=1}^{J} c_{\mathrm{mol},\,j}\epsilon_{\mathrm{mol},\,j}(\lambda) \tag{2.12}$$

where $I_0(\lambda)$ and $I(\lambda)$ are the light intensities entering and transmitting the sample at the wavelength λ, respectively (for derivation, see [11]). The characteristic molar absorptivity of the jth analyte at the wavelength λ is denoted with $\epsilon_{\mathrm{mol},\,j}(\lambda)$ (in $\mathrm{L\,mol^{-1}cm^{-1}}$), $c_{\mathrm{mol},\,j}$ is its molar concentration (in $\mathrm{mol\,L^{-1}}$) and l is the optical path length (in cm) which is equal to the sample thickness for non-scattering samples. The model assumes linear mixing such that sample matrix interactions between the analytes are negligible and the absorptivities of the analytes do not change as a function of concentration. Mass concentrations $c_{\mathrm{mass},\,j}$ (in $\mathrm{g\,L^{-1}}$) can be alternatively used in the model in which case each molar absorptivity is to be divided by the molar mass of the corresponding analyte to obtain $\epsilon_{\mathrm{mass},\,j}(\lambda)$ (in $\mathrm{L\,g^{-1}cm^{-1}}$). Mass fractions $c_{\mathrm{w},\,j}$, which obey the closure constraint $\sum_{j=1}^{J} c_{\mathrm{w},\,j} = 1$, may also be attempted to be used in the model as

$$A(\lambda) = l\left(\sum_{i=1}^{J} c_{\mathrm{mass},\,i}\right)\sum_{j=1}^{J}\frac{c_{\mathrm{mass},\,j}}{\sum_{i=1}^{J} c_{\mathrm{mass},\,i}}\epsilon_{\mathrm{mass},\,j}(\lambda) = b\sum_{j=1}^{J} c_{\mathrm{w},\,j}\epsilon_{\mathrm{mass},\,j}(\lambda)\,. \tag{2.13}$$

Here, the optical path length and the sum of the mass concentrations have been merged into the coefficient b. The sum of the mass concentrations varies between samples in which the J analytes have been mixed in different proportions unless all J chemical species have equal densities. Thus, the coefficient b is expected to generally vary between measurements made on different samples even if the optical path length remains constant, and the absorbance values are not expected to be linearly proportional to the mass fractions $c_{\mathrm{w},\,j}$.

Multichannel spectroscopic data are usually modeled according to the linear mixture model as

$$\mathbf{x}_{i,\,\mathrm{chem}} = \sum_{j=1}^{J} c_{ij}\mathbf{s}_j, \;\; i = 1, 2, \ldots, I, \tag{2.14}$$

The elements of the $L \times 1$ column vector $\mathbf{x}_{i,\,\mathrm{chem}}$ are the absorbances at L narrow contiguous wavelength bands. If the I samples are measured in a cuvette of constant thickness, the constant optical path length can be merged with the characteristic absorptivities to yield the pure analyte spectra $\mathbf{s}_j$. The units of the concentration variables c_{ij} are application dependent. In matrix notation, the measured absorbance spectra are usually transposed and presented as the rows of the $I \times L$ matrix $\mathbf{X}$ as in

$$\mathbf{X} = \mathbf{C}\,\mathbf{S}^{\mathrm{T}}, \tag{2.15}$$

where the columns of the $I \times J$ matrix $\mathbf{C}$ are the concentration profiles of the J pure analytes, and the columns of $\mathbf{S}$ are the pure analyte spectra. The superscript T denotes matrix transposition.

In the case of turbid media, such as pharmaceutical powders, the scattering of light makes the aforementioned linear mixture model invalid. Through reflection and refraction, photons are deviated from their original trajectories each time they are obliquely incident on an interface between two substances of different refractive indices. In dense fine-particulate media, the direction of a photon is constantly changed as it frequently encounters such interfaces between the suspending medium, such as air, and the particles. The distance of photon travel in the medium thus exhibits variance around a mean photon path length which is larger than the thickness of the sample. As the refractive indices are not constant with respect to wavelength, different wavelengths are refracted at varying amounts, and the distribution of the photon path length varies between wavelength channels. Furthermore, the apparent optical path length has a complex relationship with the morphology, particle size distribution and the packing density of the solid. Optical path length is proportional to the apparent absorbance, and varying optical path length hence results in a multiplicative error in the spectra.

Despite the nonlinearity induced by the scattering of light, the $\log(1/T)$ transform is often used with turbid samples. Under stable measurement conditions and small concentration ranges, the linear response between the apparent absorbance and the concentrations holds approximately. In pharmaceutical applications, NIR transmittance spectroscopy has been found useful in the analysis of the solid dosage forms due to the large mass sampled by the NIR light which permits the analysis of the whole tablet in one measurement [11]. Transmittance measurements are performed mainly in the third overtone region 750–1300 nm due to the generally less intense absorption which permits high transmissivity even for relatively thick samples.

2.2.2 Diffuse reflectance measurement mode

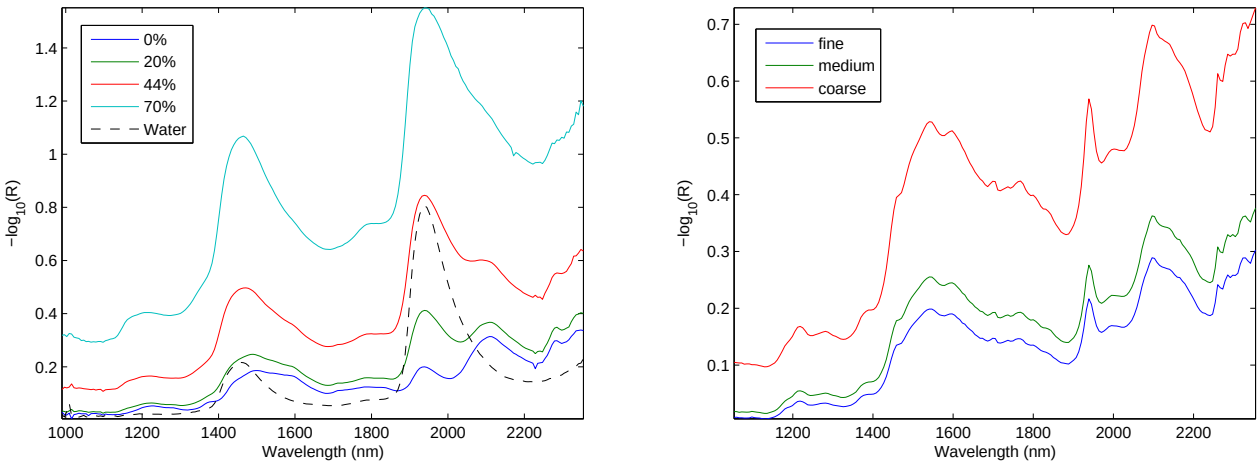
The capability to perform fast, non-contact and inexpensive measurements on turbid samples has made NIR spectroscopy the method of choice in industrial process monitoring applications. In diffuse reflectance (DR) measurement mode, the sample is

illuminated with a source of NIR radiation, such as halogen lamp, and the diffusely reflected light is either detected with a light detector or collected into optical fibers. Diffusively reflected photons interact with the interior of the sample by dipping into the material before re-emerging on the surface after multiple scattering events. Contrary to transmittance measurements where samples must be placed in cuvettes, DR measurements require minimal sample preparation which facilitates inline measurements.

The reflectance $R(\lambda)$ of a sample is defined as the ratio of intensities for light diffusively reflected from the sample, $I(\lambda)$, and light diffusively reflected from the surface of a non-absorbing reference material placed at the location of the sample, $I_0(\lambda)$. The reference material, such optical teflon, should reflect all wavelengths with equal intensity and have no absorption bands in the NIR range. Although the measurement geometry and the sample properties are very different from the assumptions of the Beer-Lambert's law, DR spectra are usually transformed into apparent absorbance units with the transformation $\log(1/R)$ [11]. Again, the apparent optical path length is wavelength-dependent and it has a complex relationship with the physical characteristics of the sample. Thus, the linearity between the apparent absorbance units and the concentration of the target analyte is often lost, and the interpretation of the measured spectra is complicated. The chemical information in the DR spectra is interfered with physical effects, resulting from changes in the light scattering properties of the sample, which might, however, provide useful information on the physical state of the sample in their own right.

The strong absorption bands of water in the NIR region makes NIR spectroscopy suitable for moisture measurements. The effect of moisture in $\log(1/R)$ spectra is demonstrated in Fig. 2.3a, where microcrystalline cellulose (MCC) powder is mixed with water at three different weight ratios. In addition to the increase of apparent absorbance at the water band locations, the baseline offset of the measured spectra is increased along with moisture. As the air in the inter-particulate spaces is replaced with water, the discontinuity in the refractive index in the interface between the suspending medium and the particles is decreased which results in diminished refraction of light [4]. The probability for a diffusely scattered photon to return to the surface of incidence is thus reduced and the baseline effect is attributed to the decrease in the total intensity of the backscattered light. The presence of moisture might also have a complex effect also on the apparent optical path length. Decrease in the refraction of light causes the backscattered photons to travel longer paths in the medium on average. On the other hand, the presence of water in the inter-particulate spaces increases the probability of absorption which favors shorter optical path lengths. The net effect of these two interactions determines the correlation between the moisture content and the spectral amplitude, i.e., the multiplicative error.

Varying particle size distribution in powders has a similar baseline effect in the $\log(1/R)$ spectra as moisture. Fig. 2.3b presents the spectra of three chemically identical anhydrous lactose powders with different particle size distributions. It is seen that coarse powder exhibits smaller intensity for the backscattered light, hence the increase in the apparent absorbance. This effect can be understood using the Mie theory for light scattering, according to which larger particles have greater tendency



(a) Effect of moisture

(b) Effect of particle size

Figure 2.3: DR measurements. (a) Mixtures of microcrystalline cellulose (MCC) and water. The mass percentages of water are given in the figure. Scaled water spectrum is shown for reference; (b) Lactose powders of three different particle size distributions. The spectra were measured with the instrumentation and materials elaborated in Chapt. 3.

to forward scatter light [16]. Similarly to the moisture effect, fewer photons are now diffusely scattered back to the direction of incidence. In addition to the baseline effect, the spectrum of the coarsest sample also has larger amplitude than the spectra of the two finer powders. This is due to the increased apparent optical path length caused by the forward-scattering tendency of the large particles.

NIR spectroscopy is routinely used in the monitoring of moisture content in wet pharmaceutical manufacturing processes, such as wet granulation (cf. Sect. 3.1). The capabilities of NIR reflectance spectroscopy in particle size analysis have also been studied extensively [17]. Linear regression models for both the median particle size [18] and the complete discretized percentage particle size distribution [19, 20] in pharmaceutical powders have been successfully constructed.

NIR DR spectroscopy is also often used to measure the content of the active pharmaceutical ingredient (API), e.g., in content uniformity analysis during the mixing of powders [21]. Understanding of the measurement scale involved and the random nature of sampling is important in such applications [22]. DR measurements gather information only from the outermost layers of the samples and high absorbtivity generally implies small sampling volume. Hence the shorter wavelengths in the third overtone region penetrate deeper into the material than the first overtone wavelengths above 2000 nm. The sampling volume, which also depends on the density of the material, is often quantified by the information depth, which is defined as the thickness of a layered sample at which the intensity of diffusely backscattered reaches its maximum [23]. As the thickness of a layered sample is increased, the intensity of backscattered light increases monotonically until the information depth is reached, since fewer

photons are lost in transmission. In a typical non-tapped MCC powder of the density $0.30\,\mathrm{g/cm^3}$, the third overtone wavelengths reach the information depth of about 2 mm, whereas the low-frequency end of the NIR spectrum is unable to penetrate deeper than 0.5 mm [23]. The small sampled volume makes NIR DR spectroscopy relatively inaccurate measurement method. Powder mixtures are usually very heterogeneous at the involved measurement scale and single-point measurements hence suffer from nonrepresentative sampling [22]. The effect of segregation, in which small particles tend to move towards the bottom layers of the mixture to fill the interparticulate voids, is a major cause of mixture heterogeneity when free-flowing powders are handled [22]. The detection limit for the mass percentage of a given chemical species is usually of the order of 0.1 % w/w in NIR DR measurements [12].

2.3 Multivariate calibration

Chemometrics is a subdiscipline of analytical chemistry which involves the use of mathematical and statistical tools in chemical and process analytical problems. Its purpose is to permit efficient collection of measurement data and extraction of relevant information from them [24]. Data analysis in chemometrics can be coarsely divided into exploratory and confirmatory methods [25]. The former involves, e.g., the application of unsupervised learning techniques in the extraction of meaningful features and qualitative attributes, such as the identity of chemical species, from the measured data. Explanatory data analysis involves the extraction of quantitative information form the measurements, e.g., by constructing calibration models.

Calibration models are extensively used in spectroscopy to relate the measured L-dimensional explanatory variable $\mathbf{x}_i$ (the discretized spectra) to an explained variable y_i, which is usually the mass fraction or concentration of a target analyte in chemical mixtures. Due to the multidimensional nature of the spectra, calibration methods are multivariate. Also, the explained variable is multidimensional in the general case of multivariate calibration [4], but only scalar values are treated in this work. The purpose of calibration is to reduce the amount of labor-intensive laboratory work, such as the use of wet chemical methods in determining the absolute concentration values, by constructing a mathematical model

$$y_i = f\left(\mathbf{x}_i\right),\ i = 1, \ldots, I, \tag{2.16}$$

which can be used to predict the value of the dependent variable from the subsequent measurements $\mathbf{x}_{\mathrm{new}}$ obtained with a fast and economical method, such as NIR DR spectroscopy.

Due to the linear mixture model (Eq. (2.14)), linear regression models are chemically interpretable [26] and they are naturally applicable for the prediction of concentration values from optical absorbance spectra. Given the calibration set, i.e., the reference values $\mathbf{y} = [y_1\, y_2 \cdots y_I]^{\mathrm{T}}$ and the measured spectra $\mathbf{X}$, a forward linear regression model [4] can be written as

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta}. \tag{2.17}$$

In multiple linear regression (MLR), the regression vector $\boldsymbol{\beta}$ is estimated in least-squares (LS) sense as

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\,\mathbf{X}\right)^{-1}\,\mathbf{X}^{\mathrm{T}}\,\mathbf{y}, \tag{2.18}$$

where the superscript -1 denotes matrix inversion. The predicted y-value for a new measurement $\mathbf{x}_{\mathrm{new}}$ is now given by

$$\hat{y}_{\mathrm{new}} = \mathbf{x}_{\mathrm{new}}^{\mathrm{T}}\,\hat{\boldsymbol{\beta}}. \tag{2.19}$$

To account for a constant term in the model, the spectral vectors are often supplied with an extra element with the constant value of unity.

Due to the general smoothness of NIR spectra, the matrix $\mathbf{X}$ often exhibits a high degree of collinearity in NIR applications. The absorbance value at one wavelength channel can often be very accurately represented as a linear combination of the absorbance values at some other neighboring channels in NIR spectra. Hence the columns of $\mathbf{X}$ are linearly dependent, and the matrix inversion in Eq. (2.18) is unstable. Moreover, MLR is very sensitive to noise and thus prone to overfitting [4].

To remove the problem of collinearity and to mitigate the effect of noise in calibration, the calibration model is often written in bilinear form as

$$\mathbf{X} = \mathbf{T}\,\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{2.20a}$$

$$\mathbf{y} = \mathbf{T}\,\mathbf{q} + \mathbf{f}. \tag{2.20b}$$

The rank of the matrix $\mathbf{X}$ is reduced by approximating each measured spectrum as a linear combination of loadings, the columns of the $L \times A$ loading matrix $\mathbf{P} = [\mathbf{p}_1 \cdots \mathbf{p}_A]$. The number of loadings A is chosen so that all significant spectral variations are explained by them. Usually A is equal to or slightly larger than the chemical rank of $\mathbf{X}$, i.e., the number of pure analytes present in the mixtures J. The variations unmodeled by the A loadings are given by the matrix $\mathbf{E}$, which is expected to contain only uninformative noise. The projections of the sample vectors onto the loading vectors, the scores, are given by the columns of the $I \times A$ matrix $\mathbf{T} = [\mathbf{t}_1 \cdots \mathbf{t}_A]$. The vector $\mathbf{q}$ can be thought as the regression coefficients of $\mathbf{y}$ on $\mathbf{T}$. Now $\mathbf{T}$ is a full-rank matrix with a stable pseudo-inverse, and the regression coefficients $\hat{\mathbf{q}}$ can be estimated similarly to Eq. (2.18). The unmodeled variations in the reference $\mathbf{y}$ are given by the vector $\mathbf{f}$. The prediction phase involves the estimation of the scores $\hat{\mathbf{T}}_{\mathrm{new}} = [\hat{t}_1 \cdots \hat{t}_A]$ for the new measurement $\mathbf{x}_{\mathrm{new}}$, after which the predicted value of the dependent variable is given by $\hat{y}_{\mathrm{new}} = \hat{\mathbf{T}}_{\mathrm{new}}\,\hat{\mathbf{q}}$.

The bilinear model is often constructed with principal component regression (PCR), where the matrix $\mathbf{X}$ is orthogonalized using singular value decomposition (SVD)

$$\mathbf{X} = \mathbf{U}\,\boldsymbol{\Sigma}\,\mathbf{V}^{\mathrm{T}}. \tag{2.21}$$

The columns of $\mathbf{U}$ and $\mathbf{V}$ are the orthonormal eigenvectors of the unitary matrices $\mathbf{X}\,\mathbf{X}^{\mathrm{T}}$ and $\mathbf{X}^{\mathrm{T}}\,\mathbf{X}$, respectively. The non-zero elements of the diagonal matrix $\boldsymbol{\Sigma}$ are square roots of the eigenvalues of the two unitary matrices sorted in decreasing order. Usually $\mathbf{X}$ is mean-centered prior to SVD to obtain $\mathbf{X} = \left[\mathbf{X} - \mathbf{1}\,\overline{\mathbf{x}}^{\mathrm{T}}\right]$, where $\overline{\mathbf{x}}$ is the

mean of all I measured spectra. In this case $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is proportional to the covariance matrix of $\mathbf{X}$, and each column of $\mathbf{V}$ accounts for a part of variance in $\mathbf{X}$ proportional to the corresponding squared diagonal element of $\mathbf{\Sigma}$. Now the SVD in Eq. (2.21) is called the principal component analysis (PCA) factorization of $\mathbf{X}$. The principal component score and loading matrices for the bilinear model in Eq. (2.20a) are obtained by collecting the contributions of $A < I$, usually the most significant, PCs as

$$\mathbf{T} = \mathbf{U}_{(:,1:A)}\,\mathbf{\Sigma}_{(1:A,1:A)} \text{ and } \mathbf{P} = \mathbf{V}_{(:,1:A)}, \tag{2.22}$$

where, mimicking MATLAB notation, the first A columns of $\mathbf{U}$ are given by $\mathbf{U}_{(:,1:A)}$. Before prediction, the scores of the new spectrum are estimated by projecting it onto the orthonormal loading vectors as $\hat{\mathbf{T}}_{\mathrm{new}} = \mathbf{x}_{\mathrm{new}}^{\mathrm{T}}\,\mathbf{P}$.

In PCR, the loadings are chosen on the basis how well they explain the variance in $\mathbf{X}$. It is difficult to conclude which combination of PCs gives best predictive performance for the calibration model, since the bilinear model does not reveal how well they correlate with the dependent variable $\mathbf{y}$. Partial least squares regression (PLSR) [4] solves this problem by utilizing the reference values $\mathbf{y}$ in the decomposition of $\mathbf{X}$ into scores and loadings. PLSR has many variations, but the standard version includes another set of loadings, the orthonormal loading weights as the columns of the $L \times A$ matrix $\mathbf{W}$. The ith unit loading weight vector $\mathbf{w}_i$ is chosen so that the covariance between the residual reference $\mathbf{y}_{i-1}$ and the ith score vector $\mathbf{t}_i = \mathbf{X}_{i-1}\,\mathbf{w}_i$ is maximized. The jth residual reference $\mathbf{y}_j$ is the original $\mathbf{y}$ from which the contributions of the j first chemical loadings $\mathbf{t}_i\,q_i$, $i = 1, \ldots, j$ have been subtracted. Similarly, the residual matrix $\mathbf{X}_j$ is the original $\mathbf{X}$ from which the contributions of the j first loadings $\mathbf{t}_i\,\mathbf{p}_i^{\mathrm{T}}$, $i = 1, \ldots, j$ have been subtracted. Due to the updating of $\mathbf{X}$, both the scores $\mathbf{T}$ and the loading weights $\mathbf{W}$ are orthogonal. The generally nonorthogonal loadings $\mathbf{P}$ needed in the bilinear model are estimated in LS sense using the previously estimated scores and the corresponding residual matrices $\mathbf{X}_j$. Similarly, the loadings $\mathbf{q}$ are estimated in LS sense using the residual reference $\mathbf{y}_{i-1}$ and the scores $\mathbf{t}_i$. Both the data matrix $\mathbf{X}$ containing the mixture spectra and the reference vector $\mathbf{y}$ are usually mean-centered prior to developing the PLSR model.

