UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

Prediction of milk quality parameters using vibrational spectroscopy and chemometrics

– opportunities and challenges in milk phenotyping

PhD thesis 2016 · Carl Emil Aae Eskildsen

Prediction of milk quality parameters using vibrational spectroscopy and chemometrics

– opportunities and challenges in milk phenotyping

PhD thesis 2016

Carl Emil Aae Eskildsen

Department of Food Science

Faculty of Science

University of Copenhagen

Title

Prediction of milk quality parameters using vibrational spectroscopy and chemometrics - opportunities and challenges in milk phenotyping

Submission

March 11th 2016

Defense

May 27th 2016

Supervisors

Assoc. Prof. Thomas Skov, PhD

Department of Food Science, Faculty of Science

University of Copenhagen, Denmark

Assist. Prof. Nina Aagaard Poulsen. PhD

Department of Food Science

Aarhus University, Denmark

Opponents

Assoc. Prof. Åsmund Rinnan, PhD (chairman)

Department of Food Science, Faculty of Science

University of Copenhagen, Denmark

Assist. Prof. Henk Bovenhuis, PhD

Animal Breeding and Genetics Group

Wageningen University, The Netherlands

Research Programme Manager Steve Holroyd, PhD

Fonterra Co-operative Group Ltd., Palmerston North, New Zealand

Cover photo by Jens V. Aae Christensen

ISBN 978-87-93476-06-6

Printed by SL grafik, Frederiksberg C, Denmark (www.slgrafik.dk)

PREFACE

This thesis is submitted in order to obtain the PhD degree from the PhD school of Science, University of Copenhagen.

The working title of the PhD project is high-throughput FT-IR based solutions for economically important milk quality characteristics.

The PhD project is financed under the InSPIRe platform – Pillar II on Process Analytical Technology – with Aarhus University, FOSS Analytical A/S and Arla Foods amba as collaborators.

The work presented in this thesis has been carried out from October 2011 to February 2016 (in the 4+4 PhD program). The work was mainly done at section for Spectroscopy & Chemometrics, Department of Food Science, Faculty of Science, University of Copenhagen. Part of the work has been carried out at Department of Chemistry, University of Rome "La Sapienza", Team Chemometric Development, FOSS Analytical A/S and Department of Food Science, Aarhus University.

The PhD project was supervised by Assoc. Prof. Thomas Skov, PhD, Department of Food Science, University of Copenhagen and Assist. Prof. Nina Aagaard Poulsen, PhD, Department of Food Science, Aarhus University.

Carl Emil Aae Eskildsen
Frederiksberg, February 2016

ACKNOWLEDGMENTS

The Danish Council for Strategic Research together with the InSPIRe platform, Arla Foods amba and FOSS Analytical A/S are acknowledged for financial support.

Acknowledgment goes to my supervisors Assoc. Prof. Thomas Skov, PhD, Department of Food Science, University of Copenhagen and Assist. Prof. Nina Aagaard Poulsen, PhD, Department of Food Science, Aarhus University. Both deserve recognition for their supervision. I specially thank Thomas and Nina for always being positive and easy-going. Thomas has encouraged me to work on my own ideas, whereas Nina has been good at keeping focus on the overall project. The effort both my supervisors have put in the project is appreciated.

Furthermore, I would like to thank the project responsible, Prof. Lotte Bach Larsen, PhD, Department of Food Science, Aarhus University and the industrial partners, Research Scientist Mikka Stenholdt Hansen, PhD, Arla Strategic Innovation Centre, Arla Foods amba and Senior Chemometrician Per Waaben Hansen, PhD, Team Chemometric Development, FOSS Analytical A/S for good collaboration throughout the project period.

Special thanks must go to Assoc. Prof. Frans van den Berg, PhD and Prof. Søren Balling Engelsen, PhD, Department of Food Science, University of Copehagen for good discussions and critical feedback.

Federico Marini, PhD, Department of Chemistry, University of Rome “La Sapienza” is thanked for hosting me during my 7 month research stay in Rome.

All co-authors are thanked for their contributions to the scientific publications.

My father Jens V. Aae Christensen is thanked for providing the cover photo.

LIST OF PUBLICATIONS INCLUDED IN THE THESIS

PAPER I

C. E. Eskildsen, F. v. d. Berg and S. B. Engelsen, 2016: Vibrational Spectroscopy in Food Processing, *Accepted for publication,* In *Encyclopedia of Spectroscopy and Spectrometry,* Editors: J. Lindon, G. Tranter, D. Koppenaal, 3rd edition, Elsevier, Oxford.

PAPER II

N. A. Poulsen, **C. E. Eskildsen**, M. Akkerman, L. B. Johansen, M. S. Hansen, P.W. Hansen, T. Skov and L. B. Larsen, 2016: Predicting hydrolysis of whey protein by mid-infrared spectroscopy, *Accepted for publication, International Dairy Journal*.

PAPER III

C. E. Eskildsen, P.W. Hansen, T. Skov, F. Marini, L. Nørgaard, 2016: Evaluation of multivariate calibration models transferred between spectroscopic instruments: Applied to near infrared measurements of flour samples*, Journal of Near Infrared Spectroscopy*, 24 (2):151-156.

PAPER IV

C. E. Eskildsen, M. A. Rasmussen, S. B. Engelsen, L. B. Larsen, N. A. Poulsen and T. Skov, 2014: Quantification of individual fatty acids in bovine milk by infrared spectroscopy and chemometrics: Understanding predictions of highly collinear reference variables, *Journal of Dairy Science*, 97:7940-7951.

PAPER V

C. E. Eskildsen, T. Skov, M. S. Hansen, L. B. Larsen and N. A. Poulsen, 2016: Quantification of bovine milk protein composition and coagulation properties using infrared spectroscopy and chemometrics: A result of collinearity among reference variables, *In review, Journal of Dairy Science.*

ADDITIONAL PUBLICATIONS NOT INCLUDED IN THE THESIS

PAPER VI

J. Stokholm, S. Schjørring, **C. E. Eskildsen**, L. Pedersen, A. L. Bischoff, N. Følsgaard, C. G. Carson, B. L. K. Chawes, K. Bønnelykke, A. Mølgaard, B. Jacobsson, K. A. Krogfelt and H. Bisgaard, 2014: Antibiotic use during pregnancy alters the commensal vaginal microbiota, *Clinical Microbiology and Infection*, 20:629-635.

PAPER VII

M. Crafack, H. Keul, **C. E. Eskildsen**, M. A. Petersen, S. Saerens, A. Blennow, M. Skovmand-Larsen, J. H. Swiegers, G. B. Petersen, H. Heimdal and D. S. Nielsen, 2014: Impact of starter cultures and fermentation techniques on volatile aroma and sensory profile of chocolate, *Food Research International*, 63:306-316.

PAPER VIII

K. Kamatara, D. Mpairwe, M. Christensen, **C. E. Eskildsen**, D. Mutetikka, J. Muyonga, D. Mushi, S. Omagor, Z. Nantongo and J. Madsen, 2014: Influence of age and method of carcass suspension on meat quality attributes of pure breed Ankole bulls, *Livestock Science*, 169:175-179.

PAPER IX

D. T. Berhe, **C. E. Eskildsen**, R. Lametsch, M. S. Hviid, F. v. d. Berg and S. B. Engelsen, 2016: Prediction of total fatty acid parameters and individual fatty acids in pork backfat using Raman spectroscopy and Chemometrics: Understanding the cage of covariance between highly correlated fat parameters, *Meat Science*, 111:18-25.

PAPER X

N. A. Poulsen, J. Lassen, **C. E. Eskildsen**, A. J. Buitenhuis, U. Sundekilde and M. K. Larsen, 2016: Variation in methane emission and importance of covariance to major milk components on methane prediction models; use of infrared, milk fatty acids and milk metabolites, *In review, Journal of Dairy Science*.

ABBREVIATIONS & NOTATION

Abbreviations

α-LA	α-Lactalbumin
β-CN	β-Casein
β-LG	β-Lactoglobulin
CFR	Curd Friming Rate
CLA	Conjugated Linoleic Acid
CN	Casein
FA	Fatty Acid
FDA	United States Food and Drug Administration
FT	Fourier Transform
FT-IR	Fourier Transform Infrared
IR	Infrared
κ-CN	κ-casein
κ-CN G	Glycosylated κ-casein
LV	Latent Variable
NIPALS	Non-linear Iterative Partial Least Squares
NIR	Near-Infrared
PAT	Process Analytical Technology
PCA	Principal Component Analysis
PLS	Partial Least Squares
RCT	Rennet Coagulation Time
RMSE	Root Mean Squared Error
RMSECV	Root Mean Squared Error of Cross-Validation
RMSEP	Root Mean Squared Error of Prediction
WPC	Whey Protein Concentrates
WPH	Whey Protein Hydrolysates
WPI	Whey Protein Isolates

Algebra notation

$\boldsymbol{X}$	In uppercase, italics and bold; matrix
$\boldsymbol{x}$	In lowercase, italics and bold; vector
$\boldsymbol{X}^T, \boldsymbol{x}^T$	Transpose of a matrix or vector
x	In lowercase and italics; scalar

ABSTRACT

Vibrational spectroscopic techniques are widely used throughout all stages of food production. The analysis of raw materials, *real-time* process control, and end-product quality evaluation are all crucial steps in food production. In order to increase production throughput there is a *need for speed* when collecting information from the different processing steps. Hence, conventional methods from analytical chemistry (like Kjeldahl digestion for protein determination) are not compatible with modern production methods.

The aim of this thesis is to show how infrared spectroscopy may and may not be used by dairies and related industries. The focus is specially on possibilities and limitations in applying Fourier transform infrared spectroscopic measurements for detailed milk composition to be used in, for example, breeding programs. Previous studies reported successful predictions of individual fatty acids, protein fractions and coagulation properties from Fourier transform infrared measurements. This thesis shows how such predictions are trapped in a *cage of covariance* with major milk constituents like total fat and protein content. The prediction models for detailed milk composition are not based on causal relationships and this may seriously compromise calibration robustness. It is not recommended to implement indirect models for detailed milk composition in milk recording or breeding programs as such model are providing information on, for example, total protein rather than the specific protein fractions.

If Fourier transform infrared based models on detailed milk composition are to be implemented in, for example, breeding programs it is recommended to decompose, for example, the individual fatty acids into functional groups, such as methyl, methylene, olefinic and carboxylic groups. The average proportions of these groups may be reliable estimated from Fourier transform infrared measurements in contrast to concentrations of individual fatty acids.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The food industry faces strong demands from consumers and regulatory bodies to produce safe and consistently high quality products. The main objective for the food industry is to increase productivity and at the same time meet consumer demands. Hence, a desired and consistent end-product quality should be reached in a cost effective, safe, traceable, environmental and sustainable responsible way. This can be achieved by e.g. assuring that the raw materials meet certain quality demands, controlling the process by *real-time* continuous measurements of core parameters, rapid final product quality evaluation, and assuring key quality attributes during packaging. Collecting the right information sufficiently fast using conventional analytical chemistry is a major challenge in the food industry. For that reason, vibrational spectroscopic techniques are widely used during all stages of food production (**PAPER I**).

At the dairy, bovine milk is processed into a number of dairy products like cheese, butter, cream, consumer milk, yoghurt and ice cream. Furthermore, bovine milk plays an important part in the production of ingredients to other food products or as nutritional supplements in, for example, infant formula and protein powders for body building or hospitalized patients. Dairy products are important components of western diets and therefore, bovine milk is an important raw material. Bovine milk mainly consists of lipids, proteins, carbohydrates and the main constituent, water.

For decades, milk pricing has focused on fat and protein content. However, in recent years more attention has been paid toward the important relations between detailed milk composition, nutritional value and technological properties (Bovenhuis et al., 2013).

From a human health point of view, the fatty acid (FA) composition is important. While some unsaturated FA in milk, for example, conjugated linoleic acids (CLA) and omega-3 FA are associated with positive health effects, it is generally recommended to minimize the intake of saturated FA (Simopoulos, 1991; German et al., 2009; Givens, 2010).

The FA composition is important for fat based dairy products like butter. Butter quality is defined by characteristics such as spreadability, which in turn depends on the FA composition (Couvreur et al., 2006). The melting point of FA decreases with reduced chain length and increased degree of unsaturation (Pratt and Cornely, 2004a). Hence, the overall FA composition determines the softness of butter. Moreover, the FA composition also affects shelf life of dairy products, where increased unsaturation reduces shelf life due to oxidation.

Bovine milk protein composition is also highly important, for example, during cheese making. Here the structure and behavior of the casein (CN) micelles and especially the presence of κ-CN and the level of glycosylated κ-CN (κ-CN G) is important for the milk coagulation process and thereby cheese yield (Frederiksen et al., 2011; Jensen et al., 2012; Bonfatti et al., 2014).

Furthermore, the dairy industry is currently making huge profits on products related to nutritional supplements. The composition of, for example, infant formula should ideally match the composition of human breast milk. However, the composition of bovine milk and human milk is not identical. For example, both the FA profile and the protein composition differ. Human milk fat contains in general more unsaturated FA than bovine milk fat (Jensen et al., 1990). For proteins, the ratio of whey proteins to CN proteins differs (60:40 in human milk and 20:80 in bovine milk). For this reason, whey protein concentrate (WPC) is added to infant formula. Nevertheless, β-Lactoglobulin (β-LG) is a major protein in bovine whey while absent in human milk. On the other hand, α-Lactalbumin (α-LA) is a major protein in human milk, but not in bovine milk. This results in differences in the composition of essential amino acids in human milk and infant formula. Moreover, α-LA is believed to, for example, facilitate infant absorption of essential minerals. Therefore, it is recommended to increase the levels of α-LA and decrease the levels of β-LG when making bovine milk based infant formula (Lien, 2003; Lönnerdal and Lien, 2003). But it is not necessarily easy to achieve this and it would be desirable if the bovine milk, used to produce infant formula, already from the cow contained higher levels of, for example, α-LA and lower levels of β-LG.

The suitability of milk for specific dairy products could be improved by changing milk fat and protein composition. Nevertheless, today milk from a large number of cows and multiple herds are pooled together and mixed in storage tanks at the dairies. Variation in milk quality from individual cows or herds is evened out and the dairy products are made from a uniform starting material (exposed to some seasonal variation). However, milk with a specific composition suitable for a particular product, could be collected separately if a simple and reliable analytical tool was available for characterizing the FA composition and protein profile. This would open possibilities for more efficient production at the dairy and possibly enhance product quality. For example, milk with an ideal composition for cheese making could be identified and used for this purpose and milk with an ideal composition for making infant formula could be used for that purpose.

Bovine milk protein composition and to some extent fat composition is affected by genetic factors (Schopen et al., 2009; Krag et al., 2013). This suggests that detailed milk composition can be changed by selective breeding. Therefore, efficient phenotyping tools are key elements in breeding programs of livestock animals. Hence, there is a need for accurate, cheap, and high-throughput predictive methods for detailed milk composition.

A high-throughput method for detailed milk composition would be valuable in breeding programs. However, such a method could also be used in, for example, the payment system to the farmers. Moreover, in a milk *fit-for-purpose* perspective, it is of course crucial for the dairies to document

that the milk meets the specific requirements for a particular product. Hence, such a high-throughput method could be used for screening milk upon arrival at the dairy.

Process Analytical Technology (PAT) is an important part of the Quality by Design paradigm (Food and Drug Administration, 2004; van den Berg et al., 2013), and a high-throughput PAT method for quantification of detailed milk composition could facilitate better product quality and economic gains. The dairy industries are already using PAT-solutions to some extent. For example, high-throughput Fourier Transform Infrared (FT-IR) based technologies are routinely applied in the dairy industry during milk standardization (fat, protein and lactose). The Infrared spectrum is highly specific and almost all chemical substances show characteristic patterns of absorption with infrared light (Coates, 2010).

Milk recording agencies and dairies routinely use FT-IR for milk analyses. The FT-IR measurements are cheap, high-throughput and information on chemical substances can accurately be retrieved. Therefore, FT-IR is attractive for providing information on detailed milk composition, including FA profile and protein composition. Traditionally, detailed milk composition is only characterized by time-consuming analytical methods such as gas chromatography or liquid chromatography.

A number of studies have already investigated the potential of using FT-IR measurements and multivariate data analysis for predicting the FA profile and the protein composition of milk. In general, FT-IR seems promising for measuring FA profile and protein composition in milk, as also outlined in the review on FT-IR spectroscopy as a phenotyping tool by De Marchi et al. (2014). Furthermore, some, from a vibrational spectroscopy point of view, more or less unorthodox parameters like cow body energy status (McParland et al., 2011), cow methane emission (Dehareng et al., 2012) and milk coagulation properties have been predicted from FT-IR measurements of milk. It would be surprising if these showed a direct absorption signal in IR measurements. However, these attempts are based on, for example, relationships between milk FA profile and methane emission (Chilliard et al., 2010) and milk protein composition and coagulation properties (Bonfatti et al., 2014).

Nevertheless, different FA and different protein fractions are, respectively, composed of more or less the same chemical groups. Hence, they are, respectively, likely to provide very similar signals in IR measurements. If calibration models for a specific FA is using spectral features of other FA (or more likely total fat content, i.e. the sum of all FA), then the predicted FA values of future samples will depend on, for example, total fat content. An increase of total fat will lead to an increase of the predicted FA concentration, even if the increase in total fat is caused by other FA than the one of interest. Such an indirect model cannot detect a change in the fatty acid independent of the total fat content. Detecting changes in the fatty acid profile independent of the total fat is what breeders and the dairy industry will need if a high-throughput method should be used for breeding programs and milk differentiation. Nevertheless, no previous studies paid much attention to this problem.

The aim of this thesis is to show how process analytical tools may and may not be used by dairies and related industries. The focus will specially be on possibilities and limitations in applying FT-IR for detailed milk composition.

CHAPTER 2

PROCESS ANALYTICAL TECHNOLOGY

The US Food and Drug Administration (FDA) officially introduced PAT in a guidance to the industry in September 2004 (Food and Drug Administration, 2004). The FDA defined PAT as,

> *"A system for designing, analyzing, and controlling manufacturing through timely measurements (i.e. during processing) of critical quality and performance attributes of raw and in-process materials and processes with the goal of ensuring final product quality."*

Thereby, FDA encouraged manufacturers to ensure product quality by process design instead of post-production quality testing. Implementation of PAT in a manufacturing process aims at controlling the manufacturing process and thereby reducing risks related to product quality and regulatory concerns and at the same time improving efficiency by, for example, reducing production cycle times.

Process understanding is the key element when implementing PAT. In order to control the manufacturing process, critical sources of variability must be identified and explained. The aim is to handle this variability by adapting the process and thereby ensuring high consistent product quality. The most important components in facilitating process understanding are experimental design, process analyzers, multivariate data analysis and process control tools. Furthermore, *theory of sampling* cannot be ignored during successful implementation of PAT at a production site.

A vital part of the PAT paradigm is moving the measurement technologies from the laboratory to the production site. This serves the purpose of obtaining rapid in-line, on-line and at-line measurements of the production operations (Food and Drug Administration, 2004; van den Berg et al., 2013). Vibrational spectroscopic techniques are often used in PAT solutions, as they are fast, cheap, informative and non-destructive measurement technologies. Coupling spectroscopic measurements with multivariate data analysis (chemometrics) is highly powerfull in process understanding and control.

Even though the FDA guidance document clearly was intended the pharmaceutical industry, the pharmaceutical industry seems hesitant (compared with the food industry) to introduce PAT in the manufacturing. There may be several reasons for this. A pharmaceutical production is strictly regulated. Hence, PAT implementation in an existing production may be an expensive task. Furthermore, drug development (from discovery to clinical trials) takes round 10 years and is highly ex-

pensive. The total costs of a pharmaceutical products reflects only to a minor extend the costs of raw materials and production. Therefore, the economic gain of implementing PAT in the production is limited. Last, a pharmaceutical production may be less exposed to variations in raw materials. Hence, consistent product quality can easier be achieved by recipe-driven production.

In the food industry, on the other hand, PAT-like solutions have existed years before the publication of the FDA guidance. Food products often consist of highly complex mixtures of fats, carbohydrates and proteins and raw materials face a high degree of natural variation. Therefore, consistent end-product quality is difficult to archive by recipe-driven productions. Furthermore, the price of food and foodstuffs reflects to a larger degree the costs of raw materials and production. Therefore, it is more attractive for the food industry to optimize the production (including use of raw materials). Moreover, implementation of PAT in the food industry may be easier compared with the pharmaceutical industry as the food industry is less regulated by authorities (van den Berg et al., 2013).

PAPER I concerns how the food industry is using vibrational spectroscopy and chemometrics in PAT like solutions. Figure 2.1 is a schematic representation of the different stages in a food production. Sorting raw materials according to certain quality demands may be a cost effective way in increasing productivity (Figure 2.1, Process inputs) and securing high and consistent end-product quality. Sorting enables that the best raw materials are used for high-value products and it eases control of down-stream processing steps.

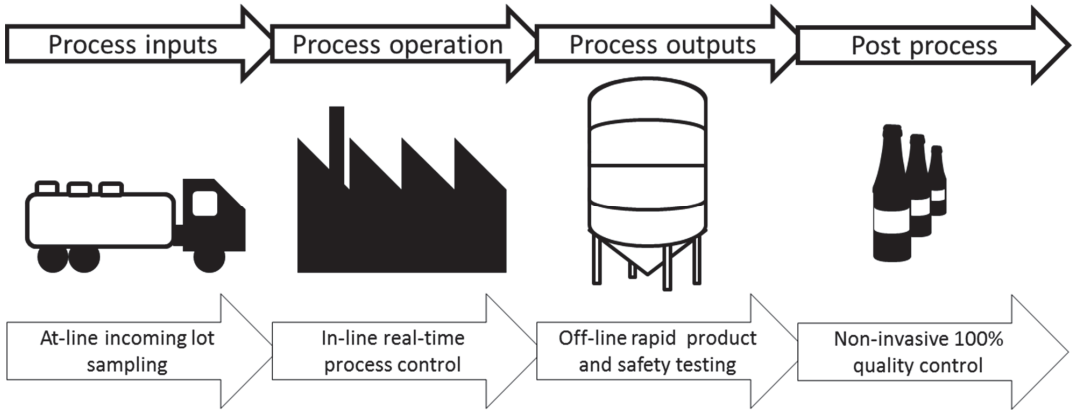


***Figure 2.1**. Schematic representation of the different stages and the analytical questions involved of a food production. Adapted from **PAPER I**.*

Even though proper sorting is done at the gate, process monitoring and control (Figure 2.1, Process operation) is often essential in increasing production throughput and reaching consistent product quality. The dairy industry has really recognized the benefits of continuous process monitoring and control. Successful examples include FT-IR for on-line standardization of milk, which optimizes the used of fat, protein and lactose (Ellen and Tudos, 2003). Another example is process

monitoring, for optimized moisture content (increased yield), of butter production by in-line near-infrared (NIR) spectroscopy (Funahashi and Horiuchi, 2008).

Safety and quality of end-products are obviously essential parameters (Figure 2.1, Process output). Due to the limited shelf life of most food products, extensive product testing by classical analysis is often not desirable. Vibrational spectroscopic methods can provide information for these important parameters, typically without any sample preparation or use of reagents. High speed spectroscopic measurement can be used in 100% quality assurance of, for example, every bottle coming by on a high speed filling line (Figure 2.1, Post process).

The benefits of successfully implementing PAT in a production are many. In summary, PAT offers optimized utilization of raw materials, reduced production time, cheap and swift measurements instead of costly and time consuming laboratory testing and it leads to less variation in final product quality (Food and Drug Administration, 2004; van den Berg et al., 2013).

2.1 Vibrational Spectroscopy

Vibrational spectroscopic techniques like infrared (IR), NIR and Raman spectroscopy are frequently used for process analytical measurements. Whereas Raman and IR measurements provide information on fundamental molecular vibrations, NIR measurements provides information on overtones and combination bands of the fundamental vibrations. There are pros and cons for each of the three spectroscopic methods and the decision of the method to use is based on factors like sample type, information required, costs and ease of implementation (Coates, 2010). This chapter concerns IR and NIR spectroscopy.

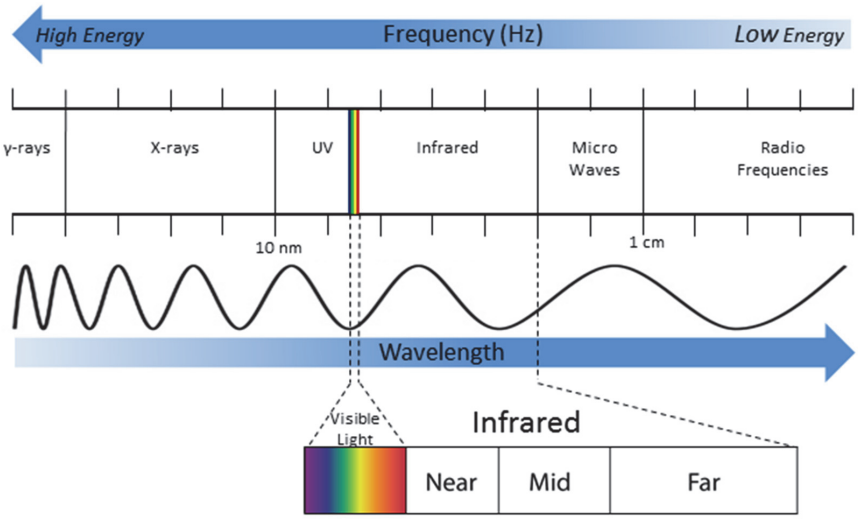


***Figure 2.2**. The electromagnetic spectrum. The infrared region is divided into three parts; the near-infreared (Near) region, the mid-infrared (Mid) region, and the far-infrared (Far) region. Modified from Pavia et al. (2001).*

Electromagnetic radiation is characterized by wavelength or frequency and can be understood as sinusoidal waves (or particles with sinusoidal wave-like behavior) travelling at the speed of light. The infrared part of the electromagnetic spectrum may be divided into the NIR region, 12,500-4,000 cm^{-1}, the mid-infrared region, 4,000-400 cm^{-1} and the far-infrared, 400-100 cm^{-1} (Figure 2.2; Dufour, 2009). The mid-infrared region will be referred to as the IR region and only the IR and the NIR regions will be considered in this thesis. For convenience IR spectra are reported in wavenumbers (units of cm^{-1}), and NIR spectra are reported in wavelengths (units of nm). Wavenumbers and wavelengths are related by Equation 2.1 (Pavia *et al.,* 2001; Bruice, 2007b; Dufour, 2009; Larkin, 2011a).

$$\bar{v} = \frac{10^7}{\lambda\,(nm)} \qquad \textbf{\textit{Equation 2.1}}$$

Where $\bar{v}$ is the given wavenumber and λ is the given wavelength.

Energy, frequency and wavelengths of electromagnetic radiation are related by Equation 2.2 (Pavia *et al.,* 2001; Dufour, 2009; Larkin, 2011a).

$$E = hv = \frac{hc}{\lambda} \qquad \textbf{\textit{Equation 2.2}}$$

Where E is energy, h is *Planck's* constant, v is the given frequency and c is the speed of light. From Equation 2.1 and Equation 2.2 it can be found that energy is directly proportional to frequency and wavenumbers but inversely proportional to wavelength.

2.1.1 Molecular Vibrations

At temperatures above the absolute zero all molecules exhibit vibrations. A chemical bond, connecting two atoms, can be viewed as a mass-less spring connecting two vibrating masses. There are two types of molecular vibrations, stretching and bending. The stretching and bending vibrations of CH_2 are shown in Figure 2.3.

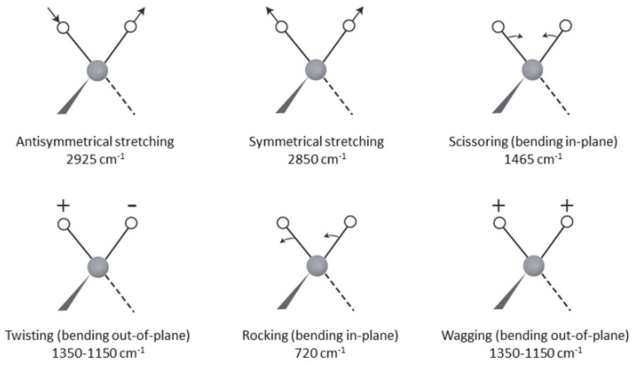


***Figure 2.3**. Stretching and bending vibrational modes for a CH_2 group. Illustration modified from Pavia et al., (2001).*

Molecular vibrations may absorb electromagnetic radiation. If the molecular vibrations express a change in the electrical dipole and this dipole is changing with the same frequency as the incoming radiation, then the molecular vibration will absorb the incoming radiation and the absorbed energy will increase the amplitude of the vibrational motions. Larger differences in electronegativity between atoms involved, results in better coupling of the changing electrical dipole of the bond and the sinusoidal changing electromagnetic field of the incoming radiation. Hence, larger differences in electronegativity between atoms will result in stronger absorption (Vidrine, 2000; Pavia et al., 2001; Coates, 2010).

Different vibrational modes of the same bond vibrate with different frequencies (Figure 2.3). Hence, similar information occurs in several regions throughout the absorption spectrum. The position of absorption bands can be estimated from the strength of the molecular bond and the masses of the atoms involved. The approximate wavenumber of the fundamental absorption band, $\bar{\nu}_0$ may be calculated from Equation 2.3 (derived from *Hooke's law*).

$$\bar{\nu}_0 = \frac{1}{2\pi c}\sqrt{\frac{K}{\mu}} \qquad \textbf{\textit{Equation 2.3}}$$

Where K is the force constant (stiffness) of chemical bond (spring) and μ is the reduced mass of the system given by Equation 2.4. Stronger molecular bonds will have larger K and stronger bonds will therefore absorb electromagnetic radiation with higher wavenumbers.

$$\mu = \frac{m_1 m_2}{m_1 + m_2} \qquad \textbf{\textit{Equation 2.4}}$$

Where m_1 and m_2 are the masses for the two bonded atoms. Hence, bonds involving lighter atoms will have smaller μ and will therefore absorb electromagnetic radiation with higher wavenumbers. The position of overtones can be found from Equation 2.5,

$$\bar{\nu}_n = n\bar{\nu}_0(1 - n\chi) \qquad \textbf{\textit{Equation 2.5}}$$

Where $\bar{\nu}_n$ is the wavenumber of the overtone corresponding to the vibrational quantum number, n and χ is the anharmonicity constant. Hence, the position of an overtone ($n > 1$) is an approximate multiple of the fundamental vibration ($n = 1$). The vibrational mode, hybridization, electron delocalization and hydrogen bonding all affect K (Equation 2.3) of a given molecular bond and thereby the position of the absorption band in the spectra.

Vibrational bending modes have lower K than stretching modes. Therefore, bending motions are found at lower wavenumbers like the scissoring (1,450 cm^{-1}), wagging (1,300 cm^{-1}), twisting (1,300 cm^{-1}) and rocking (720 cm^{-1}) modes of CH_2 groups.

The length and strength of, for example, C-H bonds are affected by hybridization of the carbon orbitals. The s orbital is closer to the nucleus than the p orbital. Therefore, increased hybridization

will decrease the strength (increase the length) of the C-H bond. The C-H stretching for sp carbon occurs at $\sim 3{,}300\ cm^{-1}$, for sp^2 carbon at $\sim 3{,}100\ cm^{-1}$ and for sp^3 carbon at $\sim 2{,}900\ cm^{-1}$ (Bruice, 2007a).

The strength of molecular bonds is also affected by electron delocalization. Consider the carbonyl group in Figure 2.4. Nitrogen is not very electronegative. Hence, if nitrogen takes the position of Z in Figure 2.4 (like an amide) the predominant effect will be electron donation. On the other hand, oxygen is very electronegative and if oxygen takes the position of Z in Figure 2.4 (like an ester), the predominant effect will be inductive electron withdrawal. Therefore, carbonyl groups in amides show more single bond character than carbonyl groups in esters. This reduces K for carbonyl in amides compared with carbonyl in esters and consequently the absorption band of the carbonyl group in an amide is found at lower wavenumbers (Pavia *et al.,* 2001; Bruice, 2007a; Bruice, 2007b).

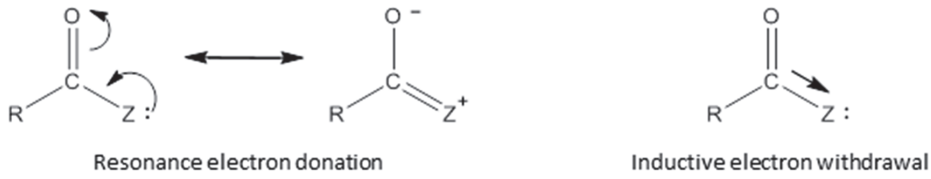


Figure 2.4*. Electron delocalization. Illustration adapted from Bruice (2007b).*

Figure 2.5 shows a hydrogen bond between two water molecules. Hydrogen bonds will decrease K of the covalent O-H bond. Therefore, molecular bonds that are affected by hydrogen bonds are found at lower wavenumbers. Hydrogen bonds vary in strength. This results in very broad absorption bands of hydrogen-bonded molecules (Bruice, 2007b).

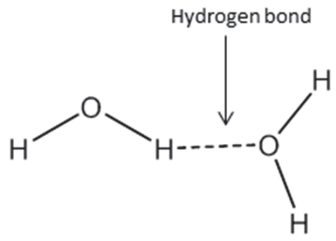


Figure 2.5*. Hydrogen bond (---) between two water molecules. Illustration adapted from Bruice (2007b).*

Different kind of information is found in specific parts of the absorption spectrum. Stretching vibrations involving hydrogen atoms (light atoms) occurs at higher wavenumbers (2,500-4,100 cm^{-1}),

for example anti-symmetric and symmetric sp^3 C-H stretching vibrations found at 2,960 ± 10 cm^{-1} and 2,870 ± 10 cm^{-1}, respectively. Stretching vibrations involving carbon, oxygen and nitrogen (middle-weight atoms) characteristically occupies the 800-1,200 cm^{-1} region like C-OH bonds of lactose found at 1,080 cm^{-1}. Double bonds and triple bonds occur in the 1,600-2,300 cm^{-1} region like carbonyl stretching vibrations of triglycerides found at 1,750 cm^{-1}. (Pavia et al., 2001; Bruice, 2007b; Larkin, 2011b).

2.1.2 Quantitative Spectroscopy

There are in general two different measurement methods for vibrational spectroscopy; Transmission (Figure 2.6a) and reflectance (Figure 2.6b). When light is absorbed by molecular vibrations, the irradiance of the beam of the light is decreased. The light with irradiance, P_0, strikes the sample and the irradiance of the beam emerging from the sample is P (Figure 2.6). If light is absorbed by the sample then $P \leq P_0$.

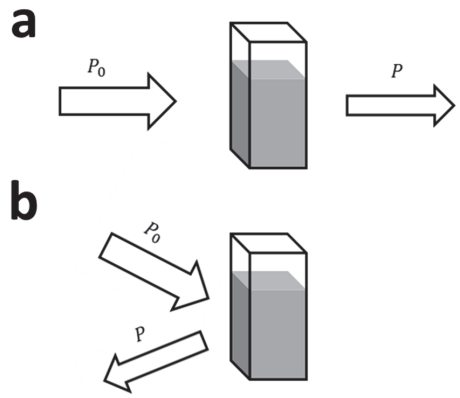


***Figure 2.6**. (a)Transmission of light through matter. (b) Reflection of light from matter. P_0 is irradiance of incoming beam. P is irradiance of the beam emerging from the sample.*

The absorbance, *A*, is defined as (Harris, 2010),

$$A = log\left(\frac{P_0}{P}\right) \qquad \textbf{Equation 2.6}$$

Beer's law (Equation 2.7) relates the absorption of light to the properties of the material through which the light is travelling (Harris, 2010).

$$A = \varepsilon l c \qquad \textbf{Equation 2.7}$$

Where *A* is absorbance, *c* is the concentration, ε is the molar absorptivity and l is the path length of the light in the sample. Both ε and l are constants and *A* is therefore directly proportional to *c*. This relationship is the heart of quantitative spectroscopy.

2.1.3 Interferometer Instruments

In **PAPER II**, **PAPER IV** and **PAPER V** MilkoScan FT2 (FOSS Analytical A/S, Hillerød, Denmark) was used for obtaining IR measurements. In **PAPER III** InfraXact (FOSS Analytical A/S, Hillerød, Denmark) and DS2500 (FOSS Analytical A/S, Hillerød, Denmark) was used for obtaining NIR measurements.

All the three instruments are based on the Fourier transform (FT) principle. The most important part of any FT instrument is the interferometer (Figure 2.7).

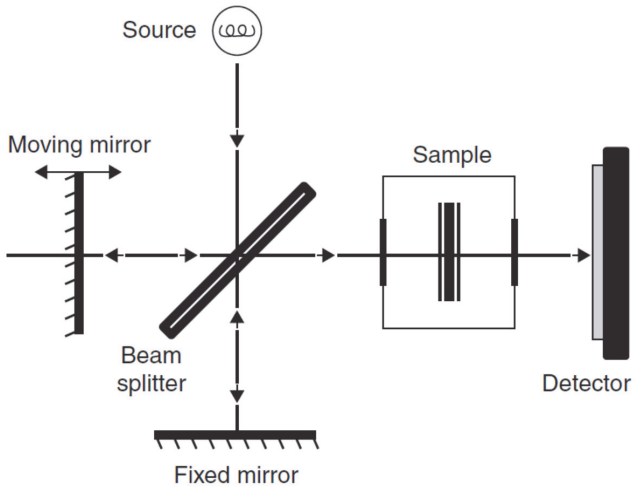


***Figure 2.7**. Schematic diagram of a Michelson interferometer. Illustration adopted from Subramanian & Rodriquez-Saona (2009).*

In a Michelson interferometer, a beam splitter divides the radiation into two equal beams. One beam is reflected towards a moving mirror and one beam is reflected towards a fixed mirror. Both beams are reflected back towards the beam splitter where they are recombined. However, the motion of the moving mirror introduces a continuous change in the optical path length between the two beams. When recombined the path length differences of the two beams causes (a changing) interference of the beams where some wavelength show constructive interference and other wavelength show destructive interference. Since the optical path difference is constantly changing, the various frequencies present in the beam are modulated at different rates. The combined beam is called the interferogram and is found in the time domain of Figure 2.8 (Pavia et al., 2001; Subramanian and Rodriguez-Saona, 2009). The generated interferogram is passed through the sample where wavelength specific parts are absorbed (Pavia et al., 2001). As mentioned, the interfereogram is found in the time domain. In order to obtain the frequency-domain spectrum from the interferogram, a Fourier transform (FT) is applied. The FT separates the individual frequencies from the interferogram (time domain) by fitting sinusoidal functions to the interferogram. Hence,

FT basically reverses the construction of the interferogram (Pavia et al., 2001). The individual frequencies are then recombined to the spectrum in the frequency domain (Figure 2.8).

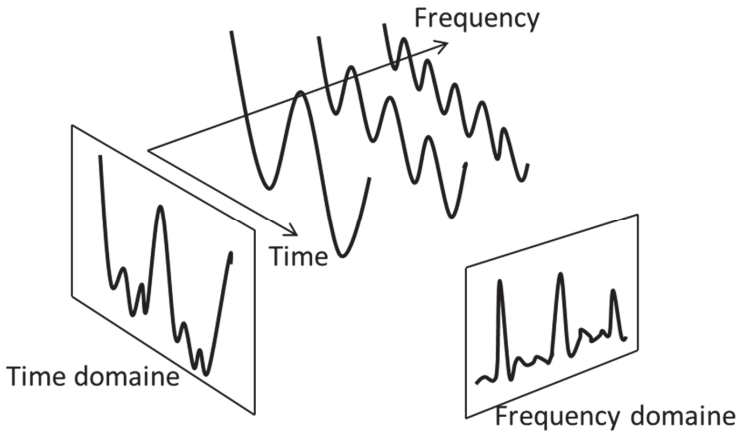


Figure 2.8*. Fourier transformations. The interferogram is found in the time domain. The interferogram is split into single frequency spectra and recombined in the frequency domain.*

The benefit of FT-IR instruments, compared to a dispersion instruments, is greater speed. Thereby more spectra can be collected and greater signal-to-noise ratio can be reached (Pavia et al., 2001). Speed and good signal-to-noise ratio are essential for applications within PAT.

2.2 Chemometrics

Spectroscopic measurements are very useful in PAT as absorption and concentration are linearly related through *Beer's law* (Equation 2.7). However, the absorption signal is not only affected by the chemical properties of the samples. Also the physical properties may affect the signal and this may corrupt *Beer's law*. Physical properties are often found as offset and slope differences. If the chemical properties are of interest, it may be necessary to remove offset and slope differences by preprocessing the spectroscopic measurements prior data analysis. A walkthrough of the most common preprocessing methods is done by Rinnan et al. (2009).

The matrix rank can be determined once the preprocessed spectroscopic measurements, $\boldsymbol{X}(n \times m)$ containing n samples and m measured variables (e.g. wavenumbers), are arranged into a data matrix. The number of linearly independent rows is equal to the number of linearly independent columns of $\boldsymbol{X}$ and this number determines the matrix rank.

$$r(\boldsymbol{X}) \leq min(n, m) \qquad \textbf{\textit{Equation 2.8}}$$

Where $r(\boldsymbol{X})$ is the rank of $\boldsymbol{X}$. The matrix is defined as full-rank if (Vandeginste *et al.,* 1998),

$$r(\boldsymbol{X}) = min(n, m) \qquad \textit{Equation 2.9}$$

Theoretically, spectroscopic measurements of multicomponent samples, $\boldsymbol{X}(n \times m)$ can be viewed as the outer product of the analyte concentration profiles, $\boldsymbol{C}(n \times q)$ containing n samples and q analytes, and the pure analyte signals, $\boldsymbol{S}(m \times q)$ containing m variables (Equation 2.10). Note that background and error terms are not taken into account.

$$\boldsymbol{X} = \boldsymbol{C}\boldsymbol{S}^T \qquad \textit{Equation 2.10}$$

This is also illustrated in Figure 2.9. In the case where all analytes give rise to unique spectral features, then $r(\boldsymbol{X})$ is determined by q (given that $q < n$ and $q < m$).

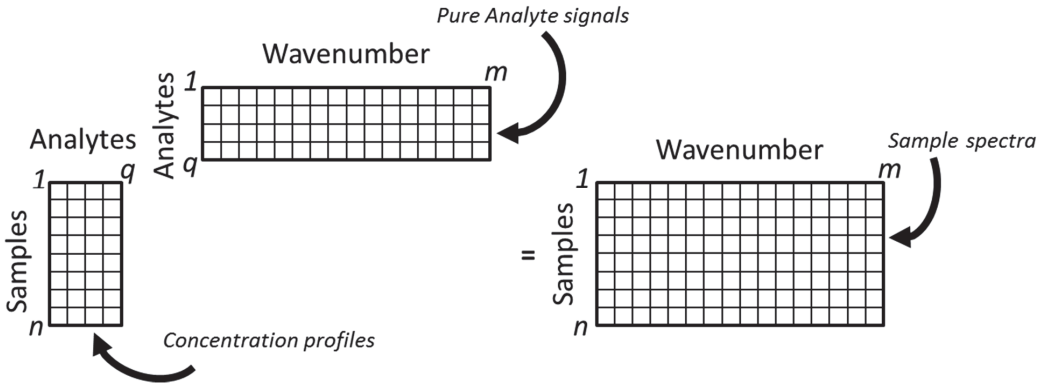


***Figure 2.9**. Multicomponent sample spectra. A function of the outer product of analyte concentration profiles and pure analyte signals. n = number of samples, q = number of analytes, m = number of variables (wavenumbers).*

If the concentration profiles are perfectly correlated or the pure analyte spectra are completely overlapping (correlated), then $r(\boldsymbol{X}) < q$. However, due to the presence of measurement noise in the spectra, the *mathematical rank* of spectroscopic measurements is often (close to) full and much higher than the *chemical rank*. Obviously, the *chemical rank* is the one of interest and in the following the term rank, $r(\boldsymbol{X})$ will denote the *chemical rank*. In quantitative analysis it is important to consider $r(\boldsymbol{X})$ as it determines the number of independent parameters (information) that can be extracted from $\boldsymbol{X}$ (Smilde et al., 2004).

2.2.1 Principal Component Analysis

Spectral measurements, $\boldsymbol{X}(n \times m)$ are fully described in a m-dimensional space. However, $\boldsymbol{X}$ consists of a systematic part and a noise part. The systematic part spans a R-dimensional (corresponding to $r(\boldsymbol{X})$) subspace of $\boldsymbol{X}$ and $\boldsymbol{X}$ can be explained in this low(er) dimensional subspace without loss of information. The R-dimensional subspace can be further decomposed into R rank one sys-

tems (Figure 2.10). This kind of matrix decomposition is the principle of Principal Component Analysis (PCA) and allows for easier interpretation and reduced influence of noise (Smilde et al., 2004).