The two sets of loadings, $\mathbf{P}$ and $\mathbf{W}$ generally closely resemble each other and both can be used in the interpretation of the PLSR model. The loadings are in decreasing order with respect to explained variance in $\mathbf{y}$. In the ideal case of pure linear mixture model (Eq. (2.14)) excluding any baseline offsets, multiplicative errors or random noise, the number of loadings, or latent variables (LVs), needed to completely explain the bilinear model equals the chemical rank of the system, J [26]. Often the number of LVs is chosen to be slightly larger than the expected chemical rank so that the deviations from the linear mixture model can be explained.

The optimal number of LVs is often selected by analyzing the root-mean-squared error of cross validation (RMSECV) of the calibration set defined as

$$\mathrm{RMSECV} = \sqrt{\frac{\sum_i^I \left(\hat{y}_i - y_i\right)^2}{I}}. \tag{2.23}$$

In leave-one-out (LOO) cross validation one sample is treated as a validation sample while the rest of the samples are used to construct a calibration model which is

used to predict the reference value of the validation sample $\hat{y}_i$. The procedure is repeated I times so that all samples have been used for validation once, and the number of selected LVs is the one that minimizes the RMSECV. In leave-block-out (LBO) cross validation, the calibration set is divided into a predetermined number of blocks consisting of approximately equal number of samples. Similarly to LOO, each block is used for validation at a time, while the others are used for calibration in the calculation of the RMSECV.

When developing a calibration model, some samples are additionally separated into an independent test set which is left out of the cross validation procedure and is used only for the evaluation of the prediction ability of the constructed calibration model. The measure of the prediction ability is usually the root-mean-squared error of prediction (RMSEP) which has the same definition as RMSECV in Eq. (2.23) with the exception that the reference values are now drawn from the test set whose size is not necessarily equal with the calibration set.

2.4 Spectral preprocessing for scattering samples

Spectral preprocessing is usually performed prior to multivariate calibration in an attempt to remove noise and spectral interferences, which do not correlate with the reference and are thus irrelevant to the analysis, from the measured spectra. In NIR DR spectrosocopy, the purpose of spectral preprocessing is to make the calibration model more robust with respect to sample-to-sample variations caused by the effects of light scattering due to, e.g., varying particle size distribution and powder packing density. Instrumental variations, such as varying measurement geometry, the use of different measurement heads or instrumental drift, can also be attempted to be removed from the measured data.

Two model-based spectral preprocessing methods, EMSC [5, 6] and OPLEC [7, 27] are briefly reviewed below. The selection of the reference and signal spectra in EMSC is discussed and an implementation for an optimized version of EMSC is proposed in Sect. 2.4.3.

2.4.1 Extended multiplicative signal correction

The physical effects in the $\log(1/R)$ or $\log(1/T)$ spectra of turbid media can be approximated with the following parameterized soft model based on the Beer-Lambert's law [5, 6]

$$\mathbf{x}_i = a_i\mathbf{1} + b_i\mathbf{x}_{i,\,\mathrm{chem}} + d_i\boldsymbol{\lambda} + e_i\boldsymbol{\lambda}^2 + \boldsymbol{\epsilon}_i, \;\; i = 1, 2, \ldots, I, \tag{2.24}$$

where $\mathbf{1}$ is a column vector of ones which accounts for the constant baseline offset. The elements of the vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^2$ follow linear and quadratic functions of the wavelength, respectively, and they attempt to explain slow curvatures in the spectral baseline. The second term on the right-hand side comprises the linear mixture model from Eq. (2.14) and a multiplicative coefficient which explains the variations in the spectral amplitude caused by varying optical path length. Due to the varying packing

density, the concept of concentration is ill-defined in the case of powders, and weight fractions are invariably used as explained variables in regression models. Hence the coefficient b_i also includes the multiplicative error b from Eq. (2.13), and the coefficients of the pure analyte spectra in Eq. (2.14) are assumed to obey the closure constraint. All unmodeled spectral variations are included in the vector $\boldsymbol{\epsilon}_i$, and I is the number of spectra in the calibration set.

If the modeled error coefficients a_i, b_i, d_i and e_i exhibit variation between measurements, the prediction ability of the constructed calibration model is deteriorated. The additive baseline terms can be interpreted as an apparent increase in the chemical rank of the system, and a PLSR model can be expected to compensate their effect with an increased number of LVs. However, the nonlinearity caused by the multiplicative error b_i cannot be handled with linear regression models, unless they are mild. If the error coefficients can be reliably estimated, and if $\boldsymbol{\epsilon}_i$ is assumed to be negligible, the linear mixture model containing only chemical information can be retrieved from

$$\mathbf{x}_{i,\,\mathrm{chem}} = \left(\mathbf{x}_i - a_i\mathbf{1} - d_i\boldsymbol{\lambda} - e_i\boldsymbol{\lambda}^2\right)/b_i. \tag{2.25}$$

If the pure analyte spectra $\mathbf{s}_j$, $j = 1, \ldots, J$ are inserted in Eq. (2.24), the model is supplied with the products $b_i\, c_{ij}$, $j = 1, \ldots, J$, and the estimation of b_i is impossible unless prior information on the mass fractions of the measured samples is available.

To mitigate this problem, the linear mixture model (Eq. (2.14)) is rewritten in EMSC as variations around a known reference spectrum $\mathbf{m}$ which has equal weight in all samples as [5, 6]

$$\mathbf{x}_{i,\,\mathrm{chem}} = \mathbf{m} + \sum_{j=1}^{G} h_{ij}\mathbf{g}_j \;\; i = 1, 2, \ldots, I. \tag{2.26}$$

The signal spectra $\mathbf{g}_j$, $j = 1, \ldots, G$ and the reference spectrum should be chosen so that together they span the same space as the pure analyte spectra $\mathbf{s}_j$, $j = 1, \ldots, J$ or at least the space where the spectral changes relevant to the prediction of the reference occur. To facilitate the matrix inversion below in Eq. (2.28), the reference spectrum $\mathbf{m}$ should not be a linear combination of the signal spectra. An original approach for the selection of these spectra is presented and discussed in Sect. 2.4.3.

The soft spectral model can be optionally augmented with chemical interferent spectra $\mathbf{f}_l$, $l = 1, \ldots, F$ to obtain

$$\mathbf{x}_i = a_i\mathbf{1} + b_i\mathbf{m} + \sum_{j=1}^{G} h_{ij}\mathbf{g}_j + \sum_{l=1}^{F} p_{il}\mathbf{f}_l + d_i\boldsymbol{\lambda} + e_i\boldsymbol{\lambda}^2 + \boldsymbol{\epsilon}_i, \;\; i = 1, 2, \ldots, I, \tag{2.27}$$

where the product $b_i\, h_{ij}$ has been renamed as h_{ij}. All vectors on the right-hand side are assumed to be known and linearly independent. The matrix

$$\mathbf{M} = \left[\mathbf{1}\,\mathbf{m}\,\mathbf{g}_1 \cdots \mathbf{g}_G\,\mathbf{f}_1 \cdots \mathbf{f}_F\,\boldsymbol{\lambda}\,\boldsymbol{\lambda}^2\right]$$

has thus full column rank and the model coefficients can be estimated in LS sense as

$$\mathbf{P} = \mathbf{X}\,\mathbf{M}\,\left(\mathbf{M}^{\mathrm{T}}\,\mathbf{M}\right)^{-1}, \tag{2.28}$$

where the ith row of $\mathbf{P}$ is $[a_i\, b_i\, h_{i1} \cdots h_{iG}\, p_{i1} \cdots p_{iF}\, d_i\, e_i]$, and that of $\mathbf{X}$ is $\mathbf{x}_i^{\mathrm{T}}$. The preprocessed spectra

$$\mathbf{z}_{\mathrm{EMSC},\,i} = \frac{1}{b_i}\left(\mathbf{x}_i - a_i\mathbf{1} - \sum_{l=1}^{F} p_{il}\mathbf{f}_l - d_i\boldsymbol{\lambda} - e_i\boldsymbol{\lambda}^2\right),\ \ i = 1, 2, \ldots, I \tag{2.29}$$

are now standardized with respect to the multiplicative effect and they contain only contributions from the chemically relevant vectors $\mathbf{m}$ and $\mathbf{g}_j$, $j = 1, \ldots, G$, as the physical baseline effects and the chemical interferent contributions have been subtracted from them. The preprocessed spectra nevertheless contain the scaled unmodeled residuals $\boldsymbol{\epsilon}_i/b_i$, whose presence can be avoided by replacing the right-hand side of Eq. (2.29) with the summed contributions of $\mathbf{m}$ and $\mathbf{g}_j$ corrected for the multiplicative effect b_i as in

$$\mathbf{z}_{\mathrm{EMSC},\,i} = \mathbf{m} + \sum_{j=1}^{G} \frac{h_{ij}}{b_i}\mathbf{g}_j,\ \ i = 1, 2, \ldots, I. \tag{2.30}$$

In the early version of the algorithm, multiplicative signal correction (MSC) [28], the signal and interferent spectra were not included in the model in Eq. 2.27. The chemical variations around the reference spectrum were then included in the unmodeled residual term. In order to avoid the exclusion of important chemical information, the subtraction similar to Eq. (2.29) was mandatory in MSC. Probably to keep the notation coherent between the algorithms, Eq. (2.29) is used also with EMSC. To the knowledge of the Author, the effect of the remaining unmodeled residuals in EMSC has not been addressed in the literature.

A calibration model for the mass fraction of the target analyte should now be constructed using the preprocessed spectra $\mathbf{z}_{\mathrm{EMSC},\,i}$, $i = 1, 2, \ldots, I$. If the unmodeled errors $\boldsymbol{\epsilon}_i$ are assumed to be negligible, the chemical rank of the preprocessed spectra is $G + 1$. Since the contribution of the reference spectrum is equal in all spectra, it can be subtracted in principle from them which lowers the sufficient number of LVs to G in the subsequent PLSR model. Prior to prediction, new measurements should be preprocessed with the Eq. (2.28) and (2.29) using the same matrix $\mathbf{M}$. EMSC preserves the information on the model parameters $\mathbf{P}$ which can be used in their own right in analyzing the physical properties of the measured samples.

2.4.2 Optical path length estimation and correction

OPLEC [7] is another spectral preprocessing method based on the model in Eq. (2.24). The additive baseline effects are now removed from the measured spectra by projecting them onto the orthogonal complement of the space spanned by the columns of the matrix

$$\mathbf{P} = \left[\mathbf{1}\, \boldsymbol{\lambda}\, \boldsymbol{\lambda}^2\right]. \tag{2.31}$$

The projected spectra are given by

$$\mathbf{z}_i = \left(\mathbf{I} - \mathbf{P}\,\left(\mathbf{P}^{\mathrm{T}}\mathbf{P}\right)^{-1}\mathbf{P}^{\mathrm{T}}\right)\mathbf{x}_i = b_i\sum_{j=1}^{J} c_{ij}\mathbf{k}_j + \boldsymbol{\epsilon}_i^*,\ \ i = 1, 2, \ldots, I, \tag{2.32}$$

where the linear mixture model in Eq. (2.14) has been substituted for $\mathbf{x}_{i,\,\mathrm{chem}}$. Here, $\mathbf{k}_j = (\mathbf{I} - \mathbf{P}\,(\mathbf{P}^{\mathrm{T}}\,\mathbf{P})^{-1}\mathbf{P}^{\mathrm{T}})\,\mathbf{s}_j$ and $\boldsymbol{\epsilon}_i^* = (\mathbf{I} - \mathbf{P}\,(\mathbf{P}^{\mathrm{T}}\,\mathbf{P})^{-1}\mathbf{P}^{\mathrm{T}})\,\boldsymbol{\epsilon}_i$. The projected spectra are forced to be zero-mean and free from any slow baseline curvatures explained by the vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^2$.

Contrary to EMSC, OPLEC assumes that the mass fractions of the target analyte c_{i1}, $i = 1, \ldots, I$ are known for the samples of calibration set already in preprocessing phase. With the knowledge of the mass fractions, and assuming the closure constraint $\sum_{j=1}^{J} c_{ij} = 1, i = 1, \ldots, I$, and that the error terms $\boldsymbol{\epsilon}_i^*$ be negligible in Eq. (2.32), the relative values of the optical path length coefficients $b_i, i = 1, \ldots, I$ are next estimated in LS sense with nonnegativity constraint as detailed in [7, 27]. In OPLEC, the information on the baseline effects is lost, but the estimated multiplicative coefficients b_i can be used in analyzing the physical properties of the samples.

Next, two linear multivariate calibration models are constructed using the projected spectra $\mathbf{z}_i$ as explanatory and both the products $c_{i1}\,b_i$ and the coefficients b_i as explained variables, as in

$$c_{i1}\,b_i = f_1\,(\mathbf{z}_i)\,,\;\; b_i = f_2\,(\mathbf{z}_i)\,,\; i = 1, \ldots, I. \tag{2.33}$$

A new measurement, $\mathbf{x}_{\mathrm{new}}$, should be projected as in Eq. (2.32) to obtain $\mathbf{z}_{\mathrm{new}}$. The mass fraction of the target analyte in the new sample can then be predicted as

$$\hat{c}_{\mathrm{new},\,1} = \frac{f_1\,(\mathbf{z}_{\mathrm{new}})}{f_2\,(\mathbf{z}_{\mathrm{new}})}. \tag{2.34}$$

2.4.3 Optimization of the chemically relevant spectra in EMSC

Proper selection of the reference spectrum $\mathbf{m}$, the signal spectra $\mathbf{g}_j$, $j = 1, \ldots, G$ and the chemical interferent spectra $\mathbf{f}_l$, $l = 1, \ldots, F$ for the soft spectral model in Eq. (2.27) is crucial for successful preprocessing. Due to the heuristic nature of the model, best selections are expected to be application dependent, and a globally applicable procedure which provides optimal selections cannot be expected to exist. However, the reference spectrum should have an equal contribution in all spectra when they are scaled to be linearly correlated to the mass fractions (cf. Eq. (2.30)). The reference spectrum should also be linearly independent of the signal and interferent spectra, whose role is to explain the chemical variations around $\mathbf{m}$. In [5], the similarity between this configuration and the PCA decomposition was noted. It was proposed that $\mathbf{m}$ be the mean spectrum of the given data set and the $J - 1$ first orthonormal PC loading vectors be used as the signal spectra, i.e.,

$$\mathbf{m} = \overline{\mathbf{x}}, \tag{2.35a}$$

$$\left[\mathbf{X} - \mathbf{1}\,\mathbf{m}^{\mathrm{T}}\right] = \mathbf{U}\,\boldsymbol{\Sigma}\,\mathbf{V}^{\mathrm{T}}, \tag{2.35b}$$

$$[\mathbf{g}_1 \cdots \mathbf{g}_{J-1}] = \mathbf{V}_{(:,1:J-1)}. \tag{2.35c}$$

In [5, 6, 29], another scenario, in which the pure analyte spectra are assumed to be available, was presented. If the reference spectrum is now calculated as their linear

combination or if it is assumed to be such, collinearity can be avoided by choosing the $J-1$ difference spectra to be the signal spectra, i.e.,

$$\mathbf{m} = \sum_{j=1}^{J} \alpha_j \, \mathbf{s}_j, \tag{2.36a}$$

$$\mathbf{g}_j = \mathbf{s}_J - \mathbf{s}_j, \; j = 1, \ldots, J-1. \tag{2.36b}$$

The chemical rank of the preprocessed spectra can be reduced by utilizing the chemical interferent spectra $\mathbf{f}_l$ in the correction (Eq. (2.29)). This reduces the number of LVs needed in regression and it might ease the interpretability of the bilinear model [5]. The chemical interferent spectra should be chosen to be systematical variations present in all samples. In [30], the effect of temperature on the water absorbance peak locations was removed by using spectral interference subtraction in EMSC. Here, the SVD loading vectors of the differences between water spectra measured at different temperatures were used as the interference spectra $\mathbf{f}_l$.

If $\mathbf{m}$ or $\mathbf{g}_j$ are calculated as linear combinations of spectra measured from scattering samples, the spectra preprocessed with Eq. (2.28) and (2.29) are not free from light scattering effects. This leads to instability as demonstrated in [7], where, using the data from [6], the mean spectrum was the mean of the pure analyte spectra and the signal vectors were calculated according to Eq. (2.36a). Different selections for the pure analyte spectra among replicates with different physical effects resulted in significantly degraded prediction ability for the subsequent calibration model. In order to use pure analyte spectra successfully in EMSC, they should be measured with identical measurement geometries so that their baseline offsets and amplitudes reflect comparable physical effects. In [31], the reference spectrum was chosen as an arbitrary but representative sample among the calibration set. The physical effects were subsequently removed from it by estimating the coefficients of the additive baseline effects in LS sense and subtracting their contribution from the spectrum. In the present work, it was noticed that good results may be obtained by forcing the reference and signal spectra to be orthogonal to some or all of the additive baseline effects in Eq. (2.27). The spectra in $\mathbf{X}$ may be, e.g., zero-meaned or projected onto the null-space of the matrix $\mathbf{P}$ (cf. Eq. (2.31)) prior to the selections in Eq. (2.35a).

If the mass fractions of the target analyte are known in the samples comprising a representative calibration set, it is possible to utilize the prior knowledge in the selection of the vectors $\mathbf{m}$, $\mathbf{g}_j$ and $\mathbf{f}_l$. With the reference values for the explained values available, the effect of the selections on the prediction ability of the constructed calibration model can be tested with cross validation already in the preprocessing phase. The drawback with this approach is, however, the risk of overlearning which is a problem with erroneous reference values and non-representative calibration sets. In [32], it was proposed that the reference, signal and interference vectors could be optimized as linear combinations of the SVD or PC loading vectors of the data set so that the RMSECV is minimized. To the best of the Author's knowledge, no scientific investigation on the subject is, however, available in the literature.

In the present work, a flexible algorithm for the optimization of the EMSC model

(Eq. (2.27)) is proposed. The objective is to perform the minimization procedure

$$\min_{\mathbf{A}} \text{RMSECV}(\mathbf{A}, \mathbf{B}, \mathbf{X}, \mathbf{y}, n, A). \tag{2.37}$$

The cost function is minimized with respect to the argument $\mathbf{A}$, a $B \times (1 + G + F)$ matrix, which is used to represent the reference, G signal and F interference vectors as linear combinations of the base vectors, the columns of the $L \times B$ matrix $\mathbf{B}$. The evaluation of the cost function $\text{RMSECV}(\mathbf{A}, \mathbf{B}, \mathbf{X}, \mathbf{y}, n, A)$ is presented in Algorithm 1.

Algorithm 1 Calculate $\text{RMSECV}(\mathbf{A}, \mathbf{B}, \mathbf{X}, \mathbf{y}, n, A)$

Input: $\mathbf{A}, \mathbf{B}, \mathbf{X}, \mathbf{y}, n, A$

Calculate the reference, G signal and F interferent spectra as linear combinations of the base vectors as $[\mathbf{m}\ \mathbf{g}_1 \cdots \mathbf{g}_G\, \mathbf{f}_1 \cdots \mathbf{f}_F] = \mathbf{B}\,\mathbf{A}$.

Preprocess $\mathbf{X}$ according to Eq. (2.28) and (2.29) using the reference, signal and interferent spectra above. Obtain $\mathbf{Z}$.

Divide $\mathbf{y}$ and the corresponding rows of $\mathbf{Z}$ into n sequential blocks of roughly the same size.

Calculate the LBO-RMSECV (Eq. (2.23)) with respect to $\mathbf{y}$ using PLSR models with A LVs.

Output: LBO-RMSECV

The matrix $\mathbf{B}$ containing the base vectors has to be determined prior to minimization. The base vectors should effectively span the vector space where the chemical variations are expected to occur. They can be chosen to be, e.g., the first B SVD loading vectors of $\mathbf{X}$, the pure analyte spectra of the chemical species present in the mixtures or the signals estimated by a BSS algorithm (cf. Sect. 2.5). The base vectors should be linearly independent and their number should equal or exceed the chemical rank of the system. To avoid collinearity in EMSC-preprocessing, B should be larger than or equal to $1 + G + F$. It might be beneficial to ensure that the base vectors do not contain light scattering induced additive baseline variations by zero-meaning them or projecting them as in Eq. (2.32).

The closed form solution for the LBO-RMSECV in Algorithm 1 is very complicated. Moreover, if an iterative version of PLSR is utilized, the function might exhibit ill-behaving features such as nonsmoothness, and the closed form solution would not even exist. Robust numerical optimization procedures should hence be considered for the minimization in Eq. (2.37). In the present work, simulated annealing (SA) [33, 34] was used in the optimization. SA is a stochastic optimization procedure which simulates the physical process of freezing in solids, in which randomly moving crystalline structure is fixed to the position of minimum energy during the descent of temperature. SA is capable of avoiding local minima and finding good sub-optimal solutions for even noisy and discontinuous functions since it permits the increase of the cost function with a finite probability during optimization. The argument of optimization, the matrix $\mathbf{A}$ in this case, is subjected to small random permutations generated with a predefined method and the new value of $\mathbf{A}$ is accepted if it leads to

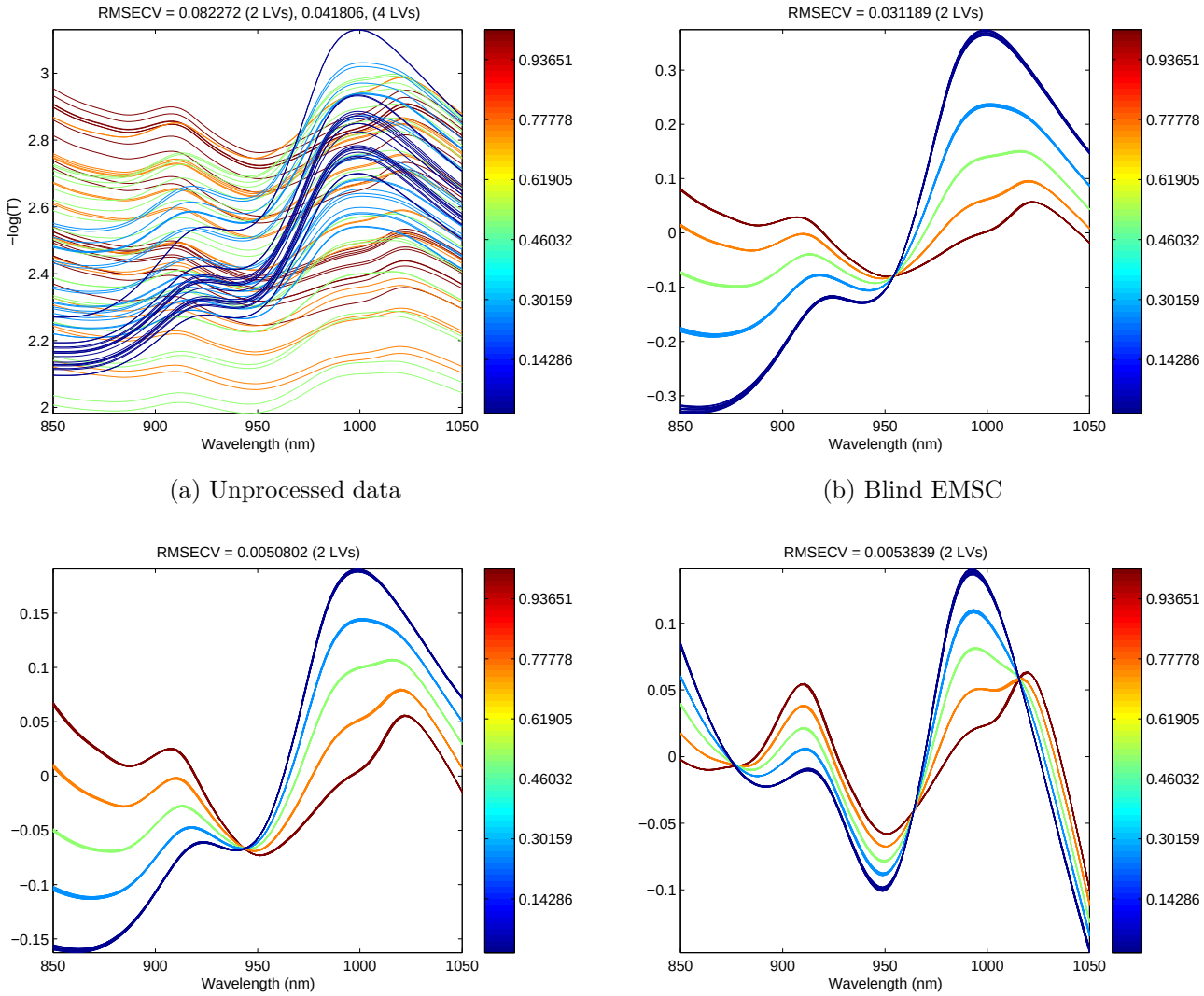
a decrease in the cost function. If the cost function increases, the new $\mathbf{A}$ is accepted with the probability

$$p = e^{-\Delta\text{RMSECV}/T}, \tag{2.38}$$

where e is the base of natural logarithm, ΔRMSECV is the change in the cost function and T is a positive temperature parameter which is often decreased in the course of the search so that the probability of accepting increased cost function values is decreased. SA is known to be rather inefficient optimization method [34], and the computational burden of the minimization procedure was reduced by using LBO-RMSECV instead of LOO-RMSECV in Algorithm 1.

The performance of the preprocessing methods is illustrated with a public data set [32, 6] in Fig. 2.4. One hundred raw $\log(1/T)$ spectra measured from binary mixtures of gluten and starch powders are presented in Fig. 2.4a. The data set comprises twenty replicates measured from each of the five unique mixtures in which the mass fraction of gluten is $\mathbf{y} = [0\ 0.25\ 0.5\ 0.75\ 1]^{\mathrm{T}}$. Each replicate has slightly different physical properties in the form of baseline offset and optical path length, as two different cuvettes were used and varying packing densities were achieved by compressing the powder [6]. The figure includes LOO-RMSECV values calculated with PLSR models with two and four LVs. The prediction ability is observed to be deteriorated by the physical effects, and the increase of the number of LVs improves the RMSECV only nominally. In Fig. 2.4b, the spectra were preprocessed with EMSC using the procedure of Eq. (2.35a) with a zero-meaned matrix $\mathbf{X}$ to select the reference and one signal spectrum. The spectra are seen to form five clusters each corresponding to a unique mixture, but the LOO-RMSECV is only slightly improved. Thus, blind selection of the reference and signal spectra did not linearize the spectra with respect to $\mathbf{y}$, the mass fraction of gluten, as is evident from the unequal spacings between the adjacent clusters.

Optimized EMSC was used in Fig. 2.4c so that the base vectors $\mathbf{B}$ were chosen to be the two first SVD loadings of the zero-meaned matrix $\mathbf{X}$, and the parameters $G = 1$, $F = 0$ and $n = 5$ were used in Algorithm 1. The minimized LBO-RMSECV is now significantly smaller than the LOO-RMSECV in the previous figure. Fig. 2.4d presents the spectra after OPLEC preprocessing, where the spectra are constrained to be orthogonal to the columns of $\mathbf{P}$ in Eq. (2.31). The spectra were standardized with respect to varying optical path length by dividing the projected spectra (Eq. (2.32)) by the estimated multiplicative coefficients b_i. The LOO-RMSECV is of the same order as with the optimized EMSC. Visually indistinguishable results with Fig. 2.4d were obtained with the optimized EMSC when the raw spectra in Fig. 2.4a were subjected to the projection in Eq. (2.32) prior to calculating the base vectors as their two first SVD loading vectors (results are not shown). The use of prior information on $\mathbf{y}$ in the optimized EMSC and OPLEC is seen to provide better preprocessing results than the blind selection of the reference and signal spectra in Fig. 2.4b. The preprocessed spectra are now linearized with respect to $\mathbf{y}$, since the adjacent clusters of spectra are equidistant from each other at each wavelength channel, corresponding to the equal distances between the mass fractions of gluten in the adjacent mixtures. The number of LVs was chosen to be two in all PLSR models in EMSC and OPLEC.



(a) Unprocessed data

(b) Blind EMSC

(c) Optimized EMSC

(d) OPLEC

Figure 2.4: NIR spectra of binary mixtures of gluten and starch powders. (a) Unprocessed $\log(1/T)$ spectra; (b) EMSC–preprocessed spectra, reference and signal vectors chosen according to Eq. (2.35a) with zero-meaned data matrix $\mathbf{X}$; (c) EMSC–preprocessed spectra, optimized reference and signal vectors constrained to be zero-mean; (d) OPLEC–preprocessed spectra. The spectra are colored according to the mass fraction of gluten.

2.5 Blind source separation

Blind source separation (BSS) constitutes a family of algorithms which attempt to resolve pure signals from their mixtures in the absence of any *a priori* information. Most BSS algorithms assume the linear mixture model (Eq. (2.15)), and they attempt to estimate the mixing matrix $\mathbf{C}$ and the pure signals $\mathbf{S}$ given only the mixtures $\mathbf{X}$. In optical spectroscopy, BSS algorithms are thus useful for analyzing so called black

systems for which no pure analyte spectra or concentration information are available. One approach to BSS, ICA, is overviewed next. The unique characteristics of BSS applied in optical spectroscopy are described in Sect. 2.5.2. Finally, the challenges of using BSS in DR measurements are addressed in Sect. 2.5.3.

2.5.1 Independent component analysis

ICA algorithms [9] comprise a myriad of different variations which all have the same basic principle; they attempt to factorize the given mixture matrix $\mathbf{X}$ into the source signals $\mathbf{S}$, i.e., the independent component (IC) loadings, which are as statistically independent as possible. The mixing process of the sources is described by the mixing matrix $\mathbf{C}$, the IC scores. When compared to factorization with PCA, where the source signals are constrained to be orthogonal, ICA has the flexibility to find a non-orthogonal basis of factors, which often accurately resemble the true underlying source signals. Often ICA algorithms involve the estimation of the unmixing matrix $\mathbf{W}$ which is used to estimate the source signals $\hat{\mathbf{S}}$ and the mixing matrix $\hat{\mathbf{C}}$ as

$$\hat{\mathbf{S}}^{\mathrm{T}} = \mathbf{W}\,\mathbf{X} \text{ and } \hat{\mathbf{C}} = \mathbf{W}^{+} \tag{2.39}$$

where the superscript + denotes the Moore-Penrose pseudoinverse of the corresponding matrix. The underlying source signals, the columns of $\mathbf{S}$, are assumed to be realizations of separate random processes drawn from unknown probability distributions. To ease the quantification of statistical independence, the source signals are assumed to be non-Gaussian. According to the central limit theorem, the probability distribution of the sum of two non-Gaussian random processes is closer to the Gaussian distribution than either one of the two original distributions. Hence, as the non-Gaussianities of the estimated source signals are maximized, the underlying source signals are expected to be estimated accurately.

Measures of non-Gaussianity include, e.g., kurtosis and negentropy [9], which are calculated for the estimated source signals using their sample statistics. Kurtosis of the random process s is defined as

$$\mathrm{kurt}(s) = E\left[s^4\right] - 3\left(E\left[s^2\right]\right)^2, \tag{2.40}$$

where $E[\cdot]$ denotes the expectation value. For random processes drawn from the Gaussian distribution, $\mathrm{kurt}(s) = 0$, whereas $\mathrm{kurt}(s) > 0$ for spiky (super-Gaussian or leptokurtic) and $\mathrm{kurt}(s) < 0$ for flat (sub-Gaussian or platykurtic) probability distributions. The negentropy is defined for the process s as the difference between the differential entropies of a Gaussian variable ν and s, viz.,

$$J(s) = H(\nu) - H(s), \text{ where } H(s) = -\int p(s)\log\left(p(s)\right)\mathrm{d}s, \tag{2.41}$$

where $p(s)$ is the probability density function (PDF) of s. The Gaussian variable ν is assumed to be of zero mean and unit variance, and the variable s is pretreated to have the same attributes. Since the differential entropy, i.e., the randomness, of

the Gaussian distribution is larger than that of any other distribution, negentropy is always nonnegative and zero only when the variable s is Gaussian.

Another approach to ICA is to minimize the mutual information

$$I\left(s_1,\, s_2, \ldots, s_J\right) = \sum_{i=1}^{J} H(s_i) - H\left(s_1,\, s_2, \ldots, s_J\right), \tag{2.42}$$

of the random processes $s_1,\, s_2, \ldots,\, s_J$. The mutual information is always nonnegative and zero only when the random processes are statistically independent by definition, i.e., their joint PDF factorizes as $p(s_1,\, s_2, \ldots, s_J) = p(s_1)\, p(s_2) \cdots p(s_J)$. Approximations of the mutual information for discrete signals are presented in [9, 35]. The entropy of a discrete random process s can be approximated with, e.g., the Shannon entropy

$$H\left(s\right) = -\sum_{l=1}^{L} p(s_l)\, \log(p(s_l)), \quad p(s_l) = \frac{|s_l|}{\sum_{l=1}^{L} |s_l|}. \tag{2.43}$$

FastICA [36] is an implementation of ICA which utilizes a fast iterative fixed-point algorithm in maximizing the negentropies of the estimated source signals (Eq. (2.39)). The negentropy is estimated as a squared difference between the expectation values of some nonquadratic function $G(\cdot)$ given the estimated signal s and the Gaussian signal ν as input, viz.,

$$J(s) \propto \left(E\left[G(s)\right] - E\left[G(\nu)\right]\right)^2. \tag{2.44}$$

The FastICA toolbox for MATLAB [37] includes four possibilities for the function $G(\cdot)$, a kurtosis-based, a Gaussian, a $\log\cosh(\cdot)$-based and one based on the measure of skewness. To ease the estimation, the input matrix is often mean-centered and whitened with a linear transformation prior to the fixed-point algorithm so that its correlation matrix becomes unity, i.e., $\mathbf{X}\,\mathbf{X}^{\mathrm{T}} = \mathbf{I}$. The whitening of the data can be done with PC decomposition during which the rank of the data can be reduced by choosing the number of retained PCs to be lower than the dimension of the matrix $\mathbf{X}$.

Denoising source separation (DSS) [38] is another variation of ICA. It involves whitening and subsequent rotation of the mixture matrix $\mathbf{X}$. It also permits the inclusion of prior knowledge on the spectral characteristics of the source signals, but it can be used for purely blind operation as well. Both FastICA and DSS assume that the number of the underlying signals, J, be known *a priori*. If the user has no prior information on it, the algorithm can be driven with several values of J, and the correct value may be determined interactively by analyzing the reconstructed signals. The determination of J can also be automated, e.g., by analyzing the magnitudes of variances explained by the few first SVD loading vectors of the original $\mathbf{X}$. The number of underlying signals can be assumed to equal the number of the first few loadings which explain most of the variance of the original data matrix. Alternatively, the number of source signals may be determined so that the error of reconstruction $||\mathbf{X} - \hat{\mathbf{C}}\,\hat{\mathbf{S}}^{\mathrm{T}}||$ is minimized [39], where $||\cdot||$ denotes the Frobenius norm of the corresponding matrix.

2.5.2 Blind source separation in optical spectroscopy

In optical spectroscopy, BSS is synonymous with multivariate curve resolution (MCR), in which the goal is to reconstruct the pure analyte spectra and their concentration profiles using only the mixture spectra and little or no additional prior information. The benefits of BSS include the identification of pure analytes present in a data set by comparing the resolved source signals to known pure analyte spectra in a data library. The phases of an industrial process or a chemical reaction can also be studied by inspecting the estimated concentration profiles [8]. BSS is thus an exploratory analysis technique which tries to decompose the matrix of mixture spectra into chemically interpretable scores and loadings. It must be noted that, without calibration, only relative concentration profiles can be estimated. The true amplitudes of the estimated pure analyte spectra cannot be determined since the corresponding concentration profile can always be divided by the inverse of the amplitude coefficient without affecting the factorization. However, in perfect factorization, where the pure signals are perfectly estimated up to a constant, the estimated IC score profiles and corresponding true underlying concentration profiles have a correlation coefficient of unity. The amplitude ambiguity in ICA also involves sign ambiguity, viz., the amplitude coefficient above can also be negative. The reversed sign of an absorption spectrum can be easily detected interactively with a human user, but the procedure can also be automated by imposing a physically meaningful nonnegativity constraint on the concentration profiles, or by ensuring that the directions of the estimated spectral vectors are similar to the directions of the rows in $\mathbf{X}$ [40].

Differentiation of the mixture spectra with respect to wavelength has been observed to enhance the separation capability of ICA algorithms [41, 42, 43]. Being a linear operation, the mixing and unmixing matrices are preserved as the mixture spectra are subjected to differentiation [43], i.e.,

$$\mathbf{X}^{(n)} = \mathbf{C}\,\left(\mathbf{S}^{\mathrm{T}}\right)^{(n)} \quad \text{and} \quad \left(\hat{\mathbf{S}}^{\mathrm{T}}\right)^{(n)} = \mathbf{W}\,\mathbf{X}^{(n)}, \tag{2.45}$$

where the superscript (n) denotes the nth order difference with respect to the discrete wavelength channels. Thus, any linear operation can be performed on the original mixture matrix $\mathbf{X}$ prior to feeding it to an ICA algorithm. The obtained unmixing matrix $\mathbf{W}$ can subsequently be used with the original $\mathbf{X}$ in Eq. (2.39) to estimate the source signals.

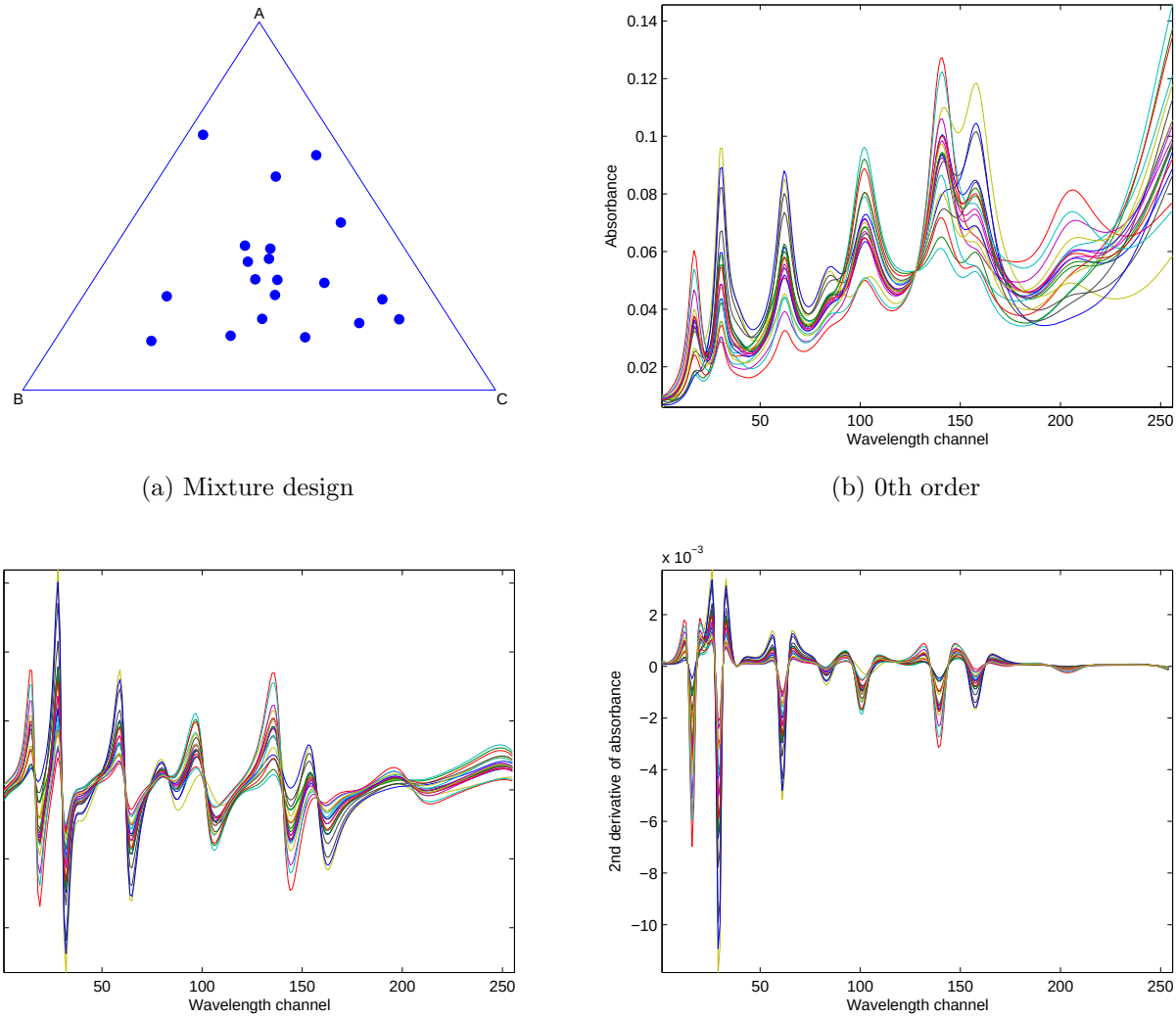
The effect of differentiation is demonstrated with synthesized data in Fig. 2.5 and 2.6. Pure analyte spectra shown as solid lines in Fig. 2.6a–2.6d were created as sums of modified Lorentz-distributions as [35]

$$\mathbf{s}_j(\lambda) = \sum_{k=1}^{K_j} \frac{a_{kj}}{(1/\lambda - 1/\lambda^{\mathrm{o}}_{kj})^2 + \gamma^2_{kj}},\ j = 1,\ 2,\ 3, \tag{2.46}$$

where the parameters K_j, a_{kj}, $\lambda^{\mathrm{o}}_{kj}$ and γ_{kj} were randomly chosen. The spectra were designed to emulate the smooth and highly autocorrelated characteristics of NIR spectra. The randomly generated mixture design of twenty mixtures shown in the ternary

diagram in Fig. 2.5a was used to produce the linearly mixed spectra in Fig. 2.5b. The mixture spectra are shown in Fig. 2.5c and 2.5d after first and second order differentiation with respect to the wavelength channel. FastICA was subsequently used to estimate the unmixing matrix $\mathbf{W}$ using the mixture spectra after 0th, 1st, 2nd and 3rd order differentiation. The function $G(\cdot)$ used in the calculation of negentropy was chosen to be $\log\cosh(\cdot)$-based, and the whitening in FastICA was performed with PCA so that the the constributions of three first PCs were retained.