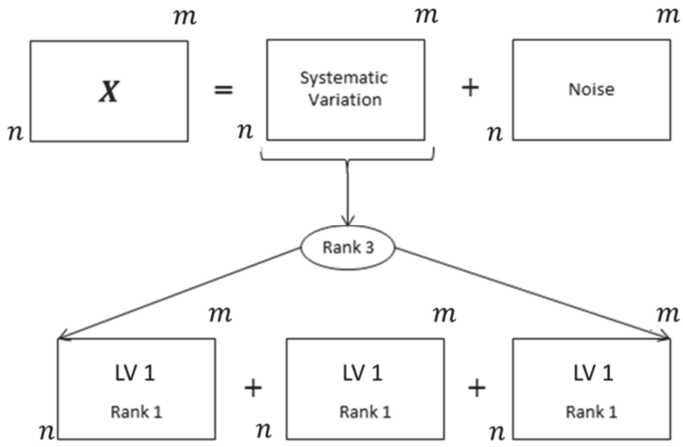


***Figure 2.10**. Decomposition of a matrix into a systematic and a noise part. The systematic part is further decomposed into rank one matrices. Illustration modified from Smilde et al., (2004).*

In PCA, $\boldsymbol{X}$ is decomposed into scores vectors, $\boldsymbol{T}(n \times R)$, loading vectors, $\boldsymbol{P}(n \times R)$ and a residuals matrix, $\boldsymbol{E}(n \times m)$, as shown in Equation 2.11.

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} \qquad \textit{Equation 2.11}$$

In PCA, $\boldsymbol{X}$ is projected onto the low(er) R-dimensional subspace spanned by orthogonal latent variables (LV). Each LV represent a rank one approximation of $\boldsymbol{X}$. The first LV is found to be the best rank one approximation of $\boldsymbol{X}$. The second LV is the best rank one approximation of $\boldsymbol{X}$ with the constraint of being orthogonal to the first LV. The third LV is the best rank one approximation of $\boldsymbol{X}$ with the constraint of being orthogonal to the first two LV, and so forth. The loading vectors are unit vectors describing the direction of each the LV in the m-dimensional space and the scores are the projections of the original data points onto the loading vectors.

Here, PCA is used for initial exploratory data analysis and most importantly estimation of $r(\boldsymbol{X})$. The *Non-linear Iterative Partial Least Squares* (NIPALS) algorithm for PCA is found in Box 1 (Wold et al., 1987; Smilde et al., 2004).

Box 1

NIPALS Algorithm for PCA

$\boldsymbol{X}(n \times m)$ is preprocessed and centered.

The algorithm is repeated for the number of latent variables $r = 1,2, \dots, R$.

1) Choose a starting guess for $\boldsymbol{t}_r(n \times 1)$ (e.g. first column of $\boldsymbol{X}$)

2) Solve $\boldsymbol{X} = \boldsymbol{t}_r \boldsymbol{p}_r{}^T$ with regards to $\boldsymbol{p}_r(m \times 1)$ and normalize $\boldsymbol{p}_r$

$$\boldsymbol{p}_r = \frac{\boldsymbol{X}^T \boldsymbol{t}_r}{\|\boldsymbol{X}^T \boldsymbol{t}_r\|}$$

3) Solve $\boldsymbol{X} = \boldsymbol{t}_r \boldsymbol{p}_r{}^T$ with regards to $\boldsymbol{t}_r$

$$\boldsymbol{t}_r = \boldsymbol{X}\boldsymbol{p}_r$$

5) Repeat 2) and 3) until convergence

6) Deflate $\boldsymbol{X}$ and start again from 1)

$$\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{t}\boldsymbol{p}^{\boldsymbol{T}}$$

2.2.2 Multivariate Calibration

The task of multivariate calibration is to find a predictive model that relates two vectorial spaces; the instrument response space and the concentration space. Here, the purpose of calibration is to model analyte concentrations, $\boldsymbol{y}(n \times 1)$ as linear combinations of absorption spectra $\boldsymbol{X}(n \times m)$, so analyte concentrations in future samples can be estimated based on the absorption spectra only.

The absorption spectrum of a given sample, $\boldsymbol{x}(m \times 1)$ is a vector represented in a m-dimensional space. Figure 2.11 shows this in a two-dimensional space, where the components of $\boldsymbol{x}$ in the base $(\boldsymbol{u}_1(2 \times 1), \boldsymbol{u}_2(2 \times 1))$ are $c_1(1 \times 1)$ and $c_2(1 \times 1)$, respectively.

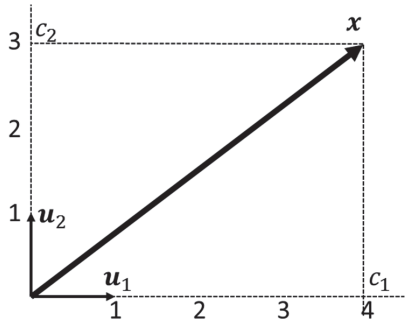


***Figure 2.11**. Absorption spectrum, $\boldsymbol{x}$ with the orthonormal base vectors, $\boldsymbol{u}_1$and $\boldsymbol{u}_2$. The components of $\boldsymbol{x}$ in the base are c_1, c_2. Modified from Sanchez and Kowalski (1988).*

The absorption spectrum, $\boldsymbol{x}$ is modelled by Equation 2.12

$$\boldsymbol{x} = \boldsymbol{u}_1 c_1 + \boldsymbol{u}_2 c_2 \qquad \textit{Equation 2.12}$$

If $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are pure unitary concentration signals of two analytes, then the projection of $\boldsymbol{x}$ onto either $\boldsymbol{u}_1$ or $\boldsymbol{u}_2$ will, in the special case where $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are orthonormal, directly give the concentration of the analytes as also shown in Figure 2.11 (note that background and error terms are not taken into account). The analyte concentrations are given by the inner product as show in Equation 2.13 and Equation 2.14. It should be noted that only the spectrum of the analyte of interest must be known to compute the concentration.

$$c_1 = \boldsymbol{x}^T u_1 \qquad \textit{Equation 2.13}$$

$$c_2 = \boldsymbol{x}^T u_2 \qquad \textit{Equation 2.14}$$

Unfortunately, spectra of pure analytes are (most often) not orthogonal, i.e. the signals are interfering. In the following, the analyte of interest (or just analyte) is a compound, having an absorption signal, for which quantification is needed. The interferent is a compound, having an absorption signal, for which quantification is not needed.

Figure 2.12 illustrates an absorption spectrum, $\boldsymbol{x}(m \times 1)$ where the base vectors are the pure analyte signal, $\boldsymbol{a}(m \times 1)$ and the pure interferent signal, $\boldsymbol{k}(m \times 1)$. Both $\boldsymbol{a}$ and $\boldsymbol{k}$ are at unitary concentration and non-orthogonal.

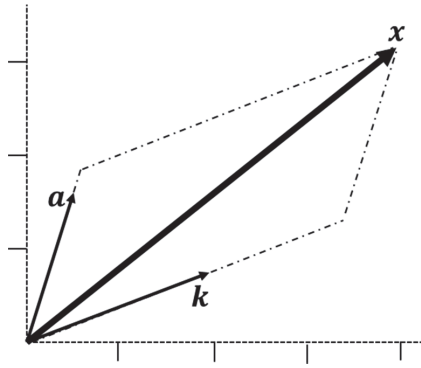


***Figure 2.12**. Absorption spectrum, $\boldsymbol{x}$ with covariant base vectors ($\boldsymbol{a}, \boldsymbol{k}$) at unitary concentrations. Modified from Sanchez and Kowalski (1988).*

The vectorial components of $\boldsymbol{x}$ in the base, $\boldsymbol{a}$ and $\boldsymbol{k}$, are simply the concentrations, $c_a(1 \times 1)$ and $c_k(1 \times 1)$. The sample spectrum, $\boldsymbol{x}$ is modelled by Equation 2.15 (background and error terms are not taken into account)

$$\boldsymbol{x} = \boldsymbol{a} c_a + \boldsymbol{k} c_k \qquad \textit{Equation 2.15}$$

However, the concentration of $\boldsymbol{a}$ can no longer be estimated by the inner product of $\boldsymbol{x}$ and $\boldsymbol{a}$, as the signal from $\boldsymbol{k}$ will interfere. Hence, the concentration of $\boldsymbol{a}$ should be estimated by the inner product of $\boldsymbol{x}$ and a specific base vector, $\boldsymbol{b}$. The new base vector, $\boldsymbol{b}$ gives the direction of $\boldsymbol{a}$ being orthogonal to $\boldsymbol{k}$ (Sanchez and Kowalski, 1988). This is shown in Figure 2.13. The base vector, $\boldsymbol{b}$ is also called the *contravariant* spectrum of $\boldsymbol{a}$.

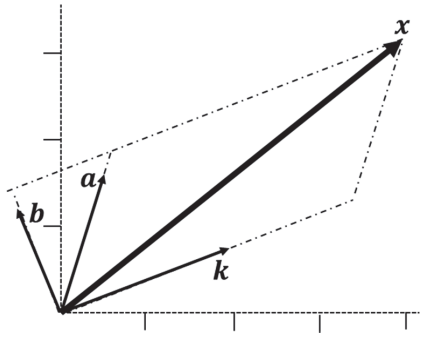


***Figure 2.13**. Absorption spectrum, $\boldsymbol{x}$ with covariant base vectors ($\boldsymbol{a}, \boldsymbol{k}$) at unitary concentrations. The contravariant spectrum of $\boldsymbol{a}$ is denoted $\boldsymbol{b}$. Modified from (Sanchez and Kowalski, 1988)*

The contravariant spectrum of an analyte depends on the pure spectra of all other interferents present in the sample because the contravariant spectrum is orthogonal to all of them (Sanchez and Kowalski, 1988). If $\boldsymbol{a}$ is at unitary concentration and $\boldsymbol{b}$ is scaled according to Equation 2.16 then it implies that $\boldsymbol{b}$ is the translation from the sample spectrum to the concentration of the analyte, $\boldsymbol{a}$. Hence, $\boldsymbol{b}$ is identified as the regression vector. It is important to notice that the regression vector for a given analyte must point in the direction of the pure analyte signal orthogonal to the signals of all interferences (Sanchez and Kowalski, 1988).

$$\boldsymbol{a}^T\boldsymbol{b} = 1 \qquad \textit{Equation 2.16}$$

In order to estimate the regression vector a calibration set, $\boldsymbol{X}(n \times m)$ is needed for which the concentrations of the analyte, $\boldsymbol{y}(n \times 1)$ is known in advance. The form of any linear regression is given in Equation 2.17

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b} \qquad \textit{Equation 2.17}$$

The concentration vector, $\boldsymbol{y}$ is known in advance and the spectra are measured. Therefore, the regression vector, $\boldsymbol{b}$ can be estimated by obtaining a pseudoinverse of $\boldsymbol{X}$, i.e. $\boldsymbol{X}^+$(Equation 2.18).

$$\hat{\boldsymbol{b}} = \boldsymbol{X}^+\boldsymbol{y} \qquad \textit{Equation 2.18}$$

Combining the equations 2.17 and Equation 2.18 one get the general first-order calibration equation (Equation 2.19),

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{X}^{+}\boldsymbol{y} \qquad \textit{Equation 2.19}$$

This is the solution to a general linear calibration model. The differences between linear calibration methods lies in the estimation of the pseudoinverse of $\boldsymbol{X}$.

2.2.3 Partial Least Squares Regression

The $r(\boldsymbol{X})$ of spectroscopic measurements (approximated by the number of analytes and interferents) is often much lower than the number of rows (samples) and columns (variables) of the spectra. This forces high degree of correlation within the variables of $\boldsymbol{X}$ and the spectroscopic measurements are said to be ill-conditioned. Solving Equation 2.17 directly (as would be the case using Multiple Linear Regression) for spectroscopic measurements would involve an underdetermined system of equations with infinitely many solutions.

Partial Least Squares (PLS) regression deals with the problem of regressing $\boldsymbol{y}$ on an ill-conditioned $\boldsymbol{X}$. In PLS, $\boldsymbol{X}$ is approximated by linear combinations like in PCA. Whereas, PCA only models $\boldsymbol{X}$, PLS compromises between modeling $\boldsymbol{X}(n \times m)$ and modeling $\boldsymbol{y}(n \times 1)$ using the same R specifically constructed LV, which are a subspace of $\boldsymbol{X}$, as shown in Equation 2.20 and Equation 2.21. A PLS model can be constructed with both univariate reference values, $\boldsymbol{y}(n \times 1)$ and multivariate reference values, $\boldsymbol{Y}(n \times j)$ as shown by Wold et al., (2001). In this thesis PLS models with univariate reference values will be considered only.

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{T} + \boldsymbol{E}_{\boldsymbol{X}} \qquad \textit{Equation 2.20}$$

$$\boldsymbol{y} = \boldsymbol{T}\boldsymbol{q} + \boldsymbol{e}_{\boldsymbol{y}} \qquad \textit{Equation 2.21}$$

Where $\boldsymbol{T}(n \times R)$ is a matrix of scores, $(m \times R)$ is a matrix of $\boldsymbol{X}$-loadings, $\boldsymbol{q}(R \times 1)$ is $\boldsymbol{y}$-loadings and $\boldsymbol{E}_{\boldsymbol{X}}(n \times m)$ and $\boldsymbol{e}_{\boldsymbol{y}}(n \times 1)$ are residuals.

The NIPALS algorithm is used for PLS in this thesis and is presented in Box 2. Please note, that the algorithm in Box 2 is non-iterative for univariate reference values. For multivariate reference values the algorithm becomes iterative (Wold et al., 2001).

Box 2

NIPALS Algorithm for PLS

$\boldsymbol{X}(n \times m)$ and $\boldsymbol{y}(n \times 1)$ are preprocessed and centered.

The algorithm is repeated for the number of latent variables $r = 1,2, \dots, R$.

1) Estimate and normalize the weights, $\boldsymbol{w}(m \times 1)$

$$\boldsymbol{w_r} = \frac{\boldsymbol{X^T y}}{\|\boldsymbol{X^T y}\|}$$

2) Estimate $\boldsymbol{t}_r(n \times 1)$

$$\boldsymbol{t_r} = \boldsymbol{X w_r}$$

3) Solve $\boldsymbol{X} = \boldsymbol{t}_r \boldsymbol{p}_r{}^T$ with regards to $\boldsymbol{p}_r(m \times 1)$

$$\boldsymbol{p_r} = \frac{\boldsymbol{X^T t_r}}{\boldsymbol{t^T}_r \boldsymbol{t_r}}$$

4) Deflate $\boldsymbol{X}$ and start again from 1)

$$\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{t_r p^T}_r$$

Step one of the PLS NIPALS algorithm projects the spectra onto the concentration vector. By repeatedly doing so each time with the new deflated $\boldsymbol{X}$, the model is forced to first span the region around the concentration vector. That in turn will span the region that is in the neighborhood of the contravariant spectrum for that analyte. The regression vector in PLS regression is estimated by Equation 2.22 (Wold et al., 2001; Andersson, 2009)

$$\widehat{\boldsymbol{b}} = \boldsymbol{W}(\boldsymbol{P}^T\boldsymbol{W})^{-1}\boldsymbol{q} \qquad \textit{Equation 2.22}$$

From Equation 2.21 it follows that,

$$\boldsymbol{q} = (\boldsymbol{T}^T\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{y} \qquad \textit{Equation 2.23}$$

2.2.4 Model Validation

When applying a calibration model to samples with unknown reference values, it is important to have an estimate of the uncertainty of the model predictions. In order to estimate this uncertainty the root mean squared error (RMSE) is often used. The RMSE is calculated between the measured reference values and the model estimated values for the calibration samples. The RMSE is given by Equation 2.24.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \qquad \textit{Equation 2.24}$$

Where $\hat{y}_i$ is the model estimated value of sample i, y_i is the reference value of sample i and n is the number of samples. Notice that the value of RMSE is in the same units and scale as the reference values. Good calibration models have small RMSE.

In order to get a reliable estimate of the RMSE a cross-validation could be performed. A cross-validation is essentially a loop, where predefined data set segments subsequently are left out from the modeling phase. A model is established on the remaining data and this model is then applied to the excluded data segment. All segments are in turn left out from the modeling stage. In each cross-validation loop predictions of the excluded samples are collected. These predictions are used (against the measured reference values) to estimate the error term. In this way the contribution from individual segments to the total error originates from models not impacted by the individual segments. For the final model all samples are used, but the root mean squared error of cross-validation (RMSECV) is reported. Different types of cross-validation exist and the main difference between methods is the way the segments are defined (Næs et al., 2002).

Even though a cross-validation is performed to give at good estimate of how good a calibration model is in predicting new samples with unknown reference values, the RMSECV may still be over-optimistic. All samples originate from the same data set, which means that they most often are collected under the same conditions. New samples may not be collected under the same conditions and this could introduce additional errors. In order to include this in the uncertainty estimate of a calibration model, an independent test set could be collected. It is of course important that the test set reflects the calibration set in terms of what the calibration model is aiming at. The calibration model is established on the calibration set while being tested with the independent test set. The root mean squared error of prediction (RMSEP) of the test set should then reflect the ability of the calibration model to predict new samples. While obtaining a more realistic estimate of the model uncertainty, the obvious drawback of test set validation is that several samples are used for testing only. These samples could potentially contribute with valuable variation to the calibration set empowering more precise estimates of the regression coefficients (Næs et al., 2002).

2.2.5 Orthogonalization

Orthogonality is an important concept in chemometrics. In, for example, multi block methods like Multi Block Variance Partitioning (Skov et al., 2008) and Sequential and Orthogonalized Partial Least Squares Regression (Næs et al., 2011) orthogonalization is an vital step. Orthogonalization is essentially a decomposition of a vector into two orthogonal subspaces. In Figure 2.14 two vectors, $\boldsymbol{a}$ and $\boldsymbol{k}$ are found in a three dimensional space. The vector $\boldsymbol{a}$ is split into a part describing the direction of $\boldsymbol{k}$, $\boldsymbol{a_k}$ and a part describing the part orthogonal to $\boldsymbol{k}$, $\boldsymbol{a_{-k}}$. The vector $\boldsymbol{a_{-k}}$ is said to be in the null space of $\boldsymbol{k}$, $\boldsymbol{N(k)}$. It follows that $\boldsymbol{a = a_k + a_{-k}}$. Every vector in $\boldsymbol{N(k)}$ will be orthogonal to any vector in the direction of $\boldsymbol{k}$ (Strang, 2006). Hence, the information described in $\boldsymbol{N(k)}$ is unrelated to the information described by $\boldsymbol{k}$.

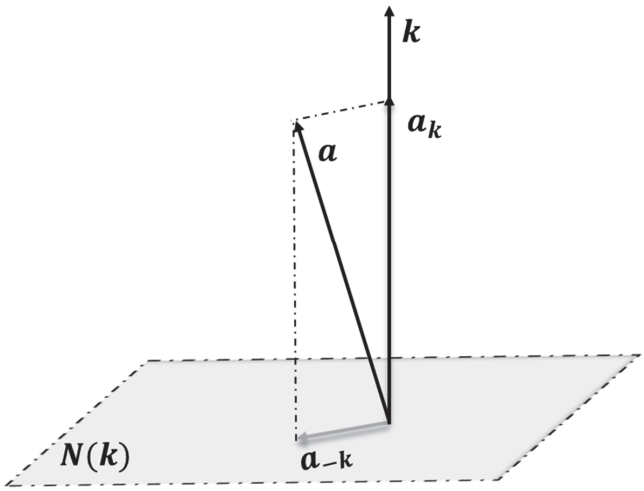


***Figure 2.14**. The vector, $\boldsymbol{a_k}$ is the projection of the vector, $\boldsymbol{a}$ onto the vector, $\boldsymbol{k}$. The vector $\boldsymbol{a_{-k}}$ is the part of $\boldsymbol{a}$ orthogonal to $\boldsymbol{k}$. Illustration modified from Strang (2006).*

The projection of $\boldsymbol{a}(n \times 1)$ onto $\boldsymbol{k}(n \times 1)$ is carried out by a projection matrix, $\boldsymbol{P}(n \times n)$ given in Equation 2.25 (Strang, 2006).

$$\boldsymbol{P} = \frac{\boldsymbol{k}\boldsymbol{k}^T}{\boldsymbol{k}^T\boldsymbol{k}} \qquad \textbf{Equation 2.25}$$

Multiplying $\boldsymbol{P}$ with any vector, $\boldsymbol{a}$ would return the projection of $\boldsymbol{a}$ onto $\boldsymbol{k}$. Hence, $\boldsymbol{a_k}(n \times 1)$ is given by Equation 2.26,

$$\boldsymbol{a_k} = \boldsymbol{P}\boldsymbol{a} = \frac{\boldsymbol{k}\boldsymbol{k}^T}{\boldsymbol{k}^T\boldsymbol{k}}\,\boldsymbol{a} \qquad \textbf{Equation 2.26}$$

Which can also be shown from Figure 2.14. The part of $\boldsymbol{a}$ orthogonal to $\boldsymbol{k}$ is obtained by subtracting $\boldsymbol{a_k}$ from $\boldsymbol{a}$ as shown in Equation 2.27,

$$\boldsymbol{a_{-k}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{a} \qquad \textbf{Equation 2.27}$$

Where $\boldsymbol{a_{-k}}(n \times 1)$ is the part of $\boldsymbol{a}$ orthogonal to $\boldsymbol{k}$, and $\boldsymbol{I}(n \times n)$ is the identity matrix.

Projecting and/or orthogonalizing a matrix, $\boldsymbol{X}(n \times m)$ on a vector, $\boldsymbol{k}(n \times 1)$ (or matrix, $\boldsymbol{K}(n \times j)$) is simply a column wise operation. Hence, Equation 2.26 and Equation 2.27 can easily be extended to include matrices.

2.3 Real-time Monitoring of Whey Protein Hydrolysis

The market is increasing for whey protein solutions, including whey protein isolates (WPI), whey protein concentrates (WPC) and whey protein hydrolysates (WPH). This is mainly due to the high nutritional value of whey proteins and derived peptides as well as functional properties as foaming and emulsifying agents. Hydrolysis changes the physicochemical and technological properties of whey proteins (Chobert et al., 1988; de Castro, Ruann Janser Soares et al., 2015). Increased degree of hydrolysis (DH%) increases the solubility of whey proteins (Svenning et al., 2000), and thus facilitates improved bioavailability of products based on WPH. This is of importance, for example, in nutrition of hospitalized patients or in sports nutrition (Madureira et al., 2007). Furthermore, WPH exhibit reduced immunological reactivities. Hence, WPH can be used in infant formulas for allergic infants (Halken and Høst, 1997). Obtaining knowledge related to DH% is therefore important for all whey products either during processing of WPC or WPI, where hydrolysis is unwanted, or as a tool for on-line process monitoring or final quality control of WPC hydrolysis.

In **PAPER II** WPC was used as substrate for FT-IR monitoring of enzymatic protein hydrolysis. The WPC consists primarily of whey protein (80-85%) and smaller amounts of lactose and fat. Figure 2.15 shows the hydrolysis of protein, where a secondary amide is hydrolysed into an acid and a primary amine.

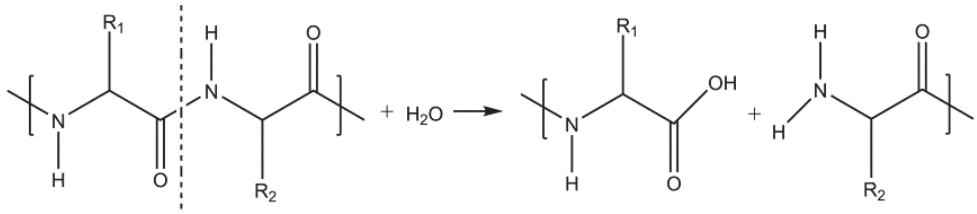


***Figure 2.15**. Hydrolysis of peptide bond.*

Hydrolysis was determined over time as the level of primary amino groups reflecting the level of free amino-terminals as Leucine equivalents. A fluorescamine assay was used as reference method (Pearce, 1979; Rollema et al., 1989). Three sample sets were prepared from the same batch of WPC. Samples consisted of aqueous WPC solutions and hydrolysis was performed using trypsin. The sample sets are presented in Figure 2.16 showing Leucine equivalents as a function of time. The first sample set (Figure 2.16a) had 5% WPC (w/w) and a trypsin:WPC ratio of 1:200. The second data set (Figure 2.16b) had also 5% WPC (w/w) but a trypsin:WPC ratio of 1:100. The third sample set (Figure 2.16c) had 8% WPC (w/w) and a trypsin:WPC ratio of 1:100. Duplicate samples were made and samples were meausered during hydrolysis at time 0, 0.5, 1, 2, 3, 4, 5, 6, 7 and 8 hrs for each sample set. Trypsin was inactivated by addition of trypsin inhibitor, as described in **PAPER II**. For one of the replicates in sample set three (Figure 2.16c) measurements at h = 6 and h = 7 were not collected resulting in a total of 58 measurements.

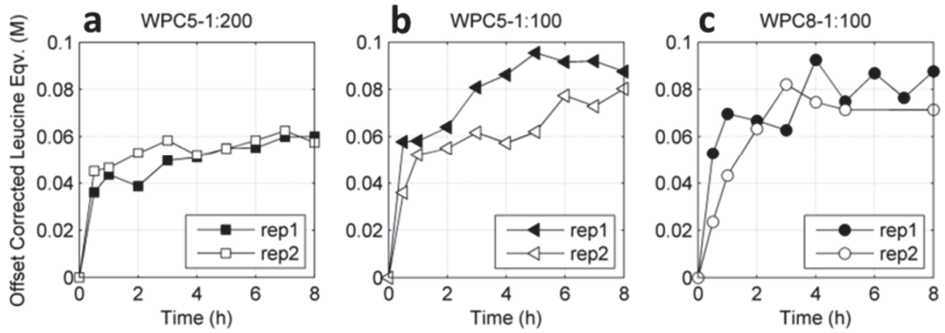


***Figure 2.16**. Fluorescamine reference method showing level of free amino terminals as leucine equivalents in whey protein concentrate solution with protein concentrations of 5 g/100 g (WPC5) at two different enzyme-to-substrate ratios (1:100, 1:200) and 8 g/100 g (WPC8) at enzyme-to-substrate ratio 1:100. Rep1 and rep2 indicate the two replicates.*

Immediately after trypsin inactivation the samples were analysed by IR spectroscopy (Milkoscan FT2). Hydrolysis of the peptide bond will introduce changes in how the molecular groups are vibrating. In the primary amine the nitrogen will donate eletrons to the carbonyl group whereas the carbonyl group in the formed acid will be exposed to the electron withdrawing O-H group. Therefore, the carbonyl bond is more constricted in the acid and this will increase vibrational frequency. The carbonyl group in a peptide will absorb at ~1,550 cm^{-1} (and thereby be hidden behind O-H bend). However, in an acid the carbonyl group will absorp just above 1,700 cm^{-1}. Furthermore, N-H absorption of a secondary amide is found ~1,540 cm^{-1}, where in a primary amine the N-H vibrations are expected ~1,600 cm^{-1} (Engelsen, 1997; Barth, 2007).

Figure 2.17a shows the FT-IR spectra shaded by leucine equivalents. The region from 5,008 to 2,980 cm^{-1} and 979 to 925 cm^{-1} was considered noise and removed from the data set. The region from 2,822 to 1,769 cm^{-1} contained no valuable information and was also removed. Furthermore, the saturated water signal (O-H bend; ~1,700 cm^{-1} to ~1,600 cm^{-1}) was removed. For spectral interpretation a narrow band from 1,704 to 1,588 cm^{-1} around the O-H bend was removed. However, for modeling a wider band from 1,715 to 1560 cm^{-1} around the O-H bend was removed. It was decided to remove a wider region around the O-H bend for modeling in order to avoid instabilities related to water absorption. Full FT-IR spectra were recorded in triplicates, and for each sample, the average spectrum was calculated and used for further analysis.

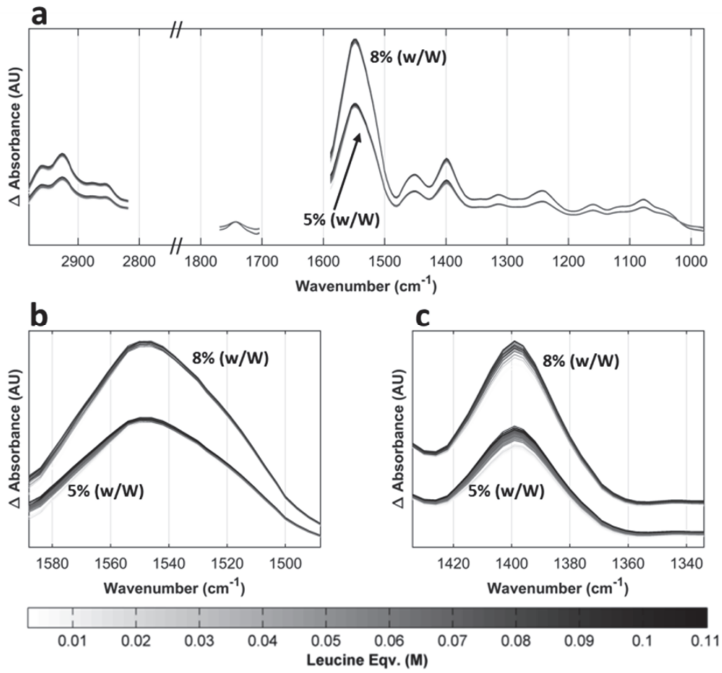


***Figure 2.17**. Fourier transform infrared (FT-IR) spectra of aqueous whey protein concentrate solution with protein concentrations of 5% (w/w) at two different enzyme-to-substrate ratios (1:100, 1:200) and protein concentrations of 8% (w/w) at enzyme-to-substrate ratio 1:100. (a) Raw spectra. (b) Zoom-in at 1550 cm^{-1}; Amide II band. (c) Zoom-in at 1400 cm^{-1}.*

It was expected that the carbonyl group from the formed acid (Figure 2.15) would show just above 1,700 cm^{-1}. However, this cannot be identified in the spectra (probably due to lack of sensitivity). From Figure 2.17b it is found that the band at ~1,540 cm^{-1}, originating primarily from the N-H bend) shifts towards higher wavenumbers, as expected. Interestingly, a peak increased remarkably at ~1,400 cm^{-1} (Figure 2.17c). More extensive analysis needs to be carried out before this peak can be assigned. However, speculations go towards changes in the amide III band or a carboxylate stretching vibration.

Figure 2.18 shows the PCA score plots of the spectra. One LV PCA models were built subsequently using the measurements consisting of 5% protein concentration and 1:200 enzyme-to-substrate ratio (Figure 2.18a), 5% protein concentration and 1:100 enzyme-to-substrate ratio (Figure 2.18b), and 8% protein concentration and 1:100 enzyme-to-substrate ratio (Figure 2.18c).

From Figure 2.18 it is observed that scores from the three different sample set look fairly much alike in the relationship with time. And compared with Figure 2.16 they seem to describe the enzymatic reaction very well. For enzymatic reactions, the reaction velocity varies with the substrate concentration (when the enzyme concentration is held constant). Over time the substrate will be

tuned into product and the reaction velocity will decrease. Hence, the transformation from substrate into product will go towards a plateau. This is very likely the phenomena seen in all three subplots of Figure 2.18 (Pratt and Cornely, 2004b).

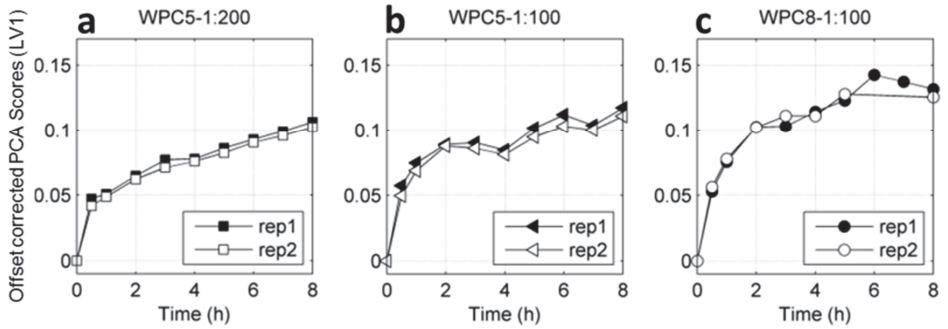


***Figure 2.18**. Score plot from Principal Component Analysis (PCA) of Fourier transform infrared (FT-IR) absorption spectra. Scores for the first latent variable (LV1) as a function of time. The PCA scores are offset corrected. The PCA model is based on FT-IR spectra of whey protein concentrate solution with protein concentrations of 5 g/100 g (WPC5) at two different enzyme-to-substrate ratios (1:100, 1:200) and 8 g/100 g (WPC8) at enzyme-to-substrate ratio 1:100. The spectra were collected over time while hydrolysis was ongoing.*

In order to explore whether the scores in Figure 2.19 contain predictive information related to the hydrolysis, the PCA scores are correlated with the measured leucine equivalents (Figure 2.16). It is found that the PCA scores obtained from the spectra are very correlated with the leucine equivalents. Hence, predictive information related to hydrolysis can be found in the spectra.

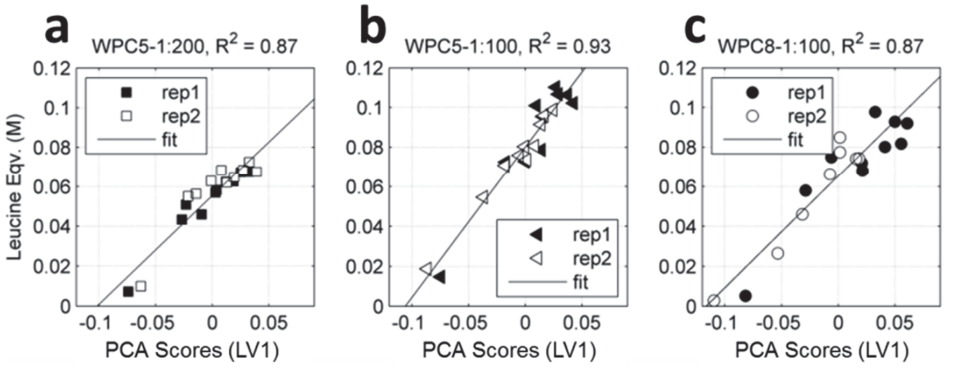


***Figure 2.19**. Correlation between leucine equivalents and Principal Component Analysis (PCA) scores of the first latent variable (LV1). The PCA model is based on FT-IR spectra of whey protein concentrate solution with protein concentrations of 5 g/100 g (WPC5) at two different enzyme-to-substrate ratios (1:100, 1:200) and 8 g/100 g (WPC8) at enzyme-to-substarte ratio 1:100.*

Knowing the enzyme-catalyzed reaction mechanism is potentially very useful for a PAT application. Obtaining spectra and calculating scores *real-time* enables fitting the parameters of the mechanistic model. This potentially gives predictive power to estimate information related to hydrolysis in future time points and thereby enable optimized process control and scheduling for the modern dairy. A similar approach has been proposed for monitoring milk coagulation during cheese making (Lyndgaard et al., 2012).

From Figure 2.19 it is observed that the relationship between the PCA scores and the leucine equivalents is not the same for the three sample sets. However, the relationship appears to be the same for the replicate samples. The enzyme and substrate concentrations are important for enzymatic reactions. Hence, it is strictly important to control these if PCA scores from spectra are to be used for estimating information related to hydrolysis in a production.

2.4 Standardization of Spectroscopic Instruments

In a PAT setting, spectroscopic measurements are often used to determine analyte concentrations. Unfortunately, the spectroscopic signal does not directly return concentrations. However, absorption spectra are, through *Beer's law*, linearly related with concentrations, as stated in **Chapter 2.1**. The translation, from absorption spectra into analyte concantrations, is contained in the regression vector by the calibration model (**Chapter 2.2**). As also highlighted in **Chapter 2.2**, the power of first order instruments (like multivariate spectroscopic instruments) is that they can deal with interferents, as long as such interferents are in the calibration set. The first order instrument can be calibrated for an analyte in the presence of interferents by estimating the analyte regression vector in the direction of the analyte contravariant spectrum, i.e. the regression vector has to be orthogonal to the signals of all interferents (**Chapter 2.2**). However, if a calibration model is established for a given analyte and a new interferent is introduced in a future sample, it is not given that the regression vector of the analyte is orthogonal to the signal of the new interferent. The consequences may be that the calibration model returns erroneous predictions of the analyte. Therefore, the quality of the calibration set is highly important when constructing high quality calibration models. It is important that all possible sources of interference are introduced in the calibration set. This may require up to hundreds or thousands of samples collected over a broad time period and all with reference values. Consequently, constructing high quality calibration models is expensive and time-consuming.

Therefore, in an industrial setting where multiple spectroscopic instruments are used for the same measurements, it could be convenient to develop the calibration model on a single instrument only but apply this calibration model to the spectroscopic measurements of all the instruments. In order to apply this successfully the translation from the instrument response space to the analyte concentration space must obviously be the same for all instruments. Hence, the instruments must provide similar output. Otherwise, the results (predictions) from the calibration model will be in-

correct. However, it is not given that two identical spectroscopic instruments are providing similar output. Therefore, a large number of methods have already been developed with the purpose of transforming instrument output, so the spectra from multiple instruments will be similar.

A calibration model transfer is often evaluated in a master/slave setting. The calibration model is developed on one instrument (the master) and tested on other instruments (slaves). The RMSEP obtained for the slave instruments is then used for evaluating the model transfer (Swierenga et al., 1998a; Swierenga et al., 1998b; Barboza and Poppi, 2003; Bergman et al., 2006; Alamar et al., 2007; Fan et al., 2008). This RMSEP value is used to verify how successful the model transfer was. The RMSEP value includes uncertainties introduced by inadequacies in the modeling steps, which can be related to differences in the instrument response space of the master and slave instruments. Nevertheless, the RMSEP value also includes uncertainties originating from the reference method. A fact, which is often forgotten. Hence, large uncertainties in the reference measurements would make the RMSEP value larger and thereby make the model transfer appear poor even in the ideal/perfect case, where the predictions of the master and the salve instrument are identical. **PAPER III** deals with how to evaluate a calibration model transfer.

Figure 2.20 illustrates how uncertainties introduced by modeling and reference method, respectively, affect the prediction error. Uncertainties in the reference method will contribute to imprecision along the horizontal direction, whereas those from the model (in the predictions) will cause the data points to be uncertain in the vertical direction. Hence, if the reference method is imprecise, the modeling will appear poor (DiFoggio, 1995; Faber and Kowalski, 1997).

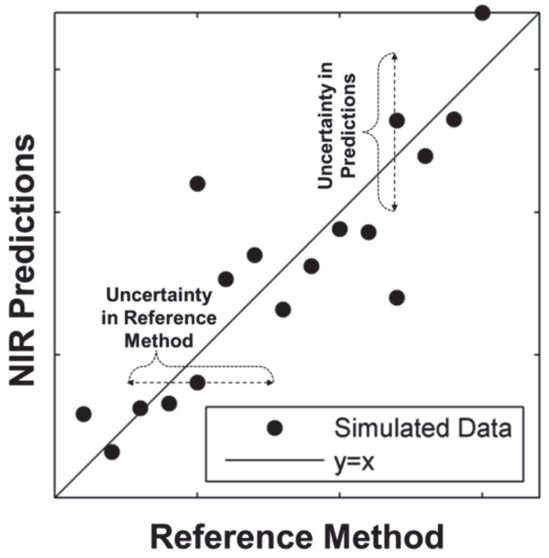


***Figure 2.20**. Influence of uncertainties originating from the reference method and near infrared predictions, respectively, on calibration model evaluation.*

In **PAPER III**, a method was proposed for evaluating model transfer between a population of spectroscopic instruments. The aim of a model transfer is to obtain the same results (i.e. predictions) on different instruments with the same samples by applying only one model. If the same predictions are obtained from different instruments for the same samples, then the instrument response space related with the analyte of interest must be the same for the instruments.

For **PAPER III**, a total number 75 flour samples were measured on ten different NIR instruments. Five NIR instruments were DS2500 instruments and five NIR instruments were InfraXact instruments. The spectra from the five InfraXact and the five DS2500 are shown in Figure 2.21a and Figure 2.21b, respectively. Protein content was quantified by Kjeldahl digestion for all 75 samples and used as reference variable during PLS modeling.

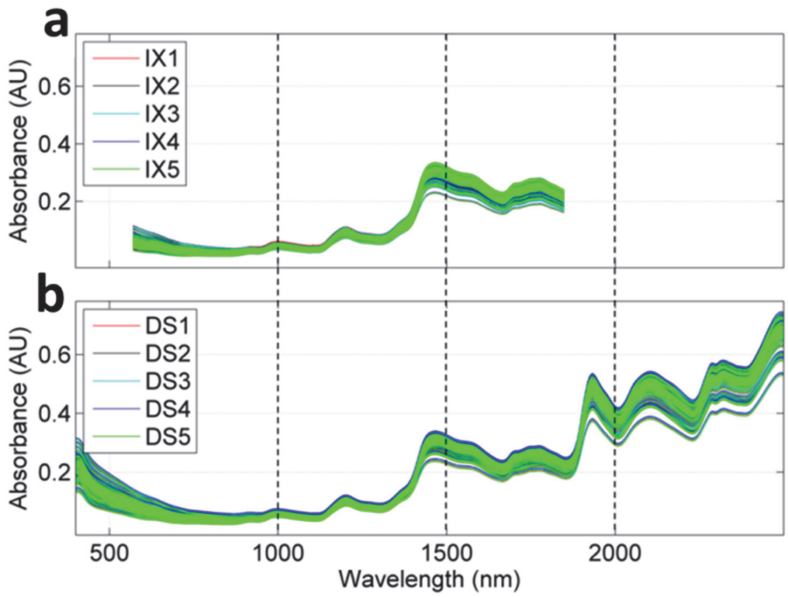


***Figure 2.21**. Raw near infrared spectra of flour samples. (a) InfraXact. (b) DS2500.*

The PLS models were validated by *Leave-One-Instrument-Out* cross-validation (**PAPER III**). Predictions obtained from cross-validation were collected and the medians across samples were found. The RMSE between these medians and the cross-validated predictions (for each instrument) was calculated. This RMSE relates to the similarities between the predicted values from the instruments and can thereby be used for evaluating model transfer. The median is used as "the true prediction value" as it is (almost) insensitive to outlying predictions, which would have affected, for example, the mean of predictions. It is important to note that the RMSE is not a measure of how well the "true" protein content can be modeled by the spectroscopic measurements. This RMSE is only a measure of how similar protein content is modeled from the measurements obtained from different instruments (i.e. the model transfer performance). In this way, the impact of

uncertainties related to the reference method will be minimized. The assumption is of course that the developed model performs at an acceptable level when benchmarked against real reference value.

In order to investigate how the methods responded to uncertainties in the spectral and the reference measurements, respectively, simulated deterioration of the data belonging to the fifth InfraXact instrument (IX5) was performed. Model performance was evaluated by traditional measured vs. predicted approach and the proposed method where predictions are compared with the median of cross-validated predictions.

First, deterioration of data was done by adding random noise to the spectroscopic measurements (i.e. changing the instrument response space) of instrument IX5, as shown in **PAPER III**. Now the original translation from instrument response space to analyte concertation space is not valid for IX5 (i.e. the model transfer is not valid). This will show up both when evaluating the model transfer by the traditional method and the new method. Errors will show up in the traditional method because the predictions of IX5 are wrong and therefore not comparable with the reference values. In the new method, errors will show up because the predictions of IX5 will be different from the predictions of the other instrument. Figure 2.22 shows the results where the spectra of IX5 are deteriorated. Instrument IX5 has a high RMSE when evaluated by both the traditional approach (Figure 2.22a) and the new approach (Figure 2.22b).