The estimated pure analyte spectra reconstructed with Eq. (2.39) are drawn in dotted lines in Fig. 2.6a–2.6d. It is seen that the separation capability of FastICA is systematically increased with the order of differentiation. This can be explained with the



(a) Mixture design

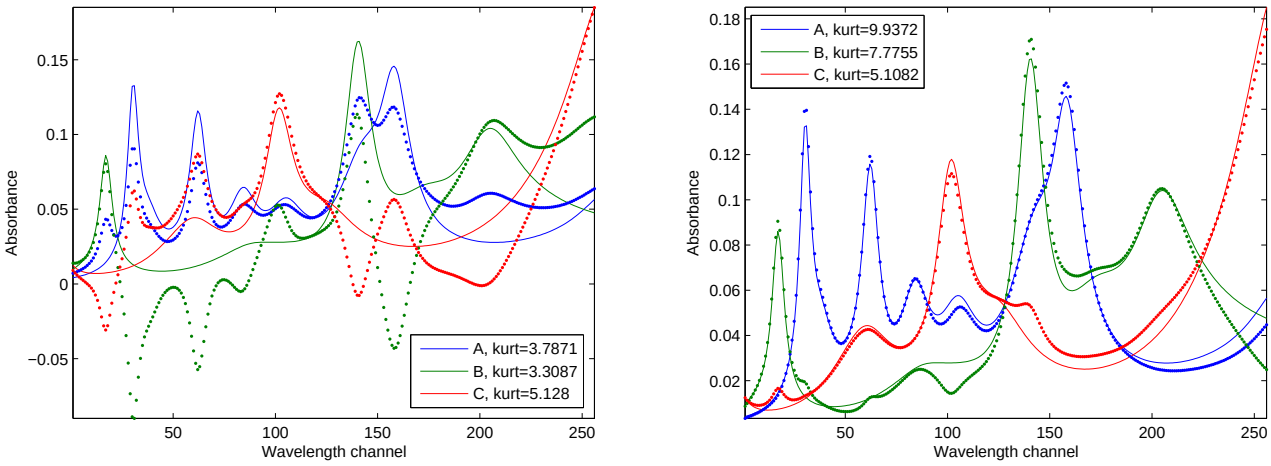
(b) 0th order

(c) 1st order

(d) 2nd order

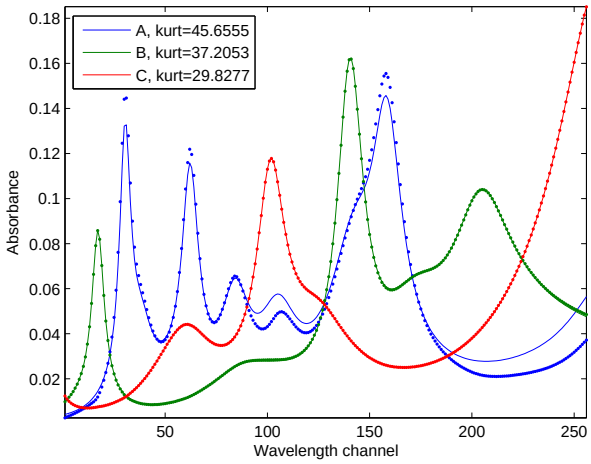
Figure 2.5: (a) The mass fraction design of the synthesized ternary mixtures; (b) The resulting mixture spectra; The mixture spectra after (c) first and (d) second order derivative taken with respect to wavelength.

increased non-Gaussianity of the differentiated pure analyte spectra, which is quantified by sample kurtosis calculated with the function `kurtosis.m` from the Statistics Toolbox in MATLAB. As shown in the figures, the kurtosis values are generally increased with the order of differentiation, i.e., the PDFs of the differentiated spectra get more spiky, or super-Gaussian, shapes. This is expected, since majority of the spectral elements get values near zero when smooth spectral signals are differentiated.

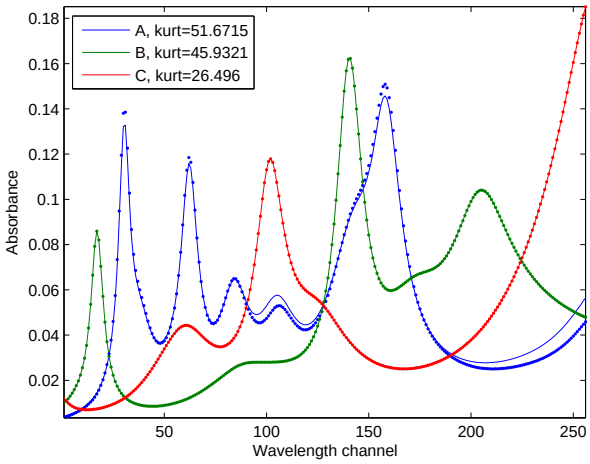


(a) 0th order

(b) 1st order



(c) 2nd order



(d) 3rd order

Figure 2.6: The pure analyte spectra A, B and C (solid lines) and the resolved IC loadings (dotted lines) using (a) 0th, (b) 1st, (c) 2nd and (d) 3rd order derivative preprocessing prior to FastICA. Sample kurtosis for each pure analyte spectrum after differentiation is given in each figure.

BSS algorithms can be made more robust to random noise by reducing the rank of the input matrix with SVD as in Eq. (2.21) and (2.22) prior to feeding the rank-

reduced loading matrix $\mathbf{P}^{\mathrm{T}}$ to the BSS algorithm [44]. To ascertain that the chosen SVD loading vectors span at least the same space as the underlying pure analyte spectra, their number should equal or slightly exceed the expected chemical rank of the system. In FastICA and DSS, rank reduction can be also done by retaining the contributions of only the first few PC loadings in the context of whitening. However, if differentiation is to be used as preprocessing prior to BSS, the denoising achieved through rank reduction should be performed already before taking the derivatives. Since differentiation is equivalent to high-pass filtering, the contribution of random high-frequency noise is enhanced in differentiated spectra. Thus, the contributions of the uninformative and noisy SVD loadings should be discarded prior to differentiation. Given the unmixing matrix $\mathbf{W}$ estimated with the SVD loading matrix $\mathbf{P}^{\mathrm{T}}$ or its differentiated version $(\mathbf{P}^{\mathrm{T}})^{(n)}$, the pure analyte spectra and the concentration profiles can now be estimated as

$$\hat{\mathbf{S}}^{\mathrm{T}} = \mathbf{W}\,\mathbf{P}^{\mathrm{T}} \text{ and } \hat{\mathbf{C}} = \mathbf{T}\,\mathbf{W}^{+} \tag{2.47}$$

using the notation in Eq. (2.22).

2.5.3 Blind source separation in diffuse reflectance

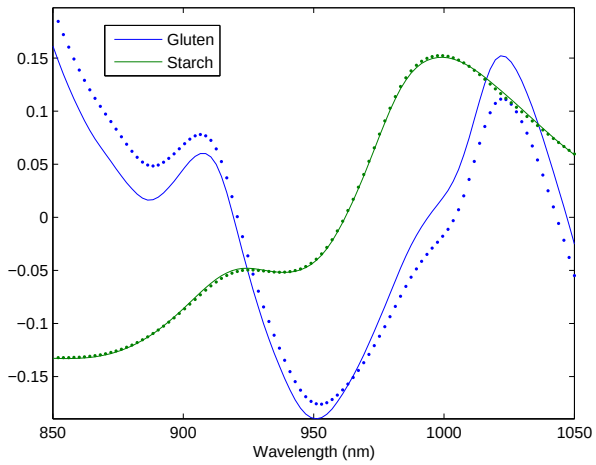
The use of BSS algorithms with $\log(1/T)$ or $\log(1/R)$ spectra of scattering samples has not been thoroughly addressed in the literature. The non-linearities introduced by light scattering can be again modeled with Eq. (2.24). To prevent the baseline offsets from entering the BSS factorization, the measured spectra should be zero-meaned or projected as in Eq. (2.32) prior to BSS. The multiplicative effect does not prevent the estimation of the pure analyte spectra but it deteriorates the linear correlation between the true underlying concentration profiles and their estimates.

EMSC can be attempted to remove the remaining multiplicative effect, but the optimal selection of the reference and signal vectors is difficult for black systems without any prior knowledge on the concentrations or mass fractions. Alternatively, normalization of the spectra may be used to standardize the apparent optical path length as

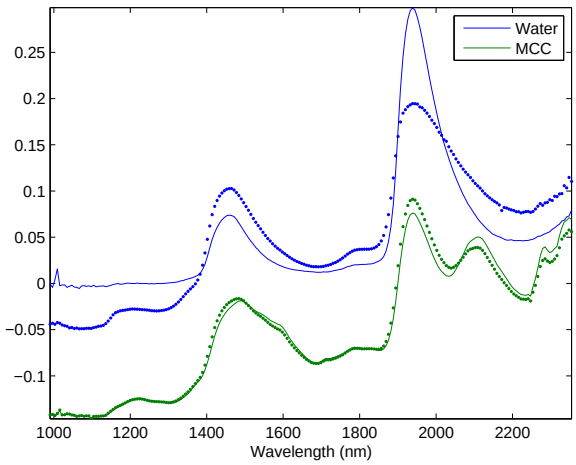
$$\mathbf{x}_i = b_i\,\mathbf{x}_{i,\,\mathrm{chem}},\;\; \mathbf{z}_i = \frac{\mathbf{x}_i}{||\mathbf{x}_i||} = \frac{\mathbf{x}_{i,\,\mathrm{chem}}}{||\mathbf{x}_{i,\,\mathrm{chem}}||},\; i = 1, 2, \ldots, I. \tag{2.48}$$

The multiplicative coefficient is thus merely replaced by the inverse of the spectral norm, and the normalized spectra cannot be expected to follow the linear mixture model with respect to the mass fractions. However, if the chemical changes are small within the data set, i.e., the norm $||\mathbf{x}_{i,\,\mathrm{chem}}||$ exhibits small variation, normalization can be expected to improve the linearity.

BSS is demonstrated with the $\log(1/T)$ spectra of binary mixtures of gluten and starch given in Fig. 2.4a. To make the setting more challenging, the 40 pure analyte spectra were excluded from the analysis. The remaining 60 mixture spectra were zero-meaned, their rank was reduced to two with SVD and they were subjected to third order differentiation before feeding them to FastICA. The estimated IC loading vectors are close to the true zero-meaned pure analyte spectra as is shown in Fig. 2.7a. The three



(a) Zero-meaning, SVD and 3rd derivative



(b) Zero-meaning, SVD and 5th derivative

Figure 2.7: BSS performed on the spectra of scattering samples. The pure analyte spectra (solid lines) and the estimated IC loadings (dotted lines) estimated from (a) the mixtures of gluten and starch Fig. 2.4a and (b) the mixtures of MCC and water in Fig. 2.3a. For visualization, the spectra are separated by constant offsets.

mixture spectra with different water fractions shown in Fig. 2.3a were preprocessed similarly prior to FastICA with the exception that 5th order derivative was used this time. As shown in Fig. 2.7b, the spectrum of MCC was correctly resolved, but the estimated water spectrum has broadened absorption peaks and tilted baseline when compared to the pure water spectrum. Since the spectra of water and MCC are highly cross-correlated, it is difficult to determine whether the discrepancies are caused by sample matrix interactions or by the insufficient separation of FastICA. The order of differentiation was determined interactively so that the spectral reconstruction was visually most accurate.

Chapter 3

Materials and methods

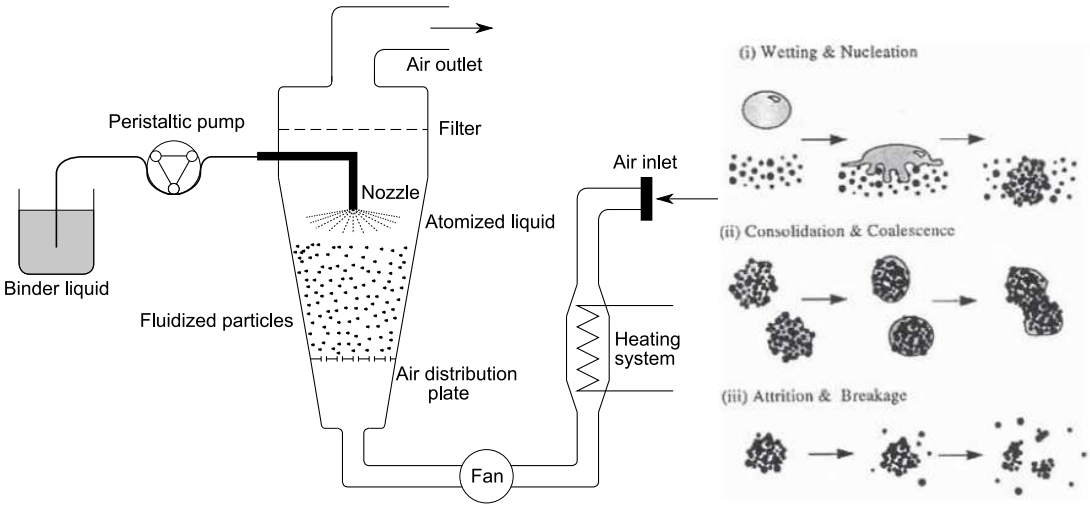
3.1 Fluid bed granulation

In pharmaceutical manufacturing of solid dosage forms, the processing of fine powder mixtures is often facilitated by agglomerating the material into granules larger than the original particle size. The purpose of this procedure is to [45]:

- improve the flow properties of the material by reducing static electricity
- densify the material
- produce uniform mixtures that do not segregate into small and large particles
- facilitate accurate dosing by ensuring homogenous mixtures
- improve the tablet compression characteristics of the material
- control the drug release rate – larger granules dissolve more slowly
- reduce the amount of dust
- improve the appearance of the product
- reduce variations between different batches of raw materials.

Along with wet massing in a high-shear mixer, FBG is an important wet granulation method frequently used in pharmaceutical industry. In both methods, the agglomeration of particles is evoked by mixing the powder with a binder liquid after which excess moisture is removed from the granules. In FBG, both the mixing of the powder mass with the binder liquid and the subsequent drying are induced by blowing air through the powder layer. As the name implies, the powder bed ideally acts as liquid or fluid while it is fluidized in the air flow which keeps the powder in constant motion by causing the formation of air bubbles. As is the case with real fluids, the surface level of the powder bed should ideally stay horizontal even if the chamber is tilted.

A diagram of a typical fluid bed processing system [46] is given in Fig. 3.1a. The air flow is usually generated by a turbine fan suction located upstream of the cone-shaped granulation chamber. The inlet air duct often includes an air conditioner, such as heating or a humidifying element, which permits the adjustment of the inlet air parameters in the case of, e.g., varying humidity of the ambient air. The gas distribution plate is designed to create an air flow pattern which optimally fluidizes



(a) Fluid bed granulator (b) Granule growth processes

Figure 3.1: (a) A diagram of a fluid bed granulator, adapted from [46]; (b) The granule growth process, edited from [46].

the product in the given granulation chamber. The binder liquid is sprayed usually on top of the powder bed through a nozzle. To prevent the powder from being blown out of the chamber, dense filters must be placed in the ceiling of the chamber.

An FBG batch process consists of three phases.

- In the mixing phase, the mixture of pharmaceutical powders, usually one API and a couple of excipients, is mixed into homogeneous mass by plain fluidization.
- In the wetting phase, the fluidized mass is sprayed with the binder liquid. This induces the formation of granules through the nucleation and coalescence processes as illustrated in Fig. 3.1b.
- In the drying phase, the granules are dried by continuing the fluidization after the ending of the liquid feed.

The granules are consolidated as the dissolved material between coalesced nuclei transform into solid bridges in the course of drying. In the case of poorly soluble powders, a binder agent, such as polyvinylpovidone (PVP), is dissolved in the aqueous binder liquid to promote the consolidation. The binder agent, which stays in the granulated product after drying, contains long polymers which prevent crumbling by adhering tightly onto the surfaces of the granules.

The behaviour of the fluidized powder is governed by a net effect of several particle-particle, particle-container and particle-gas interactions. The incorporation of the small-scale physico-chemical interactions, such as the electrostatic and van der Walls forces, into the modeling of the process is difficult if not impossible. Thus, in order to ensure the product quality through efficient control of a complex multi-factorial process, such as FBG, the concept of design space has been developed. It comprises the critical adjustable process parameters and the process variables, which have an effect on the quality of the end-product, and their mutual interactions [45]. In FBG, important process parameters are, e.g., the flow rate, humidity and temperature of the inlet

air and the feed rate and the droplet size of the sprayed binder liquid [45, 46]. The granule size distribution, moisture and temperature of the fluidized powder are important measurable process variables whose values correlate with the process parameters and the quality of the end-product. For example, excessive wetting and insufficient air flow might induce irrecoverable powder bed collapse due to excessive agglomeration during fluidization. End point determination of the drying phase is another critical task in FBG which requires the monitoring of the process variables in real time. To ensure the microbiological stability, accurate dosing and optimal compression properties of the product, excess moisture should removed from the granules. On the other hand, prolongated drying phase and excessive air flow induces the formation of fines through attrition.

The non-invasive nature of NIR spectroscopy and its sensitivity to both the particle size effects and moisture has made it an important inline monitoring method for the FBG process, as both parameters can be quantified from the same NIR measurement using different calibration models. In [47], a four-wavelength NIR system is demonstrated to be accurate in inline moisture measurement. In [48], the full spectra in the range 1350–1500 nm, measured inline during drying and preprocessed with first derivative with respect to wavelength, were used to build a multivariate calibration model with high predictive ability for the moisture of the lactose powder bed using Karl-Fischer titration as the reference method. In the same article, the NIR spectra between 1100–1900 nm were analyzed using PCA, and a PLSR model was built for the prediction of the mean granule size determined with sieve analysis. It was shown that, besides the mean granule size, the increase in the amorphous lactose content in the granules could be determined from the NIR spectra, as well.

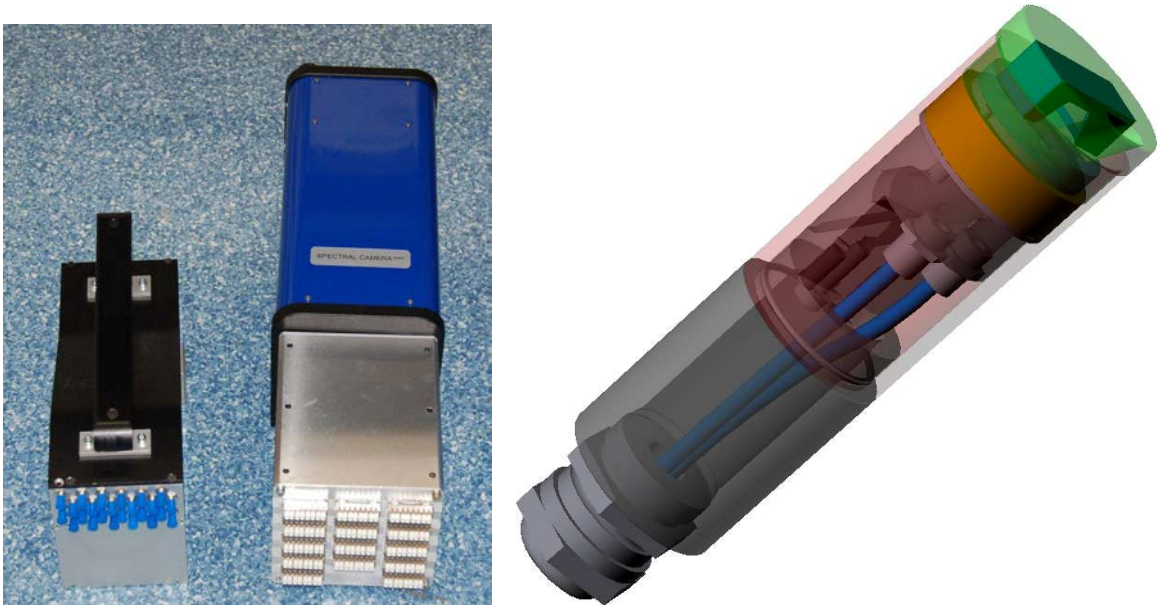
In the present work, an inline FBG process measurement was conducted using NIR spectroscopy as described in Sect. 3.3. The process data were analyzed qualitatively in Sect. 4.1, where information on the temporal moisture profile was attempted to be extracted.

3.2 The multipoint NIR instrument

All measurements in this work were conducted using the multipoint NIR instrument partly designed and implemented at VTT. It consists of a light source, a spectral camera and fiber optic probes. The system is designed to perform DR measurements simultaneously at multiple locations. The instrument is described in closer detail and its performance is evaluated in both offline and inline situations in [49].