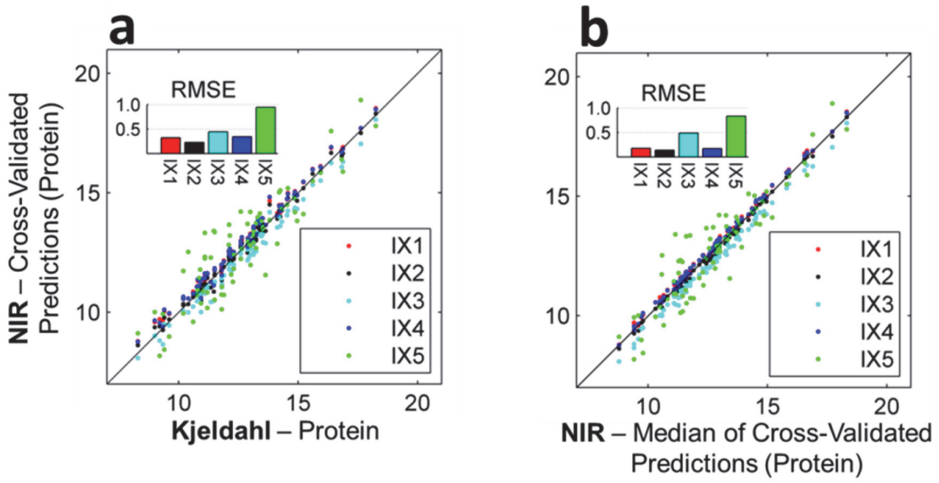


***Figure 2.22**. Results obtained with simulated deterioration of spectra from instrument IX5. (a) Traditional approach. (b) New approach.*

Secondly, the reference data belonging to IX5 were deteriorated by random noise, as shown in **PAPER III**. The model calibrated on the remaining InfraXact instruments will still be valid for IX5. The translation from the instrument response space to the analyte concentration space will be the same for all instruments. And if the instrument response space for the instruments is similar, then

the same predictions will be yielded from all instruments. Nevertheless, evaluating the model transfer by the traditional method, the predictions of IX5 will be compared with the erroneous reference values belonging to IX5. Therefore, IX5 will appear to have a large RMSE. Using the new method for model transfer evaluation, the predictions obtained from IX5 will be compared with the predictions obtained from the remaining instruments. Figure 2.23 shows the results where the reference values belonging to IX5 have been deteriorated. Figure 2.23a shows the errors obtained using the traditional approach. Here it is found that IX5 has a high RMSE. Figure 2.23b shows the errors obtained using the new method. Here the predictions are compared and it is found that the predictions obtained from the measurements of IX5 are very similar to the predictions obtained from the measurements of the remaining InfraXact instruments. Hence, Figure 2.23a indicates that the model transfer is not successful, whereas Figure 2.23b indicates that the model transfer indeed is successful.

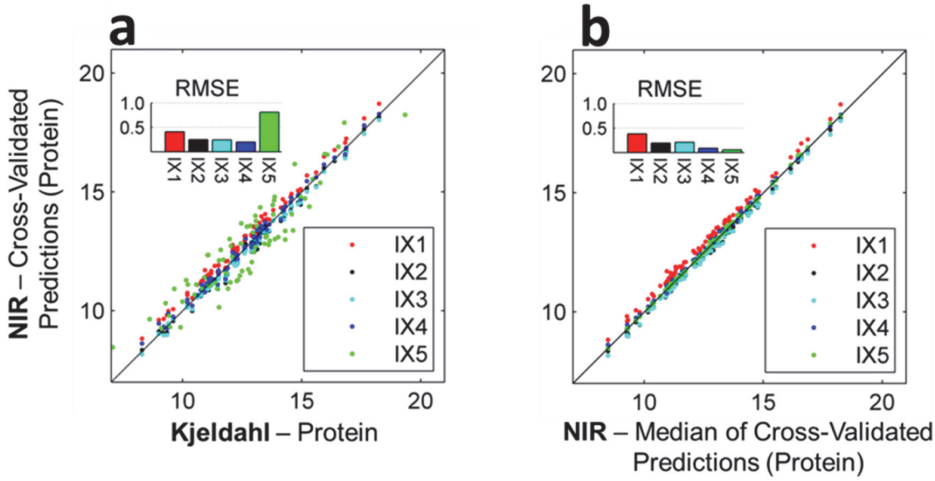


***Figure 2.23**. Results obtained with simulated deterioration of reference values belonging to instrument IX5. (a) Traditional approach. (a) New approach.*

In model transfer, a model is developed on one instrument and then applied to other instruments. The spectroscopic measurements are model input and the predictions are model output. In an industrial setting where one model is applied to the measurements of multiple instruments it is important that the predictions are identical for identical samples. Therefore, the right way of evaluating model transfer is by comparing predictions. Hence, Figure 2.23b (the new method) is giving a more direct and accurate picture of the model transfer than Figure 2.23a (the traditional method).

When using multiple instruments for the same measurements it can be convenient to do a model transfer. It is believed that the task of doing model transfer is easier to successfully overcome, if the original spectra from the multiple instruments are already providing similar results. Therefore,

it is believe that this new method can be used in a quality criterion for instruments. Figure 2.24 shows the comparison of InfraXact and DS2500 using the new method. The results for the five InfraXact instruments are shown in Figure 2.24a. The results for the five DS2500 instruments are shown in Figure 2.24b. It is found that the DS2500 instruments are providing very similar results compared with the InfraXact instruments, which are showing a minor bias. Such evaluation would not have been possible with the traditional method as the bias variation in InfraXact would be masked by the errors in the reference values.

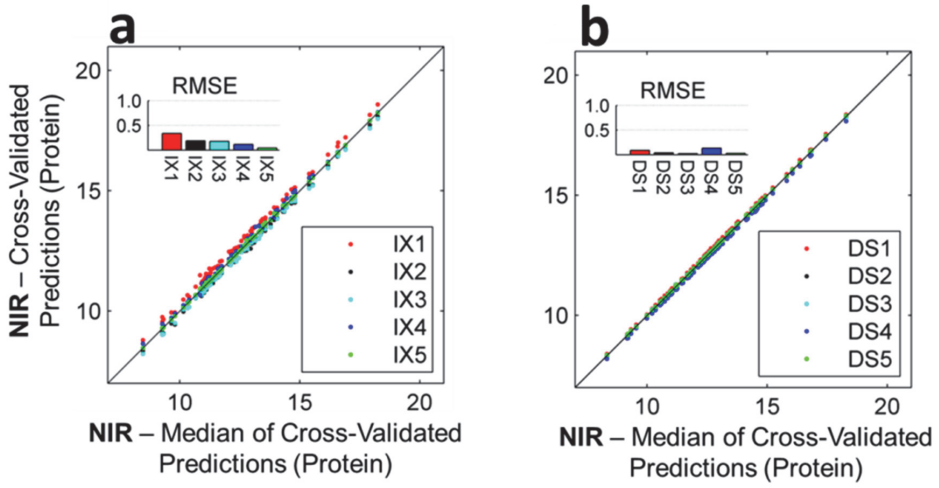


***Figure 2.24**. Evaluated by new approach. Comparison of InfraXact instruments (a) and DS2500 instruments (b).*

PAPER III deals with the important aspect that a calibration model in fact is a relation between two vectorial spaces. In order for a calibration model to be valid, the relationship between the two spaces must be constant. If the relation between the two spaces changes, then the model is not valid anymore. Changes may be introduced by a number of factors, for example the path length of the infrared light through the sample. This may change from one instrument to another, and this will change the relationship between the instrument response space and the analyte concentration space. Furthermore, models based on indirect relationships risk to be subjected to changes in the relation between the two vectorial spaces. This is exactly what was highlighted in **Chapter 2.2**. If the regression vector for a given analyte is not orthogonal to the signals of all interferents, then the translation from the instrument response space to the analyte concentration space will change with one or more interferents. The interferents will impact the part of the spectra used for predicting the analyte. Hence, the estimated concentrations of analytes will depend on the interferents.

CHAPTER 3

MILK PHENOTYPES

The expression of specific bovine milk traits, for example, specific FA and protein fractions are determined by the cow's phenotypic characteristics. The milk phenotype is influenced by both the cow's genotype and environmental factors the cow has been exposed to throughout life (including stage of lactation and number of parities).

The most interesting phenotypic traits, for the dairy industry, relate to both oligosaccharides and protein composition but also FA composition has previous received considerable focus. This thesis will however focus on protein and FA composition. Detailed information on protein composition and FA profile can be used, by the dairy industry, to ensure a good quality of milk and to differentiate milk for specific purposes. In this way, for example, milk with a protein composition maximizing the cheese yield, can be used for cheese production.

Both genetic and environmental factors, for example, lactation stage, feeding and health affect the FA and protein composition of cow milk (Palmquist et al., 1993; Bobe et al., 1999; Bauman and Griinari, 2003; Lock and Bauman, 2004; Vlaeminck et al., 2006a; Vlaeminck et al., 2006b; Craninx et al., 2008; Heck et al., 2009; Poulsen et al., 2012). The FA are either biosynthesized in the mammary gland or blood derived (Bauman and Griinari, 2003). In general, short and medium chain saturated FA (C4:0-C14:0 and a part of C16:0) are synthesized in the mammary gland, whereas long chain fatty acids (including the other part of C16:0) are blood derived (Bauman and Griinari, 2003). The blood derived fatty acids mainly originate from the cow feed. However, these FA are extensively changed by hydrogenation and desaturation in the rumen and the mammary gland, respectively (Barber et al., 1997; Pereira et al., 2003; Craninx et al., 2008). The proteins are synthesized in the mammary gland too. Free amino acids, primarily extracted from the blood, are precursors for the protein synthesis. The amino acid profile taken up by the mammary gland may be quite different from the amino acid profile of the milk, which suggests that extensive transamination and oxidation of amino acids occurs in the mammary gland (Maas et al., 1997). Therefore, it is believed that both milk fat and protein composition is determined, to a large extent, by genetic factors and less by feeding. Hence, efficient breeding programs may be able to change the detailed milk composition so, for example, the FA composition is enhanced for human health, protein composition is improved for better milk coagulation or traits meeting the requirement for nutritional supplements are increased.

Figure 3.1 shows two PCA models on bovine milk FA composition and bovine milk protein composition, respectively. The phenotypic traits are expressed as grams/100 grams of milk and were scaled to unit variance and centered (i.e. auto scaled) prior to PCA. Samples originate from a sample set consisting of a total of 892 morning milk samples from individual cows (456 Holsteins and 436 Jersey). Milk Samples originated from the Danish-Swedish Milk Genomics Initiative (www.milkgenomics.dk). The Holstein samples were collected from 20 Danish dairy herds from October to December 2009. The Jersey samples were collected from 22 Danish dairy herds from February to April 2010. All samples were from conventional herds and taken while cows were housed. The sampling strategy aimed at minimizing environmental variation but maximizing the genetic variation of the sample population (Poulsen et al., 2012). Quantification of FA was done using gas chromatrography as described by Poulsen et al. (2012). Protein fractions were quantified using liquid chromatrography as described by Jensen et al. (2012).

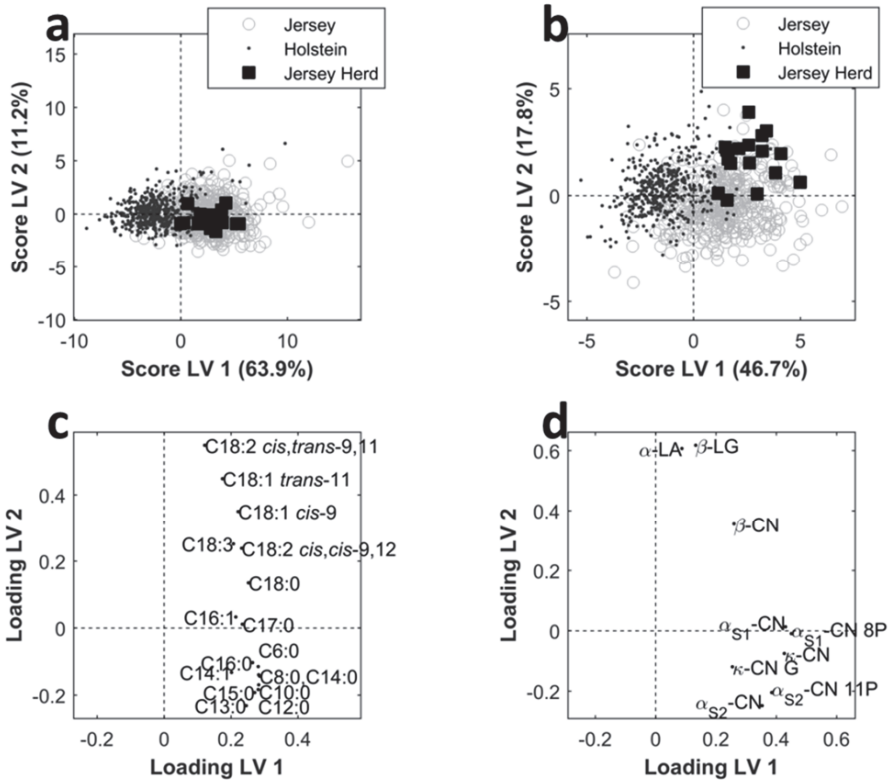


***Figure 3.1**. Principal component of detailed milk composition. (a) score plot of fatty acid composition. (b) score plot of protein composition. (c) loading plot of fatty acid composition. (d) loading plot of protein composition. Traits are expressed as grams/100 grams of milk. In the score plots the two breeds (Jersey and Holstein) are highlighted together with Jersey samples from a specific herd.*

The PCA model on FA composition is based on a subset of 890 samples (455 Holstein and 435 Jersey; two samples were excluded due to lack of reference values). This sample subset was also used in **PAPER IV**. The PCA model on protein composition is based on a subset of 832 samples (426 Holstein and 406 Jersey; 60 samples were excluded due to lack of reference values). This sample subset was also used in **PAPER V**.

The score plots are shown in Figure 3.1a and Figure 3.1b for FA and protein composition, respectively. The loading plots are shown in Figure 3.1c and Figure 3.1d for FA and protein composition, respectively. From the PCA models (Figure 3.1) it follows that the Jersey samples in general have higher content of all FA and protein fractions than Holstein samples. In the score plots, Jersey samples originating from a selected herd are highlighted. Samples from the selected herd group in the score plots. In general, samples from the same herds grouped in the score plots (data not shown). This indicates that samples from the same herds have more or less the same phenotypic characteristics.

Examining Figure 3.1b and 3.1d it appears that cows from the selected Jersey herd produce milk with higher content of the whey fractions (α-LA and β-LG) as well as β-CN. Such information could potentially be relevant in, for example, milk differentiation.

Traditionally, detailed milk composition is measured through time-consuming chromatography based analyses and this is also the case for the data presented in Figure 3.1. In order to make it feasible for animal breeders and the dairy industry to include detailed milk composition as routine quality parameters a high-throughput method is needed. As FT-IR is currently being used it is an attractive method for providing high-throughput information on detailed milk composition. Such high-throughput information is useful in relation to breeding, documentation and process control.

Phenotyping milk from individual cows for breeding purposes may not be a classical PAT example. However, as outlined in the beginning of **Chapter 2**, the key element in PAT is generally to identify and explain all critical sources of variability. In the end, this must also be the purpose of breeding programs.

CHAPTER 4

THE CAGE OF COVARIANCE – INFRARED MEASUREMENTS FOR DETAILED MILK COMPOSITION

4.1 Absorption of Infrared Radiation in Milk

Bovine milk contains in general around 4% fat, 3% protein and 5% carbohydrate. For decades IR has globally been successfully applied in quantifying these parameters in bovine milk. Bovine milk fat is mainly found as triglycerides, which are composed of three FA esterified to a glycerol backbone. More than 400 different FA have been found in milk (Jensen, 2002). Most of these FA are however found in trace amounts and only about 12 of them are present in concentrations higher than 1 % of the milk fat. Approximately 75 % of the FA are saturated and 25 % are unsaturated. The protein fractions κ-CN, β-CN, α_{s1}-CN, α_{s1}-CN, α-LA and β-LG account for about 90 % of the total protein in bovine milk. The remaining 10 % consist of minor proteins like lactoferrin. The most abundant carbohydrate in milk is lactose, which is a disaccharide consisting of galactose and glucose linked by a β1-4 glycosidic bond. Large differences in the milk composition may however exist between individual cows (Fox and McSweeney, 1998; Walstra, 1999; Bovenhuis et al., 2013).

Water molecules are very polar and therefore, water shows intense absorption in the IR spectrum. Water gives rise to absorption at approx. 3,500 cm^{-1} to 3,000 cm^{-1} (O-H stretch) and approx. 1,700 cm^{-1} to 1,600 cm^{-1} (O-H bend) (Bruice, 2007b). The hydrogen bonds between water molecules will cause the water absorption bands to be very broad. Figure 4.1 shows FT-IR absorption spectra of milk samples with intense and broad water signals.

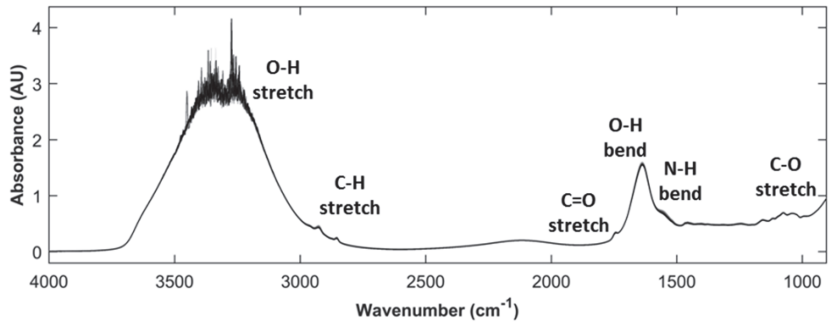


***Figure 4.1**. Infrared spectra of milk samples.*

Nevertheless, in order to suppress the water signals, a milk spectrum can be ratioed against a water spectrum (as background). This is, for example, the case when measuring milk samples using MilkoScan FT2. Figure 4.2 shows a FT-IR spectrum of a milk sample measured with water as background.

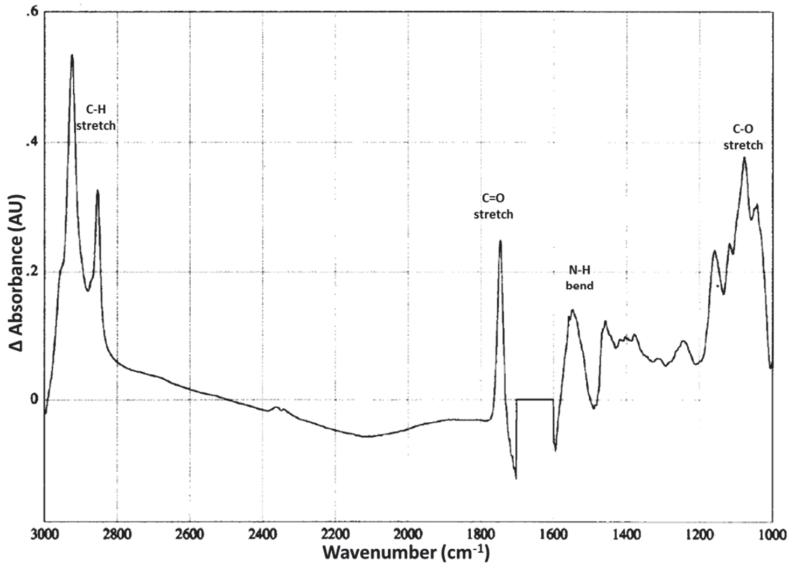


***Figure 4.2**. FT-IR milk spectrum ratioed against water as background. Modified from (Luinge et al., 1993).*

The aliphatic fat groups in milk will give rise to IR absorption as antisymmetrical and symmetrical C-H stretching vibrations at ~2,925 cm^{-1} and ~2,850 cm^{-1}, respectively, and methylene in-plane bending (scissoring) at ~1,465 cm^{-1}. Furthermore, methylene out-of-plane bending modes (twisting and wagging) and methyl umbrella deformation absorb in the wavenumber range from 1,380 cm^{-1} to 1,150 cm^{-1}. A methylene in-plane bending mode (rocking) absorb at ~720 cm^{-1} and the olefinic, =C-H stretching will occur just above 3,000 cm^{-1}. However, the latter two are not accessible in spectra obtained from MilkoScan FT2 due to lack in measurement range and poor signal to noise ratio, respectively. The carbonyl stretching of FA in triglycerides are found at ~1,745 cm^{-1} in the spectra (Engelsen, 1997). The two bonds in the O-C-O part of the ester group will absorb in the range from 1,250 cm^{-1} to 1,050 cm^{-1}. Where the C-O bond next to the carbonyl bond will show absorption in the higher end of the range, the other C-O bond will absorb at the lower end of the range. The intensity of the carbonyl signal provides approx. evidence for the molarity of the fat and the methylene signals provides approx. evidence for the weight percentage of the fat.

Proteins in milk give rise to five IR absorption signals, all related to the polypeptide backbone. The amide A band originates from the N-H stretching and is located at ~3,300 cm^{-1}. The exact frequency depends on the strength of the hydrogen bond. The amide A band can be part of a Fermi reso-

nance doublet, which occurs between the amide A and most likely an overtone of, for example, a N-H bending mode giving rise to weak absorption ~3,050 cm^{-1} (amide B). Nevertheless, the region above 3,000 cm^{-1} (and thereby the amide A and amide B bands) is rarely considered due to poor signal to noise ratio, when measurements are obtained from MilkoScan FT2. The amide I band originates primarily from the stretching vibration of the carbonyl groups in the polypeptide backbone. Due to electron donation by the nitrogen atoms, the carbonyl groups in the peptide backbone (amide I band) absorb at ~1,650 cm^{-1}. Hence, the amide I band is hidden behind the O-H bend from water, when measurements are obtained on milk. The amide II band absorb at ~1,550 cm^{-1}. This band is being used for quantification of protein in milk. The amide II band originates from the out-of-phase combination of the N-H in-plane bend and the C-N stretching vibration with smaller contributions from the carbonyl in plane bending and the C-C stretching. The amide II is hardly affected by side chain vibrations. The amide III mode of polypeptides is more complex. It originates from the in-phase bending of N-H and C-N and also depends on the side chain structure. The amide III band absorb between 1,400 cm^{-1} and 1,200 cm^{-1} (Luinge et al., 1993; Barth, 2007).

Lactose is the most abundant carbohydrate in milk. For IR quantification of lactose in milk, the bonds between the carbon atoms and the hydroxyl groups are used. This bond is characteristic for carbohydrates and absorbs energy at 1,080 cm^{-1}. Hence, lactose determination is not specific, but will include all carbohydrates (Luinge et al., 1993).

Commercial milk recording agencies and dairies have for decades been using FT-IR measurements for estimating fat, protein and lactose content in bovine milk and these parameters have been important in milk recording systems, breeding programs and pricing. However, routinely and accurately predictions of detailed milk characteristics is growing in importance, both for quality characterization of livestock products and for genetic purposes. Infrared spectroscopy is a rapid and cost effective tool and it is suggested in numerous peer-reviewed papers (Table 4.1) that information on detailed milk composition can be obtained from IR measurements. Prediction of the FA profile, protein composition and coagulation properties have by far received most attention.

***Table 4.1**. Overview of studies predicting detailed milk composition from infrared measurements*

Reference	Traits	Citations*
Soyeurt et al. (2006)	Fatty Acids	123
Soyeurt et al. (2008)	Fatty Acids	57
Rutten et al. (2009)	Fatty Acids	53
Afseth et al. (2010)	Fatty Acids	16
De Marchi et al. (2011)	Fatty Acids	32
Soyeurt et al. (2011)	Fatty Acids	77
Ferrand et al. (2011)	Fatty Acids	35
Maurice-Van Eijndhoven et al. (2013)	Fatty Acids	17
Lopez-Villalobos et al. (2014)	Fatty Acids	1
Ferrand-Calmels et al. (2014)	Fatty Acids	9
Soyeurt et al. (2007)	Protein Fractions	39
De Marchi et al. (2009b)	Protein Fractions	19
Bonfatti et al. (2011)	Protein Fractions	19
Rutten et al. (2011)	Protein Fractions	28
Soyeurt et al. (2012)	Protein Fractions	15
McDermott et al.(2016)	Protein Fractions	0
Dal Zotto et al. (2008)	Coagulation Properties	60
De Marchi et al. (2009a)	Coagulation Properties	84
De Marchi et al. (2013)	Coagulation Properties	19

*Source: Google Scholar, January, 2016

4.2 Estimation of Detailed Milk Composition from Infrared Measurements

In general, studies investigating the potential of retrieving phenotypic characteristics from IR measurements of bovine milk (Table 4.1) find that milk FA present in high concentrations are predicted with good accuracy, R^2CV around 0.9 (between measured and predicted values) and, particularly, C14:0 and C16:0 have been highlighted as FA, which could be quantified routinely from

FT-IR measurements (Soyeurt et al., 2006; Rutten et al., 2009). The prediction ability of protein fractions and coagulations properties is more moderate with R^2CV around 0.6 to 0.7 (between measured and predicted values) for the best models. Nevertheless, casein fraction appears to be better predicted than whey fractions (De Marchi et al., 2014).

For this PhD a total of 892 morning milk samples from individual cows (456 Holsteins and 436 Jersey) were measured with FT-IR using MilkoScan FT2 (the same sample set was dealt with in **Chapter 3**). The spectra of these samples are shown in Figure 4.3. Figure 4.3a shows the absorption spectra from MilkoScan FT2 in full spectral range. Figure 4.3b shows the parts of the absorption spectra used for modeling. The spectra are shaded by total fat content.

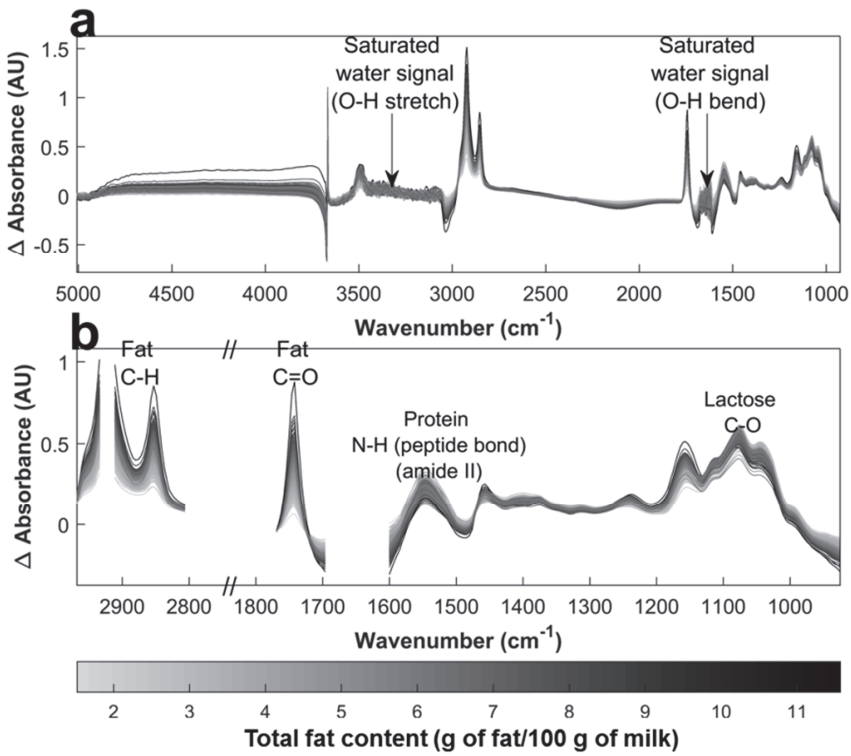


***Figure 4.3**. Fourier transform infrared absorption spectra obtained from MilkoScan FT2. (a) full spectral range. (b) spectral regions included for modeling. Spectra are shaded by total fat content.*

Figure 4.4 shows a principal component analysis of the absorption spectra. The spectra were pre-processed by Standard Normal Variate followed by mean centering. Figure 4.4a shows the score plot and Figure 4.4b shows the loading plot. The first LV explains 90.9% of the total variation in the spectra (Figure 4.4a). The first LV relates primarily to variation associated with total fat content (Figure 4.4b). The second LV explains 7.5% of the total variation (Figure 4.4b) and relates primarily to variation associated with total protein content (Figure 4.4b). From Figure 4.4 it is found that the

Jersey samples, in general, have higher content of fat and protein than the Holstein samples. This is in agreement with the information shown earlier in Figure 3.1 (**Chapter 3**).

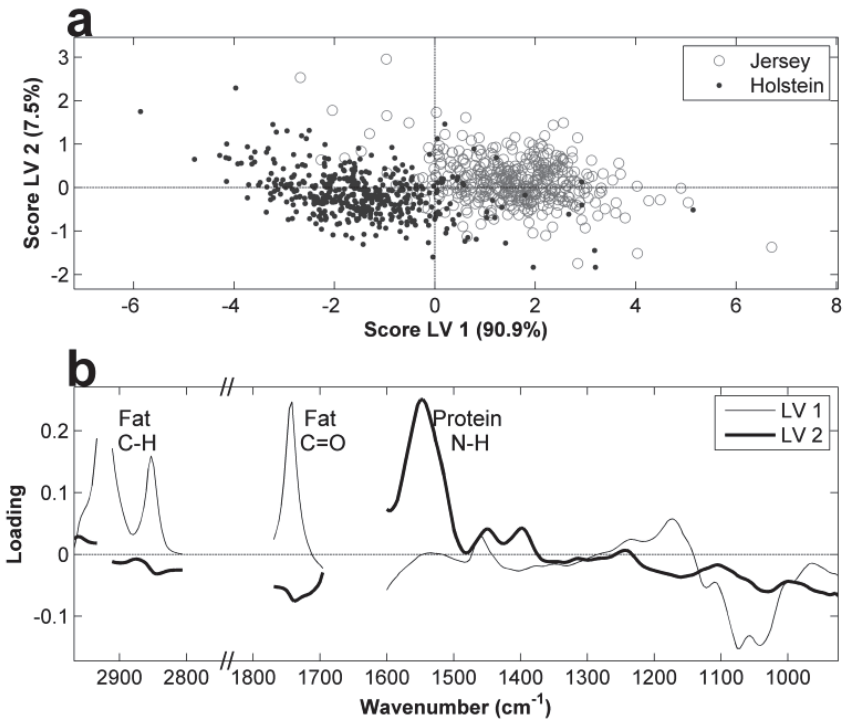


Figure 4.4*. Principal Component analysis of Fourier transform infrared measurements. (a) score plot. (b) loading plot. LV = latent variable.*

PAPER IV concerns the prediction of FA profile from FT-IR measurements using PLS as regression method. As already mentioned in **Chapter 3**, a subset of 890 samples (455 Holstein and 435 Jersey) was included in the study and quantification of FA was done using gas chromatography as described by Poulsen et al. (2012). The results of PLS models for individual FA are shown in Table 4.2. The results are, to a large extent, in agreement with already existing literature.

***Table 4.2**. Individual fatty acids. Results from partial least squares regression models based on raw milk samples*[1]

Fatty acid	Range	Mean	SD	LV	R^2CV	RMSECV
C6:0	0.03 – 0.25	0.13	0.04	5	0.88	0.01
C8:0	0.02 – 0.16	0.07	0.02	5	0.89	0.01
C10:0	0.04 – 0.36	0.16	0.05	9	0.91	0.02
C12:0	0.05 – 0.45	0.18	0.06	9	0.91	0.02
C13:0	0.00 – 0.03	0.01	0.01	2	0.60	0.01
C14:0	0.16 – 0.90	0.51	0.12	9	0.90	0.04
C14:1	0.01 – 0.09	0.04	0.01	1	0.37	0.01
C15:0	0.01 – 0.13	0.05	0.02	2	0.74	0.01
C16:0	0.41 – 3.07	1.41	0.44	7	0.91	0.14
C16:1	0.01 – 0.34	0.07	0.03	8	0.63	0.02
C17:0	0.00 – 0.09	0.02	0.01	1	0.54	0.01
C18:0	0.14 – 1.55	0.53	0.19	7	0.82	0.08
C18:1 *trans*-11	0.01 – 0.21	0.07	0.03	1	0.30	0.02
C18:1 *cis*-9	0.32 – 2.94	0.85	0.21	8	0.82	0.09
C18:2 *cis*,*cis*-9,12	0.02 – 0.25	0.07	0.02	7	0.65	0.01
C18:3	0.01 – 0.04	0.02	0.01	1	0.39	0.01
C18:2 *cis*,*trans*-9,11 (CLA)	0.01 – 0.06	0.02	0.01	7	0.37	0.01

[1]SD = Standard deviation; LV = number of latent variables; R^2CV = cross-validated coefficient of determination; RMSECV = root mean square error of cross-validation. Range, mean, SD and RMSECV are in units of g of FA/100 g of milk.

PAPER V concerns prediction of protein fractions and coagulation properties from FT-IR measurements using PLS as regression method. As already mentioned in **Chapter 3** a subset of 832 samples (426 Holstein and 406 Jersey) was included in the study and protein fractions were quantified using liquid chromatography. Milk coagulation properties, curd firming rate (CFR) and rennet coagulation time (RCT), were determined using RoeRox4 rheometer (MediRox AB, Nyköping, Sweden) as described by Jensen et al. (2012). The results of PLS models for individual protein fractions and coagulation properties are shown in Table 4.3. Also here the results are, to a large extent, in agreement with existing literature.

***Table 4.3**. Protein fractions and coagulation properties. Results from partial least squares regression models based on raw milk samples[1]*

Protein fraction or coagulation trait	Range	Mean	SD	LV	R^2CV	RMSECV
κ-CN	0.11 – 0.40	0.25	0.06	2	0.71	0.03
κ-CN G	0.01 – 0.15	0.05	0.02	4	0.20	0.02
α_{S1}-CN	0.50 – 1.64	1.05	0.18	2	0.66	0.11
α_{S1}-CN 8P	0.29 – 1.40	0.80	0.16	2	0.71	0.09
α_{S2}-CN	0.09 – 0.47	0.20	0.06	2	0.36	0.05
α_{S2}-CN 11P	0.05 – 0.32	0.13	0.04	2	0.47	0.03
β-CN	0.57 – 1.81	1.25	0.16	7	0.25	0.14
α-LA	0.02 – 0.20	0.11	0.02	4	0.06	0.02
β-LG	0.09 – 0.51	0.27	0.06	9	0.34	0.05
CFR	0 – 44.85	14.80	8.42	2	0.66	4.90
RCT	5.53 – 30.35	13.29	2.33	4	0.12	2.19

[1]SD = Standard deviation; LV = number of latent variables; R^2CV = cross-validated coefficient of determination; RMSECV = root mean square error of cross-validation; CFR = curd firming rate; RCT = rennet coagulation time. Range, mean, SD and RMSECV are in units of grams/100 grams of milk for protein fractions, units of Pa/min for CFR and units of min for RCT.

4.3 The Cage of Covariance

Infrared radiation is absorbed by exciting fundamental vibrations of molecular bonds expressing a change in the dipole moment. Therefore, in traditional spectroscopy FA would be divided into functional groups such as methane, methylene, methyl, ester, ether, olefinic, aliphatic and carboxylic groups with their own characteristic group frequencies. In the same way protein absorption would be viewed as amide A, amide B, amide I, amide II and amide III bands. Where the average proportions of these functional groups may be measured by FT-IR, it is not to be expected that individual FA nor protein fractions will show distinct absorption patterns in FT-IR measurements of a complex aqueous mixture like milk.

Even though Chapman (1965) investigated infrared absorption of pure lipids and found that differences in chain length of solid FA gave rise to minimal differences in the fingerprint region of the infrared spectra, these minimal differences are expected to disappear when FA are present in a complex liquid mixture like milk. Consider, for example, the vibrations of methylene groups in a given FA. The vibrations may be slightly affected by the distances to the carbonyl group (at least in

theory). Hence, individual methylene groups may vibrate with a slightly altered frequency depending on the position down the FA carbon chain. However, it is unlikely that these small differences can be assigned to a single given FA. Milk contains multiple different FA, all with a very similar structure. Therefore, it is very likely that the small vibrational differences between the methylene groups in, for example, C14:0 will be overlapped by the methylene groups in, for example, C12:0, C16:0 and C18:0. Hence, the signal from a given fatty acid will be covered by overlapping Gaussian shaped signals originating from multiple FA. Specific milk protein fractions are differentiated by the composition of their amino acid residues. The different amino acids are present in more or less all the protein fractions, though in varying relative contents (Table 4.4).

***Table 4.4**. Amino acid composition of major bovine milk proteins (mol/kg protein; Walstra et al., 2006)*

Component	MW	Casein	α_{s1}-Casein	α_{s2}-Casein	β-Casein	κ-Casein	α-Lactalbumin	β-Lactoglobulin
N	14.007	11.30	11.22	11.37	11.17	11.67	11.42	11.27
P	30.974	0.26	0.34	0.44	0.21	0.05	0.00	0.00
S	32.06	0.23	0.21	0.24	0.25	0.22	0.63	0.49
Glycine (Gly)	75.1	0.25	0.38	0.08	0.21	0.11	0.42	0.22
Alanine (Ala)	89.1	0.36	0.38	0.32	0.21	0.79	0.21	0.82
Valine (Val)	117.2	0.62	0.47	0.55	0.79	0.58	0.42	0.49
Leucine (Leu)	131.2	0.74	0.72	0.52	0.92	0.42	0.92	1.20
Isoleucine (Ile)	131.2	0.47	0.47	0.44	0.42	0.68	0.56	0.55
Proline (Pro)	115.1	1.02	0.72	0.40	1.46	1.05	0.14	0.44
Phenylalanine (Phe)	165.2	0.33	0.34	0.24	0.38	0.21	0.28	0.22
Tyrosine (Tyr)	181.2	0.34	0.42	0.48	0.17	0.47	0.28	0.22
Tryptophan (Trp)	204.2	0.06	0.08	0.08	0.04	0.05	0.28	0.11
Serine (Ser)	105.1	0.68	0.68	0.67	0.67	0.68	0.49	0.38
Threonine (Thr)	119.1	0.38	0.21	0.59	0.38	0.74	0.49	0.44
Cysteine (Cys)	121.2	0.02	0.00	0.08	0.00	0.11	0.56	0.27
Methionine (Met)	149.2	0.21	0.21	0.16	0.25	0.11	0.07	0.22
Arginine (Arg)	174.2	0.22	0.25	0.24	0.17	0.26	0.07	0.16
Histidine (His)	155.2	0.19	0.21	0.12	0.21	0.16	0.21	0.11
Lysine (Lys)	146.2	0.56	0.59	0.95	0.46	0.47	0.85	0.82
Asparagine (Asn)	132.1	0.31	0.34	0.55	0.21	0.42	0.56	0.27
Aspartic acid (Asp)	133.1	0.22	0.30	0.16	0.17	0.16	0.92	0.55
Glutamine (Gln)	146.1	0.74	0.59	0.63	0.83	0.74	0.42	0.49
Glutamic acid (Glu)	147.1	0.87	1.06	0.95	0.79	0.68	0.49	0.88
g protein per gram of N[a]		6.32	6.36	6.28	6.39	6.12	6.25	6.34
MW protein[a]		23192	23618	25231	23986	19026	14181	18282

[a] Calculated from the amino acid sequence of each protein, including (organic) phosphate but not carbohydrate groups.

Moreover, the FT-IR active groups in the amino acids are more or less the same (Figure 4.5). Therefore, it might also be very difficult to obtain unique FT-IR signals for specific protein fractions in milk.

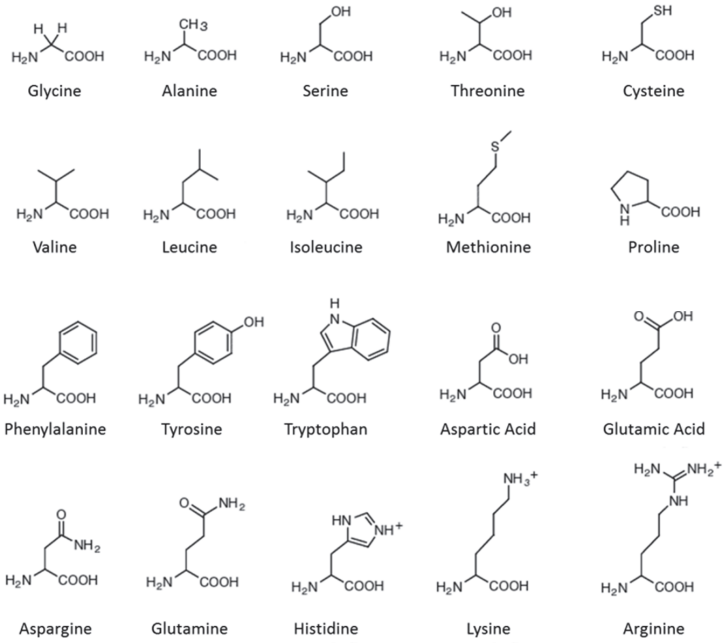


***Figure 4.5**. Overview of amino acids. Modified from (Walstra et al., 2006)*

If a given FA is to be estimated independently of other FA, then the regression vector for the FA of interest has to be orthogonal to the signals of all interferents (including all other FA), as outlined in **Chapter 2.2**. If the pure FA signals are completely overlapping in the space spanned by FT-IR measurements, unique regressions vectors cannot be obtained for individual FA and the estimations of individual FA will depend on other FA. The same issues occur for the overlapping signals of specific protein fractions.

If no unique signals for individual FA or protein fractions exist, unique models cannot be built and the prediction of the detailed milk parameters need to rely on indirect covariance structures in the calibration set. Total fat content in the milk is made up by the sum of all FA. In the same way, total protein content is made up by the sum of all protein fractions. Hence, total fat content and total protein content can (by IR measurements) be estimated from the summed signals of all FA and all protein fractions, respectively. For natural reasons, the concentrations of FA will correlate with the concentration of total fat and the concentrations of protein fractions will correlate with the concentration of total protein. Figure 4.6 shows a heat map of the concentration profile correlations (R^2). Figure 4.6a shows how FA correlate in the concentration profiles (reference values) and Figure 4.6b shows how protein fractions correlate in the concentration profiles (reference values). From Figure 4.6a it is found that a fairly large part of the FA correlates very well with total fat content in the concentration profiles. The protein fractions and coagulation properties show less cor-

relation with total protein content. However, the casein fractions correlate in general better with total protein than the whey fractions (Figure 4.6b).

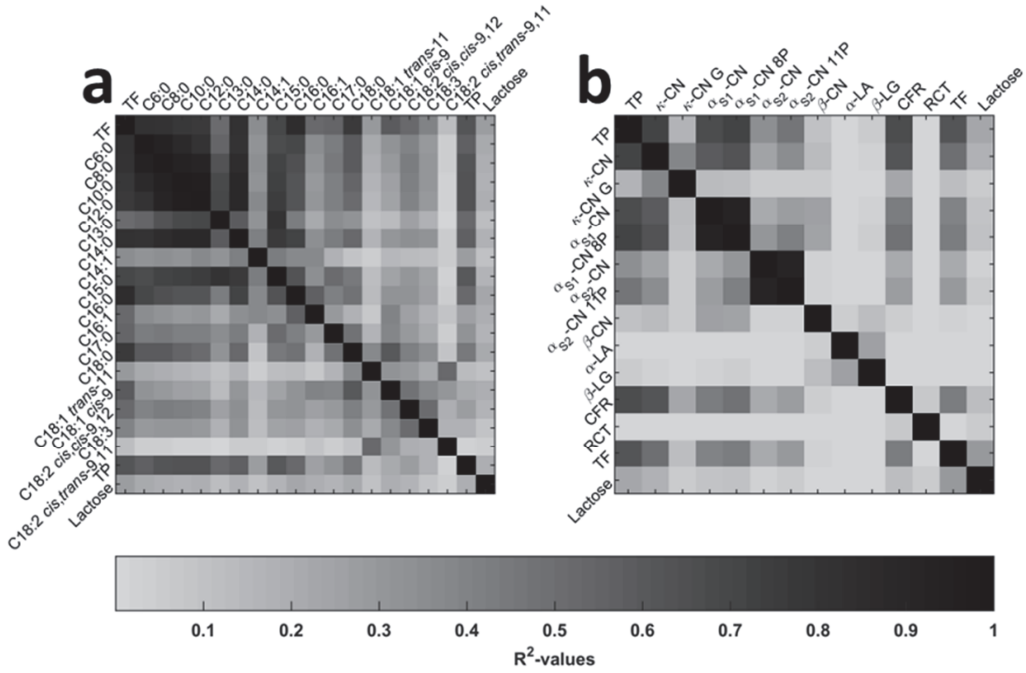


***Figure 4.6**. Heat maps showing correlations expressed as R^2-values. (a) reference values (concentrations) of fatty acids. (b) reference values (concentrations) of protein fractions. TF = total fat content; TP = total protein content.*

For obvious reasons, the major FA (like C16:0) tend to correlate better with total fat content. In the same way, the major protein fractions (casein fractions) tend to correlate better with total protein content. From this it follows that the major FA and the major protein fractions in general are "estimated" best, in a *cage of covariance*, from IR measurements. This is also what is reported in existing literature (De Marchi et al., 2014) and found in **PAPER IV** where, for example, C16:0 appears to be accurately estimated (Table 4.2) and in **PAPER V** where the casein fractions appear to be better estimated than the whey fractions (Table 4.3).

Moreover, other factors like breed, which may give specific absorption patterns in the IR spectra may also be used in predicting phenotypes. Figure 4.8 shows the prediction of CFR from IR measurements. If the CFR predictions are considered for the individual breeds separately, then the predictions are very poor. However, it appears that the Jersey samples in general are having higher CFR than the Holstein samples. Hence, when the CFR predictions are considered for the two breeds together, then the model appears to perform moderate. Nevertheless, the model is clearly providing information related to the two breeds rather than information related to CFR (Figure 4.7).

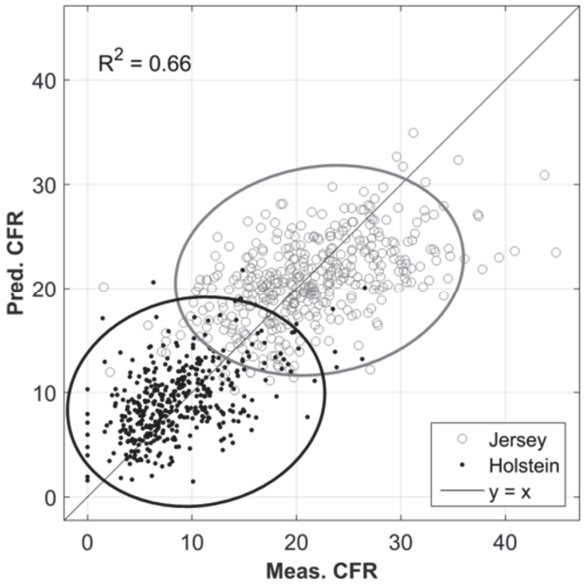


***Figure 4.7**. Prediction of curd firming rate (CFR) from infrared measurements.*

The direct (independent of major milk constituents and other factors) and the indirect (dependent of major milk constituents and other factors) prediction of a given phenotype is shown schematically in Figure 4.8.

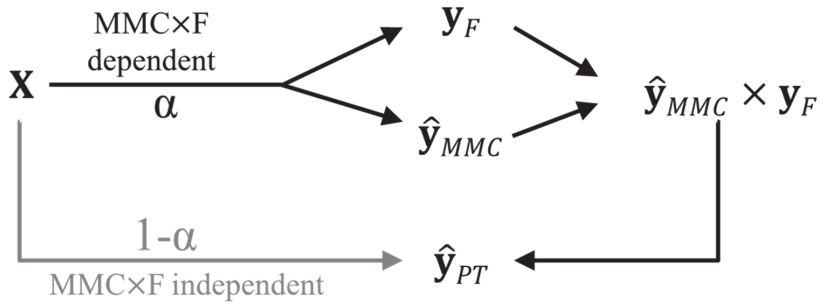


***Figure 4.8**. Prediction of a given phenotypic trait, $\hat{\mathbf{y}}_{PT}$ from infrared measurements, $\mathbf{X}$; $\hat{\mathbf{y}}_{PT}$ is partly described by variation dependent and independent on major milk constituents, $\hat{\mathbf{y}}_{MMC}$ and other factors, $\mathbf{y}_F$ (like breed) giving rise to specific absorption patterns in $\mathbf{X}$. MMC = major milk constituents; F = other factors (like breed) impacting the spectra.*

If predictions of a given phenotypic trait are indirect, then predictions of future samples are trapped in a *cage of covariance*, where initial covariance structures between the major milk constituents, factors like breed and the phenotypic trait determine the variation in the predicted values of the phenotypic trait.

The problem of detailed milk composition being modeled in an indirect fashion relates back to what was touched upon on **Chapter 2.2** and **Chapter 2.4**. The aim of establishing a calibration model is to use the model for predicting, for example, an individual FA based on the IR spectrum obtained on milk. As already discussed, calibration models use specific relationships between the analyte (i.e. FA) concentration and the measured spectrum in order to quantify the analyte (i.e. FA). Hence, if, for example, the relationship between individual FA and total fat content is used for calibration of a model, then this model will not be valid for samples with an altered FA profile. This is shown in Figure 4.9, where a model for predicting C14:0 is calibrated on the Jersey samples and tested on the Holstein samples. Figure 4.9a shows that the relationship between C14:0 and total fat content is different for the two breeds. As the calibration model is based on the relationship between C14:0 and total fat content for the Jersey samples (the *cage of covariance*), this model will not be valid for the Holstein samples, as the Holstein samples express another covariance structure. Figure 4.9b shows that the Jersey samples (the calibration set) are well predicted and that the Holstein samples (the test set) shows a bias. Hence, the model is not valid for the Holstein samples.

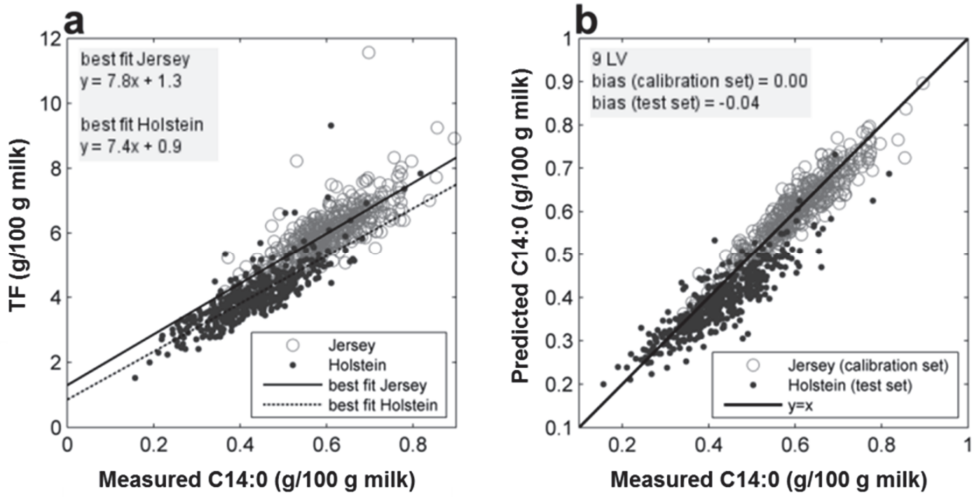


***Figure 4.9**. (a) Relationship between measured C14:0 and total fat content (TF) in raw milk samples from Jersey and Holstein cows. (b) measured versus predicted values of C14:0; C14:0 was predicted by partial least squares (PLS) regression applied to Fourier transform infrared measurements. The PLS model was calibrated on Jersey samples and tested with Holstein samples. LV = latent variables.*

Indirect relationships for calibrating e.g. PLS regression models is not a problem in itself, but may be problematic in terms of prediction accuracy and calibration robustness. If indirect relationships are used to calibrate a regression model, the model will not be valid for future samples unless the indirect relationships are conserved in the new samples. If such indirect relationships are not conserved, the model will provide incorrect predictions as shown in both **PAPER IV** and **PAPER V**. The

indirect relationships may change with factors like breed, lactation stage, feed, etc. Therefore, it is not given that indirect covariance structures are conserved in a future sample set.

4.3.1 Diagnostics – The Cage of Covariance

It is difficult to determine whether PLS models are based on direct or indirect relationships and it will be convenient to have some diagnostics indicating this. As highlighted in Figure 2.9 from **Chapter 2.2**, the multicomponent sample spectra are composed of the outer product of the analyte concentration profiles and the pure analyte signals. A *cage of covariance* may be established when the pure analyte signals are completely overlapping and the concentration profiles are well correlated. Hence, overlapping signals force indirect models and covariance between analyte concentrations enables good predictions from these indirect models.

As also presented in **PAPER IV** and **PAPER V** a heat map of R^2-values between the reference values and the (preprocessed) spectra will give an immediate and very intuitive feeling whether the phenotypes are predicted independently or not. These heat maps are shown in Figure 4.10. Figure 4.10a is the heat map for protein fraction and coagulation properties. The spectra used for Figure 4.10a are preprocessed by first derivative. Figure 4.10b is the heat map for individual FA. Here preprocessing of the spectra did not have any impact on model performance and therefore, the spectra used in Figure 4.10b are the raw spectra.

From Figure 4.10a it is found that total protein and all the protein fractions (including coagulation properties) are having a similar correlation pattern with the spectral profile. This could indicate that the protein fractions are modeled by the same wavenumbers as total protein and in turn this could give problems with indirect models. From Figure 4.10b a similar trend is found. Here the correlation pattern for individual FA is very similar to the pattern observed for total fat content. Again this could give an indication that the FA models are not independent. Note the difference in correlation pattern between Figure 4.10a and 4.10b. In Figure 4.10a, relatively high correlations are observed between 1,600 cm^{-1} and 1,500 cm^{-1} (amide II region). Where in Figure 4.10b, this region are having poor correlations.

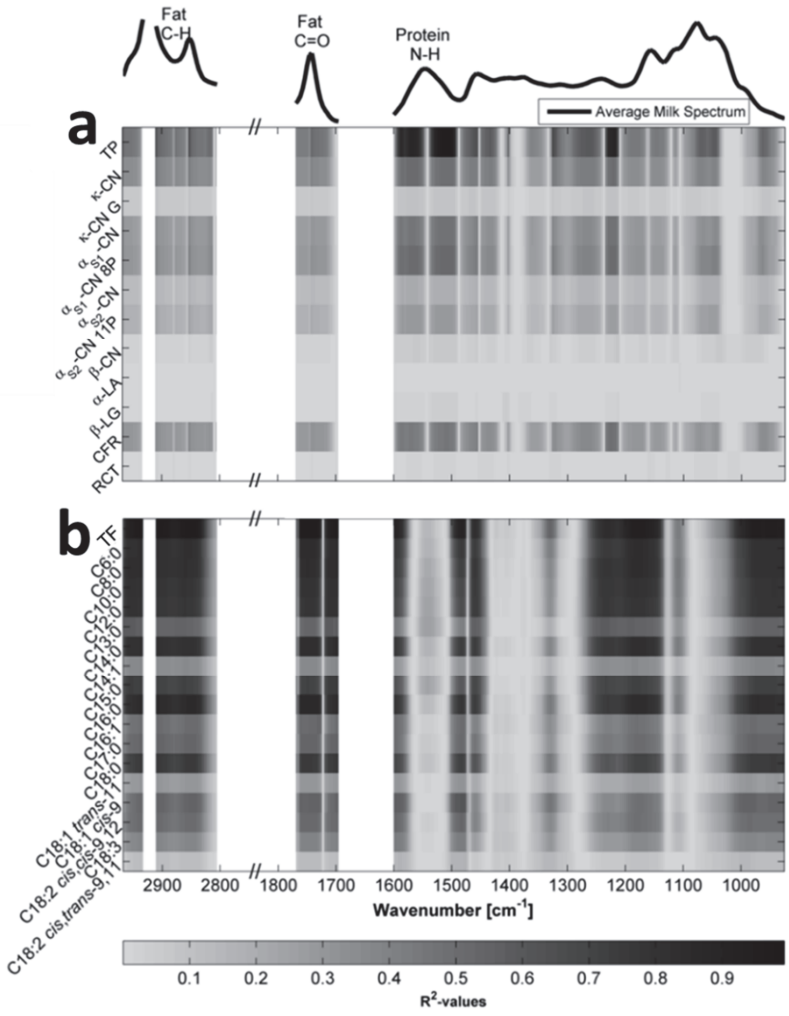


***FIGURE 4.10**. heat map of R^2-values between the reference values and the (preprocessed) IR spectra of milk. (a) protein fractions and coagulation properties. (b) fatty acids. The IR spectra are preprocessed by first derivative in (a) and raw in (b). TP = total protein; CFR = curd firming rate; RCT = rennet coagulation time; TF = total fat content*

As mentioned in **Chapter 2.2**, it is important to consider the *chemical rank* of the spectra before modeling a high number of different phenotypes from IR measurements. Here, PCA is used to give an indication of the *chemical rank*. A PCA using nine LV describes 99.9 % of the total variation in the spectra (Figure 4.11). Hence, around nine independent components (or parameters) can be extracted from the IR measurements. A total of 17 individual FA (Table 4.2) and nine protein fractions (Table 4.3) were available in **PAPER IV** and **PAPER V**. Based on Figure 4.11, it seems unlikely that all these phenotypes can be predicted independently by the IR spectra. The *chemical rank* of

the spectra is simply too small. Therefore, some models are most likely based on indirect relationships.

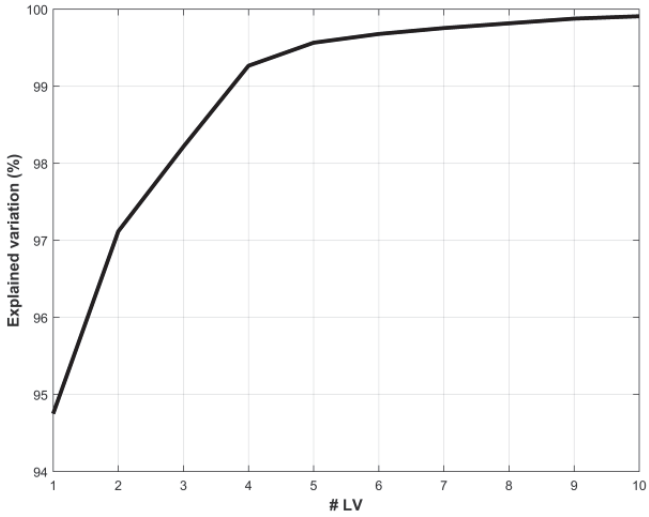


***Figure 4.11**. Output from principal component analysis on Fourier transform infrared measurements. Explained variation as a function of number of latent variables (LV)*

When building PLS calibration models proper validation is important, as outlined in **Chapter 2.2**. A simple cross-validation would in this case not be sufficient. The problem of cross-validation is that the data most often originates from one original data set. Hence, the same covariance structures are likely to be present in all samples and the *cage of covariance* will thereby not show up in the cross-validation output. However, a real test set validation, where the model is calibrated on one set and tested with another (truly) independent set consisting of samples from other breeds, lactation stage, feeding systems, etc., would most likely contain different covariance structures and thereby highlight model inadequacies. In Figure 4.9 the PLS model was calibrated with the Jersey samples and tested with the Holstein samples. The covariance structures used for calibration are valid for the Jersey samples but not the Holstein samples, which show biased predictions. In this way a test set validation was used to make the *cage of covariance* visual.

Observing how the covariance structures among the dependent variables (i.e. phenotypes) change when going from measured (i.e. reference) values to predicted values may be very powerful in visualizing the *cage of covariance*. Here this is first shown by a small simulated data set (no background or error terms are included) and later with the data set of **PAPER IV** and **PAPER V**.

The simulated data set consists of 50 samples and three analytes of interest, $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$. All analytes are mean centered. The three analytes (the reference data) are plotted against each other in Figure 4.12. The R^2 between $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ is 0.70 (Figure 4.12a), R^2 between $\boldsymbol{a}_1$ and $\boldsymbol{a}_3$ is 0.60 (Figure 4.12b) and R^2 between $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$ is 0.40 (Figure 4.12c).

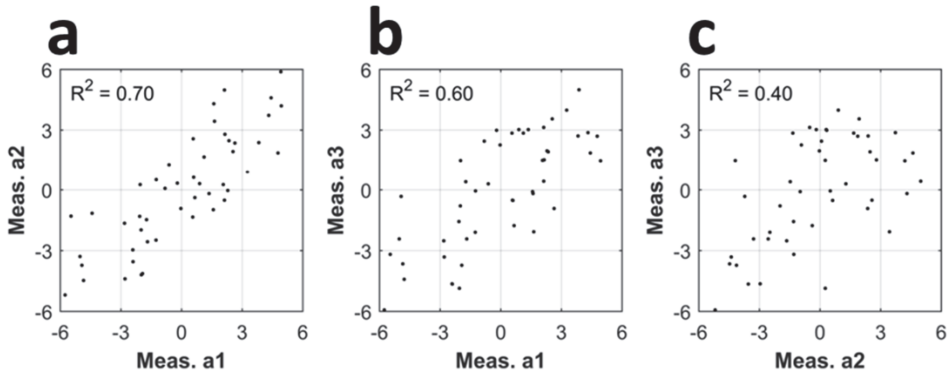


***Figure 4.12**. Reference variables. (a) correlation between $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$. (b) correlation between $\boldsymbol{a}_1$ and $\boldsymbol{a}_3$. (c) correlation between $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$. Simulated data*

The predictor data (i.e. the spectra) are plotted in Figure 4.13a. The predictor data consist of 100 mean centered variables. While $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are having a direct (and noiseless) relationship between the reference data and the predictor data, information related to $\boldsymbol{a}_3$ is not directly found in the predictor data. Hence, predictions of $\boldsymbol{a}_3$ rely on covariance with $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$. It is expected that that the predictions of $\boldsymbol{a}_3$ primarily rely on covariance with $\boldsymbol{a}_1$, as the correlation here is better (Figure 4.12).

Individual PLS models were built between predictor data and each of the three analytes. All PLS models were based on two LV (the rank of the predictor data). The regression vectors obtained for the three analytes are presented in Figure 4.13b. The regression vectors suggest that $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are predicted independently of each other, whereas $\boldsymbol{a}_3$ to a large extent is predicted by a linear combination of the $\boldsymbol{a}_1$-signal. Hence, the predicted values of $\boldsymbol{a}_3$ will be trapped in a *cage of covariance* with $\boldsymbol{a}_1$ and the predicted values of $\boldsymbol{a}_1$ and $\boldsymbol{a}_3$ will have a similar variation (i.e. they will be correlated).

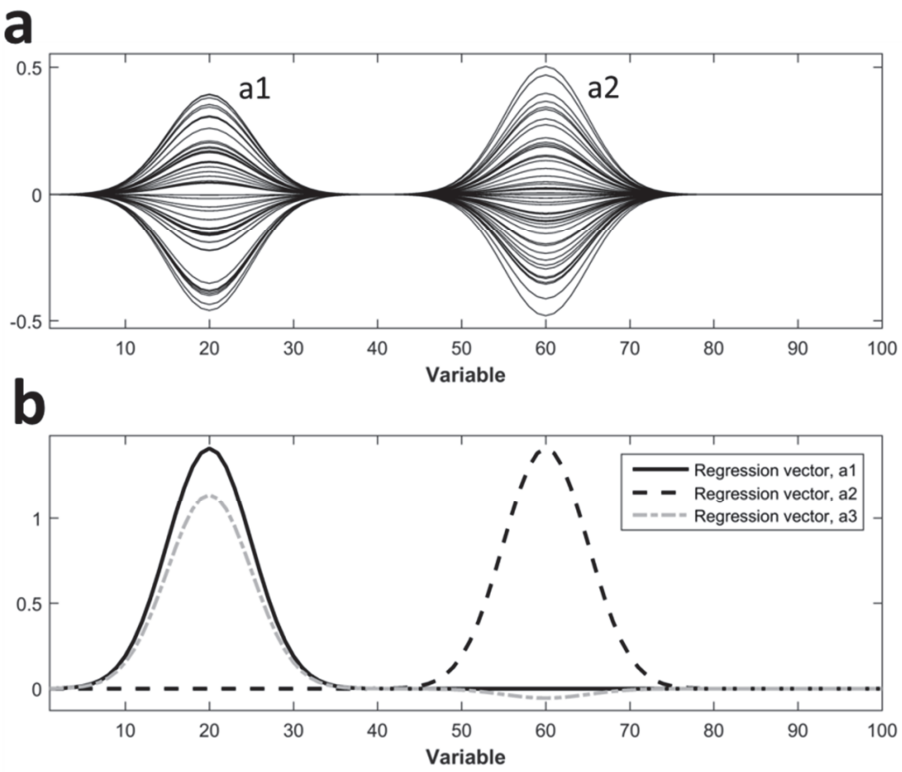


***Figure 4.13**. (a) predictor data. Information related to the two analytes, $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ is found, whereas information related to analyte $\boldsymbol{a}_3$ is not present. (b) Regression vectors for the three analytes. Simulated data.*

By examining the model performance (Figure 4.14) it is found that $\boldsymbol{a}_1$ (Figure 4.14a) and $\boldsymbol{a}_2$ (Figure 4.14b) are perfectly predicted. However, this is not very interesting. What is interesting (but not surprising) is that the model performance for $\boldsymbol{a}_3$ (Figure 4.14c) corresponds to the correlation between $\boldsymbol{a}_1$ and $\boldsymbol{a}_3$ in the raw data (Figure 4.12b).

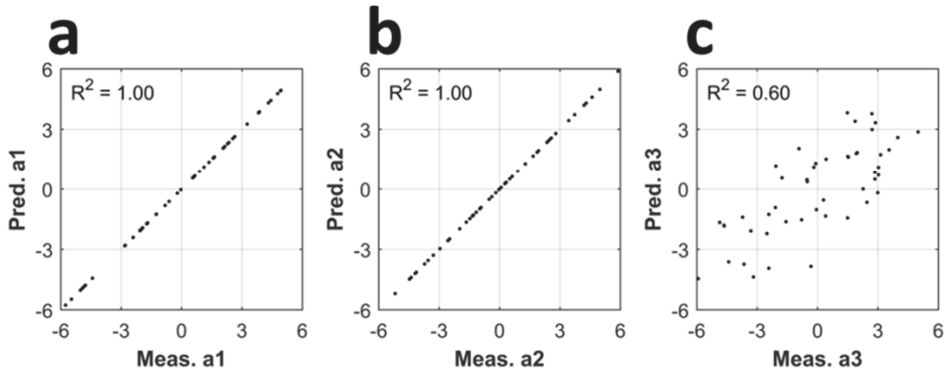


***Figure 4.14**. Partial least squares regression model performance. (a) measured vs. predicted $\boldsymbol{a}_1$. (b) measured vs. predicted $\boldsymbol{a}_2$. (c) measured vs. predicted $\boldsymbol{a}_3$. Simulated data.*

If two analytes are predicted by the same linear combination of the predictor data, then the predictions of the two analytes will be correlated, regardless of how they correlate in the raw data. On the other hand, if two analytes are predicted independently from the predictor data (and the models are good), then the predictions of the two analytes will show the same correlation as the analytes were showing in the raw data. This can be used to show indirect relationships.

In Figure 4.15 the correlations are examined between predicted $\boldsymbol{a}_1$ and measured $\boldsymbol{a}_1$ (Figure 4.15a), predicted $\boldsymbol{a}_1$ and measured $\boldsymbol{a}_2$ (Figure 4.15b) and predicted $\boldsymbol{a}_1$ and measured $\boldsymbol{a}_3$ (Figure 4.15c). This shows how the measured values of the analytes correlates with the predicted values of $\boldsymbol{a}_1$. These pattern should then be compared with how the predicted values of the analytes correlates with the predicted values of $\boldsymbol{a}_1$. This is done in Figure 4.15d for the predicted values of $\boldsymbol{a}_1$, in Figure 4.15e for the predicted values of $\boldsymbol{a}_2$ and in Figure 4.15f for the predicted values of $\boldsymbol{a}_3$. It is found that the correlation between predicted $\boldsymbol{a}_1$ and measured $\boldsymbol{a}_2$ (Figure 4.15b) is similar to the correlation between predicted $\boldsymbol{a}_1$ and predicted $\boldsymbol{a}_2$ (Figure 4.15e). This is because $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are predicted independently of each other, as also shown by the regression vectors (Figure 4.13b). However, when comparing predicted $\boldsymbol{a}_1$ and measured $\boldsymbol{a}_3$ (Figure 4.15c) with predicted $\boldsymbol{a}_1$ and predicted $\boldsymbol{a}_3$(Figure 4.15f) it is found that the correlation increases. This is because $\boldsymbol{a}_1$ and $\boldsymbol{a}_3$ are modeled by the same linear combination of the predictor variables. The predictions of $\boldsymbol{a}_3$ are trapped in a *cage of covariance* with the predictions of $\boldsymbol{a}_1$. Hence, the variation in the predicted values of $\boldsymbol{a}_3$ is determined by the variation in $\boldsymbol{a}_1$. This is of course not ideal, if the purpose of the calibration model is to obtain information related to $\boldsymbol{a}_3$. Furthermore, if the correlation between $\boldsymbol{a}_1$ and $\boldsymbol{a}_3$ is broken in a future samples set, then the calibration model built here will obviously not be valid anymore.

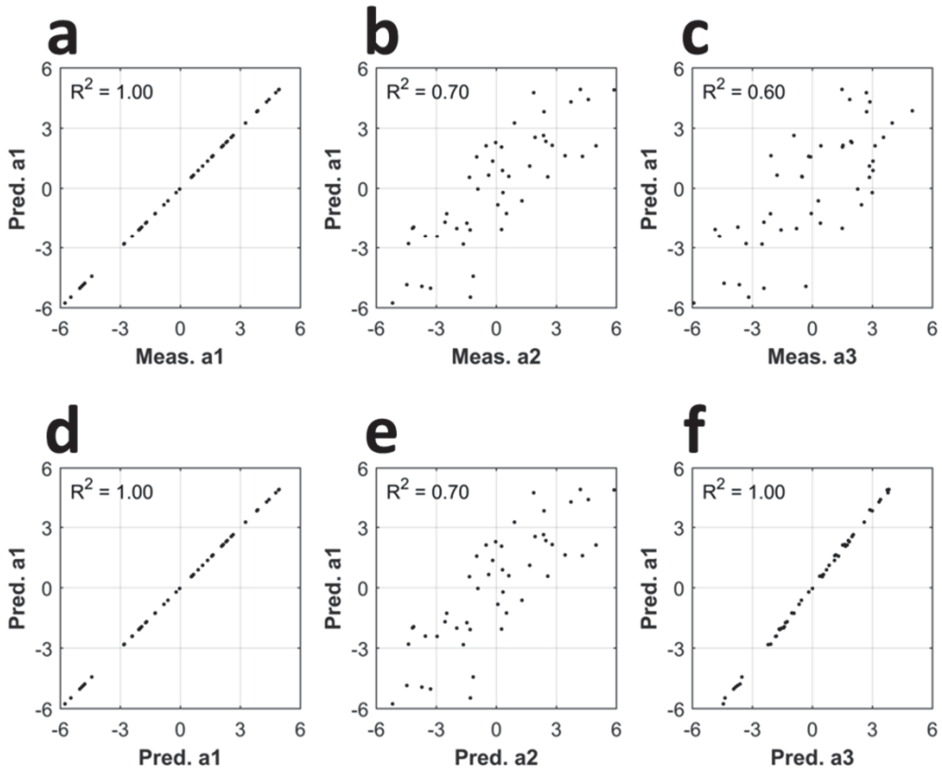


***Figure 4.15**. (a-c) correlation structures between predicted $\boldsymbol{a}_1$ and measured $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$. (d-e) correlation structures between predicted $\boldsymbol{a}_1$ and predicted $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$. Simulated data.*

For simplicity the information found in Figure 4.14 and Figure 4.15 are plotted together in a single plot (Figure 4.16). The model performance (i.e. information from Figure 4.14a-c) is plotted on the x-axis. The information found in Figure 4.15a-c is plotted on the y-axis as open squares and the information found in Figure 4.15d-f is plotted on the y-axis as filled triangles.

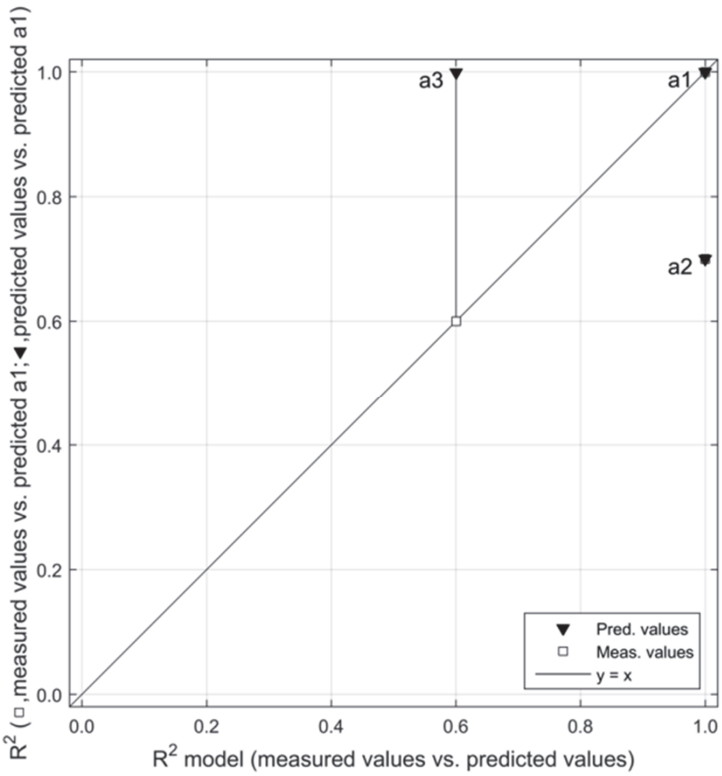


***Figure 4.16**. Partial least squares regression model performance (x-axis) and correlation with $\boldsymbol{a}_1$ for measured and predicted analyte values, respectively (y-axis). Simulated data*

The increase on the y-axis for $\boldsymbol{a}_3$ when going from measured values to predicted values (Figure 4.16) clearly indicates that $\boldsymbol{a}_1$ and $\boldsymbol{a}_3$ are modeled by the same linear combination, whereas the *steady-state* for $\boldsymbol{a}_2$ indicates that it is modeled independently of $\boldsymbol{a}_1$.

Figure 4.17a shows the results when comparing the FA with total fat content (**PAPER IV**). Note that in Figure 4.17a, total protein content and lactose are included. In Figure 4.17a total protein and lactose have the same relationship with total fat for both measured and predicted values. All the FA exhibit increased R^2-values with total fat content when going from measured to predicted values. This clearly indicates that total protein and lactose are modeled independent of total fat content, whereas the FA, to a large extent, are modeled by the same linear combination as total fat content (Figure 4.17a). Figure 4.17b tells the same story but for protein fractions (**PAPER V**). Note that in Figure 4.17b total fat content and lactose are included. In Figure 4.17b, total fat content and lactose have the same relationship with total protein content for both measured and predicted values. All protein fractions and coagulation properties exhibit increased R^2-values with total protein content when going from measured to predicted values. Hence, total fat content and lac-

tose are modeled independent of total protein, whereas all the protein fractions and coagulation properties are modeled more or less by the same linear combination as total protein content. Figure 4.17 is a very strong indication that models for individual FA are trapped in a *cage of covariance* with total fat and that protein fractions are trapped in a *cage of covariance* with total protein.

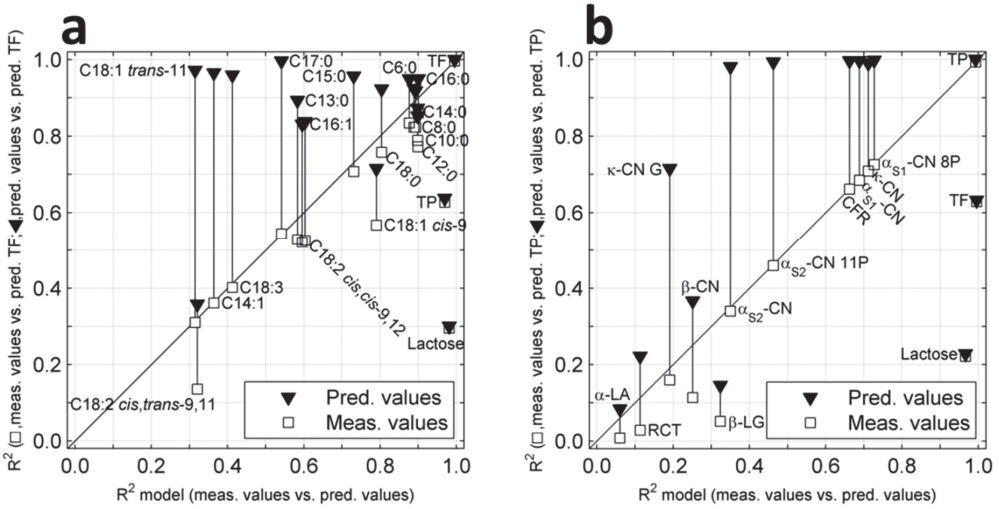


***Figure 4.17**. Partial least squares regression model performance (x-axis) and correlation with measured and predicted phenotype, respectively (y-axis). (a) Fatty acids Profile. (b) Protein composition. TP = total protein; TF = total fat content.*

In **PAPER IV** a projecting procedure was proposed for visualizing how much common variation is found in the FA predictions and the interaction of breed and total fat content. In short, PLS models for individual FA were fitted (Equation 4.1).

$$\hat{\boldsymbol{y}}_{\boldsymbol{FA}} = \boldsymbol{X} \cdot \hat{\boldsymbol{b}}_{\boldsymbol{FA}} \qquad \textit{Equation 4.1}$$

Where $\hat{\boldsymbol{y}}_{FA}$ is the estimated concentrations of a given FA, $\boldsymbol{X}$ are the preprocessed and centered FT-IR spectra and $\hat{\boldsymbol{b}}_{TF}$ is the estimated regression vector. The Idea was to split $\hat{\boldsymbol{y}}_{FA}$ into a part described by the interaction of total fat content and breed, $\hat{\boldsymbol{y}}_{FA_{TF\times Breed}}$ and a part orthogonal to the interaction, $\hat{\boldsymbol{y}}_{FA_{-TF\times Breed}}$ (Equation 4.2).

$$\hat{\boldsymbol{y}}_{\boldsymbol{FA}} = \hat{\boldsymbol{y}}_{\boldsymbol{FA_{TF\times Breed}}} + \hat{\boldsymbol{y}}_{\boldsymbol{FA_{-TF\times Breed}}} \qquad \textit{Equation 4.2}$$

This was done by introducing a split on $\boldsymbol{X}$. Using the matrix, $\boldsymbol{V}$ spanning the interaction of predicted total fat content, $\hat{\boldsymbol{y}}_{TF}$ and breed (**PAPER IV**). $\boldsymbol{X}$ was first projected onto $\boldsymbol{V}$ (Equation 4.3) to ob-

tain the part of $\boldsymbol{X}$ explained by the interaction, $\boldsymbol{X}_{TF\times Breed}$. This part was then subtracted the original $\boldsymbol{X}$ to obtain the part orthogonal to the interaction, $\boldsymbol{X}_{-TF\times Breed}$ (Equation 4.4).

$$\boldsymbol{X_{TF\times Breed} = V(V^T \cdot V)^{-1}V^T \cdot X} \qquad \textbf{Equation 4.3}$$

$$\boldsymbol{X_{-TF\times Breed} = X - X_{TF\times Breed}} \qquad \textbf{Equation 4.4}$$

By combining Equation 4.1 and Equation 4.4 the following is reached,

$$\boldsymbol{\hat{y}_{FA} = X \cdot \hat{b}_{FA} = X_{-TF\times Breed} \cdot \hat{b}_{FA} + X_{TF\times Breed} \cdot \hat{b}_{FA}} \qquad \textbf{Equation 4.5}$$

In Equation 4.5 the prediction of the fatty acid is split into a part related to total fat content and breed ($\boldsymbol{X}_{TF\times Breed} \cdot \hat{\boldsymbol{b}}_{FA}$) and a part unrelated to total fat content ($\boldsymbol{X}_{-TF\times Breed} \cdot \hat{\boldsymbol{b}}_{FA}$). The contribution of each part in the total FA variation is visualized by comparing their sum of squares (Figure 4.18).

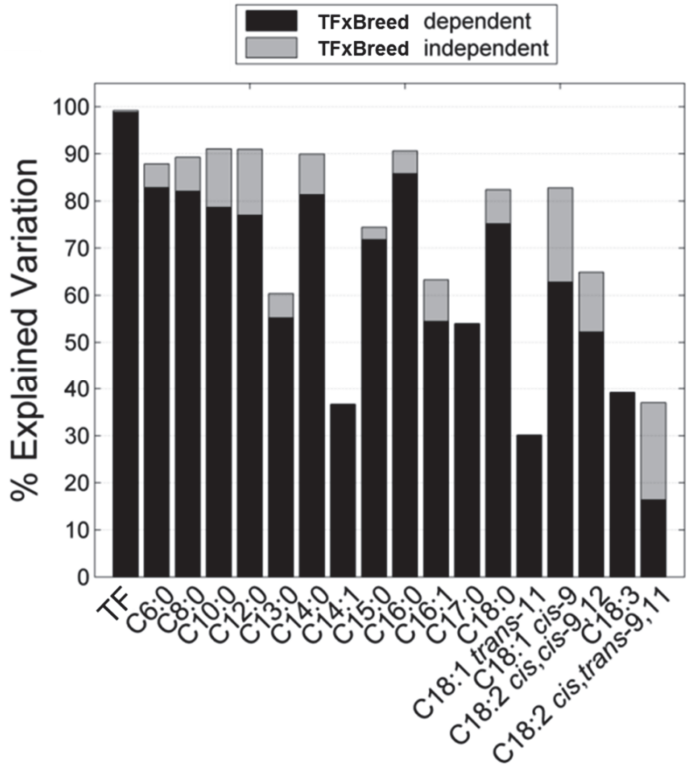


***Figure 4.18**. Prediction of FA by partial least squares regression applied to infrared measurements. The percent of explained variation is divided into a part related with interaction of total fat content and breed (TFxBreed) and a part unrelated to this interaction. TF = total fat content*

Figure 4.18 shows in a very visual way that the FA predictions contain a large amount of common variation with the interaction between breed and total fat content. However, it must be stressed that Figure 4.18 only visualizes the amount of common variation that goes into the predictions and not exactly model dependencies. The correlations between pure analyte signals are not considered in the results presented in Figure 4.18. This fact should definitely have been stressed more or given higher impact in **PAPER IV**. When applying this projection procedure it is recommended to accompany Figure 4.18 with, for example, Figure 4.17a that shows model dependencies. Otherwise the results may be over interpreted.

In order to obtain a better estimate of dependencies between models, a new projection/orthogonalization based method has been developed, which is believed to be more efficient than the projection method outlined in **PAPER IV**. The method presented in **PAPER IV** and the new method are in fact very similar but they differ in the way explained variance in the dependent part and the independent part are compared.

Consider Figure 4.19. Here, the pure analyte signal at unitary concentration is given by $\boldsymbol{a}$. Two interferents are present and their pure signals are given by $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$, respectively. The regression vector for $\boldsymbol{a}$, is orthogonal to both $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$, and is estimated by $\widehat{\boldsymbol{b}}$. The vector $\boldsymbol{a}_{-\boldsymbol{k}}$ is the part of $\boldsymbol{a}$ orthogonal to $\boldsymbol{k}_1$. Hence, $\boldsymbol{a}_{-\boldsymbol{k}}$ is in the null space of $\boldsymbol{k}_1$, $\boldsymbol{N}(\boldsymbol{k}_1)$. Furthermore, the vectors $\boldsymbol{k}_2$ and $\widehat{\boldsymbol{b}}$ are also in $\boldsymbol{N}(\boldsymbol{k}_1)$. It follows that the magnitude of $\boldsymbol{a}$ projected onto $\widehat{\boldsymbol{b}}$ is identical to the magnitude of $\boldsymbol{a}_{-\boldsymbol{k}}$ projected onto $\widehat{\boldsymbol{b}}$, as indicated by the dashed lines in Figure 4.19.

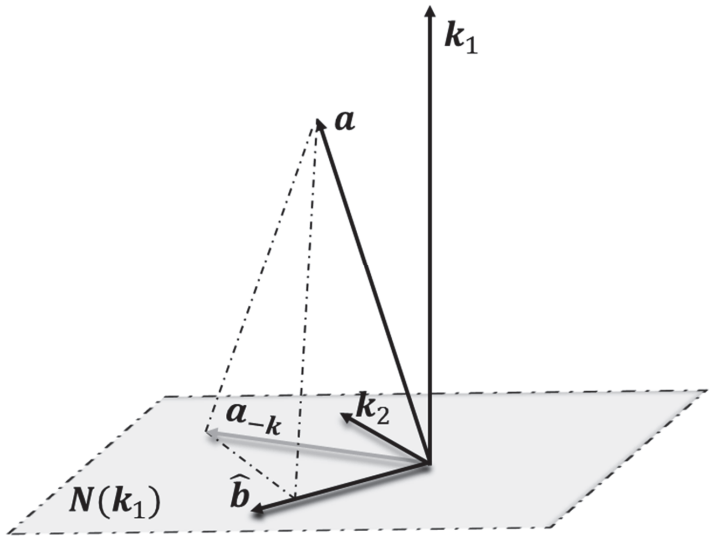


Figure 4.19*. Regression vector, $\widehat{\boldsymbol{b}}$ for analyte, $\boldsymbol{a}$. The regression vector is orthogonal to both interferents, $\boldsymbol{k}_1$and $\boldsymbol{k}_2$. $\boldsymbol{N}(\boldsymbol{k}_1) = null\ space\ of\ \boldsymbol{k}_1$.*

In Equation 2.16 (**Chapter 2.2)** it was stated that the inner product of $\boldsymbol{a}$ and $\widehat{\boldsymbol{b}}$ is equal to one. If this is true, then the inner product between $\boldsymbol{a_{-k}}$ and $\widehat{\boldsymbol{b}}$ is also equal to one. Hence, $\widehat{\boldsymbol{b}}$ has predictive power for concentrations of both $\boldsymbol{a}$ and $\boldsymbol{a_{-k}}$.

Nevertheless, if $\widehat{\boldsymbol{b}}$ is non-orthogonal to $\boldsymbol{k}_1$ (as illustrated in Figure 4.20), then $\widehat{\boldsymbol{b}}$ is not the true regression vector of $\boldsymbol{a}$, and estimated concentraions of $\boldsymbol{a}$ will also depend on $\boldsymbol{k}_1$ (outlined in **Chapter 2.2**). The projection of $\boldsymbol{a}$ onto $\boldsymbol{N}(\boldsymbol{k}_1)$ is still given by $\boldsymbol{a_{-k}}$. However, as $\widehat{\boldsymbol{b}}$ and $\boldsymbol{a}$ are both outside $\boldsymbol{N}(\boldsymbol{k}_1)$, then the magnitude of $\boldsymbol{a}$ projected onto $\widehat{\boldsymbol{b}}$ is larger than the magnitude of $\boldsymbol{a_{-k}}$ projected onto $\widehat{\boldsymbol{b}}$. The inner product of $\boldsymbol{a}$ and $\widehat{\boldsymbol{b}}$ is still one and therefore, in Figure 4.20 where $\boldsymbol{k}_1$ and $\widehat{\boldsymbol{b}}$ are non-orthogonal, the inner product of $\boldsymbol{a_{-k}}$ and $\widehat{\boldsymbol{b}}$ will be less than one. Hence, $\widehat{\boldsymbol{b}}$ will have poor predictive relevance for $\boldsymbol{a_{-k}}$. This fact can be used to highlight model dependencies on $\boldsymbol{k}_1$.

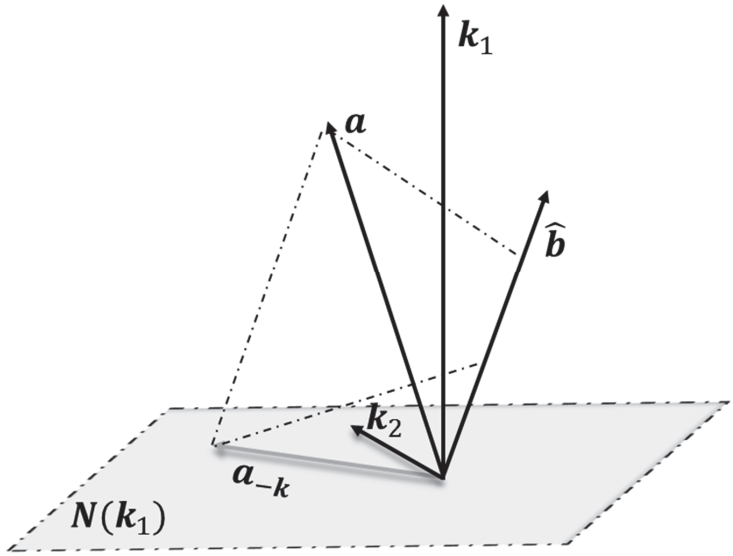


Figure 4.20*. Regression vector,* $\widehat{\mathbf{b}}$ *for analyte,* $\mathbf{a}$*. The regression vector is non-orthogonal to interferent,* $\boldsymbol{k}_1$*and orthogonal to* $\boldsymbol{k}_2$*.* $\boldsymbol{N}(\boldsymbol{k}_1) = null\ space\ of\ \boldsymbol{k}_1$.

The regression vector is the translation of the multicomponent spectral space, $\boldsymbol{X}$ into the analyte concentration space, $\boldsymbol{y_a}$. Hence, a PLS model can be fitted for a given analyte, $\boldsymbol{y_a}$ and the regression vector, $\widehat{\boldsymbol{b}}$ can be estimated. This is shown in Equation 4.6.

$$\widehat{\boldsymbol{y}}_{\boldsymbol{a}} = \boldsymbol{X} \cdot \widehat{\boldsymbol{b}} \qquad \textit{Equation 4.6}$$

Where $\widehat{\boldsymbol{y}}_{\boldsymbol{a}}$ is a good model estimate of $\boldsymbol{y_a}$. If the model is independent of an interferent, $\boldsymbol{k}$ then Equation 4.7 will give a good estimate of $\boldsymbol{y_{a-k}}$ (i.e. the part of $\boldsymbol{y}$ orthogonal to $\boldsymbol{k}$), as illustrated in Figure 4.19.

$$\hat{\boldsymbol{y}}_{\boldsymbol{a-k}} = \boldsymbol{X}_{\boldsymbol{-k}} \cdot \hat{\boldsymbol{b}} \qquad \textit{Equation 4.7}$$

Where $\boldsymbol{X}_{\boldsymbol{-k}}$ is the part of $\boldsymbol{X}$ orthogonal to $\boldsymbol{k}$. However, if the model depends on $\boldsymbol{k}$ then Equation 4.7 will be a poor estimate of $\hat{\boldsymbol{y}}_{\boldsymbol{a-k}}$, as illustrated in Figure 4.20.