The light source is illustrated in Fig. 3.2a. Using mirror optics, the light from a halogen lamp is projected into two optical fiber bundles, each comprising 12 fibers. The 24 separate optical fibers which can be accessed via subminiature A (SMA) connectors on the front panel of the case. The case contains also a chopper which can block the light from entering the fiber bundles so that either one of the bundles is illuminated at a time or both bundles are simultaneously blocked.

The fiber optics probes consist of 1.5-m long illumination and detection fibers of the diameters 400 μm and 600 μm, respectively, and a probe head which is shown in



(a) Illumination unit and spectral camera (b) Probe head

Figure 3.2: (a) The illumination unit and the SWIR spectral camera with fiber-optics input module; (b) Schematic of one probe head.

Fig. 3.2b. The probe head contains mirror optics which project the light beam from the illumination fiber onto the measured sample. To eliminate all specular reflections between a possible glass window between the probe and the sample and to ensure that only diffuse reflection is measured, the direction of the outgoing light beam is at an angle with respect to the normal of the probe plane. The detected light enters the probe in the reverse direction through the mirror optics before entering the detection fiber. The probe head has a revolving plate which permits internal reference measurement by blocking the light outlet with a reference.

Spectral camera, SWIR from Specim Ltd. (Oulu, Finland) [50] with fiber optics interface, is used as light detector. The fiber optics module, designed by VTT, permits simultaneous use of 106 aforementioned probes. Each detection channel, interfaced through an MU connector, guides the polychromatic detected light to a prism-grating-prism (PGP) spectrograph (ImSpector N25e, Specim Ltd. [51, 52]) which disperses the light onto a mercury-cadmium-telluride (MCT) matrix detector comprising 256×320 pixels. The NIR spectrum of the detected light from one channel can be observed in the wavelength range 1000–2500 nm on one 256-pixel column on the MCT detector. Light is dispersed almost linearly onto the detector so that each pixel detects light intensity from a narrow wavelength band of the width 6.3 nm at maximum. To prevent crosstalk between contiguous channels, every third of the 320 columns are utilized on the detector matrix.

The spectral camera digitizes the light intensity with the accuracy of 14 bits. The measurements can be controlled and data can be saved with a DataCube software from Specim Ltd. Important measurement parameters are the integration time, the sampling frequency and the number of averaged spectra. In this work, the measured

spectra were transformed into apparent absorbance units through the equation

$$A(\lambda) = -\log_{10}\left(\frac{I(\lambda) - D(\lambda)}{I_0(\lambda) - D(\lambda)}\right), \tag{3.1}$$

where $I(\lambda)$ denotes the analog to digital (AD) counts measured from the sample at the wavelength λ, $D(\lambda)$ is the dark current measured with blocked light source and $I_0(\lambda)$ is the reference measured either with a sheet of optical teflon (Gigahertz-Optik GmbH, Germany), as in the process measurements, or a Spectralon reflectance standard with 99 % reflectance value (Labsphere, USA), as in the laboratory measurements. The integration time was set so that the AD counts were slightly below the saturation level in the reference measurement for the channel with the lowest attenuation. Due to the strong attenuation of the optical fibers at the low-frequency end and the optical unidealities at both ends of the NIR bandwidth, the usable wavelength range was found to be 1100–2300 nm.

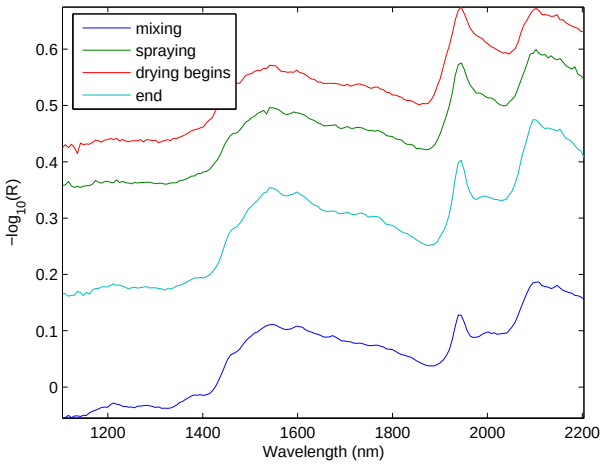
3.3 Fluid bed granulation measurements

An inline measurement of the FBG process was conducted using the multipoint NIR equipment. The granulation was performed at the Department of Pharmaceutical Technology in University of Kuopio with the Aeromatic STREA-1 fluidized bed granulator (Aeromatic-Fielder AG, Switzerland) equipped with top-spray unit and custom made granulation chamber of the height 485 mm (Fig. 3.3a).

The formulation used in the granulation comprised 80 % w/w of Lactose monohydrate (Pharmatose 200M, DMV) and 20 % w/w of Caffeine anhydrous (Scharlau, Spain) to yield the total dry mass of 200 g. The powder mass was manually mixed for 2 minutes prior to feeding it into the granulator. The binder liquid contained 16.7 % w/w of PVP (Kollidon K30, BASF, Germany) and 83.3 % w/w of purified water, and 55 grams of it was sprayed during the process. The durations of the mixing, wetting and drying phases were, 5, 12 and 16 min, respectively. The fluidization was halted briefly once during the mixing and twice during the wetting phase, and the granulator was hit with a rubber sledgehammer to dislodge stationary powder mass off the inner walls of the chamber.

Eight fiber optic probes were attached to the granulator chamber. The measurement was non-invasive as the probes looked through glass windows. NIR spectra were sampled at 3 Hz with the exposure time of 8 ms throughout the process. Three consecutive spectra were subsequently averaged to obtain one spectrum per second and a total of 2023 spectra from each of the eight channels. Four spectra measured with the lowest probe on the right in Fig. 3.3a during different process phases are presented in Fig. 3.3b. The presence of moisture is visually observed as spectral baseline offset and increase in the intensity of the water absorption band near 1936 nm.

(a) Granulator and the probes



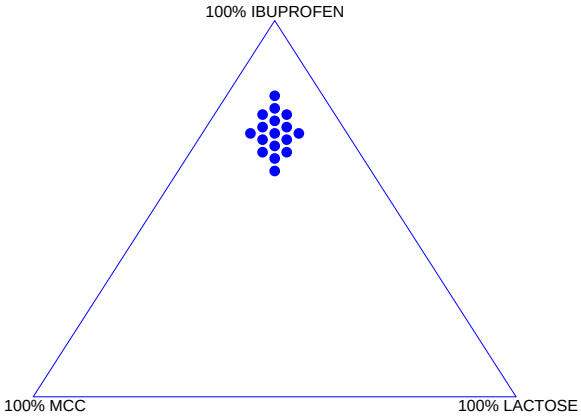
(b) Spectra measured during the process

Figure 3.3: (a) The measurement set up during granulation process. Four probes are attached to both sides of the granulation container; (b) Spectra measured with one probe at different instants during the granulation process: the mixing phase (blue), the wetting phase (green), beginning of the drying phase (red) and end of the process (cyan).

3.4 Laboratory measurements

A set of ternary powder mixtures of ibuprofen, MCC and lactose monohydrate (cf. Table 3.1) was prepared offline in laboratory. To simulate a realistic scenario in the manufacturing of a solid dosage from, the mass fraction of the API, i.e., ibuprofen, was constrained to reside between 0.6–0.8. The two excipients were then assigned mass fractions symmetrically between 0.1–0.2. The mixture design was constructed as a union of two constrained ternary simplex centroid designs [53] to efficiently span the relevant part of fraction space as illustrated in the ternary diagram in Fig. 3.4a. The 17 mixtures in the design were prepared in triplicate, one with each of the three lactose brands. Each mixture had a total mass of 20 grams. The pure powders were weighed with the accuracy of 10 μg with an electronic balance (ABJ 220-4M, Kern & Sohn GmbH, Germany). The mixing was performed in a magnetic stirrer (Arex, Velp Scientifica, Italy) for the duration of 2.5 min at 500 rpm.

For DR measurements, self-made cuvettes were constructed by attaching a black metallic ferrule (SM1V05, Thorlabs Inc.) of 1 inch in diameter onto a 1-mm thick microscope glass slide (Menzel-Gläser, Germany). The powder mixtures were dispensed with a spattle into the ferrule so that the height of the powder column was approximately 1 cm in each measurement. The cuvette was placed carefully on top of a probe so that the measurement spot was at the center of the sample (cf. Fig. 3.4b). To avoid any systematic errors caused by the instability of the spectrometer or light source and fluctuations in the ambient temperature and relative air moisture, the measurement order of the mixtures was randomized. Each mixture was measured with four probes.



(a) Mixture design

(b) Three samples and the optical reference

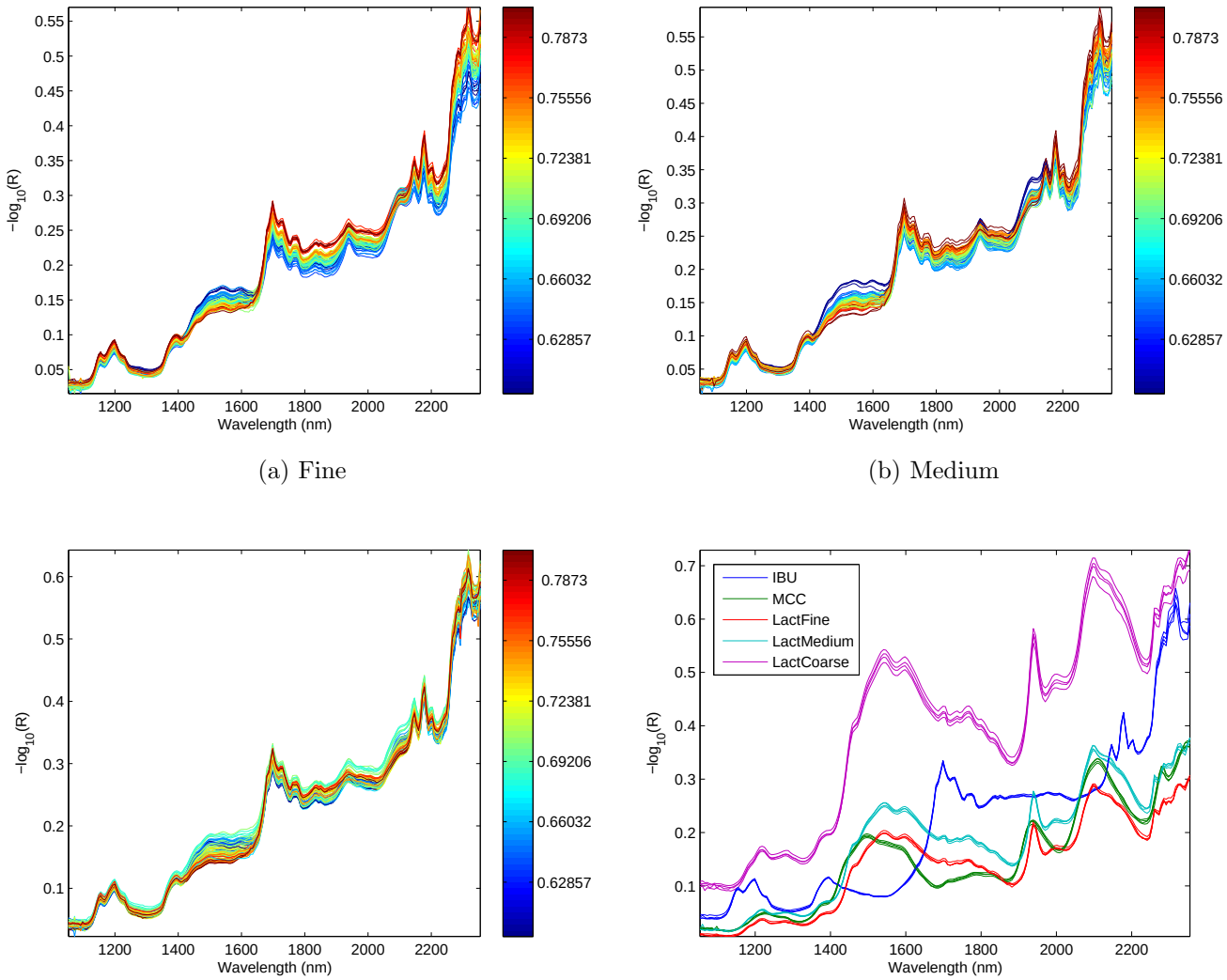
Figure 3.4: (a) The mixture design in mass fraction space; (b) The measurement setup: three powder samples in cuvettes and the optical reference placed on top of four probes.

Table 3.1: Powders in the laboratory data set.

Name	Brand	Median particle size
Ibuprofen	50FF, BASF Pharma	50 μm
Microcrystalline cellulose	VIVAPUR 101, JRS Pharma	65 μm
Lactose monohydrate, fine	Pharmatose 450M, DMV-Fonterra	20 μm
Lactose monohydrate, medium	Pharmatose 200M, DMV-Fonterra	35 μm
Lactose monohydrate, coarse	Pharmatose 80M, DMV-Fonterra	200 μm

In total, the 17 mixtures $\times$ 3 lactose brands $\times$ 4 probes amounted to 204 spectra. The $\log_{10}(1/R)$ spectra of the mixtures and the pure powders are illustrated in Fig. 3.5.

To get an estimate for the errors caused by sample heterogeneity and the use of different probes, the midpoint mixture in the set with medium-grained lactose was sampled into the cuvette 15 times in the course of the measurement. Each sample was measured with 4 probes to yield 60 replicates in total. The replicate spectra and their variance spectrum are shown in Fig. 3.6a. The variance spectrum follows the upward-sloping shape of the spectra and it has some features from the pure spectra of lactose and ibuprofen which might imply sample heterogeneity between measurements. The variance increases drastically towards the low-frequency end of the NIR spectra which is explained by the deterioration of the signal quality due to the decreased transmissivity of the fiber optics at long wavelengths. The replicate spectra were also subjected to EMSC-preprocessing in which only the physical effects were corrected. The signal and interferent spectra were excluded from Eq. (2.27), and a random spectrum among the replicates was chosen to be the reference spectrum. The amplitude of the variance spectrum decreased significantly. Its shape now lacks any resemblance to the pure analyte spectra. Physical effects may thus be concluded to be prevalent over chemical effects caused by sample heterogeneity in the errors of the



(a) Fine

(b) Medium

(c) Coarse

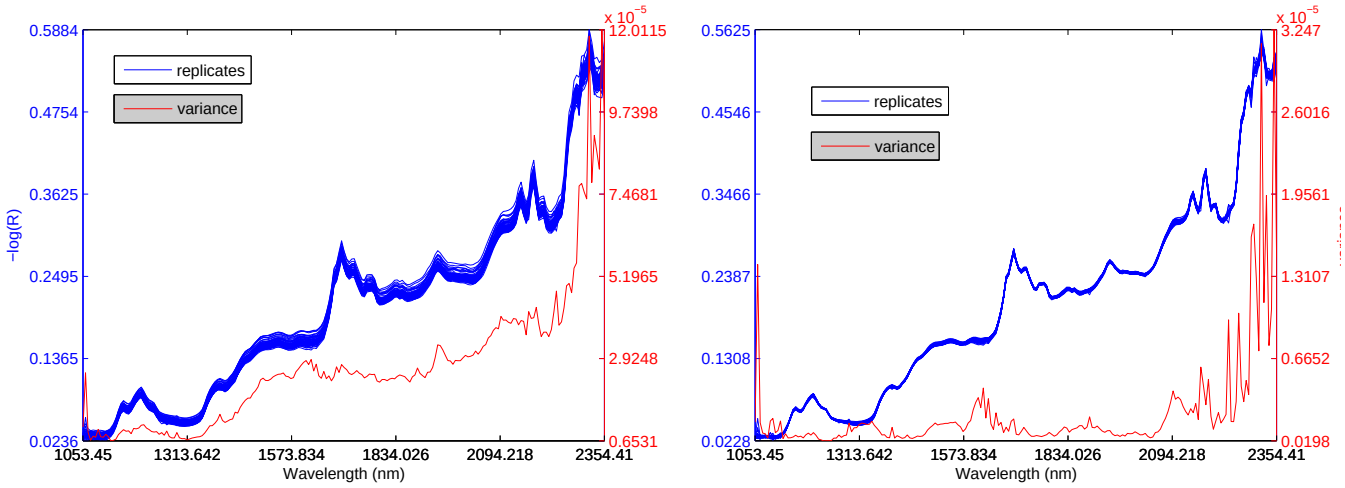
(d) Pure analyte spectra

Figure 3.5: Measured spectra of the ternary mixtures of ibuprofen, microcrystalline cellulose (MCC) and lactose with (a) fine, (b) medium-grained and (c) coarse lactose powder. The spectra are colored according to the mass fraction of ibuprofen; (d) The pure analyte spectra each measured with four probes.

laboratory data set.

3.5 Data analysis and algorithms

All computational analysis was done in MATLAB 7.7.0 R2008b environment. The PLSR version used in this work was the non-orthogonalized PLSR, implemented according to [4], which relaxes the orthogonality constraint of the scores, the columns of $\mathbf{T}$ in the bilinear model (cf. Eq. (2.20a)). The algorithm subsequently needs only



(a) Replicates and their variance

(b) EMSC-preprocessed replicates and their variance

Figure 3.6: (a) Sixty replicates (15 samplings × 4 probes) and their variance spectrum; (b) EMSC-preprocessed replicate spectra and their variance spectrum.

one set of loading vectors, the orthonormal loading weights in $\mathbf{W}$. The spectral preprocessing algorithms EMSC and OPLEC were implemented according to the Ref. [6] and [7, 27], respectively. The regular version, denoted as plain EMSC henceforth, utilized the mean spectrum of the zero-meaned data matrix $\mathbf{X}$ as the reference spectrum $\mathbf{m}$. The signal vectors $\mathbf{g}_j$ were chosen to be the $J-1$ first PC loading vectors of the zero-meaned data matrix, where J equals the chemical rank of the system. Interferent spectra were not used in this version. In the optimized EMSC, denoted as EMSCopt, the versatile publicly available implementation of SA [54] was used with default settings in the minimization in Eq. (2.37). The base vectors $\mathbf{B}$ in Algorithm 1 were chosen to be the first J SVD loading vectors of the zero-meaned mixture spectra $\mathbf{X}$. As the chemical rank of the laboratory set is three, other arguments of the Algorithm 1 were chosen to be $G = 1$, $F = 1$ and $n = 5$. Each minimization procedure was performed five times with randomly generated but sufficiently different initializations of the minimization argument $\mathbf{A}$, and the solution which provided minimum value for the cost function was chosen.

For BSS, the FastICA [37] and DSS [55] packages for MATLAB were utilized. In FastICA, the $\log\cosh(\cdot)$-based function was used for negentropy in Eq. (2.44). The whitening was performed with PCA without further rank reduction. The ICs were calculated sequentially, i.e., in the deflation mode. DSS was used with the default settings without any prior information.

Chapter 4

Results and discussion

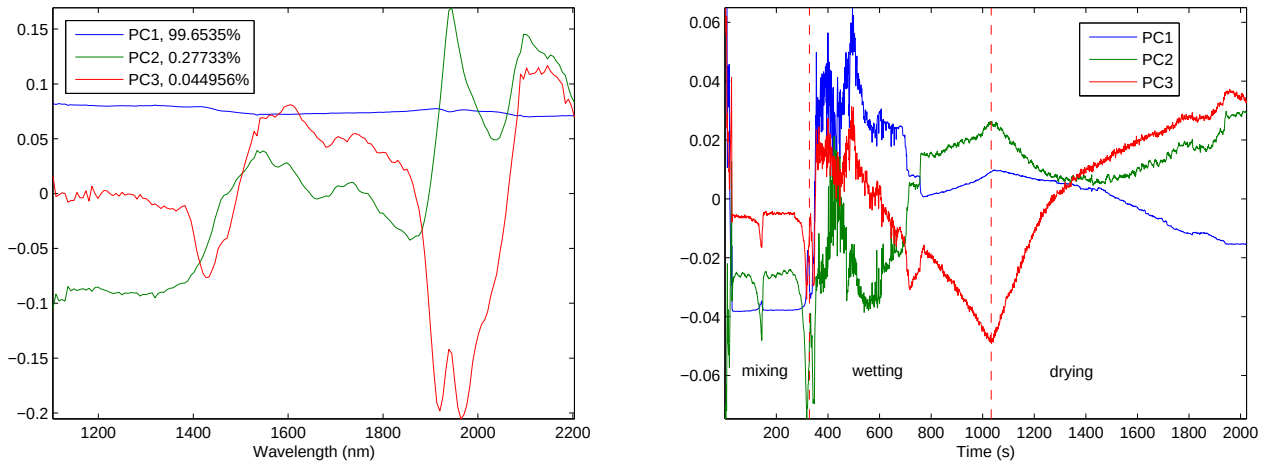
4.1 Qualitative analysis of the granulation process data

Since no reference measurements on moisture or particle size were conducted during the FBG process, the process data represents a black system which can be analyzed only qualitatively. The goal of the analysis was to find features which would provide information on the moisture and particle size distribution of the fluidized powder mass. The full multipoint characteristic of the NIR instrument was not exploited in this work, as only one channel was used in the analysis. The lowermost probe on the right in Fig. 3.3a was chosen because the effect of moisture was most prominent in this channel.

4.1.1 Principal component analysis

PCA was performed on the data as a first step in the qualitative analysis. The first three PC loading vectors are shown in Fig. 4.1a and the corresponding scores are plotted as a function of time in Fig. 4.1b. In detail:

- The first PC explains 99 % of the variance in the data and its loading vector is strictly positive, rather smooth and featureless. It can be attributed to the baseline offset which is expected to correlate with both the moisture content and the particle size distribution. Its score behaves consistently with the decreasing moisture content as it decreases monotonically during the drying phase.
- The second PC loading resembles the pure spectrum of lactose whose major absorption band near 1950 nm has been broadened due to the presence of water (cf. Fig. 3.3b). Its score contains fluctuation with a period of approximately 30 s. This effect is most probably caused by cooling system of the light source which was observed to function periodically with approximately the same frequency.
- The third PC loading vector resembles inverted pure water spectrum and it has a positive contribution from the pure spectrum of lactose, observed as the sharp



(a) PC loadings

(b) PC scores

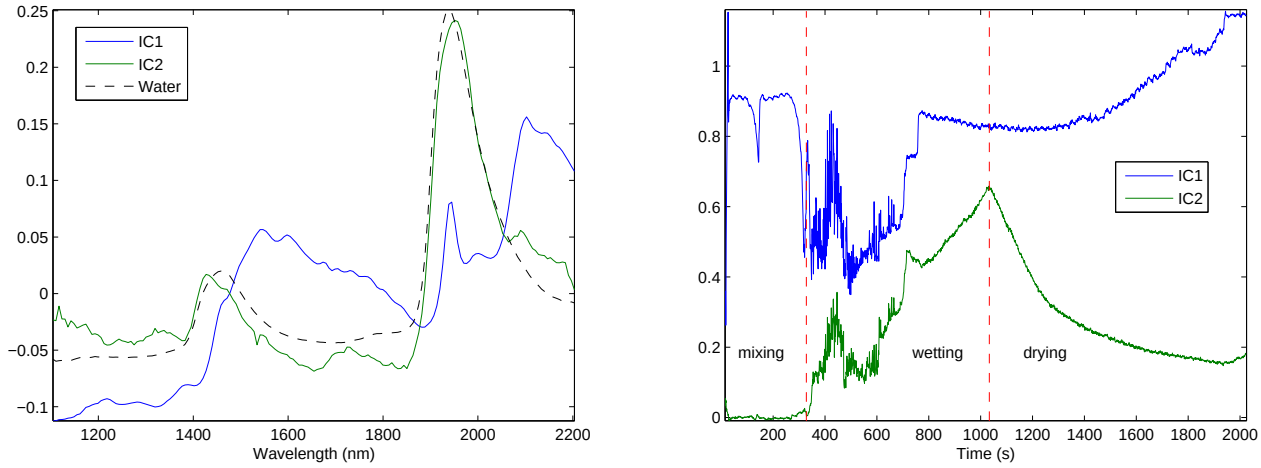
Figure 4.1: (a) Three first PC loading vectors; (b) Temporal profiles of the PC scores.

absorption band near 1950 nm. The temporal profile of its score can thus be expected to correlate negatively with the true moisture content.

In Fig. 4.1b, the transition from wetting to drying is clearly visible in all score profiles. The noisy signals at the beginning of the wetting phase are probably caused by a temporal collapse of the fluidized mass during which the probes did not see any powder. The fluidization was stopped at for a few seconds at 120, 300 and 580 s, and the two first stops can be observed as perturbations in the scores.

4.1.2 Blind source separation

The chemical composition of the time-resolved spectra was also analyzed with both FastICA and DSS. For the analysis, the spectra were preprocessed in three phases: The effect of baseline offset was initially diminished by zero-meaning them, their rank was then reduced to two by SVD to suppress random noise and the separation capabilities of the BSS algorithms were further enhanced by differentiating the spectra with respect to wavelength. It is worthwhile noticing that the chemical rank of the process spectra is four by definition, since the fluidized mass contains lactose, caffeine, water and PVP. In reality, the mixture of the lactose and caffeine powders can be assumed to stay homogeneous and thus it can be treated as one component. The least noisy IC loading vectors were obtained when the contribution of PVP was neglected, i.e., the fluidized mass was considered to consist of only powder and water. This is a reasonable assumption, since PVP was present in only trace amounts, as its mass fraction was less than 4.59 % w/w in the end product. The IC loading vectors estimated with FastICA using 2nd order derivative in preprocessing are shown in Fig. 4.2a, and their corresponding scores estimated with Eq. (2.47) are plotted as a function of time in Fig. 4.2b.



(a) IC loadings

(b) IC scores

Figure 4.2: (a) Two IC loading vectors; (b) Temporal profiles of the IC scores.

In the analysis, the optimal order of differentiation was determined interactively so that the visual resemblance between one IC loading vector and the pure water spectrum was maximized. The resolved source signals can be analyzed as follows:

- The first IC corresponds to the powder mixture, as its loading vector resembles the spectrum of lactose. The presence of caffeine can be observed as a slight increase in the apparent absorbance near 1700 nm, where its main absorption peak is located. The first IC score is not dependent on the moisture profile (the second IC score) since its temporal profile exhibits no response to the transition from wetting to drying. Thus, the resolution of FastICA can be concluded to be successful. The periodical fluctuation attributed to the instability of light source is now observed in the first IC score.
- The second IC loading vector is close to the water spectrum albeit it contains some unidentified interference near 2100 nm. The deviations from the true water spectrum may be explained by sample matrix interactions, the presence of PVP and the fact that some information on the water spectrum may have been lost in rank reduction. The temporal profile of the second IC score exhibits consistent behavior with the true underlying moisture profile of the powder mass during the process. It stays at a constant value near zero during mixing, it increases during wetting and declines monotonically during drying. The drying profile even follows the expected scenario [45]: the drying rate stays constant at the beginning, and it gradually decreases towards the end of the process.

Although some discrepancies can still be observed in the resolved spectra, the use of ICA combined with preprocessing proved to be beneficial as it improved the chemical interpretability of the system when compared to PCA.

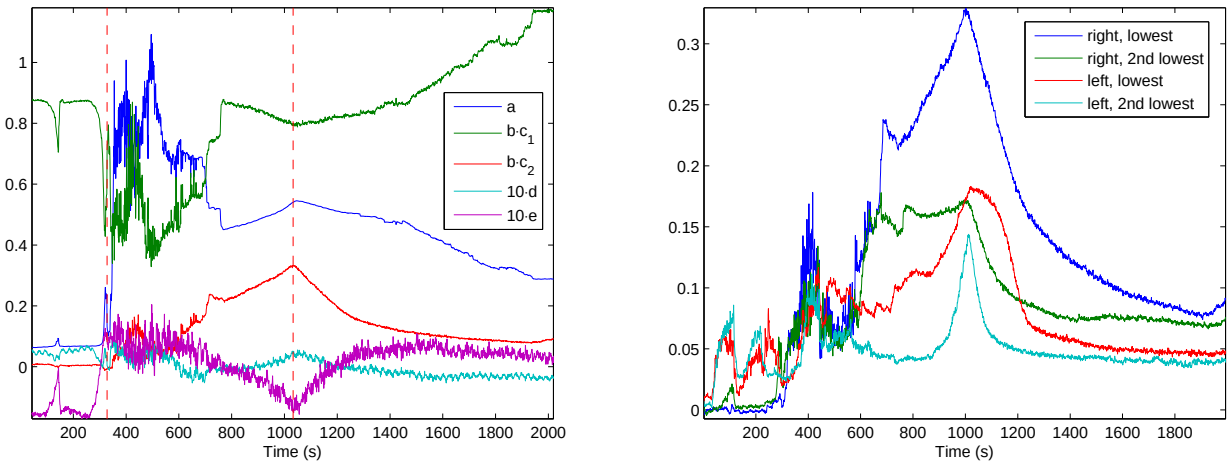
The increase in the score of the first IC during drying may be explained by the contamination of the measurement window with the sticky fluidized powder mass. Towards the end of the process, the layer of powder mass on the window increases in

thickness which results in increased intensity of the backscattered light. With a thin layer, majority of light is transmitted, but a thicker layer increases the probability that a photon is scattered back towards the probe. The photons also have longer apparent optical path lengths when the layer is thicker, hence the increase in the contribution of the first IC. At the end of the process, the inner wall of the granulator was covered with a partly hardened powder mass. Since the probes saw only the layer of stationary powder on the window, it is probable that no information on the particle size distribution was collected during the process. Furthermore, the spectra were not corrected for the multiplicative effect prior to ICA. Hence the scores are not expected to be linearly correlated with the mass fractions of powder and water.

Although the collection of the FBG data was termed to be an inline measurement, the data analysis was done offline after the process. All the measured temporally resolved spectra were preprocessed and fed to FastICA at the same time. This is one drawback of the BSS algorithms, viz., they are not easily applicable for inline use. Of course, the BSS algorithm can be driven successively using, e.g., all previously measured spectra or subsequent temporal windows of data as input. However, it cannot be guaranteed that the BSS algorithm converges such that the IC loading vectors are identical in all iterations. To develop an inline monitoring method for the FBG process, the pure analyte spectra should be known *a priori* or they should be estimated from a previous data set with the same formulation for the fluidized mass and the binder liquid. The temporal concentration profiles may then be estimated in LS sense by utilizing the measured spectra and assuming the linear mixture model. In Fig. 4.3a, the spectral model from Eq. (2.24) was utilized and the temporal profiles of the model parameters were estimated using Eq. (2.28). The two IC loading vectors from Fig. 4.2a were chosen to comprise the chemical linear mixture model $\mathbf{x}_{i,\,\mathrm{chem}}$. The elements of the sloping baseline vector $\boldsymbol{\lambda}$ were linearly spaced between -1 and 1, and $\boldsymbol{\lambda}^2$ was obtained by element-wise squaring of $\boldsymbol{\lambda}$. The multiplicative error was not corrected for and it is thus present in the estimated coefficients of the chemically relevant spectra. The estimated model parameters provide online information on both the physical and chemical properties of the fluidized powder mass.

The temporal profile of the parameter a, the weight of the constant baseline offset, resembles that of the first PC score in Fig. 4.1b. It correlates negatively with the coefficient of the powder spectrum which is probably due to the effect of layer thickness explained above. Smaller intensity of backscattered light, and hence larger baseline offset, occurs with thinner layer in the beginning of the drying phase. The baseline offset decreases as the layer thickness increases during drying. As expected, moisture content also correlates positively with the baseline parameter a, although the shape of the moisture profile is not visually detectable in it. Although their contributions are small, the coefficients of $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^2$ follow the smooth changes of the estimated moisture profile. The presence of water thus causes a slight twist in the curvature of the measured spectrum. The periodical fluctuation observed previously in the first IC score is now observed mainly in the term describing linear baseline sloping. The fluctuating light source thus exhibits minute tilting in the spectrum.

The effect of probe fouling is evident when the signals from other channels are inspected. Using the two ICs shown in Fig. 4.2a as loadings, the corresponding scores



(a) Model parameters

(b) Moisture profiles

Figure 4.3: (a) Temporal profiles of the model parameters in Eq. (2.24); (b) Moisture profiles measured with four lowest probes.

were estimated in LS sense for the zero-meaned spectra measured with the four lowest probes in Fig. 3.3a. In Fig. 4.3b, the temporal profiles of the scores corresponding to the estimated water spectrum are shown. Although the transition between wetting and drying phases is visible in all channels, the profiles are very different in shape. The probes suffer from unrepresentative sampling, as they see only the local stationary powder layers deposited on their glass windows.

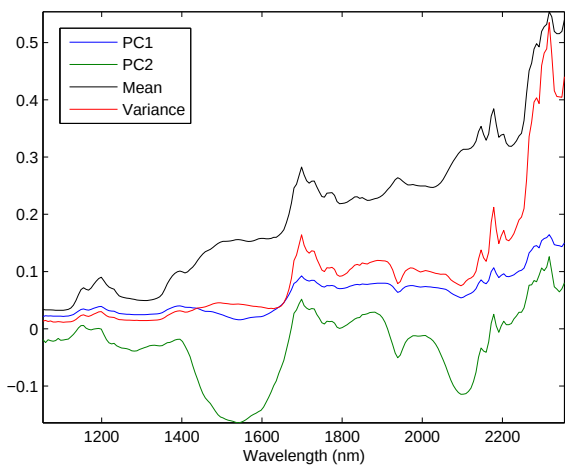
4.2 Qualitative analysis of the laboratory data

To analyze the signal quality and to detect possible input outliers in the measured laboratory data, the laboratory data set was analyzed qualitatively using PCA in Sect. 4.2.1. The proposed three-phase combination of spectral preprocessing and ICA is tested with the laboratory data in Sect. 4.2.2.

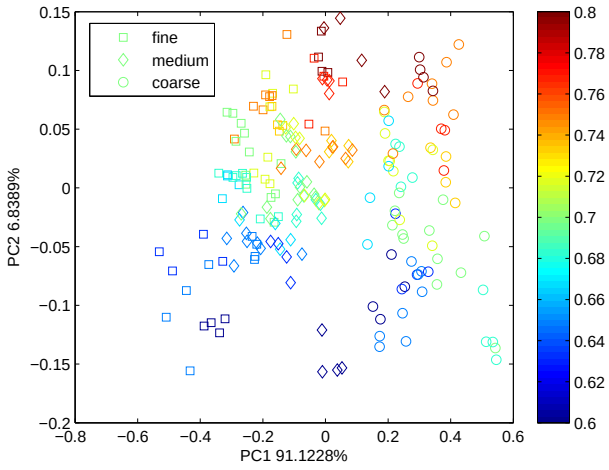
4.2.1 Principal component analysis

All 204 mixture spectra were collected into one matrix $\mathbf{X}$ and they were subjected to PCA. The two first PC loading vectors as well as the mean and scaled variance spectra are shown in Fig. 4.4a. The PC scores are shown in Fig. 4.4b, where the three mixture sets with different lactose powders are drawn with distinct symbols and they are colored according to the mass fraction of ibuprofen.

The first PC, which explains 91 % of spectral variance, is positive and relatively featureless. The mixtures with coarse lactose powder have the strongest contribution of the first PC and they form a separate cluster on the PC1 axis in Fig. 4.4b. The smoothness and the positivity of the first PC loading vector explains the constant baseline



(a) PC loadings, mean and variance spectra



(b) PC scores

Figure 4.4: (a) Two first PC loading vectors, the mean spectrum and scaled variance spectrum for all measured mixture spectra; (b) The score plot for the PCs colored according to the mass fraction of ibuprofen.

offset in $\log(1/R)$ spectra which results from the decreased intensity of backscattered light due to large particle size. The PC has some resemblance to the pure spectrum of ibuprofen, and it also has a negative contribution from the lactose spectrum, which is observed as the inverted dip at location of the strong absorption band of lactose at 1950 nm. Thus, the first PC might explain the effect of segregation in which the finer ibuprofen powder is carried towards the bottom of the cuvette to fill the inter-particulate voids between the large lactose particles. Hence the probe sees relatively more ibuprofen than lactose in the samples containing coarse lactose powder. The second PC contains positive contribution from ibuprofen and negative contributions from both lactose and MCC, as is manifested by the negative values at 1550, 1950 and 2100 nm, i.e., at the locations of the major absorption peaks of lactose and MCC. It partly explains the mass fraction of ibuprofen, as its contribution is largest in the spectra with high ibuprofen content, as is seen in the score plot.

The variance spectrum, which gives the variance of each wavelength variable in the data set, resembles the pure spectrum of ibuprofen. Most of the variance is thus attributed to chemical changes where ibuprofen expectedly dominates. The decreased signal quality at large wavelengths caused mainly by the low transmissivity of the optical fibers results in unproportionally high variance at the low-frequency end of the variance spectrum.

4.2.2 Blind source separation

The data set was also analyzed with FastICA and DSS using zero-meaning, rank reduction to three with SVD and differentiation as preprocessing. Prior to BSS, the wavelength range was reduced to 1100–2250 nm to remove the nonlinearities in the

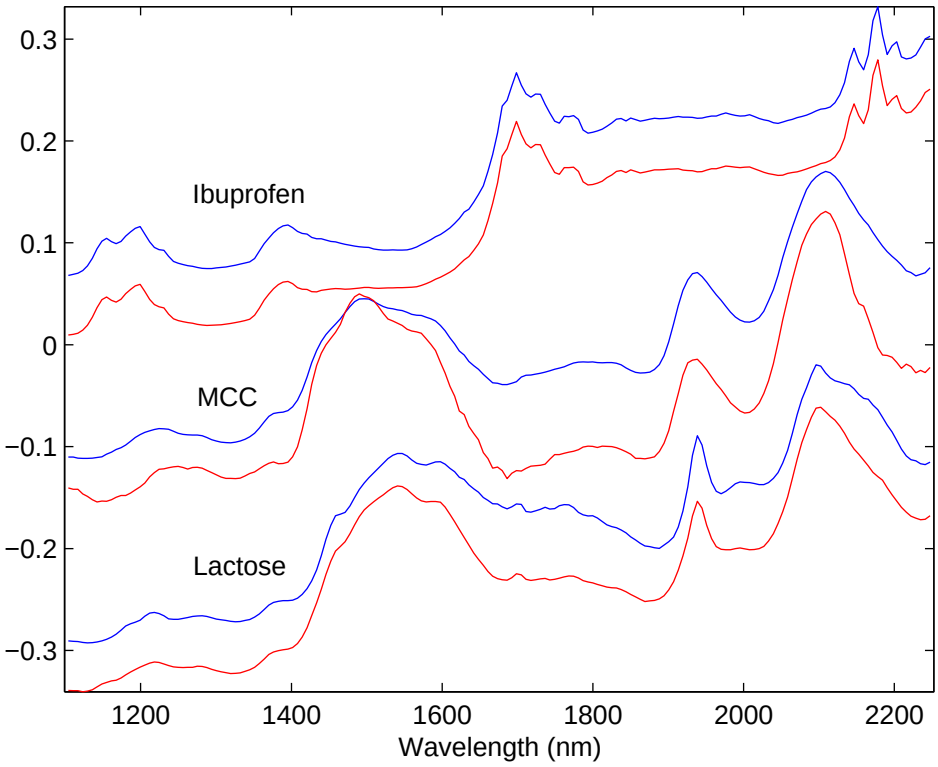
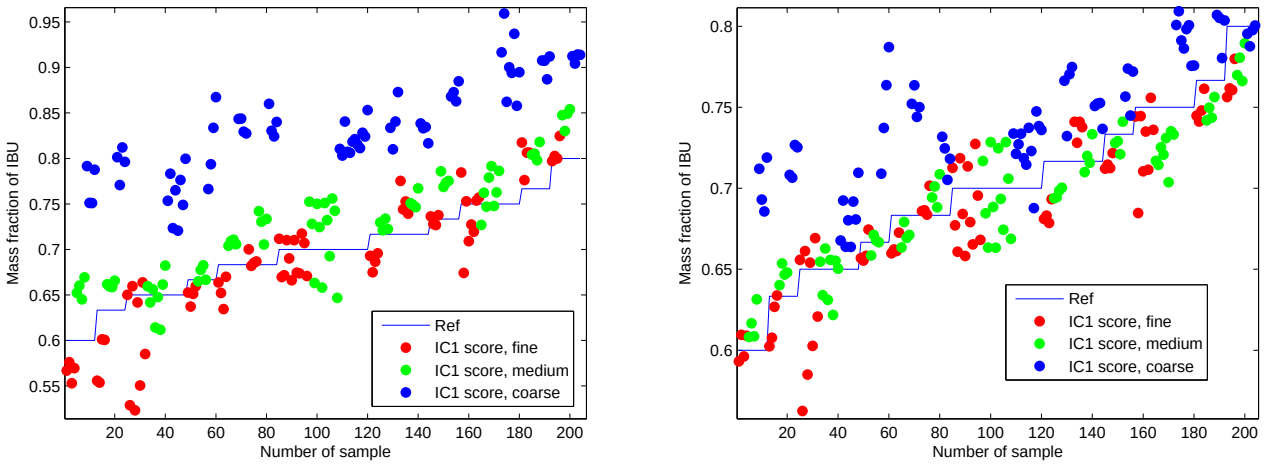


Figure 4.5: Pure analyte spectra (blue) and their corresponding IC loading vectors (red) for ibuprofen, microcrystalline cellulose and lactose (baseline offsets added for visualization).

noisy ends of the NIR spectra. The best reconstructions of the underlying pure analyte spectra shown in Fig. 4.5 were obtained with DSS using third order derivative in the preprocessing phase. Even the spectra of MCC and lactose, whose contributions were significantly smaller than that of ibuprofen in the mixtures and which are strongly cross-correlated signals, were rather accurately resolved. The scores of the IC corresponding to ibuprofen are shown in Fig. 4.6a scaled such that they are comparable with the true mass fractions of ibuprofen. As expected, the effect of increased apparent optical path length due to larger particle size is evident in the mixtures with coarse lactose powder, which is manifested by the systematically larger IC score. The effect of the multiplicative error was subsequently standardized by normalizing the zero-meaned mixture spectra (cf. (2.48)) prior to differentiation and DSS. The resolved IC loading vectors were then almost identical to those in Fig. 4.5 (results are not shown), and the correlation coefficient between the first IC scores in Fig. 4.6b and the true mass fractions of ibuprofen was increased from 0.61 to 0.76 when compared to Fig. 4.6a. The mixture spectra with coarse lactose powder have still generally largest score values for the first IC, i.e., the presence of ibuprofen is largest in them. This is congruent with the analysis of the first PC above, i.e., the segregation of powders is strongest in the mixtures with coarse lactose powder.