Figure 4.21 show the results where PLS models have been fitted individually toward specific protein fractions, coagulation properties, total fat and lactose (Equation 4.6). The preprocessed spectra were then deflated/orthogonalized with the interaction of rhe estimated total protein content and breed ($TP \times Breed$) and new predictions were obtained with the deflated spectra but the original regression vector (Equation 4.7). In order to investigate how dependent the individual models are on total protein and breed information, it was explored if predictions obtained from the non-deflated spectra (Equation 4.6) were equally good in describing the non-deflated measured analyte concentrations as the predictions from the deflated spectra (Equation 4.7) were in describing the deflated measured analyte concentrations.

The performance of the non-deflated data are the black bars where the performance of the deflated data are the gray bars in Figure 4.21.

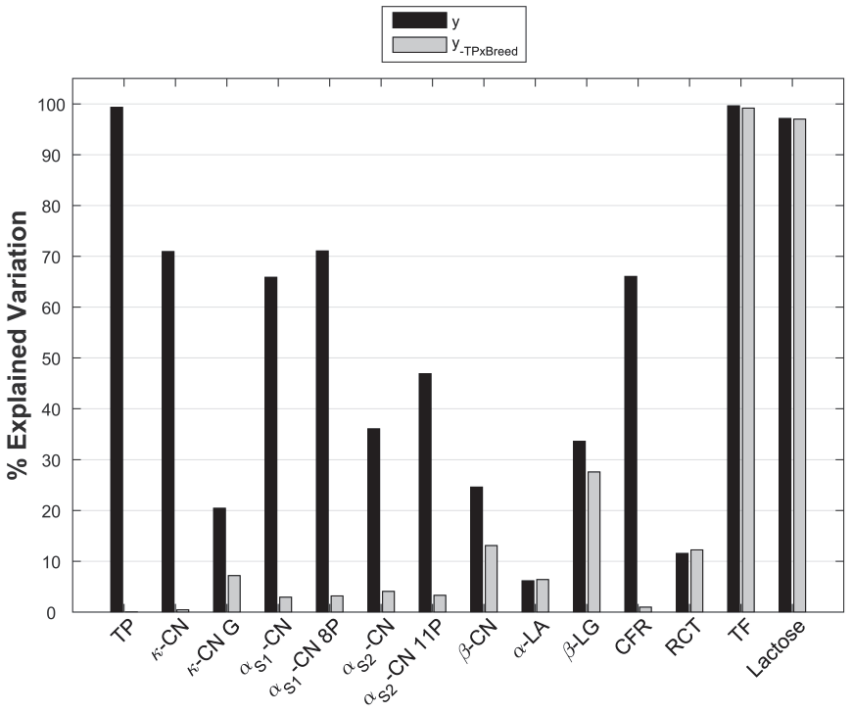


Figure 4.21*. Partial Least Squares (PLS) model performance (black) and PLS model performance independent of the interaction between total protein and breed (gray). TP = total protein, CFR = curd firming rate, RCT = rennet coagulation time, TF = total fat.*

Obviously, the model for total protein is very poor for the deflated data but all other protein fractions and coagulation properties also show remarkably worse predictions for the deflated data than for the non-deflated data (except β-LG and RCT but both are poor models and therefore not interesting). This clearly indicates that the models for all the protein fractions depend on total protein and breed. However, total fat and lactoce performs equally good in the non-deflated (Figure 4.21; black bars) and the deflated data set (Figure 4.21; gray bars). This tells that these models are independent of total protein and breed. Total fat and lactose are also known to have a very different absorption pattern than total protein. Hence, it makes sense that these are predicted independently.

The best way of ensuring that models are independent of each other is by making sure that the analyte and interferent concentrations are unrelated by experimental design. For natural reasons, this is very difficult (if not impossible) in the case of phenotypes in raw cow milk. Nevertheless, in **PAPER VI** a spiking experiment was performed. Here, three monoacid triglycerids were spiked to a similar background of skimmed milk. Pure C14:0, C16:0 and C18:1 *cis*-9 monoacid triglycerides were added to a background of commercially bought skimmed milk, 0.1 % fat. The monoacid triglycerides were added to 17 samples (including two replicates) in a three component mixture design (Montgomery, 2009) as shown in Figure 4.22. Each monoacid triglyceride was added to the skimmed milk in concentrations from 0 to 3 g. pr. 100 mL skimmed milk. In this way, concentrations of the three FA is unrelated to each other but also importantly, the samples have no variation in total fat content. Hence, predictions cannot rely on covariance structures.

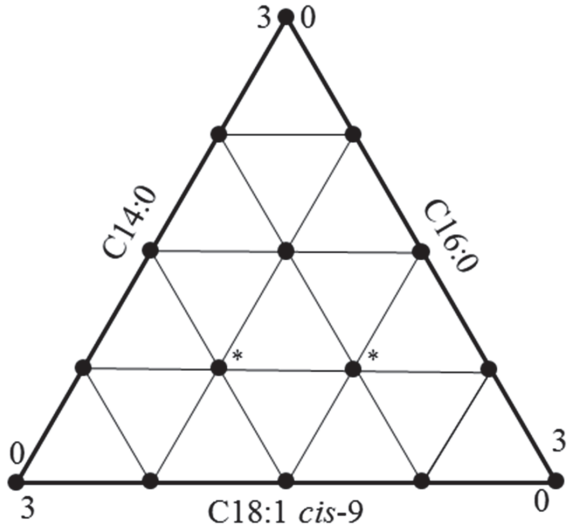


***Figure 4.22**. Three component mixture design. *duplicate samples*

FT-IR spectra were obtained from the spiked milk samples and PLS models for each FA were constructed. In theory, this is a very simple system with three sources of variation. Hence, it should be

possible to model the data set by three LV. However, as the three FA (sources of variation) are orthogonal by the experimental design, a one LV PLS model should be able to model a single FA, if the FA gives rise to a unique signal in the spectra. Figure 4.23 shows the cross-validated prediction error from PLS modeling. Note that the error at LV = 0 is the error obtained when the average FA concentration is used as estimate for all samples. Using one to three LV for PLS modeling is not very successful compared with the average (no model). However, the error decreases when (over-fitting) using around five LV. Nevertheless, the errors are still by fare larger than obtained from raw milk samples (TABLES 4.2). This indicates that the three FA do not provide unique signals in FT-IR spectra obtained from milk samples.

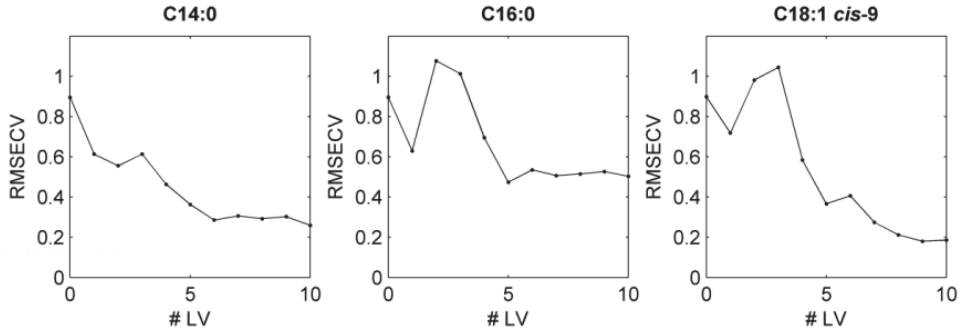


***Figure 4.23**. Results from spiking experiment. Root mean squared error of cross-validation (RMSECV) from partial least squares regression models. #LV = number of latent variables.*

3.3.2 Direct estimation of κ-casein from IR Measurements – Unpublished attempt

A small experiment was carried out in an attempt to obtain direct information related to κ-CN. An enzymatic hydrolysis reaction, similar to Figure 2.15, was followed over time. In this experiment chymosin, which is extreamly selective towards Phe_{105}-Met_{106} bond in κ-CN, was used. Chymosin was added to (heated) milk in a dose of 40 IMCU pr. liter of milk. After addition of chymosin, the milk was stirred in order to avoid gel formation. Using MilkoScan FT 2, FT-IR spectra were obtained in 5 min intervals from 0 min to 45 min (Figure 4.24). It was expected to observe changes (over time) in the absorption spectra due to hydrolysis. The spectral changes were expected to be similar to what was outlined in **Chapter 2.3**. However, the spectra did not exhibit any variation that could be assigned to the hydrolysis, most likely due to lack of sensitivity of the FT-IR instrument.

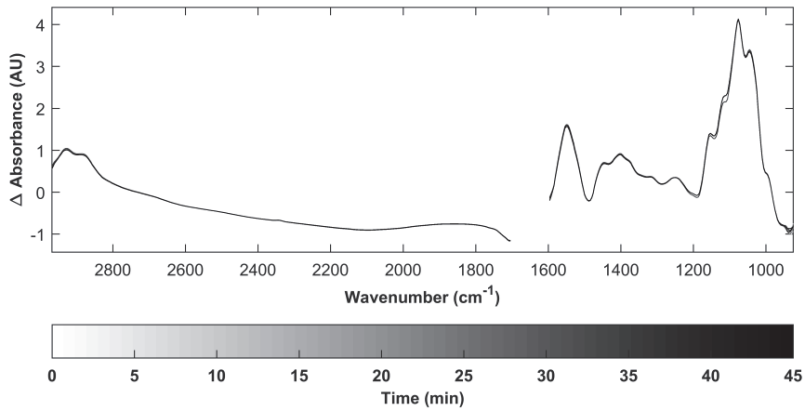


Figure 4.24. *Fourier transform infra spectra of milk samples with added chymosin measure over time.*

CHAPTER 5

OUTREACH

Collecting the right information sufficiently fast using conventional analytical chemistry is a major challenge in the food industry. Vibrational spectroscopy offers remote and non-invasive high speed measurements of quality attributes during all stages of the food production chain, it is compatible with the chemical and physical questions involved and thereby it enables process control. This is not surpassed by any other analytical method and vibrational spectroscopic techniques such as IR and NIR are therefore widely and increasingly used in the food industry.

Unfortunately, spectroscopic instruments do not directly return concentrations but reflectance or transmittance spectra. By a logarithmic transformation reflectance or transmittance spectra can be turned into absorbance spectra, which, through Beer's law, are linearly related to concentrations. A regression model can be calibrated to contain the specific relationship between absorption spectra and the concentrations of a given analyte. The translation, done by the calibration model, from the absorption spectral space to the analyte concentration space is in fact just simple correlations. Hence, if the relationship between the two spaces changes, the calibration model will no longer be valid. For this reason, the concept of interfering compounds is highly important and cannot be neglected in calibration aspects.

For zero order instruments like Kjeldahl digestion units, an interferent makes calibration impossible. Imagine extracting information on both protein and melamine content from a Kjeldahl digestion analysis, which returns a quantification of total nitrogen. A poorly designed calibration set could return good looking calibration models in terms of high R^2 and low RMSEC for both protein and melamine. The calibration model estimates will be two vectors, one for protein and one for melamine, and the two vectors can be viewed as two dimensions. However, the space spanned by the two vectors is only one dimension (total nitrogen estimates from Kjeldahl digestion). Therefore, viewing the estimates of protein and melamine as being independent of each other would be an overestimation of the chemical rank, which in this case equals one. In fact, the predictions of melamine and protein will be trapped in a *cage of covariance* with the estimates of nitrogen. The calibration models return information on nitrogen content rather than protein and melamine content. This compromises prediction accuracy and calibration robustness. Model performances will depend on how good protein and melamine correlate with nitrogen in the data set.

The power of first order instruments like IR and NIR instruments is that they can deal with interferents. The multidimensional measurements enable fitting the regression vector for an analyte in a dimension orthogonal to the interfering signals. Nevertheless, the number of dimensions available for fitting the regression vector is limited by the chemical rank of the spectra. It may happen that a regression vector for a given analyte is fitted in a direction (partly) describing the signal of an interferent. Then the analyte model is no longer independent of the interferent. The calibration model will be indirect and calibration robustness is seriously compromised by the *cage of covariance* phenomenon. In this thesis, projection/orthogonalization based methods was developed to visualize the *cage of covariance* and thereby evaluate calibration robustness.

Detailed milk composition was estimated from FT-IR measurements. Individual FA, specific protein fractions and coagulation properties were predicted and the model performances were in agreement with existing literature. In contrast to previous studies, this thesis focuses on important aspects like prediction accuracy and calibration robustness. Reported predictions of individual milk FA, protein fractions and coagulation properties from FT-IR measurements are, by projection/orthogonalization based methods, shown to rely on indirect correlations, which are confined to covariance structures in the data set rather than absorption bands directly associated with individual FA or protein fractions. Even though this is not a problem in itself, it may be problematic in terms of prediction accuracy and calibration robustness, which are important but neglected parameters when evaluating the usefulness of PLS models for predicting individual FA, protein fractions and coagulation properties from FT-IR measurements of milk. If indirect relationships are used to calibrate a PLS model, the model will not be valid for future samples unless the indirect relationships are conserved in the new sample set. If the indirect relationships are not conserved, the model will provide incorrect predictions. The indirect relationships may change with factors like breed, lactation stage, etc. Therefore, PLS models predicting detailed milk composition are not valid for milk samples of a different nature (different FA profile or protein composition) as compared with the calibration samples. This aspect is believed to limit the usefulness of applying such PLS models in e.g. breeding programs, milk recording systems and for process control.

The problems of fitting indirect regression models have been long known but tend to be forgotten. Osborne and Fearn (1986) touched on the problems in relation to NIR predictions of flour quality. They suggested that problems are likely to occur when major chemical constituents (like water and protein) show correlation. However, this thesis suggests that it is important to distinguish correlations in the analyte concentration profiles with correlations in the pure analyte signals. Correlations in the concentration profiles enables indirect models if unique spectral features originating from the analyte of interest are not present (or if the model is underfitted). However, if unique spectral features are available, the regression vector can be estimated in the direction of the part of the pure analyte signal being orthogonal to all other interferents. Hence, correlations in the concentration profiles are not a problem in itself.

The problem of predicting detailed milk composition from FT-IR measurements lies in overlapping signals. Hence, the chemical rank of the spectra will be smaller than the number of phenotypes and unique regression vectors for each phenotype cannot be obtained. Therefore, if FT-IR spectra are to be used in, for example, breeding programs it is recommended to "decompose" the detailed milk composition into functional groups and then predict the average proportions of these groups. For example, FA could be "decomposed" into, for example, methyl, methylene, olefinic and carboxylic groups. The average proportions of these groups could be predicted by reliable FT-IR based models and included in, for example, breeding programs and milk recording. Furthermore, it is recommended to increase the wavenumber range of the FT-IR measurements. In this thesis the region from 3,000 cm^{-1} to 925 cm^{-1} was available. Above 3,000 cm^{-1} the signal to noise ratio is very poor and the MilkoScan FT2 is not providing measurements below 925 cm^{-1}. Nevertheless, valuable information is found in the FT-IR spectra just above 3,000 cm^{-1} in form of sp^2 C-H bonds. Furthermore, the CH_2 rocking (bending in plane) is found at ~720 cm^{-1}. This bending motion is known to appear only if at least three CH_2 groups are present after each other in the carbon chain. Hence, this mode could potentially contain information related to conjugation, which could be useful.

Vibrational spectroscopic techniques are by far cheaper, faster and easier to operate than chromatography based analyses. When large sample sets have to be measured it would be preferable if, for example, FT-IR could provide the wanted information. However, in the case of detailed milk composition the FT-IR based predictions may not be the best solution. The predictions for FA and protein fractions are, in the multidimensional space, estimated by information pointing in the same direction as total fat content and total protein content, respectively. Therefore, the predictions cannot be regarded independent and the predictions lose value.

REFERENCES

Afseth, N. K., H. Martens, A. Randby, L. Gidskehaug, B. Narum, K. Jorgensen, S. Lien and A. Kohler. 2010. Predicting the fatty acid composition of milk: A comparison of two fourier transform infrared sampling techniques. Appl. Spectrosc. 64:700-707.

Alamar, M. C., E. Bobelyn, J. Lammertyn, B. M. Nicolaï and E. Moltó. 2007. Calibration transfer between NIR diode array and FT-NIR spectrophotometers for measuring the soluble solids contents of apple. Postharvest Biol. Technol. 45:38-45.

Andersson, M. 2009. A comparison of nine PLS1 algorithms. J. Chemometrics. 23:518-529.

Barber, M., R. Clegg, M. Travers and R. Vernon. 1997. Lipid metabolism in the lactating mammary gland. Biochimica Et Biophysica Acta-Lipids and Lipid Metabolism. 1347:101-126.

Barboza, F. D. and R. J. Poppi. 2003. Determination of alcohol content in beverages using short-wave near-infrared spectroscopy and temperature correction by transfer calibration procedures. Analytical and Bioanalytical Chemistry. 377:695-701.

Barth, A. 2007. Infrared spectroscopy of proteins. Biochimica Et Biophysica Acta (BBA) - Bioenergetics. 1767:1073-1101.

Bauman, D. E. and J. Griinari. 2003. Nutritional regulation of milk fat synthesis. Annual Reviews. 23:203-227.

Bergman, E., H. Brage, M. Josefson, O. Svensson and A. Sparén. 2006. Transfer of NIR calibrations for pharmaceutical formulations between different instruments. J. Pharm. Biomed. Anal. 41:89-98.

Bobe, G., D. C. Beitz, A. E. Freeman and G. L. Lindberg. 1999. Effect of milk protein genotypes on milk protein composition and its genetic parameter Estimates1. J. Dairy Sci. 82:2797-2804.

Bonfatti, V., G. Chiarot and R. Carnier. 2014. Glycosylation of kappa-casein: Genetic and nongenetic variation and effects on rennet coagulation properties of milk. J. Dairy Sci. 97:1961-1969.

Bonfatti, V., G. Di Martino and P. Carnier. 2011. Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and contents of protein genetic variants of individual milk of simmental cows. J. Dairy Sci. 94:5776-5785.

Bovenhuis, H., M. Visker and A. Lundén. 2013. Selection for milk fat and milk protein composition. Advances in Animal Biosciences. 4:612-617.

Bruice, P. Y. 2007a. Electronic structure and bonding, acids and bases. Pages 2-70 in Organic Chemistry. 5th ed. P. Y. Bruice ed. Pearson Education, Inc., Upper Saddle River, NJ, USA.

Bruice, P. Y. 2007b. Mass spectrometry, infrared spectroscopy and ultraviolet/visible spectroscopy. Pages 512-568 in Organic Chemistry. 5th ed. P. Y. Bruice ed. Pearson Education, Inc., Upper Saddle River, NJ, USA.

Chapman, D. 1965. Infrared spectroscopy of lipids. Journal of the American Oil Chemists Scoiety. 42:353-371.

Chilliard, Y., C. Martin, J. Rouel and M. Doreau. 2010. Milk fatty acids in dairy cows fed whole crude linseed, extruded linseed, or linseed oil, and their relationship with methane output (vol 92, pg 5199, 2009). J. Dairy Sci. 93:4997-4997.

Chobert, J. M., C. Bertrand-Harb and M. G. Nicolas. 1988. Solubility and emulsifying properties of caseins and whey proteins modified enzymically by trypsin. J. Agric. Food Chem. 36:883-892.

Coates, J. P. 2010. Infrared spectroscopy for process analytical applications. Pages 157-194 in Process Analytical Technology. 2nd ed. K. A. Bakeev ed. WILEY, United Kingdom.

Couvreur, S., C. Hurtaud, C. Lopez, L. Delaby and J. L. Peyraud. 2006. The linear relationship between the proportion of fresh grass in the cow diet, milk fatty acid composition, and butter properties. J. Dairy Sci. 89:1956-1969.

Craninx, M., A. Steen, H. Van Laar, T. Van Nespen, J. Martin-Tereso, B. De Baets and V. Fievez. 2008. Effect of lactation stage on the odd- and branched-chain milk fatty acids of dairy cattle under grazing and indoor conditions. J. Dairy Sci. 91:2662-2677.

Dal Zotto, R., M. De Marchi, A. Cecchinato, M. Penasa, M. Cassandro, P. Carnier, L. Gallo and G. Bittante. 2008. Reproducibility and repeatability of measures of milk coagulation properties and predictive ability of mid-infrared reflectance spectroscopy. J. Dairy Sci. 91:4103-4112.

de Castro, Ruann Janser Soares, M. P. Bagagli and H. H. Sato. 2015. Improving the functional properties of milk proteins: Focus on the specificities of proteolytic enzymes. Current Opinion in Food Science. 1:64-69.

De Marchi, M., C. C. Fagan, C. P. O'Donnell, A. Cecchinato, R. Dal Zotto, M. Cassandro, M. Penasa and G. Bittante. 2009a. Prediction of coagulation properties, titratable acidity, and pH of bovine milk using mid-infrared spectroscopy. J. Dairy Sci. 92:423-432.

De Marchi, M., M. Penasa, A. Cecchinato, M. Mele, P. Secchiari and G. Bittante. 2011. Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of brown swiss bovine milk. Animal. 5:1653-1658.

De Marchi, M., V. Toffanin, M. Cassandro and M. Penasa. 2014. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. J. Dairy Sci. 97:1171-1186.

De Marchi, M., V. Toffanin, M. Cassandro and M. Penasa. 2013. Prediction of coagulating and noncoagulating milk samples using mid-infrared spectroscopy. J. Dairy Sci. 96:4707-4715.

De Marchi, M., V. Bonfatti, A. Cecchinato, G. Di Martino and P. Carnier. 2009b. Prediction of protein composition of individual cow milk using mid-infrared spectroscopy. Italian Journal of Animal Science. 8:399-401.

Dehareng, F., C. Delfosse, E. Froidmont, H. Soyeurt, C. Martin, N. Gengler, A. Vanlierde and P. Dardenne. 2012. Potential use of milk mid-infrared spectra to predict individual methane emission of dairy cows. Animal. 6:1694-1701.

DiFoggio, R. 1995. Examination of some misconceptions about near-infrared analysis. Appl. Spectrosc. 49:67-75.

Dufour, E. 2009. Principles of infrared spectroscopy. Pages 2-26 in Infrared Spectroscopy for Food Quality Analysis and Control. 1st ed. D. Sun ed. Elsevier, New York, USA.

Ellen, G. and A. J. Tudos. 2003. On-line measurements of product quality in dairy processing. Pages 263-291 in Dairy Processing Improving Quality. 1st ed. G. Smit ed. CRX Press, FL,USA.

Engelsen, S. 1997. Explorative spectrometric evaluations of frying oil deterioration. Journal of the American Oil Chemists Society. 74:1495-1508.

Faber, K. and B. R. Kowalski. 1997. Improved prediction error estimates for multivariate calibration by correcting for the measurement error in the reference values. Appl. Spectrosc. 51:660-665.

Fan, W., Y. Liang, D. Yuan and J. Wang. 2008. Calibration model transfer for near-infrared spectra based on canonical correlation analysis. Anal. Chim. Acta. 623:22-29.

Ferrand, M., B. Huquet, S. Barbey, F. Barillet, F. Faucon, H. Larroque, O. Leray, J. M. Trommenschlager and M. Brochard. 2011. Determination of fatty acid profile in cow's milk using mid-infrared spectrometry: Interest of applying a variable selection by genetic algorithms before a PLS regression. Chemometrics Intellig. Lab. Syst. 106:183-189.

Ferrand-Calmels, M., I. Palhiere, M. Brochard, O. Leray, J. M. Astruc, M. R. Aurel, S. Barbey, F. Bouvier, P. Brunschwig, H. Caillatt, M. Douguet, F. Faucon-Lahalle, M. Gele, G. Thomas, J. M. Trommenschlager and H. Larroque. 2014. Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. J. Dairy Sci. 97:17-35.

Food and Drug Administration. 2004. Guidance for Industry: PAT - A Framework for Innovative Pharmaceutical Development, Manufactoring and Quality Assurance. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Veterinary Medicine, Office of Regulatory Affairs, http://www.fda.gov/downloads/Drugs/Guidances/ucm070305.pdf.

Fox, P. F. and P. L. H. McSweeney. 1998. Dairy Chemnistry and Biochemistry. Blackie Academic & Professional Publishers, London, UK.

Frederiksen, P. D., K. K. Andersen, M. Hammershoj, H. D. Poulsen, J. Sorensen, M. Bakman, K. B. Qvist and L. B. Larsen. 2011. Composition and effect of blending of noncoagulating, poorly coagulating, and well-coagulating bovine milk from individual danish holstein cows. J. Dairy Sci. 94:4787-4799.

Funahashi, H. and J. Horiuchi. 2008. Characteristics of the churning process in continuous butter manufacture and modelling using an artificial neural network. Int. Dairy J. 18:323-328.

German, J. B., R. A. Gibson, R. M. Krauss, P. Nestel, B. Lamarche, W. A. van Staveren, J. M. Steijns, L. C. P. G. M. de Groot, A. L. Lock and F. Destaillats. 2009. A reappraisal of the impact of dairy foods and milk fat on cardiovascular disease risk. Eur. J. Nutr. 48:191-203.

Givens, D. I. 2010. Milk and meat in our diet: Good or bad for health? Animal. 4:1941-1952.

Halken, S. and A. Høst. 1997. How hypoallergenic are hypoallergenic cow's milk-based formulas? Allergy. 52:1175-1183.

Harris, D. C. 2010. Fundamentals of spectrophotometry. Pages 393-418 in Quantitative Chemical Analysis. 8th ed. D. C. Harris ed. W. H. Freeman and Company, NY, USA.

Heck, J. M. L., A. Schennink, H. J. F. van Valenberg, H. Bovenhuis, M. H. P. W. Visker, J. A. M. van Arendonk and A. C. M. van Hooijdonk. 2009. Effects of milk protein variants on the protein composition of bovine milk. J. Dairy Sci. 92:1192-1202.

Jensen, R. G., A. M. Ferris, C. J. Lammi-Keefe and R. A. Henderson. 1990. Lipids of bovine and human milks: A comparison. J. Dairy Sci. 73:223-240.

Jensen, H. B., N. A. Poulsen, K. K. Andersen, M. Hammershøj, H. D. Poulsen and L. B. Larsen. 2012. Distinct composition of bovine milk from jersey and holstein-friesian cows with good, poor, or noncoagulation properties as reflected in protein genetic variants and isoforms. J. Dairy Sci. 95:6905-6917.

Jensen, R. 2002. The composition of bovine milk lipids: January 1995 to december 2000. J. Dairy Sci. 85:295-350.

Krag, K., N. A. Poulsen, M. K. Larsen, L. B. Larsen, L. L. Janss and B. Buitenhuis. 2013. Genetic parameters for milk fatty acids in danish holstein cattle based on SNP markers using a bayesian approach. BMC Genet. 14:79-2156-14-79.

Larkin, P. 2011a. Basic principles. Pages 7-25 in Infrared and Raman Spectroscopy: Principles and Spectral Interpretation. 1st ed. P. Larkin ed. Elsevier, Amsterdam, The Netherlands.

Larkin, P. 2011b. Origin of group frequencies. Pages 63-72 in Infrared and Raman Spectroscopy: Principles and Spectral Interpretation. 1st ed. P. Larkin ed. Elsevier, Amsterdam, The netherlands.

Lien, E. L. 2003. Infant formulas with increased concentrations of alpha-lactalbumin. Am. J. Clin. Nutr. 77:1555S-1558S.

Lock, A. and D. Bauman. 2004. Modifying milk fat composition of dairy cows to enhance fatty acids beneficial to human health. Lipids. 39:1197-1206.

Lönnerdal, B. and E. L. Lien. 2003. Nutritional and physiologic significance of α-lactalbumin in infants. Nutr. Rev. 61:295-305.

Lopez-Villalobos, N., R. J. Spelman, J. Melis, S. R. Davis, S. D. Berry, K. Lehnert, S. E. Holroyd, A. K. MacGibbon and R. G. Snell. 2014. Estimation of genetic and crossbreeding parameters of fatty acid concentrations in milk fat predicted by mid-infrared spectroscopy in new zealand dairy cattle. J. Dairy Res. 81:340-349.

Luinge, H., E. Hop, E. Lutz, J. Vanhemert and E. Dejong. 1993. Determination of the fat, protein and lactose content of milk using fourier-transform infrared spectrometry. Anal. Chim. Acta. 284:419-433.

Lyndgaard, C. B., S. B. Engelsen and F. W. J. van den Berg. 2012. Real-time modeling of milk coagulation using in-line near infrared spectroscopy. J. Food Eng. 108:345-352.

Maas, J., J. France and B. McBride. 1997. Model of milk protein synthesis. A mechanistic model of milk protein synthesis in the lactating bovine mammary gland. J. Theor. Biol. 187:363-378.

Madureira, A. R., C. I. Pereira, A. M. Gomes, M. E. Pintado and F. X. Malcata. 2007. Bovine whey proteins–overview on their main biological properties. Food Res. Int. 40:1197-1211.

Maurice-Van Eijndhoven, M. H. T., H. Soyeurt, F. Dehareng and M. P. L. Calus. 2013. Validation of fatty acid predictions in milk using mid-infrared spectrometry across cattle breeds. Animal. 7:348-354.

McDermott, A., G. Visentin, M. De Marchi, D. Berry, M. Fenelon, P. O'Connor, O. Kenny and S. McParland. 2016. Prediction of individual milk proteins including free amino acids in bovine milk using mid-infrared spectroscopy and their correlations with milk processing characteristics. J. Dairy Sci.

McParland, S., G. Banos, E. Wall, M. P. Coffey, H. Soyeurt, R. F. Veerkamp and D. P. Berry. 2011. The use of mid-infrared spectrometry to predict body energy status of holstein cows1. J. Dairy Sci. 94:3651-3661.

Montgomery, D. C. 2009. Response surface methodology. Pages 417-485 in Design and Analysis of Experiments. 7th ed. D. C. Montgomery ed. John Wiley & Sons, Inc., Asia.

Næs, T., T. Isaksson, T. Fearn and T. Davies. 2002. Validation. Pages 155-176 in A User-Friendly Guide to Multivariate Calibration and Classification. 1st ed. T. Næs, T. Isaksson, T. Fearn and T. Davies eds. NIR Publications, West Sussex, UK.

Næs, T., O. Tomic, B. -. Mevik and H. Martens. 2011. Path modelling by sequential PLS regression. J. Chemometrics. 25:28-40.

Osborn, B. G. and T. Fearn. 1986. Secondary correlations. Pages 99-100 in Near Infrared Spectroscopy in Food Analysis. Secondary correlations. Longman Scientific & Technical, Essex, England.

Palmquist, D., A. Beaulieu and D. Barbano. 1993. Feed and animal factors influencing milk-fat composition. J. Dairy Sci. 76:1753-1771.

Pavia, D. L., G. M. Lampman and G. S. Kriz. 2001. Infrered spectroscopy. Pages 13-101 in Introduction to Spectroscopy: A Guide for Students of Arganic Chemistry. 3rd ed. D. L. Pavia, G. M. Lampman and G. S. Kriz eds. Harcourt College Publishers, London, UK.

Pearce, K. 1979. Use of fluorescamine to determine the rate of release of the caseino-macropeptide in rennet-treated milk. New Zealand Journal of Dairy Science and Technology.

Pereira, S. L., A. E. Leonard and P. Mukerji. 2003. Recent advances in the study of fatty acid desaturases from animals and lower eukaryotes. Prostaglandins, Leukotrienes and Essential Fatty Acids. 68:97-106.

Poulsen, N. A., F. Gustavsson, M. Glantz, M. Paulsson, L. B. Larsen and M. K. Larsen. 2012. The influence of feed and herd on fatty acid composition in 3 dairy breeds (danish holstein, danish jersey, and swedish red). J. Dairy Sci. 95:6362-6371.

Pratt, C. W. and K. Cornely. 2004a. Biological membrans. Pages 232-275 in Essential Biochemistry. 1st ed. C. W. Pratt and K. Cornely eds. John Wiley & Sons, Inc., USA.

Pratt, C. W. and K. Cornely. 2004b. Enzyme kinetics and inhibition. Pages 198-231 in Essential Biochemistry. 1st ed. C. W. Pratt and K. Cornely eds. John Wiley & Sons, Inc., USA.

Rinnan, A., F. van den Berg and S. B. Engelsen. 2009. Review of the most common pre-processing techniques for near-infrared spectra. Trac-Trends in Analytical Chemistry. 28:1201-1222.

Rollema, H., R. McKellar, T. Sorhaug, G. Suhren, J. Zadow, B. Law, J. Poll, L. Stepaniak and G. Vagias. 1989. Comparison of different methods for the detection of bacterial proteolytic enzymes in milk. Milchwissenschaft. 44:491-496.

Rutten, M. J. M., H. Bovenhuis, J. M. L. Heck and J. A. M. van Arendonk. 2011. Predicting bovine milk protein composition based on fourier transform infrared spectra. J. Dairy Sci. 94:5683-5690.

Rutten, M. J. M., H. Bovenhuis, K. A. Hettinga, H. J. F. van Valenberg and J. A. M. van Arendonk. 2009. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. J. Dairy Sci. 92:6202-6209.

Sanchez, E. and B. R. Kowalski. 1988. Tensorial calibration: I. first-order calibration. J. Chemometrics. 2:247-263.

Schopen, G. C. B., J. M. L. Heck, H. Bovenhuis, M. H. P. W. Visker, H. J. F. van Valenberg and J. A. M. van Arendonk. 2009. Genetic parameters for major milk proteins in dutch holstein-friesians. J. Dairy Sci. 92:1182-1191.

Simopoulos, A. 1991. Omega-3-fatty-acids in health and disease and in growth and development. Am. J. Clin. Nutr. 54:438-463.

Skov, T., D. Ballabio and R. Bro. 2008. Multiblock variance partitioning: A new approach for comparing variation in multiple data blocks. Anal. Chim. Acta. 615:18-29.

Smilde, A., R. Bro and P. GELADI. 2004. Two-way component and regression models. Pages 35-56 in Multi-Way Analysis with Applications in the Chemical Sciences. 1st ed. A. Smilde, R. Bro and P. GELADI eds. John Wiley & Sons, Ltd, West Sussex, UK.

Soyeurt, H., C. Bastin, F. Colinet, V. Arnould, D. P. Berry, E. Wall, F. Dehareng, H. Nguyen, P. Dardenne and J. Schefers. 2012. Mid-infrared prediction of lactoferrin content in bovine milk: Potential indicator of mastitis. Animal. 6:1830-1838.

Soyeurt, H., F. G. Colinet, V. M. -. Arnould, P. Dardenne, C. Bertozzi, R. Renaville, D. Portetelle and N. Gengler. 2007. Genetic variability of lactoferrin content estimated by mid-infrared spectrometry in bovine milk. J. Dairy Sci. 90:4443-4450.

Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. J. Dairy Sci. 89:3690-3695.

Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. P. Berry, M. Coffey and P. Dardenne. 2011. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. J. Dairy Sci. 94:1657-1667.

Soyeurt, H., F. Dehareng, P. Mayeres, C. Bertozzi and N. Gengler. 2008. Variation of delta(9)-desaturase activity in dairy cattle. J. Dairy Sci. 91:3211-3224.

Strang, G. 2006. Orthogonality. Pages 141-200 in Linear Algebra and its Applications. 4th ed. G. Strang ed. Thomson Learning, Inc., Canada.

Subramanian, A. and L. Rodriguez-Saona. 2009. Fourier transform infrared (FTIR) spectroscopy. Pages 145-178 in Infrared Spectroscopy for Food Quality Analysis and Control. 1st ed. D. Sun ed. Elsevier, New York, USA.

Svenning, C., J. Brynhildsvold, T. Molland, T. Langsrud and G. E. Vegarud. 2000. Antigenic response of whey proteins and genetic variants of β-lactoglobulin—the effect of proteolysis and processing. Int. Dairy J. 10:699-711.

Swierenga, H., P. De Groot, A. De Weijer, M. Derksen and L. Buydens. 1998a. Improvement of PLS model transferability by robust wavelength selection. Chemometrics Intellig. Lab. Syst. 41:237-248.

Swierenga, H., W. Haanstra, A. De Weijer and L. Buydens. 1998b. Comparison of two different approaches toward model transferability in NIR spectroscopy. Appl. Spectrosc. 52:7-16.

van den Berg, F., C. B. Lyndgaard, K. M. Sørensen and S. B. Engelsen. 2013. Process analytical technology in the food industry. Trends Food Sci. Technol. 31:27-35.

Vandeginste, B. G. M., D. L. Massart, L. M. C. Buydens, S. De Jong, P. J. Lewi and J. Smeyers-Verbeke. 1998. Vectors, matrices and operations on matrices. Pages 7-56 in Handbook of Chemometrics and Qualimetrics: Part B. 1st ed. B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. De Jong, P. J. Lewi and J. Smeyers-Verbeke eds. Elsevier, Amsterdam, Netherlands.

Vidrine, D. W. 2000. Mid-infrared spectroscopy in chemical process analysis. Pages 96-138 in Spectroscopy in Process Analysis. first ed. J. M. Chalmers ed. Sheffield Academic Press, Sheffield, England.

Vlaeminck, B., V. Fievez, A. R. J. Cabrita, A. J. M. Fonseca and R. J. Dewhurst. 2006a. Factors affecting odd- and branched-chain fatty acids in milk: A review. Anim. Feed Sci. Technol. 131:389-417.

Vlaeminck, B., V. Fievez, S. Tamminga, R. J. Dewhurst, A. van Vuuren, D. De Brabander and D. Demeyer. 2006b. Milk odd- and branched-chain fatty acids in relation to the rumen fermentation pattern. J. Dairy Sci. 89:3954-3964.

Walstra, P., J. T. M. Wouters and T. J. Geurts. 2006. Dairy Science and Technology. 2nd ed. CRC Press, Taylor & Francis Group, FL, USA.

Walstra, P. 1999. Casein sub-micelles: Do they exist? Int. Dairy J. 9:189-192.

Wold, S., K. Esbensen and P. Geladi. 1987. Principal component analysis. Chemometrics Intellig. Lab. Syst. 2:37-52.

Wold, S., M. Sjostrom and L. Eriksson. 2001. PLS-regression: A basic tool of chemometrics. Chemometrics Intellig. Lab. Syst. 58:109-130.

PUBLICATIONS

PAPER I

Vibrational Spectroscopy in Food Processing

C. E. Eskildsen, F. v. d. Berg and S. B. Engelsen

Accepted for publication, In *Encyclopedia of Spectroscopy and Spectrometry,* Editors: J. Lindon, G. Tranter, D. Koppenaal, 3rd edition, Elsevier, Oxford, 2016.

Vibrational Spectroscopy in Food Processing

CE Eskildsen, F van den Berg, and SB Engelsen, University of Copenhagen, Copenhagen, Denmark

Abbreviations

ATR	Attenuated total reflection	**NIR**	Near-infrared
G	Guluronic acid	**NIT**	Near-infrared transmission
IR	Infrared	**NMR**	Nuclear magnetic resonance
IV	Iodine value	**PE**	Polyester
M	Mannuronic acid	**PES**	Polyethersulfone

Introduction

The food and food ingredients industry faces strong demands from consumers and regulatory bodies to produce safe and consistently high-quality products. An ever increasing size of scale in the food industry also makes process optimization an economic necessity. From a compositional point of view, most foods and food-stuffs can be viewed as multifactorial systems consisting of mainly water, fats, proteins, and carbohydrates. Food samples are typically highly heterogeneous and present in a complex matrix (eg, amorphous solids, aqueous liquids, gels, macromolecules, macro-organelles, or cells). Furthermore, raw materials hold a high degree of natural variation including seasonal, geographical, and genetic. In order to produce food products of consistent quality, it is a necessity to sort incoming raw materials, control the process between tight limits, and continuously verifying the final-product quality.

The main objective for the food industry is to increase productivity and find ways to achieve a desired and consistent end-product quality in a cost effective, safe, traceable, and environmentally responsible way. This can be achieved by, for example, assuring that the raw materials meet certain quality demands, controlling the process by *real-time* continuous measurements of core parameters, rapid final-product quality evaluation, and assuring key quality attributes during packaging. Collecting the right information sufficiently fast using conventional analytical chemistry is a major challenge in the food industry. For this reason, vibrational spectroscopic techniques such as Raman, infrared (IR), and near-infrared (NIR) spectroscopy are widely used during all stages of food processing. Spectroscopy enables remote and noninvasive control at all the production steps (Figure 1) and is compatible with the chemical and physical questions involved.

Food is a challenging (but rewarding) subject to work with, especially due to the raw materials. Imagine, for example, the grain harvest of one farmer, a truck with pooled milk from different dairy farms, or the catch of a fishing trawler; it is easy to realize that sorting and grading of incoming raw materials (Figure 1, process inputs) is an essential step in securing quality and increasing economic gain in many food productions. Proper sorting of lots arriving *at the gate* — for example, sorting grain by protein content using an at-line NIR spectrometer — will ensure that incoming materials are of sufficient quality for the next downstream processing step and thus contribute in fulfilling the promised end-product quality. Hence, sorting not only facilitates that the best raw materials can be used to produce high-quality and high-value products, but it can also ease the control during further processing avoiding unwanted quality deviations in end-products. Traditionally, sorting is done by manual inspection, on a statistical (*by chance*) nature, but this is no longer acceptable in a modern food processing industry.

Even though sorting of raw materials is performed *at the gate*, process monitoring and control (Figure 1, process operation) is often needed in securing consistent product quality. Furthermore, process monitoring and control may increase production throughput. In terms of process dynamics and kinetics, many food processes are more challenging than, for example, processes in the chemical industry. As an example, milk coagulation, in the first stages of cheese production, cannot be modeled from mechanistic or first principles only and there is no natural end-point like a maximum conversion. Moreover, the materials used (milk composition, starter culture performance, etc.) and the quality of the end-product (the cheese) are magnified during this production step with no possibility of correction afterwards. Hence, *in-line* spectroscopy can play a central role in process optimization and scheduling for the modern dairy.

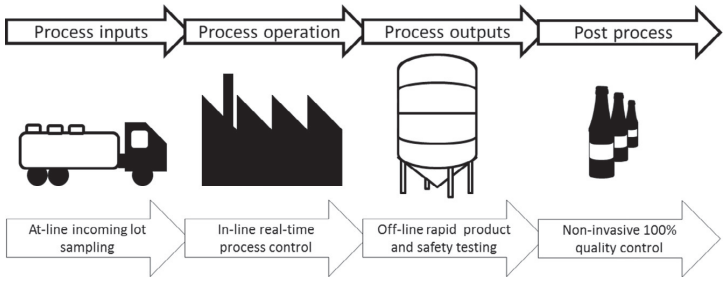


Figure 1 Schematic representation of the different stages and the analytical questions involved of a food production.

The food industry shares one important component with the pharmaceutical industry, namely that humans and animals are at the end of the chain. This means that safety and quality of the end-product are key parameters (Figure 1, process output). Extensive product testing by classical analysis is not desirable for foods due to their limited shelf life. *Off-line* vibrational spectroscopic methods, situated in *walk-in laboratories*, can provide swift answers for product-relevant parameters, typically without any sample preparation or use of reagents. Moreover, high-speed measurement can be utilized at the exit (Figure 1, post process). In principle, spectroscopic techniques are capable of *100% quality assurance* by *noninvasive* analysis of, for example, every bottle coming by on a high-speed filling line.

Using the different production stages and analytical questions symbolized in Figure 1 as guideline, the remainder of this chapter will give examples on how vibrational spectroscopy may be used in the diverse world of the food and food ingredients industry.

Process Inputs

For many years carcass grading in the meat industry has been based on volumetric measures. However, porcine carcass fat composition is an important quality parameter for grading and should thus be evaluated at the front-end of an abattoir. The porcine fat quality varies considerably between genotypes, farmers, seasons, and feeding regimes. During further processing (eg, slicing), the hardness of the fat is a critical attribute; if the fat is too soft it causes problems due to lack of cohesiveness between muscle and adipose tissue. The iodine value (IV), that is, degree of fatty acid unsaturation, relates to the hardness of the fat. Furthermore, IV relates to fat flavors, mouth feel, as well as nutritional issues. The IV of fat is thus an important quality parameter to be evaluated and can be used when sorting the carcasses at the abattoir, and this sorting has to be done *real-time* for logistic reasons. Product orders are placed (typically overnight) requesting a specific quality. Hence, for economic optimization and operational reasons, the fat composition would ideally have to be known immediately after slaughter so carcasses can be sorted according to quality on dedicated lines in the chiller rooms. Sørensen et al.[1] used NIR transmission (NIT) spectroscopy for predicting IV of fat in porcine carcasses. The NIR equipment consists of two knife-tipped probes, which are inserted into the carcass by an operator. The incision is done after veterinary inspection and the equipment fulfill abattoir cleaning and HACCP requirements. During the motorized extraction of the NIR probes, transmission spectra are obtained as a function of penetration depth (Figure 2A). Going from left to right in Figure 2A, the NIR spectra are first from the outer fat layer and then the inner fat layer. The two fat layers are clearly divided by the gab in the NIR spectral landscape (Figure 2A). The NIR spectra can be calibrated to predict the IV using gas chromatography as a laboratory reference method (IV predictions are shown in Figure 2B). The NIR method is able to predict IV in the fat of carcasses as a function of penetration depth at full production speed (approximately 1000 carcasses per hour). By this strategy, NIR can be used for *real-time* sorting of carcasses *at the gate* of the abattoir. Additionally, this information can be used in the payment system and, using the information as feed back to the farmers, this method may assist in optimizing breeding and feeding programs.

Another example where sorting of raw materials is highly beneficial is when milling wheat kernels into flour. Large variations in quality are present between grains due to, for example, genetic variations, weather conditions, soils characteristics, and growing practice. However, quality differences between wheat kernels also exist within the field and even within a single plant. High-speed sorting of single seeds before milling can result in considerable quality improvements of *flours fit for purpose* and thereby economic gain. Protein content is the most important quality attribute of wheat flour. In, for example, bread making high-protein wheat flour is preferred, whereas low-protein flour is used for, for example, crackers. Traditionally, wheat quality evaluation was performed on bulk samples using wet chemistry. When analyzing bulk samples the characteristics of the individual kernels is obviously lost and optimal sorting cannot be performed. Nielsen et al.[3] developed a nondestructive screening method for single seed protein determination. The method is based on NIT spectroscopy in the short wave near IR region, which have sufficient energy and corresponding low molecular extinction coefficients to penetrate whole seeds. The kernels were placed in a sample cassette with slots for individual kernels. The sample cassette was rotated and NIR measurements of the individual kernels were obtained. The raw and preprocessed NIR spectra are shown in Figure 3A **and B**, respectively. In Figure 3B, the spectra are preprocessed by second

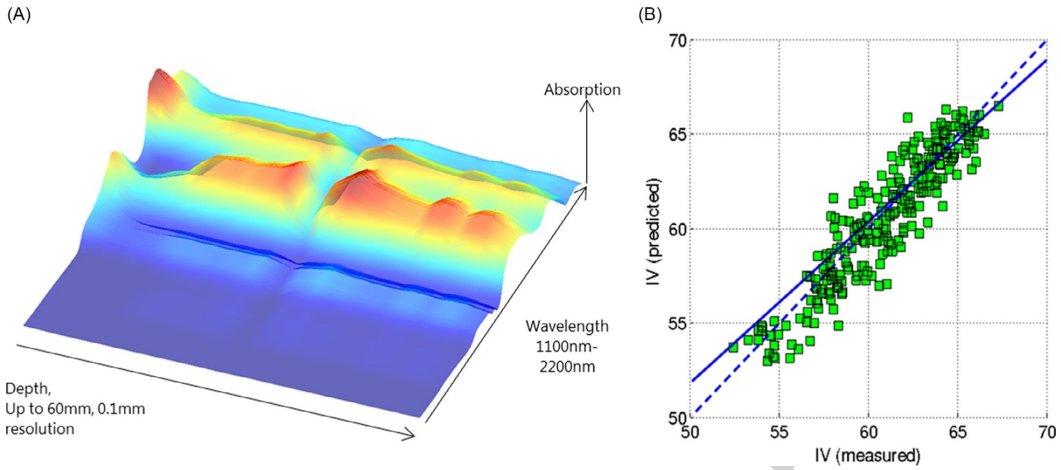


Figure 2 Carcass grading. (A) Near-infrared absorption across fat layers. (B) Measured Iodine value (IV) versus predicted IV. Modified from van den Berg, F.; Lyndgaard, C. B.; Sørensen, K. M.; Engelsen, S. B. Process analytical technology in the food industry. *Trends Food Sci. Technol.* **2013**, *31*, 27–35.

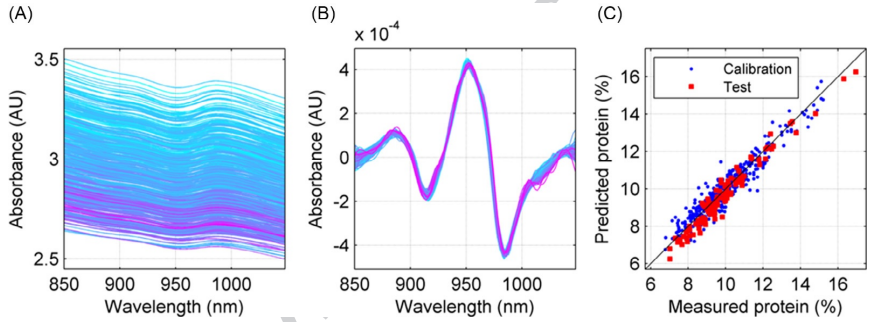


Figure 3 Prediction of single seed protein contents from near-infrared (NIR) measurements. (A) Raw NIR spectra colored by protein concentration. (B) Preprocessed (second derivative and multiplicative scatter correction) NIR spectra colored by protein concentration. (C) Predicted versus measured protein values; (blue) calibration set, (red) test set.

derivative followed by multiplicative scatter correction.[4] The protein information is found round the absorption band of the second overtone of N–H stretching vibrations. Even though the raw spectra (Figure 3A) seem to lack complex information, using five components in partial least squares regression toward a Kjeldahl reference method, protein content in single kernels can be predicted with an acceptable low error (Figure 3C). The method described by Nielsen et al.[3] has been further developed to run at high speed. At high speed, the seeds are fed into a rotating cylinder, which causes the kernels to pass one-by-one over an NIR probe. Measurements are obtained and used for sorting the seeds pneumatically. In this way, high-speed sorting of single kernels *at the gate* allows for production of flours for specific purposes, price differentiation between products and waste reduction.

Process Operation

While sorting of raw materials can be beneficial for food producers, *real-time* measurements of an on-going process are essential to optimize quality, yield, and production throughput. A critical time point in cheese production is cutting the coagulum. Cutting too early will result in loss of fat, protein, moisture, and thereby decreased yield. Longer cutting times result in smaller losses, but lower production throughput. Furthermore, the higher moisture content will cause undesirable sensory and microbiological effects. *Real-time* monitoring of milk coagulation can help producing cheeses of consistent quality, fat, and protein content as well as optimize production throughput. Additionally, batches *out of specification* can be discovered at an earlier stage allowing for the necessary

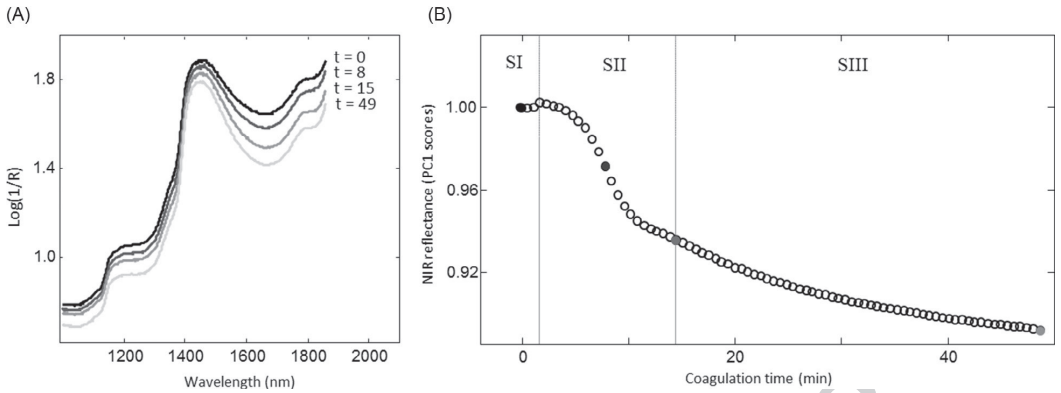


Figure 4 *Real-time* monitoring of milk coagulation. (A) Sequence of near-infrared (NIR) spectra measured at four different time points during milk coagulation. (B) The first principal component scores as a function of coagulation time. Solid circles correspond to the spectra in (A). Marking of stages is based on qualitative assessment: (SI) κ-casein proteolysis, (SII) paracasein aggregation, and (SIII) gel network formation/hardening. Modified from Lyndgaard, C. B.; Engelsen, S. B.; van den Berg, F. W. J.; Real-time modeling of milk coagulation using in-line near infrared spectroscopy. *J. Food Eng.* **2012**, *108*, 345–352.

actions to be taken more rapidly. Rennet induced milk coagulation consists of three stages: (SI) enzymatic proteolysis of κ-casein, (SII) aggregation of the altered casein micelles (paracasein), and (SIII) network formation. Using NIR reflection spectroscopy (Figure 4A), Lyndgaard et al.[5] extracted real-time information of the coagulation process. The main spectral development during the coagulation is baseline changes. The NIR spectra were decomposed using the principal component analysis and the component scores as a function of time were used to establish a kinetic model clearly describing the three coagulation stages (Figure 4B). This estimation can be used to establish automatic cutting time determination and thereby to optimize final-product quality and consistency.

Equipment cleaning is an integrated part of a process operation in the food industry. IR spectroscopy in combination with mathematical and statistical data analysis is a valuable tool in cleaning verification where a high frequency of IR measurements can be performed without too much sample preparation or cost in chemicals. This makes IR spectroscopy an ideal method in industrial detective work. Jensen et al.[6] investigated residual fouling on an ultrafiltration membrane used for 5 months in a full-scale dairy fractionation operation. The destructive analysis was performed to describe the protein pattern on a membrane leave of 1 m^2 from a spirally wound module, collected after a cleaning-in-place. For this purpose, a total of 100 spatial measurements were performed using attenuated total reflection (ATR) measurements. From the spectra (Figure 5A), the membrane material and protein fouling can be probed. A major challenge in the quantification of protein residuals is the presence of unknown grafting materials added by the membrane manufacturers (propriety knowledge), and the multi-layer nature of these system (Figure 5B). Using multivariate curve resolution as a chemometric modeling method, the spectra could be translated into a fouling pattern (Figure 5C). This information can be used to design new cleaning strategies and to develop new membrane modules.

Process Outputs

Vibrational spectroscopy is also useful in end-product quality control in the food and ingredients industry. Digestibility is an important parameter of fish meal used for animal feed. The value is often guaranteed by the producer and may determine the end price. Consequently, the fish meal is not released from the producer before results from the digestibility test are present. The most direct method for determining digestibility is biological tests in minks, taking 6–8 weeks to perform. Fish meal is produced in batches of tons and storing fish meal in the warehouse during the testing period is thus expensive. It is therefore profitable to develop an approximate but less time-consuming and reproducible method to determine the mink digestibility of fish meals. Dahl et al.[7] investigated whether NIR reflectance spectroscopy could be used here. In the study it was found that mink digestibility was negatively correlated with oil and ash content. The NIR measurements of the fish meals are shown in Figure 6A, and the predictions of mink digestibility from NIR measurements are shown in Figure 6B. Even though mink digestibility is an unorthodox and complex parameter to be predicted by NIR measurements, an indirect relationship can be established. Obviously the predictions are not perfect. However, considering the expensive and time-consuming biological tests in mink, which are subjected to large biological variations, NIR measurements could well function as a rapid and inexpensive alternative for predicting mink digestibility of fish meals.

Measuring the functional properties of alginate powders is a second example of end-product quality control of a highly complex bio-material. Alginates are hydrocolloids extracted from brown seaweed. They are used as gelling agents in the food industry.

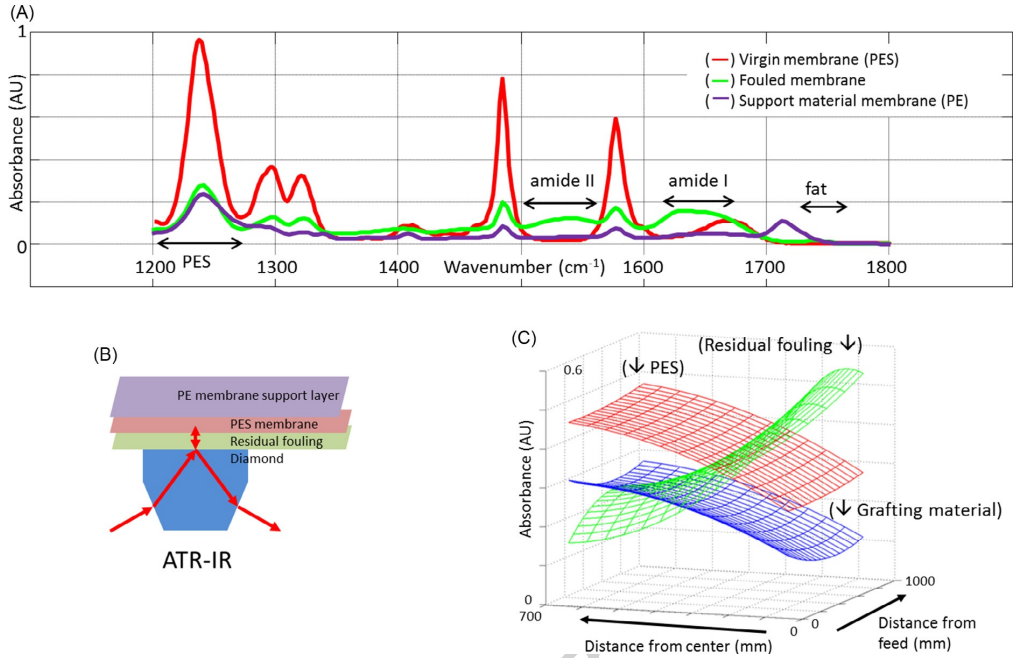


Figure 5 Residual fouling on ultrafiltration membranes. (A) Infrared spectra of virgin polyethersulfone (PES) membrane material, a membrane fouled during fractionation in the dairy industry and polyester support material (PE). (B) ATR-IR measurements of a multi-layer membrane system. (C) Estimated concentration contours over the membrane leave of PES, protein residual fouling, and membrane grafting material.

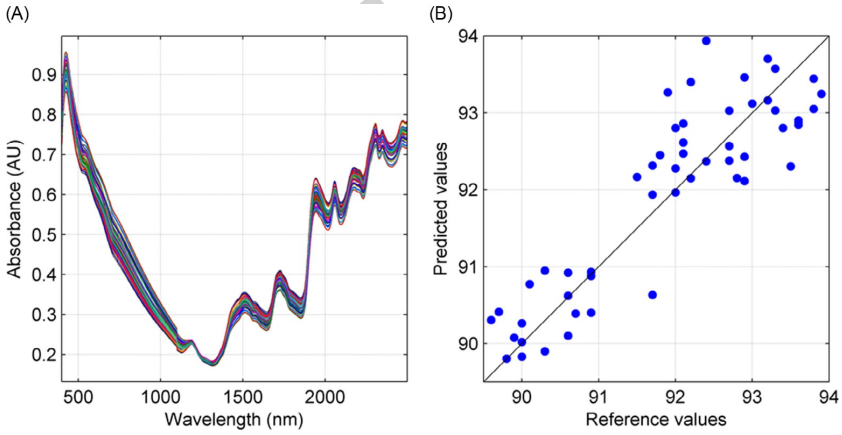


Figure 6 Digestibility of fish meal. (A) Near-infrared spectra preprocessed by multiplicative scatter correction. (B) Measured digestibility versus predicted digestibility.

Alginates are polysaccharides consisting of (1–4) linked β-D-mannuronic acid (M) and its C-5 epimer α-L-guluronic acid (G), and the functional properties of alginates are related to the M/G ratio. The composition of alginates, and thereby the functional properties, varies according to season, age of the seaweed population, species, and geographic harvesting location. Nuclear magnetic resonance (NMR) spectroscopy is the state-of-the-art method for the compositional and structural analysis of alginates. However, NMR requires sample preparation as well as a high degree of know-how during operation. It would therefore be beneficial for the industry to have a fast and simple method for measuring the alginate composition that can be done by the

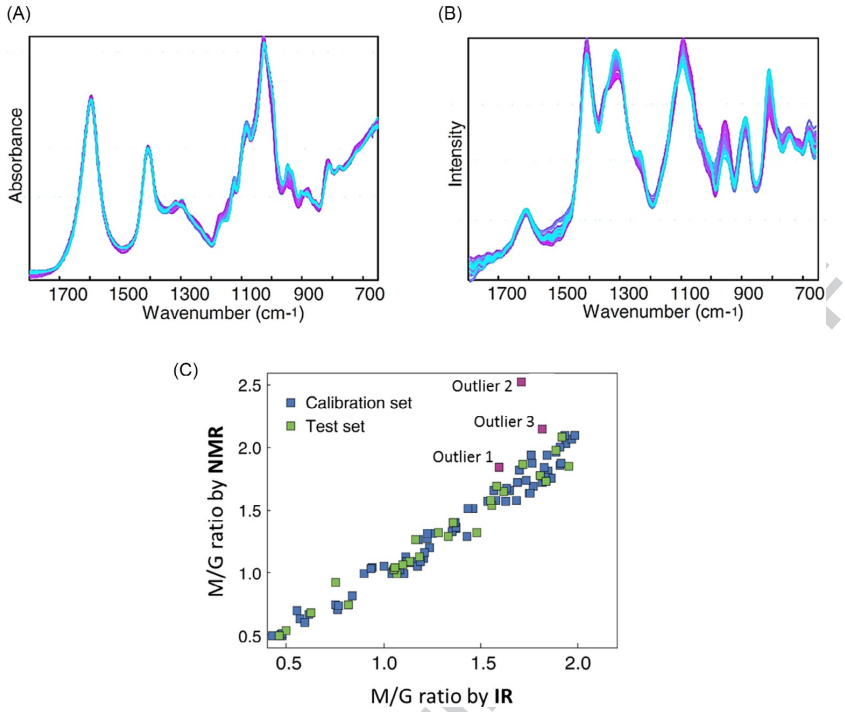


Figure 7 Functional properties of alginate powders. (A) Infrared (IR) spectra colored by the ratio of mannuronic acid (M) and guluronic acid (G) — M/G ratio. (B) Raman spectra colored by M/G ratio. (C) Estimated M/G ratio by IR versus estimated M/G ratio by nuclear magnetic resonance (NMR). Modified from Jensen, H. M.; Larsen, F. H.; Engelsen, S. B. Characterization of alginates by nuclear magnetic resonance (NMR) and vibrational spectroscopy (IR, NIR, Raman) in combination with chemometrics. *Natural Products from Marine Algae: Methods and Protocols*, 2015; pp. 347–363.

factory operators, for example, in an on-site *walk-in laboratory*. Salomonsen et al.[8] compared ATR-IR (Figure 7A), Raman (Figure 7B), and NIR reflectance (spectra not shown) for the determination of the M/G ratio in commercial sodium alginate powders. The spectra in Figure 7A **and B** are colored by the M/G ratio and it is clear that the fingerprint region contains information on this parameter. It was also found that all spectroscopic methods yielded good prediction models with errors compatible with the NMR reference method. In fact vibrational spectroscopy demonstrated to be superior and a more general method to predict the alginate M/G ratio. Vibrational spectroscopy can be applied directly on the raw alginate powders, whereas NMR needs a diluted and hydrolyzed solution of the powders. Moreover, vibrational spectroscopy measures all molecules in the sample in contrast to NMR, which in case of high calcium content samples underestimate the G content (causing overestimation of the M/G ratio). This is because the junction zone guluronic egg-box fragments becomes rigid and thus not visible to the NMR methods. This is shown in Figure 7C where three outliers with high calcium content are highlighted. Examination by wet chemistry showed that the IR estimation of the M/G ratio was the correct one. Vibrational spectroscopy is thus very suited for at-line quality control of alginate powders as it can drastically reduce the analysis cost and time.

Postprocess

Postprocess and shelf life are other important food quality attributes that can be measured by vibrational spectroscopy. Softness as an example is an important sensory quality parameter of barrel salted herring which relates to ripening quality. The ripening period of barrel salted herring is several months. During ripening the proteins degrade and the degradation relates to the ripening characteristics. Estimating the ripening quality by sampling the fish would introduce problems as the fish is heterogeneous. However, throughout the ripening period, the degraded proteins are extracted into the surrounding brine. The protein content in the brine therefore can be used as an indicator of the ripening process. Svensson et al.[10] used NIT spectroscopy to predict the protein content in brine from barrel salted herring. NIR spectroscopy proved a good alternative to the time-consuming Kjeldahl analysis normally used. The results indicate that NIR spectroscopy can be used as a fast and noninvasive method for assessing the protein content (Figure 8), which may be used as an indicator for the ripening quality of barrel salted herring.

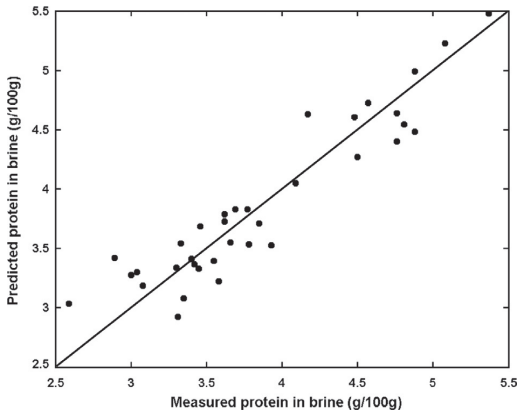


Figure 8 Prediction of protein content in brine of barrel salted herrings from near-infrared measurements. Modified from Svensson, V.; Nielsen, H.; Bro, R. Determination of the protein content in brine from salted herring using near-infrared spectroscopy. *LWT-Food Sci. Technol*. **2004**, *37*, 803–809.

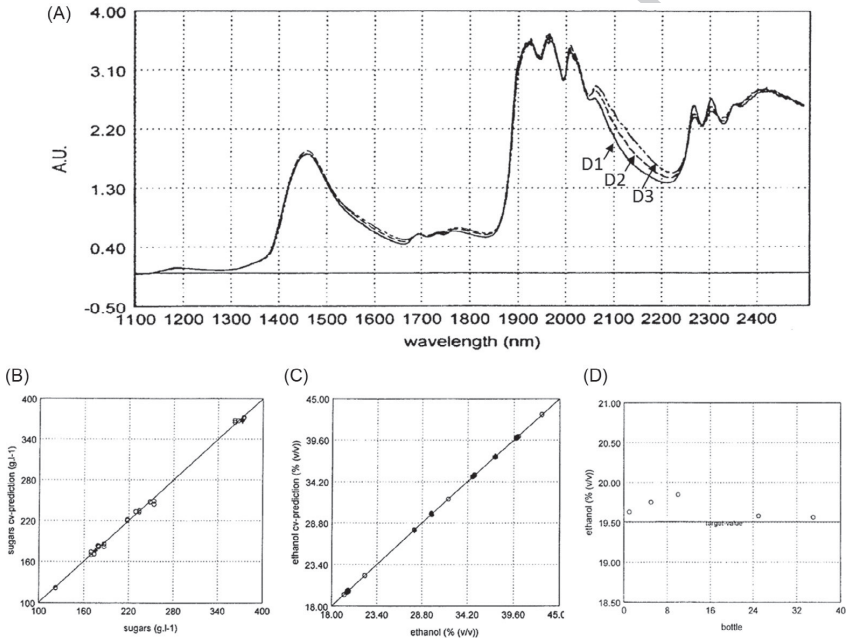


Figure 9 NIR spectroscopy of distils. (A) NIR spectra of three different distils (ethanol %(v/v)/sugar g L^{-1}): (D1) whisky (39.97/0.63), (D2) Dutch brandy (27.88/179.81), and (D3) fruit liquor (19.98/374.60). (B) Measured sugar content versus predicted sugar content. (C) Measured ethanol content versus predicted ethanol content. (D) Predicted ethanol content for a sequence of samples from a bottling line. Modified from van den Berg, F.; van Osenbruggen, W.; Smilde, A. Process analytical chemistry in the distillation industry using near-infrared spectroscopy. *Process Control Qual*. **1997**, *9*, 51–57.

In another feasibility study, van den Berg et al.[11] used NIT measurements to predict ethanol content, sugar content, and density of distils. Ethanol, natural sugars, and water are all NIR active. Together with trace amounts of flavor and color ingredients, these are the three basic components of alcoholic beverages. The NIR spectra of three different alcoholic beverages are shown in Figure 9A. In the study, information after 1830 nm was considered unreliable due the insufficient detector response. However, the absorption bands from 1330 to 1630 nm are caused by water and functional groups in natural sugars, and the band from 1630 to 1830 nm contains alcohol information. It is supposed that the density information is found in the combination of these absorption bands.

Using NIR measurements of the distils, sugar content (Figure 9B) and ethanol content (Figure 9C) were predicted with high accuracy. Hence, NIR spectroscopy could, for example, be used in *100% quality control* during bottling for legislative documenting. During product switch the filling machine is routinely cleaned with detergents followed by an ethanol washing. A memory effect in the pipelines of the filling machine is likely to occur. Hence, the first bottles after a product switch are expected to have higher ethanol content. This was also found in the study by van den Berg et al.[11] where ethanol concentration was plotted for a sequence of five bottles picked from the bottling line after the washing (Figure 9D). The plot demonstrates the memory effect and shows that the first few bottles have an above legal/target limit ethanol content.

Outro

Using six examples, some of the unique features of vibrational spectroscopy in combination with multivariate data analysis in relation to food processing have been demonstrated. The ability to remotely and noninvasively measure the bulk parameters of food in its multiple phases and at very high speed is not surpassed by any other analytical technique. Vibrational spectroscopy and food process analysis are matching just like "herring and snaps."

References

1. Sørensen, K. M.; Petersen, H.; Engelsen, S. B. *Appl. Spectrosc.* **2012**, *66*, 218–226.
2. Sørensen, K. M.; Christensen, M.; Engelsen, S. B. *NIR News* **2013**, *24*, 9–11.
3. Nielsen, J.; Pedersen, D.; Munck, L. *Cereal Chem.* **2003**, *80*, 274–280.
4. Rinnan, A.; van den Berg, F.; Engelsen, S. B. *Trac. Trends Anal. Chem.* **2009**, *28*, 1201–1222.
5. Lyndgaard, C. B.; Engelsen, S. B.; van den Berg, F. W. J. *J. Food Eng.* **2012**, *108*, 345–352.
6. Jensen, J. K.; Ottosen, N.; Engelsen, S. B.; van den Berg, F. *Int. J. Food Eng.* **2015**, *11*, 1–15.
7. Dahl, P.; Christensen, B.; Munck, L.; Larsen, E.; Engelsen, S. *J. Sci. Food Agric.* **2000**, *80*, 365–374.
8. Salomonsen, T.; Jensen, H. M.; Stenbaek, D.; Engelsen, S. B. *Carbohydr. Polym.* **2008**, *72*, 730–739.
9. Jensen, H. M.; Larsen, F. H.; Engelsen, S. B. *Characterization of alginates by nuclear magnetic resonance (NMR) and vibrational spectroscopy (IR, NIR, Raman) in combination with chemometrics. Natural Products from Marine Algae: Methods and Protocols.* **2015**, pp. 347–363.
10. Svensson, V.; Nielsen, H.; Bro, R. *LWT-Food Sci. Technol.* **2004**, *37*, 803–809.
11. van den Berg, F.; van Osenbruggen, W.; Smilde, A. *Process Control Qual.* **1997**, *9*, 51–57.

Further Reading

1. Skibsted, E.; Engelsen, S. B. Spectroscopy for process analytical technology (PAT). In *Encyclopedia of Spectroscopy and Spectrometry*; Lindon, J.; Tranter, G.; Koppenaal, D., Eds.; 2nd ed.; Elsevier: Oxford, **2010**; vol. 3, pp 2651–2661.
2. van den Berg, F.; Lyndgaard, C. B.; Sørensen, K. M.; Engelsen, S. B. *Trends Food Sci. Technol.* **2013**, *31*, 27–35.

Non-Print Items

Abstract:

The analysis of raw materials, *real-time* process control, and end-product plus post process quality evaluations are all crucial steps in food production. In order to increase production throughput, there is a *need for speed* when collecting information from the different processing steps. Hence, conventional methods from analytical chemistry (such as Kjeldahl for protein determination) are not compatible with modern production methods. By examples, this chapter provides an overview of some of the unique features of vibrational spectroscopy in relation to food processing analysis. Vibrational spectroscopy offers noninvasive high-speed measurements of quality attributes during all stages of the food production chain. This is not surpassed by any other analytical method, and vibrational spectroscopic techniques such as Raman, infrared, and near-infrared spectroscopy are therefore widely and increasingly used in the food industry.

Keywords: Food production; Infrared; Near-infrared; Noninvasive analysis; Process analytical technology; Raman; Rapid quality control; Remote analysis; Vibrational spectroscopy

Biographical Sketch

Carl Emil Eskildsen studied process analytical technology at University of Copenhagen and received his M.Sc. degree in 2013. He is currently occupying a position as PhD fellow at University of Copenhagen. His work focuses on vibrational spectroscopy and multivariate data analysis (chemometrics). He has a particular interest in the interface between academia and industry concerning industrial applications of vibrational spectroscopy and chemometrics. (ResearchGate: http://www.researchgate.net/profile/Carl_Eskildsen).

Frans van den Berg has an appreciable experience in data analysis, specifically the application of chemometrics, statistics, and mathematics in (process) data collection, integration, and interpretation/evaluation. After a basic training as laboratory technician, he continued with a civil engineering education in laboratory information and automation. This education was supplemented with a university degree in general chemistry and a PhD in process analytical chemistry with main focus on systems/control theory. After his education he was hired as assistant professor at The Royal Veterinary and Agricultural University in Denmark, and in 2004 as associate professor at the Department of Food Science, Faculty of Science, University of Copenhagen, with specialty in process analytical chemistry and technology, and chemometrics. The focus of his research is on measurements systems (primarily vibrational spectroscopy) for process monitoring and control, and on statistical and mathematical modeling of process-related data challenges. (ResearchGate: http://www.researchgate.net/profile/Frans_Berg).

Søren Balling Engelsen has a PhD in molecular modeling from The Technical University of Denmark and is professor in biospectroscopy at University of Copenhagen. His primary research topic is development and application of high-throughput quantitative spectroscopic methods (NIR, IR, Raman, and NMR) for biological samples in quality control, process analytical technology, foodomics, and metabolomics. A key issue in this line of research is the development of multivariate chemometric methods (exploration, regression, and classification) for optimal spectral information extraction. The second research topic is molecular modeling of polysaccharide structures and their interaction with water and affinity to bioactive metabolites and in turn how to exploit these interactions to tailor functionality and health-promoting effects of food. He is author or coauthor of more than 180 published peer reviewed papers, two patents, and numerous book chapters. Søren is most recently awarded the Tomas Hirschfeld Award by the International Council for Near Infrared Spectroscopy (ICNIRS) for longstanding achievement in NIR spectroscopy. (ResearchGate: https://www.researchgate.net/profile/Soren_Engelsen).

PAPER II

Predicting hydrolysis of whey protein by mid-infrared spectroscopy

N. A. Poulsen, C. E. Eskildsen, M. Akkerman, L. B. Johansen, M. S. Hansen, P.W. Hansen, T. Skov and L. B. Larsen

Accepted for publication, International Dairy Journal, 2016.

Predicting hydrolysis of whey protein by mid-infrared spectroscopy

Nina A. Poulsen[a,*], Carl E. Eskildsen[b], Marije Akkerman[a], Lene B. Johansen[c], Mikka S. Hansen[c], Per W. Hansen[d], Thomas Skov[b], Lotte B. Larsen[a]

[a] *Department of Food Science, Aarhus University, Blichers Allé 20, DK-8830 Tjele, Denmark*

[b] *Department of Food Science, University of Copenhagen, Rolighedsvej 26, DK-1958 Frederiksberg, Denmark*

[c] *Arla Strategic Innovation Centre, Rørdrumvej 2, DK-8220 Brabrand, Denmark*

[d] *FOSS Analytical A/S, Foss Allé 1, DK-3400 Hillerød, Denmark*

***** Corresponding author. Tel.: + 45 87157997

E-mail address: Nina.Poulsen@food.au.dk (N. A. Poulsen)

ABSTRACT

The possibility of using mid-infrared spectroscopy for prediction of hydrolysis of whey protein was explored. Two different protein concentrations (5% and 8%) were trypsinated at different enzyme-to-substrate ratios (1:100 and 1:200) and followed over time (0–8 h). Degree of hydrolysis was determined by use of a fluorescamine assay, whereas size exclusion chromatography (SEC) was used to quantify the loss of intact protein. The results show spectral changes as hydrolysis proceeded over time. Increased absorbance intensities in the region from 1588–1540 cm^{-1} as well as increased absorbance intensities around 1400 cm^{-1} were observed with increasing hydrolysis. These changes were assigned to formation of primary amines as well as an increase in carboxyl groups as hydrolysis proceeded. As the level of free amino terminals correlates with spectral changes, infrared spectroscopy is a promising tool for prediction of degree of hydrolysis of whey proteins.

1. Introduction

There is an increasing market for whey protein products, including whey protein isolates (WPI), concentrates (WPC) and hydrolysates (WPH), given the high nutritional value of whey proteins and derived peptides as well as their excellent functional properties as foaming and emulsifying agents. Disruption of the tertiary structure and reduced molecular weight by enzymatic hydrolysis of whey proteins change their physicochemical and technological properties (Chobert, Bertrand-Harb, & Nicolas, 1988; de Castro, Bagagli, & Sato, 2015). The solubility of whey proteins is known to increase overall with increasing degree of hydrolysis (DH%) (Svenning, Brynhildsvold, Molland, Langsrud, & Vegarud, 2000), and thus improved uptake as a result of enhanced bioavailability of pre-hydrolysed whey protein products are of importance for nutrition in segments of hospitalised patients as well as in sports nutrition (Madureira, Pereira, Gomes, Pintado, & Malcata, 2007). Furthermore, WPH exhibit reduced immunological reactivities (Heyman, 1999), which can be exploited in infant formulas for allergic infants (Halken & Høst, 1997). Predicting DH% is therefore important for all whey products either during processing of WPC or WPI, where hydrolysis is unwanted, or as a tool for quality control for finite degrees of enzymatic hydrolysis in WPH.

Fourier transform infrared spectroscopy (FT-IR) is commonly used in the dairy industry for rapid detection of major milk components, and thus the method is an attractive choice for process control at the dairies due to its ease of operation, rapid output and non-destructive sample preparation (Andersen, Hansen, & Andersen, 2002). For milk, promising FT-IR prediction models have been generated for milk fatty acid composition (Eskildsen et al., 2014; Soyeurt et al., 2006), protein profiling (Rutten, Bovenhuis, Heck, & van Arendonk, 2011) and

coagulation properties of raw milk (Cecchinato, De Marchi, Gallo, Bittante, & Carnier, 2009), which suggest that more detailed information on milk composition could be revealed from FT-IR.

Proteins have three major absorption bands in the mid-infrared (MIR) region, the amide A, the amide I and amide II bands. The amide A band originates from the N–H stretching vibration of the peptide backbone, which is found at approximately 3300 cm^{-1}. However, due to a strong water absorption between 4000 and 3000 cm^{-1}, the amide A band is rarely used. The amide I band primarily arises from the stretching vibration of the carbonyl groups in the peptide backbone. However, this absorption is found at approximately 1660 cm^{-1} and is masked by the other water signal at approximately 1560–1715 cm^{-1} (Bruice, 2007). The amide II band is mainly due to a coupled N–H bending vibration and C–N stretching vibration, which absorbs at approximately 1540 cm^{-1}. The exact vibrational frequencies of the amide I and amide II bands depend on whether the nitrogen of the N–H and the carbon of the C=O bonds are engaged in peptide bonds or not, i.e., hydrolysed. When hydrolysing the peptide bond, a carboxylic acid from the C-terminus and a primary amine from the N-terminus will be formed. Inside the peptide, prior to hydrolysis, the carbonyl group will be exposed to electron donation from the nitrogen atom. However, after peptide bond cleavage the C-terminal carbonyl group will no longer be exposed to this donation and the vibrational frequency of the carbonyl stretching vibration will move from the 1640 cm^{-1} region to approximately 1700 cm^{-1}. Furthermore, during hydrolysis the amide II band should decrease as the C–N bond is cleaved. Moreover, the N–H bend of a primary amide is found at approximately 1600 cm^{-1}. Hence, it is expected that increased absorbance intensities in the region around 1600 cm^{-1} will be observed as hydrolysis proceeds. Apart from the N- and C-termini, in intact proteins the only free amino and carboxylic

residues come from the side chains of the Asp, Glu and Lys residues. However, as hydrolysis proceeds, the amount of free amino and carboxylic residues increases, which is expected to give a change in the infrared spectrum. Furthermore, changes in secondary structure of proteins affect the absorbance pattern and forced changes due to pH can result in significant changes of IR absorption bands (Güler, Dzafic, Vorob'ev, Vogel, & Mantele, 2011).

Van der Ven et al. (2002) analysed different casein and whey hydrolysates produced using various commercial enzymes and found that FT-IR could effectively discriminate between casein and whey hydrolysates and between different enzymatic treatments. However, development in the spectra as hydrolysis proceeded was not followed. From an industry perspective, the main interest will be to monitor the level of intact protein and DH%. DH can be defined as the percentage of peptide bonds cleaved in relation to the total number of peptide bonds in the protein solution (Adler-Nissen, 1986). Various fluorescent methods based on fluorimetric *o*-phthaldialdehyde (OPA), fluorescamine or trinitrobenzene sulphonic acid (TNBS) have been developed for monitoring the level of free amino groups, and thereby free amino terminals, in protein solutions. These measurements have been found to correlate to the cleavage of protein bonds during hydrolysis (Spellman, McEvoy, O'Cuinn, & FitzGerald, 2003). Likewise, size exclusion chromatography (SEC) could be used to monitor changes in the hydrodynamic volume of proteins and thus, changes in the molecular weight distribution, which is affected by hydrolysis.

The objective of the present study was to explore the use of FT-IR for prediction of hydrolysis in whey protein. Trypsin was used to hydrolyse peptide bonds involving carboxylic groups of arginine and lysine residues at different time intervals. Thereby the content of intact versus hydrolysed protein could be followed over time. Fluorescamine assay and SEC were used

as reference methods for quantification of DH% and prediction models were developed based on the recorded FT-IR spectra.

2. Material and methods

2.1. Materials

Whey protein concentrate with 80–85% protein, 5–9% lactose, max 8% fat and max 3.5% ash based on dry weight (Lacprodan®-80, Arla Foods Ingredients, Nr. Vium, Denmark), produced from whey by ultrafiltration and spray drying, was used as substrate for monitoring hydrolysis. Aqueous WPC solutions of 5 and 8% (w/w) total protein were prepared and stored at 4 °C until required. Initial protein content (g 100 g^{-1} solution), together with fat, lactose and dry matter content, of WPC solutions was validated by infrared spectroscopy using Milkoscan™ FT2 (Foss, Hillerød, Denmark) (Table 1).

2.2. Enzymatic hydrolysis

Enzymatic hydrolysis was performed in 50 mL tubes. Trypsin from porcine pancreas with 13,000–20,000 BAEE units mg^{-1} protein (T0303, Sigma-Aldrich, St. Louis, MO, USA) in phosphate buffered saline (PBS) (0.125 M $NaH_2PO_4 \cdot H_2O$, 0.1 M NaCl, pH 7.4) was added to the samples in a ratio of 5 or 10 g trypsin per kg of soluble protein resulting in enzyme to substrate ratios (E/S) of 1:200 or 1:100 (w/w). After trypsin addition, samples were incubated at 37 °C in time intervals from 0–8 h. pH was monitored for each sample, but was not adjusted throughout

the enzymatic process. Trypsin was inactivated by addition of trypsin inhibitor (TT2011, Sigma-Aldrich) in 0.067 M sodium phosphate buffer (pH 7.6) in 4.7×10^{-3} M excess of trypsin. Duplicate samples were taken at t = 0, 0.5, 1, 2, 3, 4, 5, 6, 7 and 8 h for each protein concentration (5 and 8%, w/w) and E/S (1:100, w/w, for both protein concentrations and 1:200 for 5%). For one of the 8% (w/w) replicates at 1:100 E/S, samples at t = 6 and t = 7 h were not collected resulting in a total of 58 samples. Immediately after trypsin inactivation, the samples were analysed by infrared spectroscopy (Milkoscan™ FT2, Foss, Hillerød, Denmark) and recorded full FT-IR spectra were saved for further analysis. In addition, 3mL of each sample was frozen at –20 °C for further analysis of DH% using fluorescamine assay and SEC as reference methods.

2.3. Fluorescamine assay

Enzymatic hydrolysis was monitored by a modified fluorescamine assay (Pearce, 1979; Rollema et al., 1989). WPC samples were mixed with 25% trichloroacetic acid (TCA) solution in ratio of 1:1 and left on ice to precipitate the intact proteins. Subsequently, samples were centrifuged at 19,400 × *g* for 20 min at 4 °C to sediment intact proteins. The soluble phase was diluted (×100) with 1 M HCl, and 37.5 µL of the dilution was mixed with 1130 µL 0.1 M sodium-borate buffer (pH 8.0) along with 375 µL 20% fluorescamine in a dried acetone solution. Hereafter, 250 µL of this mixture was transferred to white opaque microtitre plates (Corning Inc. New York, NY, USA) and subsequently fluorescence was determined with a Synergy2 spectrophotometer (Biotek Instruments Inc., Winooski, VT, USA) using excitation wavelengths at 400 nm and emission wavelengths at 485 nm at 25 °C. Each sample was measured in

triplicates. The level of primary amino groups reflecting the level of free amino-terminals as L-leucine equivalents (Leucine eqv.) was calculated according to a standard curve in the range of 0.5-30 mM Leucine eqv.

DH% was calculated as:

$$\frac{Leucine\ eqv_{\ sample}}{Leucine\ eqv_{\ total\ hydrolysis}} * 100\% \qquad (1)$$

where Leucine eqv. of total hydrolysis was calculated according to that total hydrolysis of a definite whey protein solution, which has been estimated to be 8.8 mEq g^{-1} of protein for whey protein (Adler-Nissen, 1986). The theoretical trypsin specific digestions of α-LA and β-LG were calculated using PeptideMass from ExPASy (www.expasy.org). Resulting DH% were 11.4% for α-LA and 11.1% for β-LG.

2.4. Size exclusion chromatography

For peptide and protein separation, SEC was performed by use of a 1260-Infinity semi-preparative liquid chromatography (LC) system (Agilent Technologies, Santa Clara, CA, USA). Prior to analysis samples were defrosted and 1 mL denaturing and reducing buffer containing 6 M urea, 0.1 M trisodium citrate and 20 µL 0.5 M dithioerythritol (DTE) was added to 200 µL sample. The samples were incubated at 37 °C for 60 min while stirring to denature and reduce inter- and intra-molecular disulphide bonds. Hereafter the samples were centrifuged at 9300 × *g* for 10 min at 4 °C, and the supernatant was collected for further analysis. 25 µL was injected on a BioSEC-3 Column (3 µm, 7.8 × 300 mm, 300 Å, Agilent Technologies) at 30 °C. The mobile phase comprised 0.2 M sodium phosphate buffer, 0.2 M sodium chloride, pH 7.0. The column separates protein and peptides in the molecular range 5–1250 kDa. A flow rate of 1 mL min^{-1}

was applied and UV signals were detected at 214 nm. All samples were analysed in single determinations. SEC of individual samples were separated into ten regions integrating major peaks, which were used for further analysis (3.1–4.8; 4.8–5.7; 5.7–6.1; 6.1–6.4; 6.4–6.9; 6.9–7.1; 7.1–7.5;7.5–8.7; 8.7–8.8 and 8.8–11.8 min). Mass spectrometry was used to identify the major protein and peptide contents of the WPC using LC coupled with electrospray injection mass spectrometry (ESI-MS) according to the method described by Rauh et al. (2015).

2.5. Fourier transform infrared spectroscopy

FT-IR spectra were obtained in transmittance mode in the range from 5008 to 925 cm^{-1}, with a total of 1060 data points for each sample. To obey Beer's law, the spectra were transformed from transmittance into absorbance before modelling. The region from 5008 to 2980 cm^{-1} and 979 to 925 cm^{-1} was considered noise and removed from the data set. The region from 2822 to 1769 cm^{-1} contained no valuable information and was also removed. Furthermore, the saturated water signal (O–H bend) was removed. For spectral interpretation, a narrow band from 1704 to 1588 cm^{-1} around the O–H bend was removed. However, for modelling, a wider band from 1715 to 1560 cm^{-1} around the O–H bend was removed to avoid instabilities related to water absorption. Full FT-IR spectra were recorded in triplicate for each sample, and the average spectrum was calculated and used for further analysis.