(a) Reference and the IC1 scores

(b) Reference and the IC1 scores, normalization

Figure 4.6: The laboratory reference, i.e., mass fraction of ibuprofen, and the scores of its corresponding IC (a) without and (b) with normalization of the zero-meaned mixture spectra prior to BSS (scaled for visualization).

4.3 Performance of spectral preprocessing methods

The laboratory data set was used to test the performance of EMSC, EMSCopt and OPLEC in improving the robustness of subsequent linear calibration models against the physical spectral artefacts. Specifically, the analyzed problem was the estimation of the mass fraction of ibuprofen from the measured spectra using the calibration models. First, the laboratory data were divided into calibration and test sets, as described in Sect. 4.3.1. Second, linear calibration models were developed using the calibration set with and without spectral preprocessing, and the predictive performances of the models were compared using the test set (cf. Sect. 4.3.1). The PLSR model developed without preprocessing is analyzed in detail in Sect. 4.3.2. Besides providing similar analysis of the PLSR model after preprocessing the spectra with EMSCopt, Sect. 4.3.3 also contains interpretation of the information retrieved from the preprocessing methods EMSCopt and OPLEC.

4.3.1 Comparison of spectral preprocessing methods

The mixture spectra containing coarse lactose powder exhibited increased optical path lengths observed as multiplicative errors (cf. Fig. 4.6a) which would have provided a suitable testing ground for the spectral preprocessing methods. However, the effect of segregation was also largest in them (cf. Fig. 4.6b and 4.4), which made the observations unsuitable for the purposes involving calibration. In fact, in initial tests, calibration against these erroneous laboratory reference values resulted in poor predic-

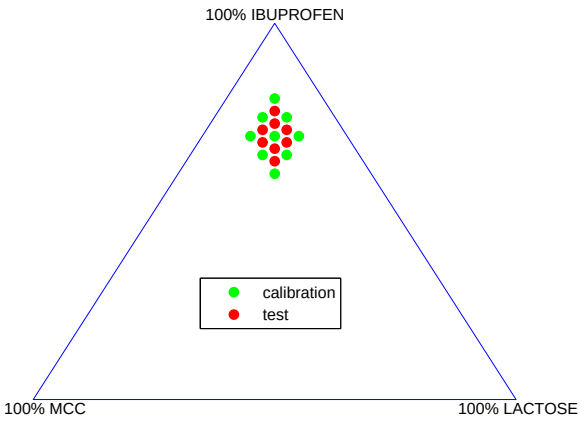
tion ability and badly interpretable results (for compactness, results are not shown). Hence the coarse set was excluded from the analysis.

Although the differences in the apparent optical path length were not significant due to the similar particle size distributions in the two remaining sets of mixture spectra, physical effects were still expected to be present due to the use of different measurement probes. For the analysis of the spectral preprocessing methods, the remaining two sets of spectra were divided into calibration and test sets as is shown in Fig. 4.7a. The calibration set spans efficiently the relevant part of the fraction space and no extrapolation is expected to be needed in the prediction phase. Each mixture point in the ternary diagram contains 8 replicates (2 particle sizes $\times$ 4 probes). Since the measurement procedure was well standardized, the signal quality was approximately equal in all measurements and no input outliers were detected when the calibration set was analyzed with PCA in a similar way as in Fig. 4.4 (results are not shown). However, the effect of segregation was noted to be evident, and two output outliers were detected and subsequently removed from the calibration set as they exhibited abnormally large errors in cross validation.

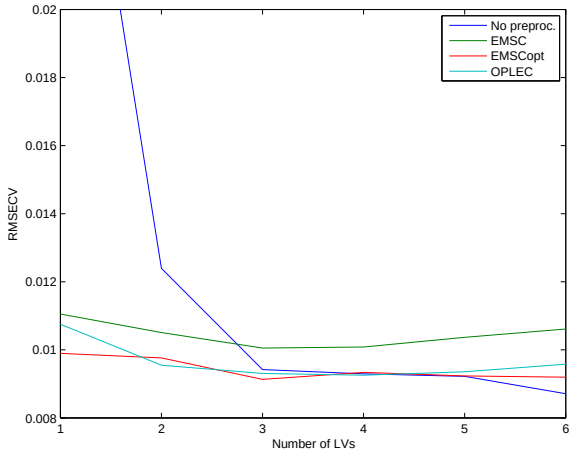
Calibration models were constructed for the prediction of the mass fraction of ibuprofen in the mixtures using PLSR. The models were optimized for a number LVs ranging from 1 to 6. To validate the effects of spectral preprocessing, the PLSR models were first built with the raw data. Then the spectra were preprocessed with EMSC, EMSCopt and OPLEC prior to calibration. The optimal number of LVs for the calibration models was analyzed using cross validation, and the leave-one-out root-mean-squared errors of cross validation (LOO-RMSECVs) of plain PLSR, EMSC and OPLEC as well as the optimized leave-block-out RMSECV of EMSCopt are plotted as a function of the number of PLSR LVs in Fig. 4.7b.

Based on the experimental results, all three spectral preprocessing methods resulted in approximately equal prediction performances which was comparable to that of plain PLSR when three or more LVs were utilized. The root-mean-squared errors of prediction (RMSEPs) are shown in Fig. 4.7c. However, spectral preprocessing methods are observed to reduce the number of LVs needed as their corresponding calibration models exhibit good prediction ability already with one and two LVs. The usefulness of model-based spectral preprocessing using EMSC and OPLEC is thus confirmed by fact that, when the measurements are affected by physical effects, its application is not only theoretically correct but also of fundamental practical importance as it allows the development of parsimonious calibration models (using one or two LVs instead of three) still capable of achieving the requested accuracy.

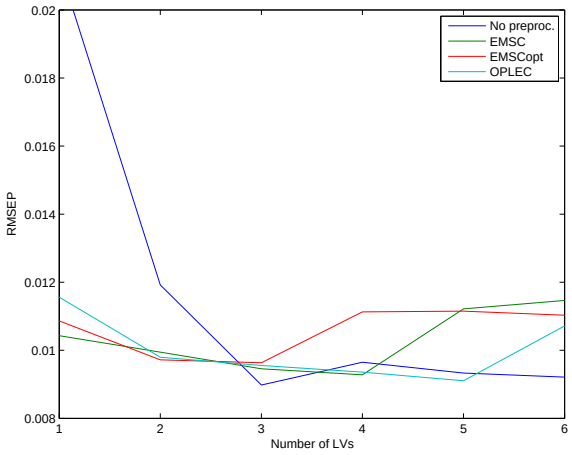
Since MCC and lactose have weak presence when compared to ibuprofen, and since the variation in their mass fractions is also small, the variation in the mass fraction of ibuprofen is a dominant chemical effect in the spectra. The first LV in the PLSR models developed after preprocessing is thus expected to explain most of the variation attributed to the changes in the presence of ibuprofen. The following two LVs then compensate for the minor perturbations caused by the presence of MCC and lactose. Furthermore, since MCC and lactose have strongly crosscorrelated spectra, the PLSR models might have difficulties in distinguishing between the two. The two excipients can be thus roughly approximated as a single chemical component, which reduces the



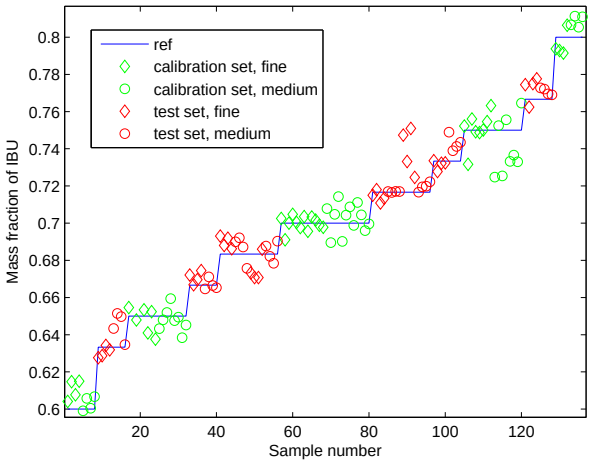
(a) Calibration and test sets



(b) RMSECV



(c) RMSEP



(d) Predicted values with PLSR

Figure 4.7: (a) The division of the data into calibration and test sets; (b) RMSECVs and (c) RMSEPs for four methods plotted as a function of the number of LVs; (d) The predicted ibuprofen mass fractions for both calibration and test sets after using PLSR with 3 LVs without preprocessing.

apparent chemical rank of the system and consequently the number of LVs needed. This is the reason for the fact that prediction performance only slightly improves as the number of LVs is increased from one to three in the cases where preprocessing was used.

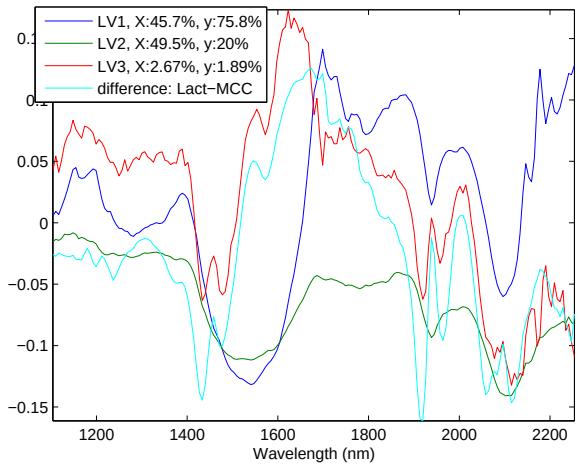
In the absence of preprocessing, the plain PLSR model is expected to model also the physical effects which do not correlate with the mass fraction of ibuprofen. This is observed as the significantly larger RMSEP with one and two LVs when compared to the cases where preprocessing was used. The physical effects are thus modeled by the first three LVs. Again, the decrease in the apparent chemical rank of the system is

obvious, as the plain PLSR has to utilize only three LVs to achieve prediction performance comparable with cases where preprocessing was used, and the performance is not improved with four or more LVs. As mentioned in Sect.2.3, the required number of LVs to completely model the spectral variations in a noiseless set of absorbance spectra equals the chemical rank of the system. It is thus expected that, with the apparent increase in the chemical rank induced by the physical baseline effects, the optimal number of LVs would be at least four in the present case. However, due to the similarity between the pure analyte spectra of MCC and lactose, the apparent chemical rank caused by chemical variations is closer to two for the system. The inclusion of physical baseline effects thus increases the apparent chemical rank to three. After the third LV, all PLSR models begin to incorporate random noise in the calibration model and the RMSEPs start to increase due to overlearning.

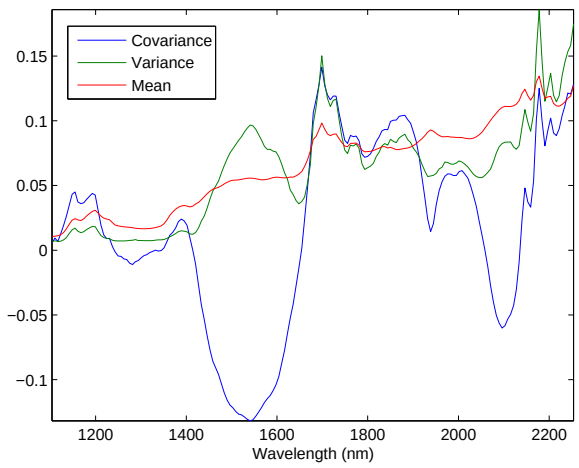
4.3.2 Analysis of the PLSR model

The predicted mass fractions of ibuprofen are shown with the reference values for both the calibration and test sets in Fig. 4.7d, where the plain PLSR was used with three LVs. When inspected visually, there seems not to be systematical differences in the prediction performance between the sets containing fine and medium lactose powder. The PLSR model thus sufficiently accounts for the possible differences in the particle size distributions in the data matrix $\mathbf{X}$ with three LVs. The first three loading weight vectors $\mathbf{w}_i$ of the model are plotted in Fig. 4.8a, and the variances explained by them in both $\mathbf{X}$ and $\mathbf{y}$ in the calibration set are given. Fig. 4.8b presents the covariance spectrum between the wavelength channels and the reference $\mathbf{y}$, the variance spectrum denoting the observed variance at each wavelength and the mean spectrum of the calibration set. Since PLSR finds loading weight vectors $\mathbf{w}_i$ which maximize the covariance between $\mathbf{X}_{i-1}\,\mathbf{w}_i$ and $\mathbf{y}$, the first vector $\mathbf{w}_1$ is always directly proportional the covariance vector between $\mathbf{X}$ and $\mathbf{y}$, as is seen in the figures.

The first loading weight vector has positive contribution from the pure spectrum of ibuprofen, as expected, and negative contributions from the spectra of MCC and lactose. It explains most of the variance (75.8 %) in $\mathbf{y}$ but less than half of the variance (45.7 %) in $\mathbf{X}$. The second loading weight vector is completely negative in sign. Due to the constant sign, it thus may account for constant baseline offset in $\mathbf{X}$. It also explains the perturbations caused by changes in the contents of MCC and lactose as it looks like a linear combination of their pure analyte spectra but lacks any resemblance to the pure spectrum of ibuprofen. On the other hand, the third loading weight vector is hard to interpret because it is noisy and thus probably encodes very little information. However, it resembles the difference between the pure analyte spectra of lactose and MCC also shown in Fig. 4.8a, and it may thus be interpreted to explain the small variations between the two excipient spectra. For example, it has an interesting feature at 1950 nm where both lactose and MCC have an absorption band. The feature might act as a way to distinguish between MCC and lactose. If the third vector is summed with the second, the inverted lactose peak in the second vector at 1950 nm becomes wider and starts to resemble the inverted absorption peak of MCC. Since the third loading weight vector explains only trace amounts of variance



(a) Three LVs, no preprocessing



(b) Covariance, variance and mean spectra, no preprocessing

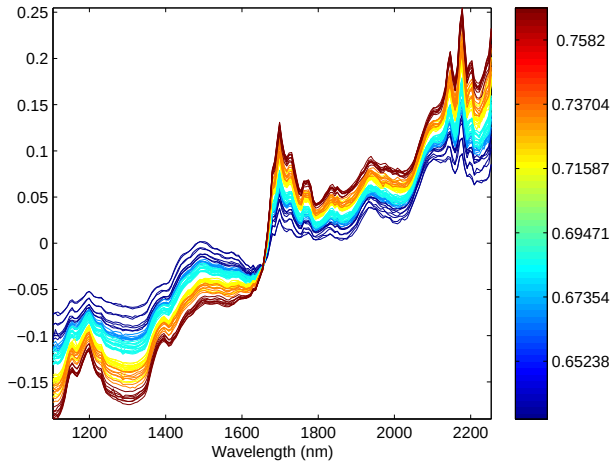
Figure 4.8: (a) Three first LV weight vectors and (b) the covariance, variance and mean spectrum of the data without preprocessing.

(2.7 %) in $\mathbf{X}$, the previous analysis regarding the decrease in the apparent chemical rank of the system due to the similarity between the MCC and lactose spectra can be verified to be realistic. That is, the two excipient spectra can be very accurately approximated as a single chemical species.

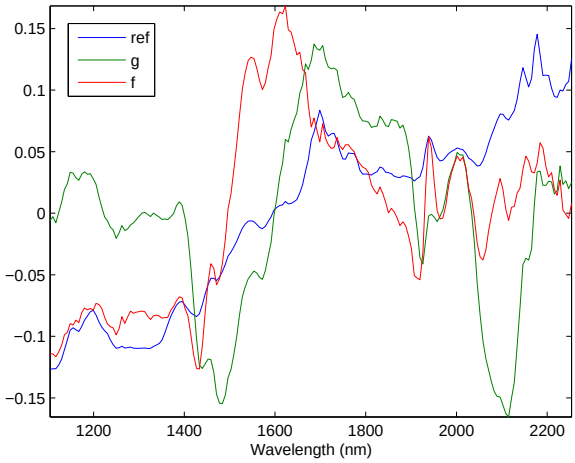
4.3.3 Analysis using spectral preprocessing methods

Because of the segregation, the use of optimized EMSC was not expected to produce accurate results when applied to the laboratory data. This is mainly due to the fact that optimizing the LBO-RMSECV can lead to overlearning. The results are here discussed for the case where one signal and one interferent vector was estimated. The chemical rank of the preprocessed spectra was thus reduced to two. As is seen in Fig. 4.9a, the spectra preprocessed with EMSCopt are approximately linearized with respect to the laboratory reference values despite the variance between replicates. However, the interpretability of the reference, signal and interferent spectra (Fig. 4.9b) is compromised because they have been optimized to model inaccurate reference values. The covariance, variance and mean spectra of the calibration set are given in Fig. 4.9b. The reference spectrum $\mathbf{m}$ is close to the mean spectrum of the calibration set after preprocessing, which is natural as $\mathbf{m}$ should have equal contribution in all spectra. Since the single signal vector explains all variations around the reference spectrum, it is obvious that it is now almost identical to both the first loading weight vector and the covariance spectrum. Also, the variance spectrum has high resemblance to the signal vector for the same reason. The second loading weight vector is already very noisy and contains little information. As stated in Sect. 2.4.1, the number of LVs needed with EMSCopt is G, one in this case, and the second LV

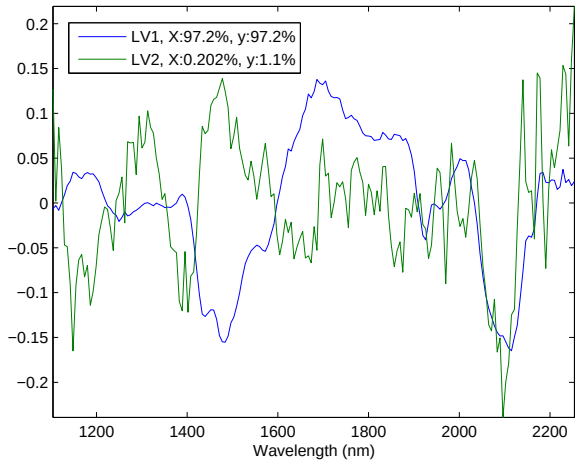
contains information only from the unmodeled residuals in Eq. (2.24).



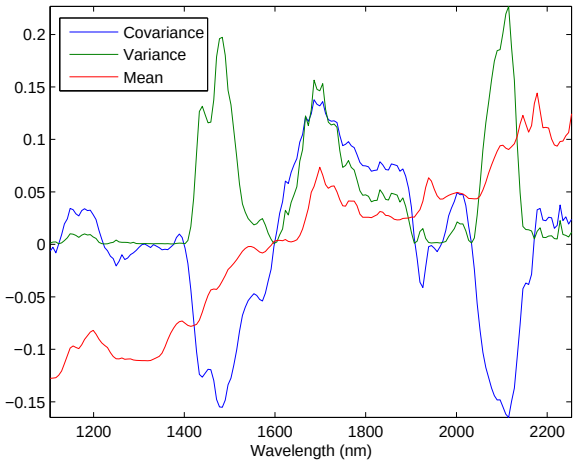
(a) EMSC-preprocessed calibration spectra



(b) Reference, signal and interferent spectra



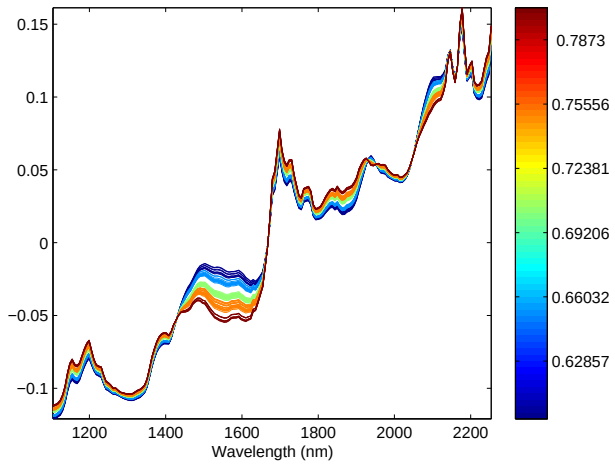
(c) Two first LVs after optimized EMSC



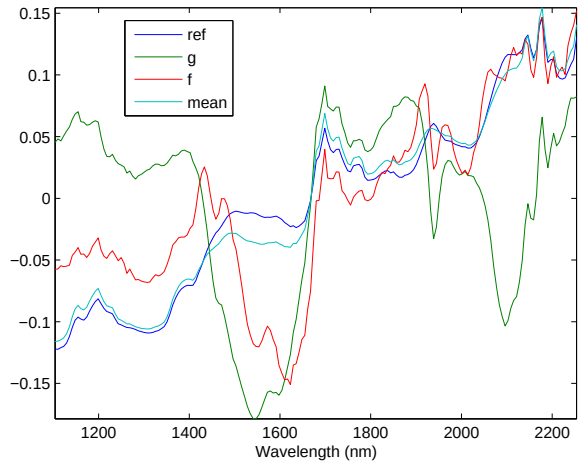
(d) Covariance, variance and mean spectra after optimized EMSC

Figure 4.9: (a) EMSC-preprocessed calibration spectra; (b) Reference, signal and interferent spectra in the EMSC model; (c) Two first LVs and (d) the covariance, variance and mean spectrum of the data after optimized EMSC preprocessing.