2.6. Data analysis

Data were analysed using MATLAB version R2014a (8.3.0.532, MathWorks Inc., Natick, MA, USA) and PLS_Toolbox, version 7.5 (Eigenvector Research Inc., Manson, WA, USA). The mean centred spectra were modelled by principal component analysis (PCA) (Wold, Sjöström, & Eriksson, 2001). Modelling by PCA was done subsequently using the samples consisting of 5% protein concentration and 1:200 E/S, 5% protein concentration and 1:100 E/S, and 8% protein concentration and 1:100 E/S. Furthermore, partial least square (PLS) regression (Wold et al., 2001) was performed on the combined data set. The FT-IR measurements were used as independent variables and Leucine eqv. was used as dependent variable. The PLS model was cross-validated with respect to protein and enzyme concentration. Prior to PLS modelling, the spectra were orthogonalised with respect to protein concentration. This was done to remove offset differences in the spectra related to initial protein concentration. The orthogonalisation was done as shown in the equation below,

$$X_{-k} = (I - k(k^T k)^{-1} k^T)X \quad (2)$$

where X_{-k} is the FT-IR spectra (centred) without variation (offsets) related to protein concentration, I is the identity matrix, k is a vector containing information on the protein content (centred) and X is the centred FT-IR measurements. Superscript T indicates transpose.

Preprocessing (Savitzky-Golay, Standard Normal Variate or Multiplicative Scatter Correction) of spectra did not improve PCA or PLS models.

3. Results and discussion

3.1. Fluorescamine assay and degree of hydrolysis

Fig. 1 shows the increase of free amino acid terminals measured by fluorescamine assay over the 8 h period. Each time point is offset corrected by subtracting the amount of free amino acid terminals at time zero. The increase varied between treatments, but for all samples the largest increase in free amino terminals occurred within the first hour followed by a slower increase up to 2 h, after which the content to some extent levelled off. After 8 h, the lowest content of amino acid terminals was found for 5% WPC solutions with the lowest E/S (1:200), whereas the high E/S (1:100) increased similarly independent of protein concentration. The fluorescamine assay has been successfully applied in other studies targeting proteolysis in high cell count milk (Larsen, Rasmussen, Bjerring, & Nielsen, 2004; Larsen et al., 2010), or storage quality of UHT milk (Jansson et al., 2014). Application to whey protein solutions, such as WPC, did seem to be more challenging, however, as also reflected in the variable results for duplicate measurements, especially for 5% whey protein (1:100), despite the clear general trends. This might be related to whey protein aggregates or eventually to modifications of the whey proteins during processing and storage through modification of the lysine side chains by, e.g., Maillard reactions.

End point DH%, calculated from the level of free amino acid terminals in the samples, varied among samples relative to protein content and enzyme concentration. Thus DH% after 8 h of hydrolysis was 15% for the 5% protein solution with low E/S, while it was 22% for 5% protein and 12% for 8% protein for the high E/S. Longer hydrolysis time would probably have resulted in minor DH% increases as the change in fluorescamine had not fully reached its plateau after the 8 h.

The obtained DH% for the 5% protein solution at low E/S and for the 8% solution at high E/S were close to the theoretical DH% of 11% for α-LA and β-LG, whereas at high E/S for the

5% protein solutions, DH% was higher than expected. This could suggest that more unspecific cleavages by trypsin is going on at high E/S for the 5% protein solutions in contrast to the 8% substrate solution, and may indicate higher accessibility of the whey proteins in the low protein solutions as the DH% would otherwise have been the same with similar high E/S irrespective of WPC protein substrate concentration. Commercial whey hydrolysates can have up to 30% DH. High DH% in WPH can be unwanted due to increasing number of bitter peptides, which can limit the utilisation of whey powder solutions (Spellman, O'Cuinn, & FitzGerald, 2009), and monitoring the development in DH% is of high value in relation to process control in production of whey protein based ingredients. In a sensory study by Liu, Jiang, and Peterson (2014), WPH bitterness was assigned to principally four main peptides, which on the other hand also possessed bioactive properties and removal therefore could also affect the health potential of WPH.

3.2. Molecular mass distribution

In Fig. 2A, a representative SEC molecular mass distribution profile of the intact WPC solution and the corresponding hydrolysed samples at different time points is given for the 5% protein and 1:100 E/S samples. Mass identification of the peak regions confirmed that protein aggregates eluted first in region 1 and 2, followed by intact β-LG (variant A and B) and α-LA, in region 3, large peptides, mainly from β-CN in region 4, followed by peptides with decreasing size in region 5–7 as and finally as buffers and amino acids in region 8–10. The intact WPC sample has a molecular mass distribution that is centred to the left, reflecting larger aggregates and intact protein and only to a lesser degree contains smaller peptides. All the hydrolysed samples show similar profiles with a distribution towards smaller molecules, reflecting generated

peptides as a result of hydrolysis. Fig. 2B shows scores for the first principal component (PC1) as a function of time based on the PCA for the ten SEC regions. Generally, the total absorbance area was higher for all 8% samples, which explains the separation of 8% and 5% by the first principal component through the whole elution (Fig. 2B). However, the main separation was between the unhydrolysed samples and the hydrolysed samples. The unhydrolysed samples had higher content of fractions with high molecular mass (elution time below 6 min), whereas the hydrolysed samples had higher content of the low molecular mass regions reflecting the disappearance of intact protein and aggregates. Enzyme concentration only affected the SEC profiles to a minor degree even though some differences seem to be apparent after 4 h of hydrolysis. The low enzyme concentration resulted in a slightly higher total area, which might reflect a loss of minor peptides at high enzyme concentration, thereby lowering the total area.

3.3. FT-IR modelling of hydrolysis

Unprocessed FT-IR spectra obtained from the enzymatic hydrolysis of milk samples with varying protein and enzyme contents are presented in Fig. 3A. Samples with 8% and 5% protein content are clearly separated with almost parallel spectra across all wavenumbers, reflecting not only higher protein content but also higher levels of fat and lactose in the 8% sample (Table 1). Hydrolysis time clearly affected the unprocessed FT-IR spectral output, and changed absorbance pattern was observed for several spectral regions. Fig. 3B shows that absorbance intensities increases in the region from 1588 cm^{-1} to ~1540 cm^{-1} at increasing number of Leucine eqv. (increasing hydrolysis time) and thereby at increasing DH%. This is due to the formation of primary amines resulting from the hydrolysis. Furthermore, it is found that the band at ~1400

cm^{-1} is increasing at increased hydrolysis level (Fig. 3C). It is not clear what is driving the observed changes in this spectral region, but it may be related to the increase in carboxyl groups at the C-terminus of the generated peptides, and changes in the side-chains of the residues becoming more exposed and thereby visible in the spectrum and/or a result of changed hydrogen bonding as hydrolysis proceeds. Expected changes in the amide II band in the region around 1600 cm^{-1} were not observed. Consequently, hydrolysis directed changes in the C–N stretching are not evident, which may be due to dominating N–H bending vibrations within this band (Bruice, 2007), which are becoming stronger at hydrolysis. It was expected to observe hydrolysis related changes in the 1700 cm^{-1} region due to increase in terminal carbonyl groups, being part of the new carboxyl groups after hydrolysis. However, these changes were not observed in the spectra. 1700 cm^{-1} is on the edge of the water signal (O–H bend) and this probably masks the carbonyl signal. Separation by FT-IR of hydrolysates generated from different protein sources and classes of proteolytic enzymes by van der Ven (2002), was mainly due to changes in the spectral regions around 1743 and 1705 cm^{-1}, around 1585 cm^{-1}, and around 1400 cm^{-1}. The spectral region around 1400 cm^{-1} was assigned to carboxylate (O–C–O) stretching vibrations. Byler and Farrell (1989) observed changes in the same region upon binding of Ca^{2+}, mainly to glutamic and aspartic acid residues in the caseins, which have side chains with carboxylic acid groups. This confirms that the spectral region around 1400 cm^{-1} could be related to carboxylic acid groups. Apart from the hydrolysis derived spectral changes, these differences could also be related to pH differences affecting the presence of ionised and non-ionised carboxyl groups, but neutralisation of pH did not remove the spectral differences (results not shown). Therefore we do not believe that the small pH differences observed as a function of hydrolysis time (pH decreased from 6.4 to 6.25) explain the spectral changes observed.

Fig. 4 shows scores from the first principal component (PC1) of a PCA model as a function of hydrolysis time. The PCA scores are offset corrected. The PCA model is based on the FT-IR spectra. The figure shows a development in the scores (and thereby the spectra) as a function of hydrolysis time. Dramatic changes in scores are observed within the first 30 min of hydrolysis comparable with the observations on the development in the level of free N-terminals and SEC data. Likewise, no clear plateau is reached after 8 h of hydrolysis, which suggests that there are still some changes occurring in the spectra. To check whether these changes were due solely to time effects, a control sample without added trypsin was monitored over 8 h. No changes in the spectra (data not shown) were observed, confirming that our results are derived from hydrolysis.

In Fig. 5 the scores from PC1 are plotted as a function of Leucine eqv. Expected correlations are observed. Furthermore, a 1 component (latent variables) PLS model was constructed to predict Leucine eqv. from the FT-IR measurements. The PLS model was based on the FT-IR spectra orthogonalised with respect to protein concentration. The measured versus predicted (cross-validated) values are shown in Fig. 6. Both the correlations between the PCA scores and the Leucine eqv. (Fig. 5) and the PLS predictions (Fig. 6) indicate that the FT-IR spectra contain information related to hydrolysis.

It should be noted that Leucine eqv. are used as a measure for hydrolysis. The standard unit for hydrolysis is DH%. However, as DH% is a relative measure it will not have a linear relationship with absorbance. On the other hand, Leucine eqv. describes the level of primary amines and will therefore have a linear relationship with absorbance through Beer's law.

3.4. Reference method for degree of hydrolysis

Use of the fluorescamine assay as reference method for DH still needs to be optimised to obtain more consistent results. However, as can be seen in Fig. 5, there is a good correlation between the PCA scores for the first latent variable from the FT-IR and L-leucine equivalents (R^2 = 0.87–0.93), and further the increase in free amino acid terminals and the spectral changes as a course of hydrolysis time are similar (Figs. 1 and 4). In this way, FT-IR measurements in combination with PCA can be used to describe the hydrolysis. Consequently, a good PLS model could also be constructed using the FT-IR measurements to predict the Leucine eqv. (Fig. 6). Compared with IR-prediction of detailed milk composition in raw milk, which might be driven by indirect covariance structure to, e.g., fat and protein rather than direct spectroscopic fingerprints (Eskildsen et al., 2014), the spectroscopic fingerprint of hydrolysis seems to be direct.

Fluorescamine and, e.g., the TNBS method target the free amino groups being generated during hydrolysis, and in experiments where several small peptides are generated, these methods can be superior compared with gel electrophoresis, which is better for targeting large fragments (Chove, Grandison, & Lewis, 2011). Furthermore, the fluorescamine assay is very sensitive at low levels of proteolysis, which can be important for quality control in bulk milk. However, as a tool to distinguish changes in WPH over time, FT-IR spectroscopy demonstrates a promising perspective, due to a strong sensitivity of hydrolysis derived changes even in a complex matrix, like milk.

The fluorescamine method has a good potential to be reference method, but still needs some optimisation for WPC solutions, which, compared with raw milk, are more complex due to potential modifications of the whey protein powders during storage as well as batch to batch variations.

4. Conclusions

The aim of the study was to explore the use of FT-IR for prediction of hydrolysis in WPC products. The results show spectral changes as hydrolysis proceeded over time, resulting in increased absorbance intensities in specific regions of the FT-IR spectra with increasing hydrolysis. These changes were assigned to formation of primary amines from new amino terminals as well as an increase in carboxyl groups as hydrolysis proceeds. As the level of free amino terminals correlates with spectral changes, we conclude that infrared spectroscopy is a promising tool for prediction of DH% of whey proteins. FT-IR can easily be applied for quality control at industry scales; however, the prediction model still needs to be validated for industry application including more complex enzyme cocktails and higher DH%.

Acknowledgements

The Danish Council for Strategic Research is acknowledged for financial support to the project under the inSPIRe consortium (Copenhagen, Denmark). Furthermore, Arla Foods amba (Viby J, Denmark) and FOSS Analytical A/S (Hillerød, Denmark) is acknowledged for financial support to the project.

References

Adler-Nissen, J. (1986). A review of food protein hydrolysis specific areas. In J. Adler-Nissen (Ed.), *Enzymic hydrolysis of food proteins* (pp. 57–131). Barking, Essex, UK: Elsevier.

Andersen, S. K., Hansen, P. W., & Andersen, H. V. (2002). Vibrational spectroscopy in the analysis of dairy products and wine. In J. M. Chalmer, & P. R. Grifiths (Ed.), *Handbook of vibrational spectroscopy* (pp. 3672–3681). West Sussex, UK: John Wiley & Sons Ltd.

Bruice, P. Y. (2007). Mass spectrometry, infrared spectroscopy and ultraviolet/visible spectroscopy. In P. Y. Bruice (Ed.), *Organic chemistry* (pp. 512–568). Upper Saddle River, NJ, USA: Pearson Education, Inc.

Byler, D. M., & Farrell, H. M., Jr. (1989). Infrared spectroscopic evidence for calcium ion interaction with carboxylate groups of casein. *Journal of Dairy Science*, *72*, 1719–1723.

Cecchinato, A., De Marchi, M., Gallo, L., Bittante, G., & Carnier, P. (2009). Mid-infrared spectroscopy predictions as indicator traits in breeding programs for enhanced coagulation properties of milk. *Journal of Dairy Science*, *92*, 5304–5313.

Chobert, J. M., Bertrand-Harb, C., & Nicolas, M. G. (1988). Solubility and emulsifying properties of caseins and whey proteins modified enzymically by trypsin. *Journal of Agricultural and Food Chemistry*, *36*, 883–892.

Chove, L. M., Grandison, A. S., & Lewis, M. J. (2011). Comparison of methods for analysis of proteolysis by plasmin in milk. *Journal of Dairy Research*, *78*, 184–190.

de Castro, R. J. S., Bagagli, M. P., & Sato, H. H. (2015). Improving the functional properties of milk proteins: focus on the specificities of proteolytic enzymes. *Current Opinion in Food Science*, *1*, 64–69.

Eskildsen, C. E., Rasmussen, M. A., Engelsen, S. B., Larsen, L. B., Poulsen, N. A., & Skov, T. (2014). Quantification of individual fatty acids in bovine milk by infrared spectroscopy

and chemometrics: Understanding predictions of highly collinear reference variables. *Journal of Dairy Science*, *97*, 7940–7951.

Güler, G., Dzafic, E., Vorob'ev, M. M., Vogel, V., & Mantele, W. (2011). Real time observation of proteolysis with Fourier transform infrared (FT-IR) and UV-circular dichroism spectroscopy: Watching a protease eat a protein. *Spectrochimica Acta Part A – Molecular and Biomolecular Spectroscopy*, *79*, 104–111.

Halken, S., & Høst, A. (1997). How hypoallergenic are hypoallergenic cow's milk-based formulas? *Allergy*, *52*, 1175–1183.

Heyman, M. (1999). Evaluation of the impact of food technology on the allergenicity of cow's milk proteins. *Proceedings of the Nutrition Society, 58,* 587-592.

Jansson, T., Clausen, M. R., Sundekilde, U. K., Eggers, N., Nyegaard, S., Larsen, L. B., et al. (2014). Lactose-hydrolyzed milk is more prone to chemical changes during storage than conventional ultra-high-temperature (UHT) milk. *Journal of Agricultural and Food Chemistry*, *62*, 7886–7896.

Larsen, L. B., Hinz, K., Jorgensen, A. L. W., Moller, H. S., Wellnitz, O., Bruckmaier, R. M., et al. (2010). Proteomic and peptidomic study of proteolysis in quarter milk after infusion with lipoteichoic acid from *Staphylococcus aureus*. *Journal of Dairy Science*, *93*, 5613–5626.

Larsen, L. B., Rasmussen, M. D., Bjerring, M., & Nielsen, J. H. (2004). Proteases and protein degradation in milk from cows infected with *Streptococcus uberis*. *International Dairy Journal*, *14*, 899–907.

Liu, X. W., Jiang, D. S., & Peterson, D. G. (2014). Identification of bitter peptides in whey protein hydrolysate. *Journal of Agricultural and Food Chemistry*, *62*, 5719–5725.

Madureira, A. R., Pereira, C. I., Gomes, A. M. P., Pintado, M. E., & Malcata, F. X. (2007). Bovine whey proteins - Overview on their main biological properties. *Food Research International*, *40*, 1197–1211.

Pearce, K. N. (1979). Use of fluorescamine to determine the rate of release of the caseinomacropeptide in rennet-treated milk. *New Zealand Journal of Dairy Science and Technology*, *14*, 233–239.

Rauh, V. M., Johansen, L. B., Bakman, M., Ipsen, R., Paulsson, M., Larsen, L. B. et al. (2015). Protein lactosylation in UHT milk during storage measured by liquid chromatography–mass spectrometry and quantification of furosine. *International Journal of Dairy Technology*, *68*, 4876–494.

Rollema, H. S., McKellar, R. C., Sorhaug, T., Suhren, G., Zadow, J. G., Law, B. A., et al. (1989). Comparison of different methods for the detection of bacterial proteolytic-enzymes in milk. *Milchwissenschaft*, *44*, 491–496.

Rutten, M. J. M., Bovenhuis, H., Heck, J. M. L., & van Arendonk, J. A. M. (2011). Predicting bovine milk protein composition based on Fourier transform infrared spectra. *Journal of Dairy Science*, *94*, 5683–5690.

Soyeurt, H., Dardenne, P., Dehareng, F., Lognay, G., Veselko, D., Marlier, M., et al. (2006). Estimating fatty acid content in cow milk using mid-infrared spectrometry. *Journal of Dairy Science*, *89*, 3690–3695.

Spellman, D., McEvoy, E., O'Cuinn, G., & FitzGerald R. J. (2003). Proteinase and exopeptidase hydrolysis of whey protein: Comparison of the TNBS, OPA and pH stat methods for quantification of degree of hydrolysis. *International Dairy Journal*, *13*, 447–453.

Spellman, D., O'Cuinn, G., & FitzGerald, R. J. (2009). Bitterness in *Bacillus* proteinase hydrolysates of whey proteins. *Food Chemistry*, *114*, 440–446.

Svenning, C., Brynhildsvold, J., Molland, T., Langsrud, T., & Vegarud, G. E. (2000). Antigenic response of whey proteins and genetic variants of β-lactoglobulin – the effect of proteolysis and processing. *International Dairy Journal*, *10*, 699–711.

van der Ven, C., Muresan, S., Gruppen, H., de Bont, D. B. A., Merck, K. B., & Voragen, A. G. J. (2002). FTIR spectra of whey and casein hydrolysates in relation to their functional properties. *Journal of Agricultural and Food Chemistry*, *50*, 6943–6950.

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

Figure legends

Fig. 1. Fluorescamine reference method showing level of free amino terminals as L-leucine equivalents (Leucine eqv.) in whey protein concentrate solution with protein concentrations of 5% (WPC5) at two different enzyme:substrate ratios (1:100, 1:200) and 8% (WPC8) at enzyme:substrate ratio of 1:100. Rep1 (closed symbols) and rep2 (open symbols) indicate the two replicates. Error bars represent SD for the triplicate fluorescamine measurements. Fluorescamine values are offset-corrected.

Fig. 2. Representative size exclusion chromatography (SEC) chromatograms for the enzymatic hydrolysis of WPC solution with 5% protein content and E/S of 1:100 followed for up to 8 h (A) and first principal component (PC1) as a function of time (B) based on ten SEC fractions, integrating major peaks of the three time series (3.1–4.8, 4.8–5.7, 5.7–6.1, 6.1–6.4, 6.4–6.9, 6.9–7.1, 7.1–7.5, 7.5–8.7, 8.7–8.8 and 8.8–11.8 min); protein concentrations of 5% (w/w) at two different enzyme:substrate ratios (1:100, 1:200) and protein concentrations of 8% (w/w) at enzyme:substrate ratio 1:100. All samples are in dublicates.

Fig. 3. Full Fourier transform infrared (FT-IR) spectra of aqueous whey protein concentrate solution with protein concentrations of 5% (w/w) at two different enzyme:substrate ratios (1:100, 1:200) and protein concentrations of 8% (w/w) at enzyme:substrate ratio 1:100: (A) raw spectra; (B) zoom-in at 1550 cm^{-1}, amide II band; (C) zoom-in at 1400 cm^{-1}.

Fig 4. Score plot from principal component analysis (PCA) showing scores for the first principal component (PC1) as a function of time. The PCA scores are offset corrected. The PCA model is based on FT-IR spectra of whey protein concentrate solution with protein

concentrations of 5 g 100 g^{-1} (WPC5) at two different enzyme-to-substrate ratios (1:100, 1:200) and 8 g 100 g^{-1} (WPC8) at enzyme:substrate ratio 1:100. The spectra were collected over time while hydrolysis was ongoing. Rep1 (closed symbols) and rep2 (open symbols) indicate the two replicates.

Fig 5. Correlation between leucine equivalents and principal component analysis (PCA) scores of the first principal component (PC1). The PCA model is based on FT-IR spectra of whey protein concentrate solution with protein concentrations of 5 g 100 g^{-1} (WPC5) at two different enzyme:substrate ratios (1:100, 1:200) and 8 g 100 g^{-1} (WPC8) at enzyme:substarte ratio 1:100. Rep1 (closed symbols) and rep2 (open symbols) indicate the two replicates.

Fig 6. Results from partial least squares regression: measured L-leucine equivalents (Leucine eqv.) versus cross-validated predicted Leucine eqv. LV = number of latent variables; RMSECV = root mean square error of cross-validation; R2 = coefficient of determination.

Table 1

Composition of WPC solutions based on Milkoscan measurements. [a]

WPC solution	Protein	Fat	Lactose	Dry matter
5%, 1:100	5.12	0.73	2.42	8.30
5%, 1:100	5.06	0.72	2.43	8.25
5%, 1:200	4.99	0.73	2.43	8.17
5%, 1:200	5.03	0.73	2.43	8.20
8%, 1:100	7.85	0.91	2.32	11.58
8%, 1:100	7.74	0.90	2.32	11.43

[a] WPC solutions of 5 and 8% (w/w) total protein were prepared with trypsin added to result in enzyme to substrate ratios of 1:200 or 1:100 (w/w). Composition was determined before hydrolysis (0 h); data are expressed in g 100 g^{-1} milk.

Figure 1 – grey-scale for print edition

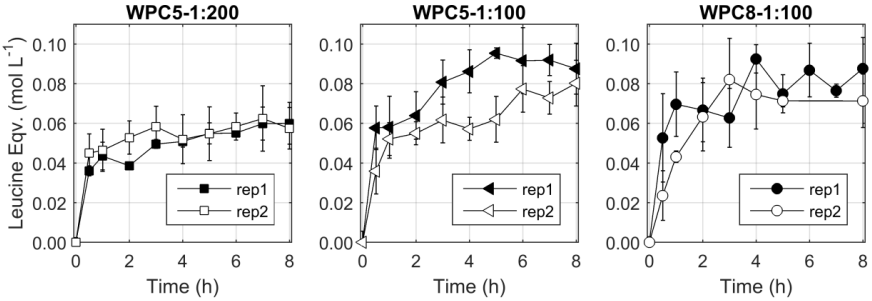
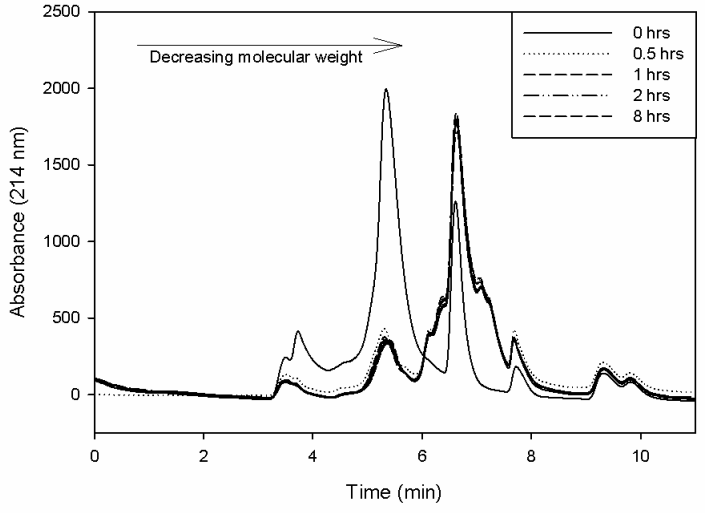


Figure 2

A


Decreasing molecular weight
0 hrs
0.5 hrs
1 hrs
2 hrs
8 hrs
Absorbance (214 nm)
Time (min)

B

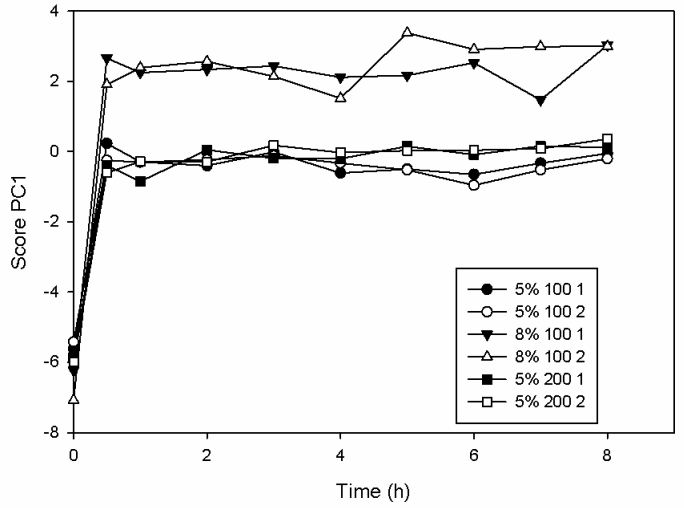

5% 100 1
5% 100 2
8% 100 1
8% 100 2
5% 200 1
5% 200 2
Score PC1
Time (h)

Figure 3 – grey-scale for print edition

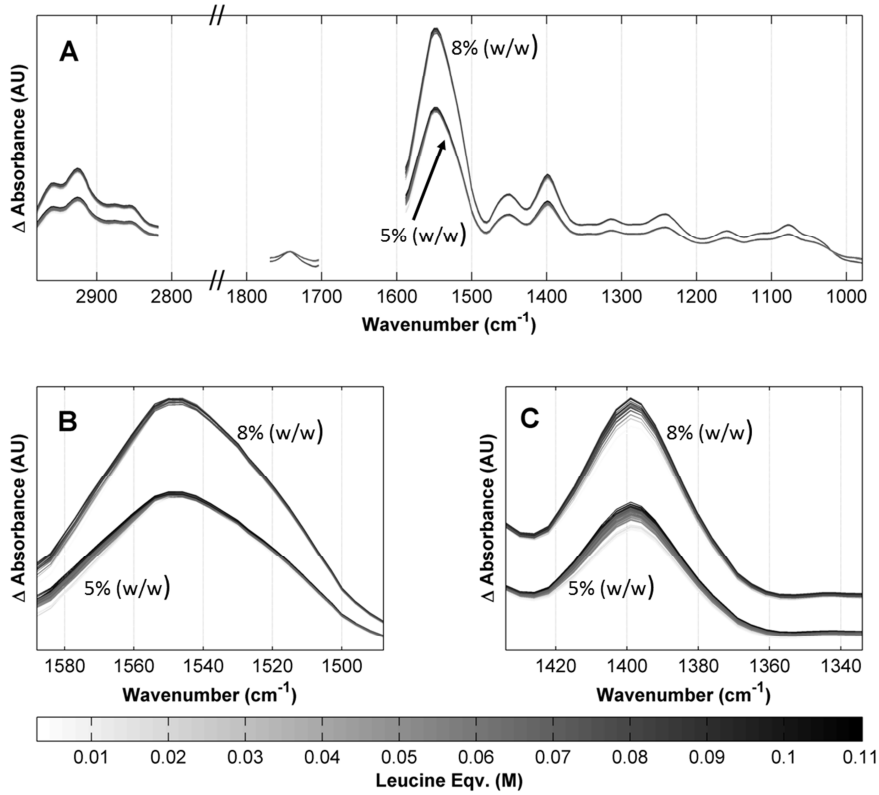


Figure 4 – grey-scale for print edition

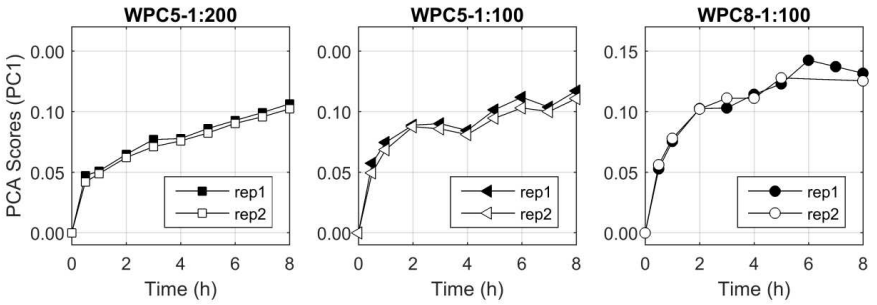


Figure 5 – grey-scale for print edition

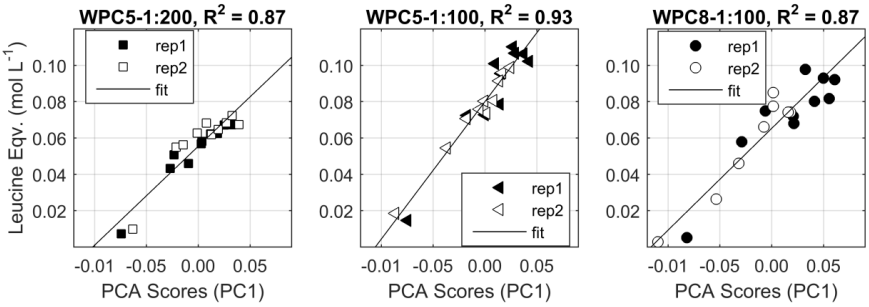
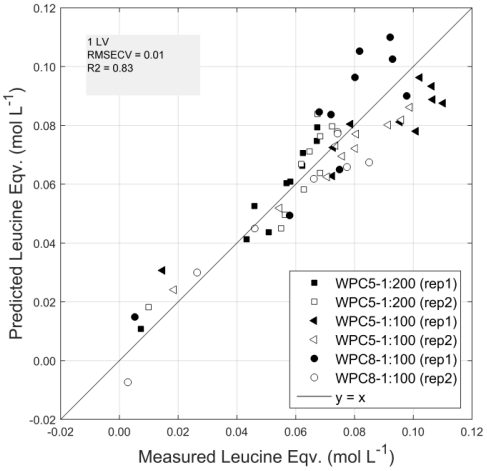


Figure 6 – grey-scale for print edition



PAPER III

Evaluation of multivariate calibration models transferred between spectroscopic instruments: Applied to near infrared measurements of flour samples

C. E. Eskildsen, P.W. Hansen, T. Skov, F. Marini, L. Nørgaard

Journal of Near Infrared Spectroscopy, 24 (2):151-156, 2016.

Received: 6 December 2015 ■ *Revised: 21 April 2016* ■ *Accepted: 27 April 2016* ■ *Publication: 9 May 2016*

JNIRS JOURNAL OF NEAR INFRARED SPECTROSCOPY

Virtual Issue: Papers Presented at NIR-2015, October 2015, Foz do Iguassu, Brazil

Evaluation of multivariate calibration models transferred between spectroscopic instruments: applied to near infrared measurements of flour samples

Carl E. Eskildsen,[a*] Per W. Hansen,[b] Thomas Skov,[a] Federico Marini[c] and Lars Nørgaard[a,b]

[a]Department of Food Science, University of Copenhagen, DK-1958 Frederiksberg, Denmark. E-mail: carle@food.ku.dk

[b]FOSS Analytical A/S, DK-3400 Hillerød, Denmark

[c]Department of Chemistry, University of Rome "La Sapienza", I-00185 Rome, Italy

In a setting where multiple spectroscopic instruments are used for the same measurements it may be convenient to develop the calibration model on a single instrument and then transfer this model to the other instruments. In the ideal scenario, all instruments provide the same predictions for the same samples using the transferred model. However, sometimes the success of a model transfer is evaluated by comparing the transferred model predictions with the reference values. This is not optimal, as uncertainties in the reference method will impact the evaluation. This paper proposes a new method for calibration model transfer evaluation. The new method is based on comparing predictions from different instruments, rather than comparing predictions and reference values. A total of 75 flour samples were available for the study. All samples were measured on ten near infrared (NIR) instruments from two instrumental platforms, five NIR instruments from each platform. Protein content was quantified for all 75 samples and used as the reference variable during modelling by partial least squares regression. By adding artificial noise to first the spectroscopic measurements and then the reference values, this paper highlights the problems of including reference values in the evaluation of a model transfer, as uncertainties in the reference method impact the evaluation. At the same time, this paper highlights the power of the proposed model transfer evaluation, which is based on comparing predictions obtained from the different instruments. In this way, the impact of uncertainties originating from the reference method is minimised.

Keywords: calibration, evaluation, model transfer, near infrared spectroscopy

Introduction

Multivariate calibration models are useful for extracting quantitative information from spectroscopic measurements. However, constructing a high quality multivariate calibration model using partial least squares (PLS) regression may require up to hundreds or thousands of samples with reference values and collected over a long time period. Consequently, constructing high quality calibration models is an expensive and time-consuming task.[1,2]

The PLS method estimates one or several dependent variables, **Y**, by means of a set of predictor variables (spectroscopic measurements), **X**. The regression vector provides the direct link from the (pre-processed) spectroscopic measurements to the estimated values of **Y**.[3,4] In an industrial setting where multiple spectroscopic instruments are used for the same measurements, it could be convenient to develop the calibration model on a single instrument only but apply this

ISSN: 0967-0335
doi: 10.1255/jnirs.1210

calibration model to the spectroscopic measurements of all the instruments. In this case, it is clearly important that all the instruments provide similar output. Otherwise, the results (predictions) from the calibration model will be incorrect.[5] Likewise, it is important that the instruments provide comparable results over time. A large number of methods have already been developed for transferring calibration models from one instrument to another as well as keeping calibration models valid over time.[1,2] The more similar the responses between instruments are, the more likely it is that the calibration transfer is successful.

Traditionally, calibration transfers are done in a master/slave setting. The calibration model is developed on one instrument (the master) and tested on other instruments (the slaves). The root mean squared errors of prediction (*RMSEP*) obtained for the slave instruments are then sometimes used to evaluate the model transfer.[6–11] This *RMSEP* value is used to verify how successful the model transfer was. The *RMSEP* value includes uncertainties introduced by inadequacies in the modelling steps, which can be related to differences in the spectroscopic measurements of the master and slave instruments. Nevertheless, the *RMSEP* value also includes uncertainties originating from the reference method. This is a fact which is often forgotten. Hence, large uncertainties in the reference measurements would make the *RMSEP* value larger and thereby make the model transfer appear poor even in the ideal/perfect case, where the predictions of the master and the salve instrument are identical.

Figure 1 illustrates how uncertainties introduced by modelling and reference method affect the prediction error. Uncertainties in the reference method will contribute to imprecision along the horizontal direction, whereas those from the model (in the predictions) will cause the data points to be uncertain in the vertical direction. Hence, if the reference method is imprecise, the modelling will appear poor.[12,13]

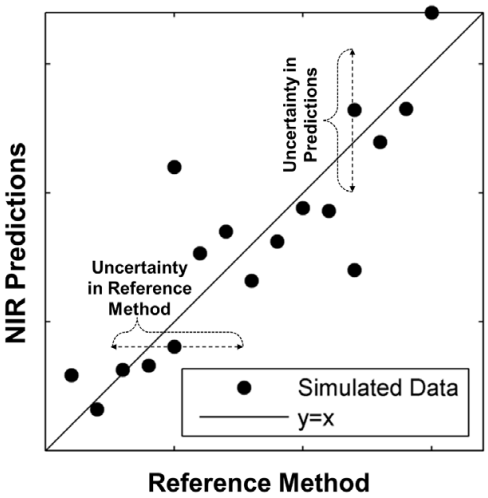


Figure 1. Influence of uncertainties originating from the reference method and near infrared predictions, respectively, on calibration model evaluation.

In this paper, we show that the success of a model transfer should not be determined based on a relationship with the reference values. The aim of a model transfer is to obtain the same results (i.e. predictions) on different instruments with the same samples by applying only one model, as was also done by, for example, Yoon *et al.*[14] In this paper we propose a new method for model transfer evaluation. This model transfer evaluation will also be based on similarities in predictions obtained from different instruments. In this way, the impact of uncertainties related to the reference method will be minimised. The assumption is, of course, that the developed model performs at an acceptable level when benchmarked against real reference values.

Materials and methods

Near infrared measurements

A total of 75 flour samples were measured on ten different near infrared (NIR) instruments. Five NIR instruments were DS2500 (FOSS Analytical A/S, Hillerød, Denmark) models and five were InfraXact (FOSS Analytical A/S, Hillerød, Denmark) models. The spectra from the five DS2500 instruments were obtained in the wavelength range from 400 nm to 2499.5 nm, with a total of 4200 data points. The spectra from the five InfraXact instruments were obtained in the wavelength range from 570 nm to 1848 nm, with a total of 640 data points. All spectra were obtained in reflectance mode. To obey Beer's law, all spectra were transformed from reflectance into pseudo-absorbance before modelling. The absorption spectra from the five InfraXact and the five DS2500 are shown in Figure 2A and Figure 2B, respectively.

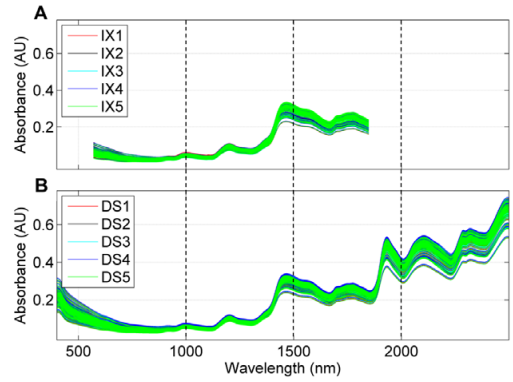


Figure 2. Raw near infrared spectra of flour samples. (A) InfraXact. (B) DS2500.

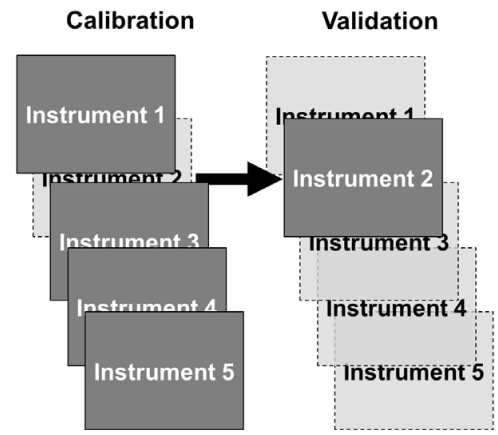


Figure 3. Leave-One-Instrument-Out cross-validation.

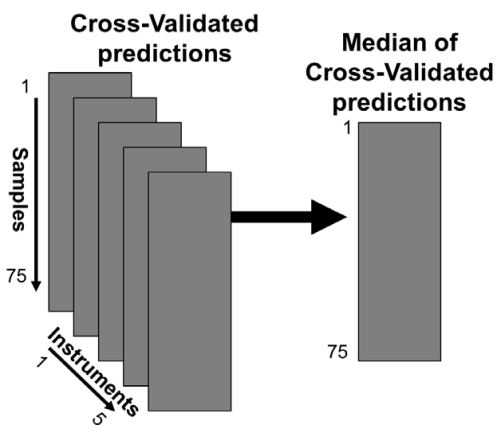


Figure 4. Median of cross-validated predictions.

Protein content was quantified by Kjeldahl digestion for all 75 samples and used as the reference variable during PLS modelling.[4]

Data analysis

Data were analysed using MATLAB version R2014a (8.3.0.532, The MathWorks Inc., Natick, MA, USA) and PLS_Toolbox version 7.5 (Eigenvector Research Inc., Manson, WA, USA). Prior to PLS modelling, spectra were pre-processed by standard normal variate and mean centring.[15]

Model transfer evaluation

To evaluate model transfer performance two PLS models were calculated. One PLS model was based on the five InfraXact instruments and one on the five DS2500 instruments. Each PLS model was validated by *Leave-One-Instrument-Out* cross-validation (Figure 3). Predictions obtained from cross-validation were collected and the medians across samples were found (Figure 4). The root mean squared error (*RMSE*) between these medians and the cross-validated predictions (for each instrument) was calculated. This *RMSE* relates to the similarities between the predicted values from the instruments and can thereby be used for evaluating model transfer. The median is used as "the true prediction value" as it is (almost) insensitive to outlying predictions, which would have affected, for example, the mean of predictions. It is important to note that the *RMSE* is not a measure of how well the "true" protein content can be modelled by the spectroscopic measurements. This *RMSE* is only a measure of how similar protein content is modelled from the measurements obtained from different instruments (i.e. the model transfer performance).

Simulation of error

In order to investigate how the methods responded to uncertainties in the spectral and the reference measurements, respectively, simulated deterioration of the data belonging to the fifth InfraXact instrument (IX5) was performed. First, deterioration of data was done by adding random noise to the spectroscopic measurements of instrument IX5. Figure 5 shows the deterioration of the spectra from IX5. Second, the reference data belonging to IX5 were deteriorated by random noise. Figure 6 shows the deterioration of the reference values belonging to IX5.

Results and discussion

This new method for evaluating model transfer should only be based on inadequacies in the modelling. Therefore, it is based on comparing predictions (i.e. model output) of the

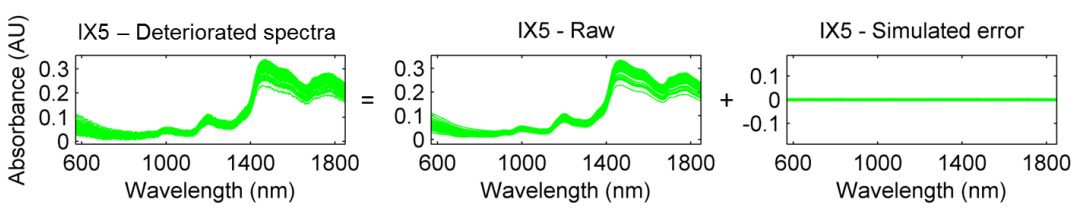


Figure 5. Simulated deterioration of spectra from instrument IX5 (InfraXact).

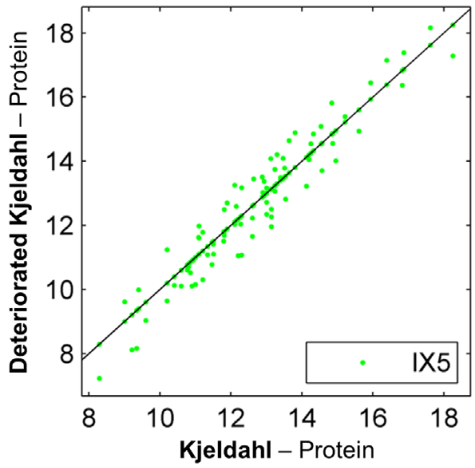


Figure 6. Simulated deterioration of reference values belonging to instrument IX5 (InfraXact).

same sample set. In order to highlight the power of this new method, it will be compared below with model transfer evaluation by means of comparing reference values with predictions.

First, the spectra of IX5 have been deteriorated by random noise. This deterioration will change the relationship between the spectroscopic measurements of IX5 and the reference values. Hence, a model calibrated on the remaining InfraXact instruments will not be valid for IX5. Applying this model to IX5 will return erroneous predictions. This will show up both when evaluating the model transfer by means of comparing transferred model prediction with reference values and when evaluating by the new method. In the first case, errors will show up because the predictions of IX5 are wrong and therefore not comparable with the reference values. In the new method, errors will show up because the predictions of IX5 will be different from the predictions of the other instrument. Figure 7 shows the results when the spectra of IX5 are deteriorated. Instrument IX5 has a high *RMSE* when evaluated both by comparing transferred model predictions with reference values (Figure 7A) and by the new approach (Figure 7B).