The version of EMSC in which the effects of unmodeled residuals were removed by using Eq. (2.30) instead of Eq. (2.29) was also tested. Now the preprocessed spectra were constrained to be in the space spanned by the reference and signal spectra, and they appear to be smoother (cf. Fig. 4.10a, one LV is used). The prediction performance stayed approximately the same (results are not shown), but the reference, signal and interferent spectra in Fig. 4.10b are now easier to interpret. The signal vector has positive contribution from the pure spectrum of ibuprofen and negative contributions from MCC and lactose. The interferent spectrum might have the same



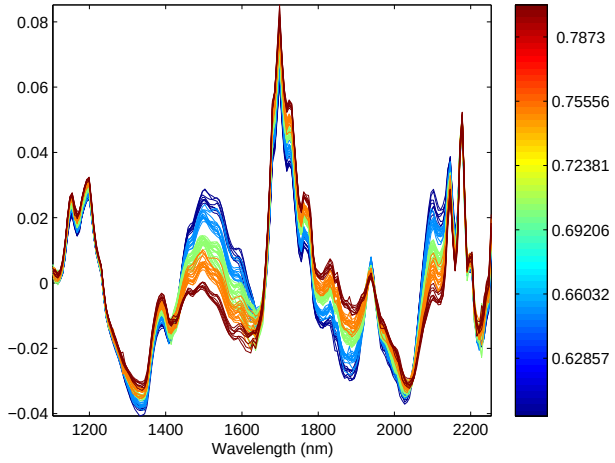
(a) EMSC-preprocessed calibration spectra



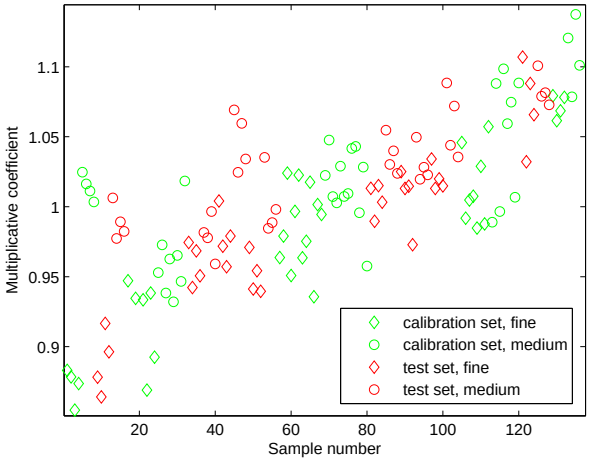
(b) Reference, signal and interferent spectra

Figure 4.10: (a) EMSC-preprocessed calibration spectra; (b) Reference, signal, interferent and mean spectra. (Eq. (2.30) used instead of Eq. (2.29))

role as the third LV in Fig. 4.8a, i.e., to explain the small differences between the MCC and lactose spectra. For example, the feature near 1950 nm widens the lactose peak when the spectrum is summed with the reference spectrum. The reference and mean spectra again resemble each other. The use of the modified version of EMSC might be thus beneficial, since it provides better interpretability for the EMSC model since it discards the effects of the unmodeled residual errors.



(a) OPLEC-preprocessed calibration spectra



(b) Estimated multiplicative coefficients

Figure 4.11: (a) OPLEC-preprocessed calibration spectra; (b) Estimated multiplicative coefficients in OPLEC.

OPLEC is also vulnerable to the erroneous reference values as it utilizes them in esti-

mating the multiplicative coefficients b_i which are present in Eq. (2.32). In Fig. 4.11a, the OPLEC-preprocessed spectra are seen to be well linearized with respect to the reference. Three LVs were used in both PLSR models in Eq. 2.33. The shapes of the spectra are distorted due to the projection in Eq. (2.32). OPLEC was also modified by replacing the projection by zero-meaning. The preprocessed spectra had then the familiar shapes, but the RMSEP was notably increased with the two first LVs (results are not shown). The measured spectra thus contain additive baseline curvatures which increase the number of needed LVs. The estimated optical path length factors b_i are shown in Fig. 4.11b for both calibration and test measurements ordered as in Fig. 4.7d. The mixtures with medium lactose powder generally exhibit slightly longer optical path lengths than the set with fine lactose powder, as is expected. The upward trend of the plot indicates that the multiplicative coefficient increases with the mass fraction of ibuprofen, which is probably due to the fact that the median particle size increases with the ibuprofen content (cf. Table 3.1).

Chapter 5

Conclusions

Near-infrared (NIR) spectra measured on solid samples in the diffuse reflectance (DR) mode are sensitive to the physical characteristics, such as the packing density and the particle size distribution, of the material. The variations caused by light scattering effects evoke nonlinearity in apparent absorbance spectra, which are often modeled to be linearly correlated with the concentrations of the chemical constituents present in the samples. The prediction ability of linear calibration models is degraded by these unmodeled effects and, due to the deviations from the linear mixture model, the use of BSS methods with NIR DR spectra becomes complicated. However, the physical light scattering effects may be taken into account and standardized with model-based spectral preprocessing methods, such as extended multiplicative signal correction (EMSC) and optical path length estimation and correction (OPLEC).

In this work, an optimized version of EMSC was proposed, in which the chemically relevant spectral vectors in the EMSC model were optimized as linear combinations of a given set of base vectors so that the root-mean-squared error of cross validation (RMSECV) of the subsequently built partial least squares regression (PLSR) model was minimized. Due to the possible ill-behaving nature of the cost function, the minimization was executed stochastically with simulated annealing. Ternary powder mixtures of ibuprofen, microcrystalline cellulose and lactose with three different particle sizes were prepared in laboratory and the method was tested with their NIR DR spectra. In the data analysis phase, the measured samples were noted to exhibit physical heterogeneity due to the segregation of powders. This resulted in a poor correspondence between the weighed mass fractions of ibuprofen and the measured NIR spectra. However, optimized EMSC resulted in a prediction accuracy comparable to regular EMSC and OPLEC. Moreover, the preprocessing methods lowered the number of LVs needed for accurate prediction of the mass fraction of ibuprofen using a PLSR model. The data set contained only mild physical variations, since a regular PLSR model constructed with unprocessed data was able to model them, as it gave comparable prediction performance.

The application of blind source separation (BSS) methods on NIR DR spectra is complicated by the presence physical spectral interferences which deflect the measured signals from the linear mixture model. The importance of appropriate spectral preprocessing was investigated in the problem of enhancing the separation capabilities

of BSS algorithms. A three-phase preprocessing method designed for the application of independent component analysis (ICA) on NIR DR spectra was proposed and discussed in this work. The variations in the constant baseline offset in the spectra was first removed by zero-meaning them. Then, the effect of random noise was decreased by discarding the least informative components in the SVD of the data set. Finally, the smooth spectral signals were made more suitable for ICA by differentiating them with respect to wavelength, which increased their non-Gaussianity. The combination of preprocessing and ICA was tested both with a process data set measured during fluid bed granulation (FBG) and the previously mentioned laboratory data set. When compared to principal component analysis (PCA), the proposed method permitted easier interpretation of the factorization, as the resolved signals were chemically more meaningful than the principal component loading vectors.

In the analysis of the FBG data, the pure spectrum of water was satisfactorily resolved, and the temporal shape of its score exhibited similarity to the expected moisture profile. The process phases were clearly distinguishable and the shape of the drying profile agreed with the theory on the drying of powders. The proposed method has thus potential in inline FBG process monitoring applications. Simultaneous resolution of both chemical and physical properties from the FBG data in real time was demonstrated using the modified Beer-Lambert's law utilized in EMSC and OPLEC. The parameters of the model, which were estimated in least squares sense for each measured spectrum, provided information both on the presence of the chemical constituents and the additive baseline effects which can in principle contain information on the granule size distribution. In the current FBG data, however, no information on granule size was expected to exist, since most of the measured signal originated from a stationary layer of powder mass deposited on the measurement window. In the analysis of the laboratory data, all three estimated pure analyte spectra were visually recognizable. The estimated scores for the independent component (IC) whose loading vector most resembled the pure analyte spectrum of ibuprofen correlated positively with the weighed mass fractions of ibuprofen. The multiplicative effect of increased optical path length due to large particle size was evident in the systematically larger IC scores corresponding to samples containing coarse powder. The segregation of powders was also evident, as the IC scores remained systematically larger even after the multiplicative effect was standardized by normalizing the mixture spectra prior to BSS.

The soft spectral model utilized in EMSC and OPLEC is heuristic and simplified. The model was revised in [56, 57] and [29], where the smooth wavelength dependent terms were modified to correspond better with the theories of Rayleigh and Mie scattering of light, respectively. In [58], the measured spectra were orthogonalized to the reduced scattering coefficient spectra which were in turn estimated using the Monte Carlo method. Closer exploration of the physics-based methods in the context of pharmaceutical powders could be the field of future research. Monte Carlo simulations combined with inversion calculation can be utilized in the estimation of the absorption and scattering coefficients of solid materials [14]. The application of BSS methods in the resolution of pure analyte absorption coefficient spectra from those estimated from powder mixtures could also be studied in the future.

Bibliography

[1] H. W. Siesler, Y. Ozaki, S. Kawata, and H. M. Heise, eds., *Near-Infrared Spectroscopy*. Wiley-VCH Verlag GmbH, 2002.

[2] J. Workman, "Review of process and non-invasive near-infrared and infrared spectroscopy: 1993–1999," *Applied Spectroscopy Reviews*, vol. 34, no. 1, pp. 1–89, 1999.

[3] P. Gemperline, ed., *Practical Guide to Chemometrics*. CRC Press, 2006.

[4] H. Martens and T. Næs, *Multivariate Calibration*. John Wiley & Sons Ltd., 1989.

[5] H. Martens and E. Stark, "Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy," *J. Pharm. Biomed. Anal.*, vol. 9, no. 8, pp. 625–35, 1991.

[6] H. Martens, J. P. Nielsen, and S. B. Engelsen, "Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures," *Anal. Chem.*, vol. 75, no. 3, pp. 394–404, 2003.

[7] Z. Chen, J. Morris, and E. Martin, "Extracting chemical information from spectral data with multiplicative light scattering effects by optical path-length estimation and correction," *Anal. Chem.*, vol. 78, no. 22, pp. 7674–7681, 2006.

[8] G. Wang, Q. Ding, and Z. Hou, "Independent component analysis and its applications in signal processing for analytical chemistry," *Trends in Analytical Chemistry*, vol. 27, no. 4, pp. 368–376, 2008.

[9] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[10] G. Reich, "Near-infrared spectroscopy and imaging: basic principles and pharmaceutical applications," *Advanced drug delivery reviews*, vol. 57, no. 8, pp. 1109–1143, 2005.

[11] H. G. Brittain, ed., *Spectroscopy of Pharmaceuticals Solids*. Taylor & Francis Group, 2006.

[12] C. Pasquini, "Near infrared spectroscopy: fundamentals, practical aspects and analytical applications," *J. Braz. Chem. Soc.*, vol. 14, no. 2, pp. 198–219, 2003.

[13] S. Chandrasekhar, *Radiative Transfer*. Courier Dover Publications, 1960.

[14] I. Yaroslavsky, A. Yaroslavsky, T. Goldbach, and H. Schwarzmaier, "Inverse hybrid technique for determining the optical properties of turbid media from integrating-sphere measurements," *Applied Optics*, vol. 35, no. 34, pp. 6797–6809, 1996.

[15] T. J. Farrell, M. S. Patterson, and B. Wilson, "A diffusion theory model of spatially resolved, steady-state diffuse reflectance for the noninvasive determination of tissue optical properties in vivo," *Medical Physics*, vol. 19, pp. 879–888, 1992.

[16] C. Bohren and D. Huffman, *Absorption and Scattering of Light by Small Particles*. J. Wiley & Sons, New York, 1983.

[17] M. Pasikatan, J. Steele, C. Spillman, and E. Haque, "Near infrared reflectance spectroscopy for online particle size analysis of powders and ground materials," *J. Near Infrared Spectrosc*, vol. 9, no. 3, pp. 153–164, 2001.

[18] P. Frake, C. Luscombe, D. Rudd, I. Gill, J. Waterhouse, and U. Jayasooriya, "Near-infrared mass median particle size determination of lactose monohydrate, evaluating several chemometric approaches," *The Analyst*, vol. 123, no. 10, pp. 2043–2046, 1998.

[19] A. O'Neil, R. Jee, and A. Moffat, "Measurement of the cumulative particle size distribution of microcrystalline cellulose using near infrared reflectance spectroscopy," *The Analyst*, vol. 124, no. 1, pp. 33–36, 1999.

[20] A. O'Neil, R. Jee, and A. Moffat, "Measurement of the percentage volume particle size distribution of powdered microcrystalline cellulose using reflectance near-infrared spectroscopy," *The Analyst*, vol. 128, no. 11, pp. 1326–1330, 2003.

[21] O. Berntsson, L. Danielsson, B. Lagerholm, and S. Folestad, "Quantitative in-line monitoring of powder blending by near infrared reflection spectroscopy," *Powder Technology*, vol. 123, no. 2, pp. 185–193, 2002.

[22] F. Muzzio, P. Robinson, C. Wightman, and D. Brone, "Sampling practices in powder blending," *International journal of pharmaceutics*, vol. 155, no. 2, pp. 153–178, 1997.

[23] O. Berntsson, T. Burger, S. Folestad, L. Danielsson, J. Kuhn, and J. Fricke, "Effective sample size in diffuse reflectance near-IR spectrometry," *Anal. Chem*, vol. 71, no. 3, pp. 617–623, 1999.

[24] P. Hopke, "The evolution of chemometrics," *Analytica Chimica Acta*, vol. 500, no. 1-2, pp. 365–377, 2003.

[25] L. Mutihac and R. Mutihac, "Mining in chemometrics," *Analytica Chimica Acta*, vol. 612, no. 1, pp. 1–18, 2008.

[26] B. Nadler and R. Coifman, "Partial least squares, Beer's law and the net analyte signal: statistical modeling and analysis," *Journal of Chemometrics*, vol. 19, no. 1, pp. 45–54, 2005.

[27] Z. Chen and J. Morris, "Improving the linearity of spectroscopic data subjected to fluctuations in external variables by the extended loading space standardization," *The Analyst*, vol. 133, no. 7, pp. 914–922, 2008.

[28] P. Geladi, D. MacDougall, and H. Martens, "Linearization and scatter-correction for near-infrared reflectance spectra of meat," *Applied Spectroscopy*, vol. 39, no. 3, pp. 491–500, 1985.

[29] A. Kohler, C. Kirschner, A. Oust, and H. Martens, "Extended multiplicative signal correction as a tool for separation and characterization of physical and chemical information in Fourier transform infrared microscopy images of cryo-sections of beef loin," *Applied Spectroscopy*, vol. 59, no. 6, pp. 707–716, 2005.

[30] H. Martens, S. Bruun, I. Adt, G. Sockalingum, and A. Kohler, "Pre-processing in biochemometrics: correction for path-length and temperature effects of water in FTIR bio-spectroscopy by EMSC," *Journal of Chemometrics*, vol. 20, no. 8–10, pp. 402–417, 2006.

[31] A. Kohler, J. Sulé-Suso, G. Sockalingum, M. Tobin, F. Bahrami, Y. Yang, J. Pijanka, P. Dumas, M. Cotte, D. van Pittius, G. Parkes, and H. Martens, "Estimating and correcting Mie scattering in synchrotron-based microscopic Fourier transform infrared spectra by extended multiplicative signal correction," *Applied Spectroscopy*, vol. 62, no. 3, pp. 259–266, 2008.

[32] The documentation in the EMSC toolbox for MATLAB by Harald Martens. *http://www.models.kvl.dk/source/EMSCtoolbox/index.asp*, 10.10.2008.

[33] S. Kirkpatrick, C. Gelatt, and M. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[34] L. Ingber, "Simulated annealing: practice versus theory," *Mathematical and computer modelling*, vol. 18, no. 11, pp. 29–58, 1993.

[35] E. Visser and T. Lee, "An information-theoretic methodology for the resolution of pure component spectra without prior information using spectroscopic measurements," *Chemometrics and Intelligent Laboratory Systems*, vol. 70, no. 2, pp. 147–155, 2004.

[36] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[37] The FastICA package for MATLAB. *http://www.cis.hut.fi/projects/ica/fastica/index.shtml*, 28.4.2009.

[38] J. Särelä and H. Valpola, "Denoising source separation," *Journal of Machine Learning Research*, vol. 6, pp. 233–272, 2005.

[39] G. Wang, W. Cai, and X. Shao, "A primary study on resolution of overlapping GC-MS signal using mean-field approach independent component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 82, no. 1-2, pp. 137–144, 2006.

[40] R. Bro, E. Acar, and T. Kolda, "Resolving the sign ambiguity in the singular value decomposition," *Journal of Chemometrics*, vol. 22, no. 2, pp. 135–140, 2008.

[41] J. Chen and X. Wang, "A new approach to near-infrared spectral data analysis using independent component analysis," *Journal of Chemical Information and Computer Sciences*, vol. 41, no. 4, pp. 992–1001, 2001.

[42] N. Pasadakis and A. Kardamakis, "Identifying constituents in commercial gasoline using Fourier transform-infrared spectroscopy and independent component analysis," *Analytica Chimica Acta*, vol. 578, no. 2, pp. 250–255, 2006.

[43] S. Astakhov, H. Stogbauer, A. Kraskov, and P. Grassberger, "Spectral mixture decomposition by least dependent component analysis," *Arxiv preprint physics/0412029*, 2004.

[44] J. Gómez Martín, P. Spietz, J. Orphal, and J. Burrows, "Principal and independent components analysis of overlapping spectra in the context of multichannel time-resolved absorption spectroscopy," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 60, no. 11, pp. 2673–2693, 2004.

[45] T. Lipsanen, *Process Analytical Technology Approach on Fluid Bed Granulation and Drying – Identifying Critical Relationships and Constructing the Design Space*. PhD thesis, University of Helsinki, 2008.

[46] P. Hede, *Fluid Bed Particle Processing*. Ventus Publishing Copenhagen, 2006.

[47] J. Rantanen, S. Lehtola, P. Ramet, J. Mannermaa, and J. Yliruusi, "On-line monitoring of moisture content in an instrumented fluidized bed granulator with a multi-channel NIR moisture sensor," *Powder Technology*, vol. 99, no. 2, pp. 163–170, 1998.

[48] F. Nieuwmeyer, M. Damen, A. Gerich, F. Rusmini, K. van der Voort Maarschalk, and H. Vromans, "Granule characterization during fluid bed drying by development of a near infrared method to determine water content and median granule size," *Pharmaceutical Research*, vol. 24, no. 10, pp. 1854–1861, 2007.

[49] A. Heikkilä, "Multipoint-NIR -measurements in pharmaceutical powder applications," Master's thesis, University of Oulu, 2008.

[50] Specim Ltd., *Spectral camera SWIR*, Data sheet, 2008. *http://www.specim.fi/media/pdf/specam-datasheets/nir-specam-ver4-08.pdf*, 20.4.2009.

[51] Specim Ltd., *ImSpector NIR and SWIR*, Data sheet, 2008. *http://www.specim.fi/media/pdf/imspector-datasheets/nir-swir-imspectors-ver1-2007.pdf*, 20.4.2009.

[52] M. Aikio, *Hyperspectral prism-grating-prism spectrograph.* PhD thesis, VTT, Technical research centre of Finland, 2001.

[53] NIST/SEMATECH, *e-Handbook of Statistical Methods*, 2006. *http://www.itl.nist.gov/div898/handbook/*, 20.4.2009.

[54] Anneal.m – Implementation of simulated annealing by Joachim Van de Kerckhove. *http://pages.stern.nyu.edu/∼acollard/anneal.m*, 28.4.2009.

[55] The DSS package for MATLAB. *http://www.cis.hut.fi/projects/dss/package/*, 28.4.2009.

[56] S. N. Thennadil and E. B. Martin, "Empirical preprocessing methods and their impact on NIR calibrations: a simulation study," *Journal of Chemometrics*, vol. 19, no. 2, pp. 77–89, 2005.

[57] S. N. Thennadil, H. Martens, and A. Kohler, "Physics-based multiplicative scatter correction approaches for improving the performance of calibration models," *Applied Spectroscopy*, vol. 60, no. 3, pp. 315–321, 2006.

[58] Z. Shi and C. Anderson, "Scattering orthogonalization of near-infrared spectra for analysis of pharmaceutical tablets," *Analytical Chemistry*, vol. 81, no. 4, pp. 1389–1396, 2009.