However, comparing the two approaches for model transfer evaluation becomes interesting when adding noise to the reference values belonging to IX5. The model calibrated on the remaining InfraXact instruments will still be valid for IX5. The link between the spectroscopic measurements of the calibration set and the estimation of the reference values (i.e. the predictions) will also be valid for the (undeteriorated) measurements from IX5. Nevertheless, evaluating the model transfer by means of comparing transferred model prediction with erroneous reference values, the predictions of IX5 will appear to have a large *RMSE*. Using the new method for model transfer evaluation, the predictions obtained from IX5 will be compared with the predictions obtained from the remaining instruments. Figure 8 shows the results where the reference values belonging to IX5 have been deteriorated. Figure 8A shows the errors obtained when evaluating by means of comparing transferred model prediction with reference values. Here it is found that IX5 has a high *RMSE*. Figure 8B shows the errors obtained using the new method. Here the predictions are compared and it is found that the predictions obtained from the measurements of IX5 are very similar to the predictions obtained from the measurements of the remaining

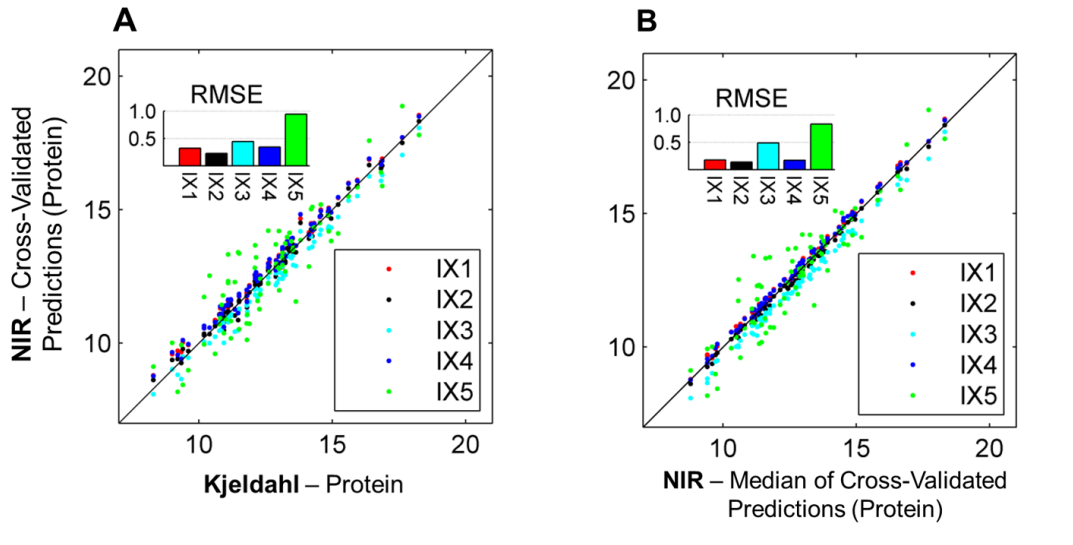


Figure 7. Results obtained with simulated deterioration of spectra from instrument IX5. (A) Model predictions vs reference values. (B) Model predictions vs median of model predictions (new approach).

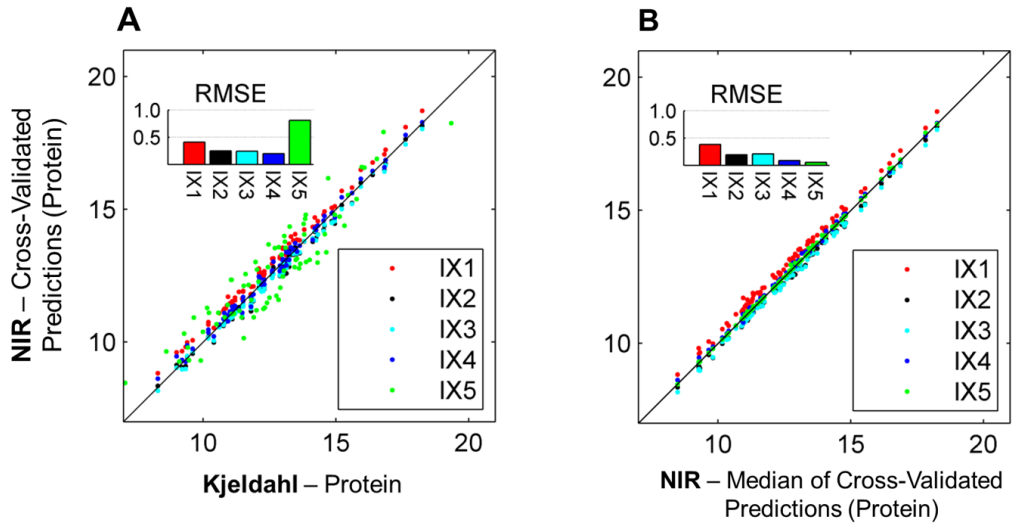


Figure 8. Results obtained with simulated deterioration of reference values belonging to instrument IX5. (A) Model predictions vs reference values. (B) Model predictions vs median of model predictions (new approach).

InfraXact instruments. Hence, Figure 8A indicates that the model transfer is not successful, whereas Figure 8B indicates that the model transfer indeed is successful.

In model transfer, a model is developed on one instrument and then applied to other instruments. The spectroscopic measurements are the model input to the predictions, which are the model output. In an industrial setting where one model is applied to the measurements of multiple instruments it is important that the predictions are identical for identical samples. Therefore, the right way of evaluating model transfer is by comparing predictions. Hence, Figure 8B (the new method) gives a more direct and accurate picture of the model transfer than Figure 8A (comparing transferred model predictions with reference values).

When using multiple instruments for the same measurements it can be convenient to do a model transfer. We believe

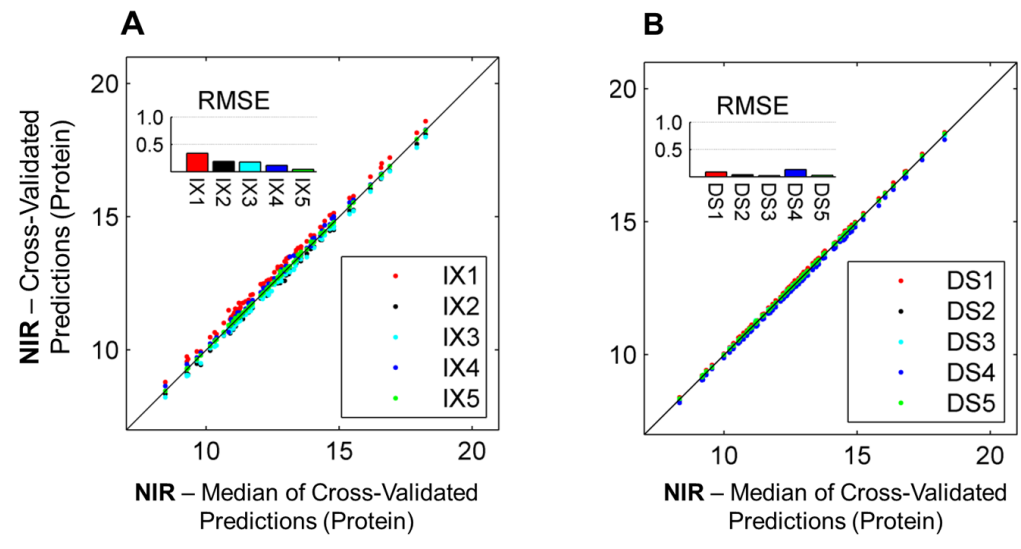


Figure 9. Evaluated by new approach. Comparison of InfraXact instruments (A) and DS2500 instruments (B).

that the task of model transfer is easier to overcome successfully if the original spectra from the multiple instruments are already providing similar results. Therefore, we also believe that this new method can be used as a quality criterion for instruments. Figure 9 shows the comparison of InfraXact and DS2500 using the new method. The results for the five InfraXact instruments are shown in Figure 9A. The results for the five DS2500 instruments are shown in Figure 9B. It was found that the DS2500 instruments provide very similar results compared with the InfraXact instruments, which show a minor bias. Such evaluation would not have been possible if the evaluation was performed by comparing the transferred model predictions with the reference values, as the bias variation in InfraXact would be masked by the errors in the reference values.

Conclusions

This paper shows how model transfer evaluation may be influenced by uncertainties in the reference method. In this paper a new method for evaluating model transfer is proposed. The new method is based on comparing predictions from multiple instruments. Therefore, the new method is insensitive to uncertainties in the reference values. For that reason, this new method gives a direct measure for model transfer performance.

References

1. F. van den Berg and Å. Rinnan, "Calibration transfer methods", in *Infrared Spectroscopy for Food Quality Analysis and Control*, Ed by D. Sun. Academic Press, San Diego, Ch. 5, pp. 105–118 (2009). doi: http://dx.doi.org/10.1016/B978-0-12-374136-3.00005-5
2. O.E. de Noord, "Multivariate calibration standardization", *Chemometr. Intell. Lab. Syst.* **25,** 85–97 (1994). doi: http://dx.doi.org/10.1016/0169-7439(94)85037-2
3. M. Andersson, "A comparison of nine PLS1 algorithms", *J. Chemometr.* **23,** 518–529 (2009). doi: http://dx.doi.org/10.1002/cem.1248
4. S. Wold, M. Sjostrom and L. Eriksson, "PLS-regression: A basic tool of chemometrics", *Chemometr. Intell. Lab. Syst.* **58,** 109–130 (2001). doi: http://dx.doi.org/10.1016/S0169-7439(01)00155-1
5. E. Sanchez and B.R. Kowalski, "Tensorial calibration: I. first-order calibration" *J. Chemometr.* **2,** 247–263 (1988). doi: http://dx.doi.org/10.1002/cem.1180020404
6. M.C. Alamar, E. Bobelyn, J. Lammertyn, B.M. Nicolaï and E. Moltó, "Calibration transfer between NIR diode array and FT-NIR spectrophotometers for measuring the soluble solids contents of apple", *Postharvest Biol. Technol.* **45,** 38–45 (2007). doi: http://dx.doi.org/10.1016/j.postharvbio.2007.01.008
7. F.D. Barboza and R.J. Poppi, "Determination of alcohol content in beverages using short-wave near-infrared spectroscopy and temperature correction by transfer calibration procedures" *Anal. Bioanal. Chem.* **377,** 695–701 (2003). doi: http://dx.doi.org/10.1007/s00216-003-2128-2
8. E. Bergman, H. Brage, M. Josefson, O. Svensson and A. Sparén, "Transfer of NIR calibrations for pharmaceutical formulations between different instruments", *J. Pharm. Biomed. Anal.* **41,** 89–98 (2006). doi: http://dx.doi.org/10.1016/j.jpba.2005.10.042
9. W. Fan, Y. Liang, D. Yuan and J. Wang, "Calibration model transfer for near-infrared spectra based on canonical correlation analysis", *Anal. Chim. Acta* **623,** 22–29 (2008). doi: http://dx.doi.org/10.1016/j.aca.2008.05.072
10. H. Swierenga, W. Haanstra, A. De Weijer and L. Buydens, "Comparison of two different approaches toward model transferability in NIR spectroscopy", *Appl. Spectrosc.* **52,** 7–16 (1998). doi: http://dx.doi.org/10.1366/0003702981942528
11. H. Swierenga, P. De Groot, A. De Weijer, M. Derksen and L. Buydens, "Improvement of PLS model transferability by robust wavelength selection", *Chemometr. Intell. Lab. Syst.* **41,** 237–248 (1998). doi: http://dx.doi.org/10.1016/S0169-7439(98)00055-0
12. R. DiFoggio, "Examination of some misconceptions about near-infrared analysis", *Appl. Spectrosc.* **49,** 67–75 (1995). doi: http://dx.doi.org/10.1366/0003702953963247
13. K. Faber and B.R. Kowalski, "Improved prediction error estimates for multivariate calibration by correcting for the measurement error in the reference values", *Appl. Spectrosc.* **51,** 660–665 (1997). doi: http://dx.doi.org/10.1366/0003702971941061
14. J. Yoon, B. Lee and C. Han, "Calibration transfer of near-infrared spectra based on compression wavelet coefficients", *Chemometr. Intell. Lab. Syst.* **64,** 1–14 (2002). doi: http://dx.doi.org/10.1016/S0169-7439(02)00042-4
15. A. Rinnan, F. van den Berg and S.B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra", *TrAC-Trends Anal. Chem.* **28,** 1201–1222 (2009). doi: http://dx.doi.org/10.1016/j.trac.2009.07.007

PAPER IV

Quantification of individual fatty acids in bovine milk by infrared spectroscopy and chemometrics: Understanding predictions of highly collinear reference variables

C. E. Eskildsen, M. A. Rasmussen, S. B. Engelsen, L. B. Larsen, N. A. Poulsen and T. Skov

Journal of Dairy Science, 97:7940-7951, 2014.

J. Dairy Sci. 97:7940–7951
http://dx.doi.org/10.3168/jds.2014-8337

Quantification of individual fatty acids in bovine milk by infrared spectroscopy and chemometrics: Understanding predictions of highly collinear reference variables

C. E. Eskildsen,*[1] M. A. Rasmussen,* S. B. Engelsen,* L. B. Larsen,† N. A. Poulsen,† and T. Skov*
*Department of Food Science, University of Copenhagen, DK-1958 Frederiksberg, Denmark
†Department of Food Science, Aarhus University, DK-8830 Tjele, Denmark

ABSTRACT

Predicting individual fatty acids (FA) in bovine milk from Fourier transform infrared (FT-IR) measurements is desirable. However, such predictions may rely on covariance structures among individual FA and total fat content. These covariance structures may change with factors such as breed and feed, among others. The aim of this study was to estimate how spectral variation associated with total fat content and breed contributes to predictions of individual FA. This study comprised 890 bovine milk samples from 2 breeds (455 Holstein and 435 Jersey). Holstein samples were collected from 20 Danish dairy herds from October to December 2009; Jersey samples were collected from 22 Danish dairy herds from February to April 2010. All samples were from conventional herds and taken while cows were housed. Moreover, in a spiking experiment, FA (C14:0, C16:0, and C18:1 *cis*-9) were added (spiked) to a background of commercial skim milk to determine whether signals specific to those individual FA could be obtained from the FT-IR measurements. This study demonstrated that variation associated with total fat content and breed was responsible for successful FT-IR–based predictions of FA in the raw milk samples. This was confirmed in the spiking experiment, which showed that signals specific to individual FA could not be identified in FT-IR measurements when several FA were present in the same mixture. Hence, predicted concentrations of individual FA in milk rely on covariance structures with total fat content rather than absorption bands directly associated with individual FA. If covariance structures between FA and total fat used to calibrate partial least squares (PLS) models are not conserved in future samples, these samples will show incorrect and biased FA predictions. This was demonstrated by using samples of one breed to calibrate and samples of the other breed to validate PLS models for individual FA. The 2 breeds had different covariance structures between individual FA and total fat content. The results showed that the validation samples yielded biased predictions. This may limit the usefulness of FT-IR–based predictions of individual FA in milk recording as indirect covariance structures in the calibration set must be valid for future samples. Otherwise, future samples will show incorrect predictions.
Key words: bovine, fatty acid, infrared spectroscopy, milk, quantification

INTRODUCTION

Traditionally, detailed milk composition is measured through time-consuming analyses such as gas chromatography (GC). A high-throughput method is needed to make it feasible for the dairy industry to measure detailed milk composition as a routine quality parameter. Fourier transform infrared (**FT-IR**) spectroscopy is currently used by commercial milk recording agencies and dairies, which makes the method attractive for providing high-throughput information on detailed milk composition, including FA profile. Such high-throughput information would be useful in relation to documentation, process control, and breeding.

Bovine milk contains, on average, 4% fat with more than 400 different FA (Jensen, 2002). Most fat in milk is found as triglycerides. In the current study, the term "FA" includes FA bound as triglycerides. The major FA found in milk are palmitic acid (C16:0), stearic acid (C18:0), oleic acid (C18:1 *cis*-9), and myristic acid (C14:0; Jensen, 2002). The FA profile in milk depends on several factors, including breed, feeding, stage of lactation, and yield (Jensen, 2002). The content and composition of FA in milk are important as quality characteristics of dairy products (Couvreur et al., 2006) as well as in human health (German et al., 2009); therefore, a high-throughput method measuring individual FA is of interest.

Several studies have investigated the potential of predicting the FA profile of milk from FT-IR measurements (Soyeurt et al., 2011; De Marchi et al., 2014;

Received May 7, 2014.
Accepted August 14, 2014.
[1]Corresponding author: carle@food.ku.dk

Ferrand-Calmels et al., 2014). In general, these studies conclude that milk FA present in high concentrations are predicted with good accuracy; in particular, C14:0 and C16:0 have been highlighted as FA that could be quantified routinely from FT-IR measurements.

Nevertheless, in traditional spectroscopy, FA are divided into functional groups such as methane, methylene, methyl, ester, ether, olefinic, aliphatic, and carboxylic groups with their own characteristic group frequencies. Whereas the average proportions of these functional groups may be measured by FT-IR, it is not likely that individual FA would show distinct absorption patterns in FT-IR measurements of a complex mixture. Chapman (1965) investigated infrared absorption of pure lipids and found that differences in chain length of solid FA yielded minimal differences in the fingerprint region of the infrared spectra. However, these minimal differences are expected to disappear when FA are present in a complex liquid mixture such as milk.

Therefore, reported predictions of individual milk FA from FT-IR measurements are likely to rely on indirect correlations, which are confined to covariance structures in the data set rather than absorption bands directly associated with individual FA. Even though this is not a problem in itself, it may be problematic in terms of prediction accuracy and calibration robustness, which are important but neglected parameters when evaluating the usefulness of partial least squares (**PLS**) regression models for predicting individual FA from FT-IR measurements of milk. If indirect correlations are used to calibrate a PLS model, the model will not be valid for future samples unless the indirect correlations are conserved for these samples. If the indirect correlations are not conserved, the model will provide incorrect predictions.

Individual FA may be highly correlated with total fat content (**TF**), which in turn is easily quantified from FT-IR measurements (Luinge et al., 1993). Hence, FA may be predicted by covariation with TF. Furthermore, if factors known to alter the FA profile (such as different breeds) provide significant absorption patterns in the FT-IR measurements, the possibility of FA being predicted by interactions of TF and one or more of these factors must be taken into account. Figure 1 illustrates how a given FA ($\hat{\mathbf{y}}_{FA}$) may be predicted from FT-IR measurements ($\mathbf{X}$). Here, the prediction is split into that part predicted by the indirect correlations between the FA and the interaction of TF and breed (TF × Breed) and the part predicted by information independent of the TF × Breed interaction (Figure 1). The more shared variation between $\mathbf{X}$ and the TF × Breed interaction used for predicting a specific FA ($\hat{\mathbf{y}}_{FA}$), the more the prediction will depend on the indirect correlations between FA and the TF × Breed interaction. In Figure 1, α defines how dependent the FA prediction is on the TF × Breed interaction. The FA prediction is more dependent on the TF × Breed interaction as α increases toward 1.

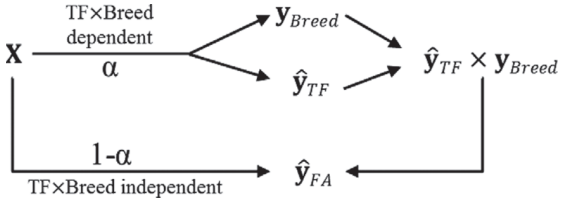


Figure 1. Prediction of a given fatty acid ($\hat{\mathbf{y}}_{FA}$) from Fourier transform infrared measurements ($\mathbf{X}$); $\hat{\mathbf{y}}_{FA}$ is partly predicted by variation described by the interaction of total fat content (TF) and breed (TF × Breed) and partly by variation independent of this interaction; α is a number between 0 and 1 and expresses how much each part contributes to $\hat{\mathbf{y}}_{FA}$; TF is given by $\hat{\mathbf{y}}_{TF}$ and breed is given by $\mathbf{y}_{Breed}$.

This study investigated how PLS models for predicting individual FA depend on spectral variation associated with the TF × Breed interaction in raw milk samples from 2 Danish dairy breeds. Furthermore, FA (C14:0, C16:0, and C18:1 *cis*-9) were added to an identical background of skim milk in a spiking experiment to investigate the contribution of individual FA to the FT-IR absorption pattern of milk.

MATERIALS AND METHODS

Raw Milk Samples

A total number of 890 a.m. milk samples from individual cows (435 Jersey and 455 Holsteins) were included in this study. Samples originated from the Danish-Swedish Milk Genomics Initiative (www.milkgenomics.dk). The sampling strategy aimed to minimize environmental variation but maximize the genetic variation of the sample population. The Holstein samples were collected from 20 Danish dairy herds from October to December 2009. The Jersey samples were collected from 22 Danish dairy herds from February to April 2010. All samples were from conventional herds and taken while cows were housed (Poulsen et al., 2012).

Quantification of FA was done using GC as described by Poulsen et al. (2012). Full FT-IR spectra were recorded on all samples using MilkoScan FT2 (Foss Analytical A/S, Hillerød, Denmark). Spectra were obtained from fresh whole milk and TF was determined using MilkoScan FT2; samples were measured in triplicate. For each FT-IR measurement, a FT-IR water spectrum was subtracted and the difference spectrum was

obtained. For further analysis, the average difference spectrum of each sample (across the 3 replicates) was calculated and used.

Spiked Milk Samples

Pure C14:0, C16:0, and C18:1 *cis*-9 monoacid triglycerides (Sigma-Aldrich, Brøndby, Denmark) were added to a background of commercial skim milk containing 0.1% fat (Arla Foods amba, Viby J, Denmark). The monoacid triglycerides were added to 17 samples (including 2 replicates) in a 3-component mixture design (Montgomery, 2009), as shown in Figure 2. Each monoacid triglyceride was added to the skim milk in concentrations from 0 to 3 g/100 mL of skim milk, following the procedure of Kaylegian et al. (2009). Cold skim milk and monoacid triglycerides were heated to 70°C to ensure that C16:0 triglyceride was melted before mixing. Immediately after mixing, the skim milk and monoacid triglycerides were stirred for 20 to 30 s using a high speed stirrer (Ultra-Turrax T18, IKA Works Inc., Wilmington, NC) operating at 18,000 rpm. After stirring, the mixture was homogenized (EmulsiFlex-C5, Avestin Inc., Ottawa, ON, Canada) at approximately 20,000 kPa. Homogenization was repeated twice, as proposed by Kaylegian et al. (2009). Samples were cooled to 5°C in glass bottles. The stability of the samples was visually assessed, and no phase separation was found in any of the samples.

Additional background samples were prepared to detect whether sample preparation (heating, stirring, homogenizing, or cooling) would introduce variation in the samples. No monoacid triglyceride was added to the background samples. Stirring, homogenization, and cooling did not affect the sample background. In contrast, heating led to changes in the FT-IR spectra (data not shown). Thus, the time each sample spent in the water bath could introduce variation in the spectra.

The FT-IR full spectra were recorded for all samples (including the background samples) using the MilkoScan FT2. Samples were measured twice. For each FT-IR measurement, a FT-IR water spectrum was subtracted to obtain the difference spectrum. For each sample, the average difference spectrum was calculated and used for further analysis.

Data Analysis

Data were analyzed using Matlab version R2013b (8.2.0.701, MathWorks Inc., Natick, MA) and PLS_Toolbox version 7.3.1 (Eigenvector Research Inc., Manson, WA).

The FT-IR spectra were obtained in transmittance mode in the range from 5,008 to 925 cm^{-1}, with a total of 1,060 data points for each sample. To obey Beer's law, the spectra were transformed from transmittance into absorbance before modeling. The difference spectra of the raw milk samples are depicted in Figure 3a. The region from 2,968 to 5,008 cm^{-1} was considered as noise and removed from the data set. The region from 1,773 to 2,802 cm^{-1} contained no valuable information and was also removed, together with the saturated water signal (O–H bend) from 1,692 to 1,604 cm^{-1}. Having saturation of the single beam spectra in mind (i.e., no signal going through to the detector), we decided to remove the narrow wavenumber region from 2,915 to 2,930 cm^{-1}. In Figure 3b, the spectral parts of raw milk samples included for modeling are shown. The same wavenumber regions were included for modeling of both the raw milk samples and spiked milk samples.

No obvious slope (multiplicative) effects were found in the FT-IR measurements of raw (Figure 3) or spiked milk samples (data not shown). However, minor offset (additive) effects were found in the spectra. Offset differences between spectra may affect the PLS models and thus these were removed before PLS modeling by preprocessing the spectra of both raw and spiked milk samples using Savitzky-Golay first derivative (window size of 9 points and second-order polynomial) followed by mean centering. In order to obtain easier interpretable loadings, the spectra were preprocessed by standard normal variate method followed by mean centering before principal components analysis (**PCA**). Differ-

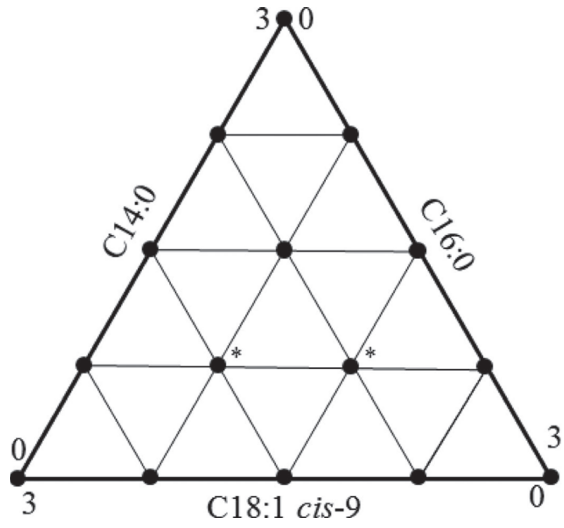


Figure 2. Three-component mixture design: each monoacid triglyceride (C14:0, C16:0, and C18:1 *cis*-9) was added (spiked) to skim milk in concentrations from 0 to 3 g/100 mL of skim milk; * indicates samples made in duplicate.

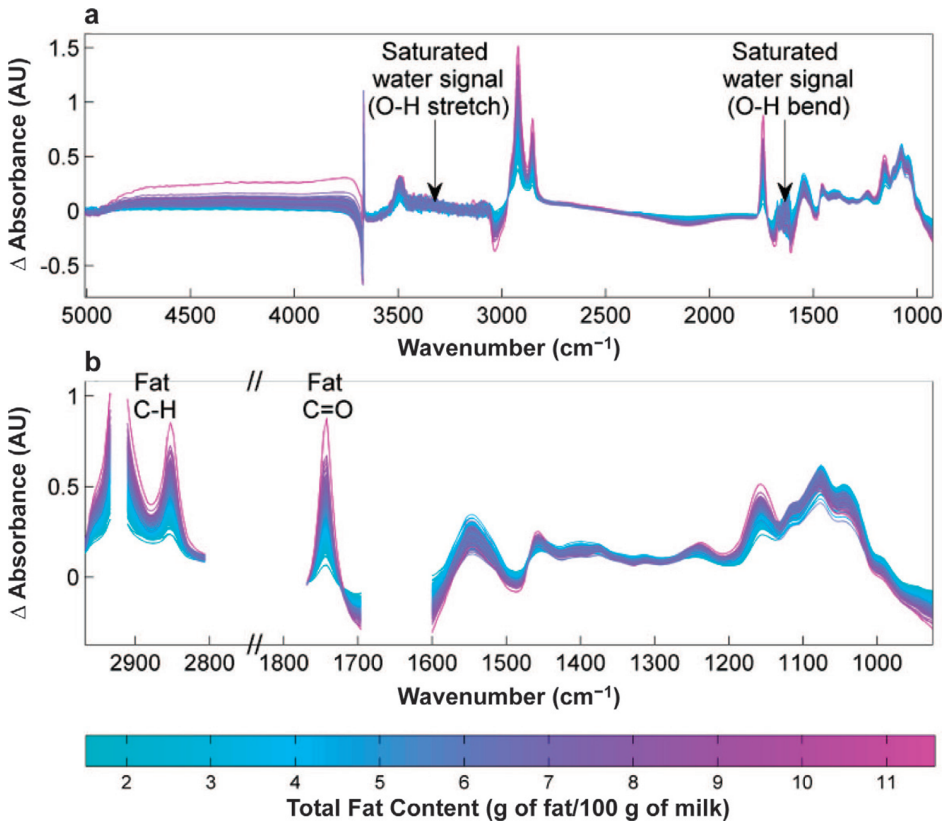


Figure 3. (a) Fourier transform infrared (FT-IR) difference spectra (water spectrum subtracted from each milk measurement); (b) parts of the FT-IR difference spectra included for PLS modeling. Spectra in both plots were from raw milk samples and are shaded (colored) by total fat content. AU = arbitrary units. Color version available in the online PDF.

ent preprocessing methods (multiplicative scatter correction, standard normal variate, and Savitzky-Golay with varying window size, polynomial order and derivative) had only minor effects on model performance in both data sets (data not shown). Variable selection by interval PLS (Nørgaard et al., 2000) did not improve models and further variable selection (compared with Figure 3b) was not performed in either experiment. All PLS models were built with a mean centered univariate response variable.

The PLS models calibrated using raw milk samples were cross-validated by the venetian blind method with 10 data splits (i.e., each validation set is determined by selecting every 10th sample in the data set, starting at sample 1 through 10). Model parameters (coefficient of determination, R^2, and root mean squared error of cross-validation, **RMSECV**) are reported and were used to choose the number of latent variables (**LV**). The PLS models calibrated using spiked milk samples were validated by leave-one-out cross-validation.

Predicting FA—Contribution from TF and Breed

A given FA is predicted by PLS regression as shown in Equation [1], where $\hat{\mathbf{y}}_{FA}$ is a vector containing the predicted values of the FA, $\mathbf{X}$ are the preprocessed FT-IR spectra, and $\mathbf{b}_{FA}$ is the regression vector. The general idea is to split the FA prediction $\left(\hat{\mathbf{y}}_{FA}\right)$ into a part explained by the TF × Breed interaction $\left(\hat{\mathbf{y}}_{FA_{TF\times Breed}}\right)$ and a part orthogonal (unrelated) to the TF × Breed interaction $\left(\hat{\mathbf{y}}_{FA_o}\right)$, as shown in Equation [2]:

$$\hat{\mathbf{y}}_{FA} = \mathbf{X} \cdot \mathbf{b}_{FA}, \quad [1]$$

$$\hat{\mathbf{y}}_{FA} = \hat{\mathbf{y}}_{FA_{TF\times Breed}} + \hat{\mathbf{y}}_{FA_o}. \quad [2]$$

In Equation [3], TF is estimated by PLS, where $\hat{\mathbf{y}}_{TF}$ is a vector containing the predicted TF values, $\mathbf{X}$ are the preprocessed FT-IR spectra, and $\mathbf{b}_{TF}$ is the regression vector. In this study, $\hat{\mathbf{y}}_{TF}$ was obtained from the MilkoScan FT2:

$$\hat{\mathbf{y}}_{TF} = \mathbf{X} \cdot \mathbf{b}_{TF}. \quad [3]$$

To obtain the variation derived from the TF × Breed interaction, a matrix $\mathbf{V}$ was constructed (Equation 4). The first 2 columns of $\mathbf{V}$ consist of a vector of ones (offset) and $\hat{\mathbf{y}}_{TF}$ (slope) for Holstein samples and zeroes for Jersey samples. The subsequent 2 columns of $\mathbf{V}$ consist of zeroes for Holstein samples and a vector of ones (offset) and $\hat{\mathbf{y}}_{TF}$ (slope) for Jersey samples. Hence, the matrix $\mathbf{V}$ spans the variation of the TF × Breed interaction:

$$\mathbf{V} = \begin{bmatrix} 1 & \hat{y}_{TF,1} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \hat{y}_{TF,455} & 0 & 0 \\ 0 & 0 & 1 & \hat{y}_{TF,456} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & \hat{y}_{TF,890} \end{bmatrix} \begin{matrix} \\ \\ \textit{Holstein samples} \\ \\ \hline \\ \textit{Jersey samples} \\ \\ \end{matrix}. \quad [4]$$

As shown by Strang (2006), the spectra, $\mathbf{X}$, may (column-wise) be projected onto a space spanned by $\mathbf{V}$. This projection is done in Equation [5]. Hence, every column of $\mathbf{X}_{TF\times Breed}$ is a linear combination of $\mathbf{V}$. Then, $\mathbf{X}_O$ (the part of $\mathbf{X}$ orthogonal to $\mathbf{V}$) is obtained by subtracting $\mathbf{X}_{TF\times Breed}$ from $\mathbf{X}$, as shown in Equation [6]. Every column of $\mathbf{X}_O$ will then be orthogonal to $\mathbf{V}$. Hence, $\mathbf{X}_O$ solely contains information not related to TF and breed. The projection and orthogonalization carried out in Equations [5] and [6] are commonly used in methods such as external parameter orthogonalization PLS (Roger et al., 2003), multiblock variance partitioning (Skov et al., 2008), and sequential and orthogonalized PLS (Næs et al., 2011):

$$\mathbf{X}_{TF\times Breed} = \mathbf{V}(\mathbf{V}^T \cdot \mathbf{V})^{-1}\mathbf{V}^T \cdot \mathbf{X}, \quad [5]$$

$$\mathbf{X}_O = \mathbf{X} - \mathbf{X}_{TF\times Breed}. \quad [6]$$

Note that $\mathbf{X} = \mathbf{X}_{TF\times Breed} + \mathbf{X}_O$. Hence, the FA prediction $\left(\hat{\mathbf{y}}_{FA}\right)$ may be split into a part explained by the TF × Breed interaction $\left(\hat{\mathbf{y}}_{FA_{TF\times Breed}} = \mathbf{X}_{TF\times Breed} \cdot \mathbf{b}_{FA}\right)$ and a part orthogonal to the TF × Breed interaction $\left(\hat{\mathbf{y}}_{FA_O} = \mathbf{X}_O \cdot \mathbf{b}_{FA}\right)$ without loss of information. This is shown in Equation [7] and, thereby, the requirement presented in Equation [2] is formalized:

$$\hat{\mathbf{y}}_{FA} = \mathbf{X} \cdot \mathbf{b}_{FA} = \mathbf{X}_{TF\times Breed} \cdot \mathbf{b}_{FA} + \mathbf{X}_O \cdot \mathbf{b}_{FA}. \quad [7]$$

RESULTS AND DISCUSSION

The FT-IR measurements obtained on raw milk samples are presented in Figure 3. The absorption band located just above 3,000 cm^{-1} (Figure 3a) remains from the olefinic, =C-H stretch and should correlate to the degree of unsaturated fat in the milk samples. However, in this study, it was not possible to establish such a relationship, most likely because water absorption caused saturation of the single beam spectrum in this wavenumber region.

A PCA was performed on the combined data set of FT-IR spectra of both raw and spiked milk samples. Wavenumber regions included in the PCA are found in Figure 3b and the results from PCA are presented in Figure 4. The first LV of the PCA explained 90.9% of the total variation. The score plot (Figure 4a) shows that Holstein and Jersey samples were separated, to a large extent, by the first LV. The loading plot (Figure 4b) shows that the first LV was related to the TF of the samples. The second LV of the PCA explained 7.2% of the total variation. The loading plot (Figure 4b) shows that the second LV was related to total protein content of the samples. The spiked milk samples are shown on the left in the score plot (negative score on the first LV; Figure 4a). This indicates that the spiked milk samples were low in TF compared with the raw milk samples. However, this was expected because FA were spiked to match concentrations of the individual FA in raw milk samples rather than the TF of raw milk samples. Figure 4a reveals that the spiked milk samples did not show FT-IR absorption remarkably different from that of the raw milk samples.

Raw Milk Samples

The FA reference values obtained from GC were converted into quantities per unit of milk (g of FA/100 g of milk). The FA quantities per unit of fat are not directly comparable with FT-IR absorption in milk and resulted, not surprisingly, in poor predictions (data not shown). In accordance with previous studies, the major FA fractions were C14:0, C16:0, C18:0, and C18:1 *cis*-9.

The results of PLS models for individual FA are shown in Table 1. Both breeds were combined in Table 1 to increase variation in the data set, even though the FA profiles from Jersey and Holstein cows show distinct characteristics (Poulsen et al., 2012). Our results in Table 1 are in agreement with previous studies, which revealed that the major FA, especially, are predicted well (Soyeurt et al., 2008; De Marchi et al., 2011; Maurice-Van Eijndhoven et al., 2013).

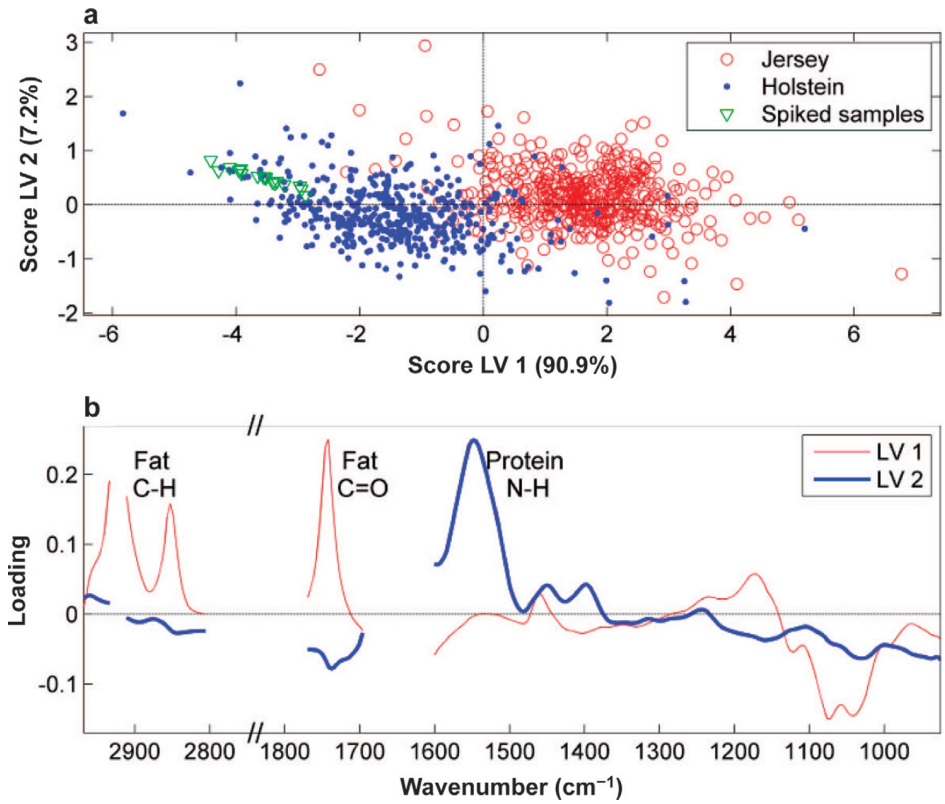


Figure 4. (a) Principal component analysis showing score plot of Fourier transform infrared (FT-IR) measurements, and (b) loading plot of FT-IR measurements. The FT-IR spectra were preprocessed by standard normal variate followed by mean centering. LV = latent variable. Color version available in the online PDF.

However, the validity and usefulness of PLS models estimating individual FA should not be based solely on evaluating model performance parameters such as R^2 and RMSECV. As mentioned, absorption patterns specific to individual FA are very unlikely to be found from FT-IR measurements in a complex mixture such as milk. Therefore, FA predictions must rely on indirect covariance structures. Several studies investigating the usefulness of FT-IR for predicting FA in milk have found that the most important spectral regions for predicting FA are the aliphatic C–H stretches (~2,900 cm^{-1}) and the carbonyl stretch (~1,745 cm^{-1}) (Rutten et al., 2009; Ferrand et al., 2011; De Marchi et al., 2011). However, these are also the regions associated with prediction of TF (Luinge et al., 1993; Kaylegian et al., 2009). If the same spectral regions are used for predicting both individual FA and TF, the predictions cannot be independent. Increased TF of new samples will then result in increased predictions of an individual FA, even though other FA are responsible for the increase in TF. Hence, if the FA-to-TF ratio found in the calibration set is not conserved in future samples, future samples will show incorrect and biased predictions.

To emphasize this point, the relationship between TF and C14:0 was examined for Jersey and Holstein samples and shown not to be the same (Figure 5a). A PLS model calibrated on the Jersey samples would be valid for the Holstein samples if the PLS model is based on direct relationships. However, if the model is based on the indirect correlation between C14:0 and TF in Jersey samples, the model would not be valid for Holstein samples. From Figure 5b, we found that concentrations of C14:0 of Holstein samples were predicted with a bias toward lower values, which suggests that the C14:0 predictions were based on an indirect correlation with TF. This finding illustrates the problem of models based on

Table 1. Results from partial least squares models based on raw milk samples[1]

Fatty acid	Range	Mean	SD	CV	LV	R^2CV	RMSECV	Relative error
C6:0	0.03–0.25	0.13	0.04	0.29	5	0.88	0.01	0.10
C8:0	0.02–0.16	0.07	0.02	0.32	5	0.89	0.01	0.11
C10:0	0.04–0.36	0.16	0.05	0.43	9	0.91	0.02	0.11
C12:0	0.05–0.45	0.18	0.06	0.35	9	0.91	0.02	0.11
C13:0	0.00–0.03	0.01	0.01	0.54	2	0.60	0.01	0.35
C14:0	0.16–0.90	0.51	0.12	0.24	9	0.90	0.04	0.08
C14:1	0.01–0.09	0.04	0.01	0.30	1	0.37	0.01	0.30
C15:0	0.01–0.13	0.05	0.02	0.35	2	0.74	0.01	0.18
C16:0	0.41–3.07	1.41	0.44	0.31	7	0.91	0.14	0.10
C16:1	0.01–0.34	0.07	0.03	0.37	8	0.63	0.02	0.23
C17:0	0.00–0.09	0.02	0.01	0.37	1	0.54	0.01	0.36
C18:0	0.14–1.55	0.53	0.19	0.35	7	0.82	0.08	0.16
C18:1 *trans*-11	0.01–0.21	0.07	0.03	0.34	1	0.30	0.02	0.34
C18:1 *cis*-9	0.32–2.94	0.85	0.21	0.24	8	0.82	0.09	0.11
C18:2 *cis*-9,*cis*-12	0.02–0.25	0.07	0.02	0.28	7	0.65	0.01	0.18
C18:3	0.01–0.04	0.02	0.01	0.26	1	0.39	0.01	0.26
C18:2 *cis*-9,*trans*-11 (CLA)	0.01–0.06	0.02	0.01	0.28	7	0.37	0.01	0.24

[1]LV = number of latent variables; R^2CV = cross-validated coefficient of determination; RMSECV = root mean squared error of cross-validation; relative error = RMSECV/mean. Range, mean, SD, CV, and RMSECV are in units of grams of FA/100 g of milk.

indirect correlations. In this study, all FA showed biased predictions when models were calibrated on one breed and validated on the other breed (data not shown). In a study by Rutten et al. (2009), samples were collected in both winter and summer, and it was found that calibrating PLS models with summer samples and predicting winter samples, or vice versa, resulted in prediction bias. In that study, different feedings were used during summer and winter. Different feedings are known to alter the FA profile (Jensen, 2002). Hence, the indirect correlation between FA and TF in summer samples are probably not conserved for winter samples. This is similar to the example illustrated in Figure 5: when indirect correlations in the calibration set are not conserved for the test set, incorrect and biased predictions are obtained.

In a study based on 49 samples (using approximately 10 LV per PLS model), Soyeurt et al. (2006) found that the correlation coefficients (r) obtained from PLS models (measured vs. predicted FA) were greater than those between the measured FA and TF. Based on these findings, Soyeurt et al. (2006) suggested that predicted concentrations of FA were based on infrared absorption bands specific to individual FA rather than to TF. As TF (estimated from MilkoScan) is already a linear combination of the FT-IR measurements, correlations obtained from PLS models (measured vs. predicted FA) will always be greater than or equal to the correlation between measured FA and TF. If correlations from PLS models are greater, this indicates that information additional to TF is used for predicting the FA. However, this additional information is not necessarily the major contributor to the FA predictions. Furthermore, the additional information cannot be assigned as absorption directly associated with FA without further investigations. In the current study, R^2 from PLS models (measured vs. predicted FA) were compared with R^2 between measured FA and TF (Figure 6a). Results similar to those obtained by Soyeurt et al. (2006) were found. However, when comparing R^2 from PLS models with R^2 between predicted FA and TF (Figure 6b), we found that the degree of correlation between predicted FA and TF exceeded model performance for most of the FA. In particular, the increased R^2 values between most FA and TF from Figure 6a to Figure 6b is an indication of the FA being primarily modeled (indirectly) by their relationship to TF.

Jersey and Holstein cows show distinct characteristics in their FA profiles (Poulsen et al., 2012). When information on breed differences is found in the FT-IR measurements, this information can likely be used to refine FA predictions derived from TF. To investigate how much spectral variation associated with the TF × Breed interaction contributes to predictions of individual FA, the projection procedure outlined in Equations [3] to [7] was applied. In Figure 7, the percentage of explained variation of each individual FA is presented. The percentage of explained variation is split into a part that is a linear combination of the TF × Breed interaction and a part that is independent of this interaction. The results in Figure 7 are from cross-validated models (as explained in the Appendix) and show that all FA (except C18:2 *cis*-9,*trans*-11) are mainly predicted as a linear combination of the TF × Breed interaction; C18:2 *cis*-9,*trans*-11 appears to be predicted more by information independent of the TF × Breed interaction, but the overall prediction of C18:2 *cis*-9,*trans*-11 is poor and therefore of no interest. The

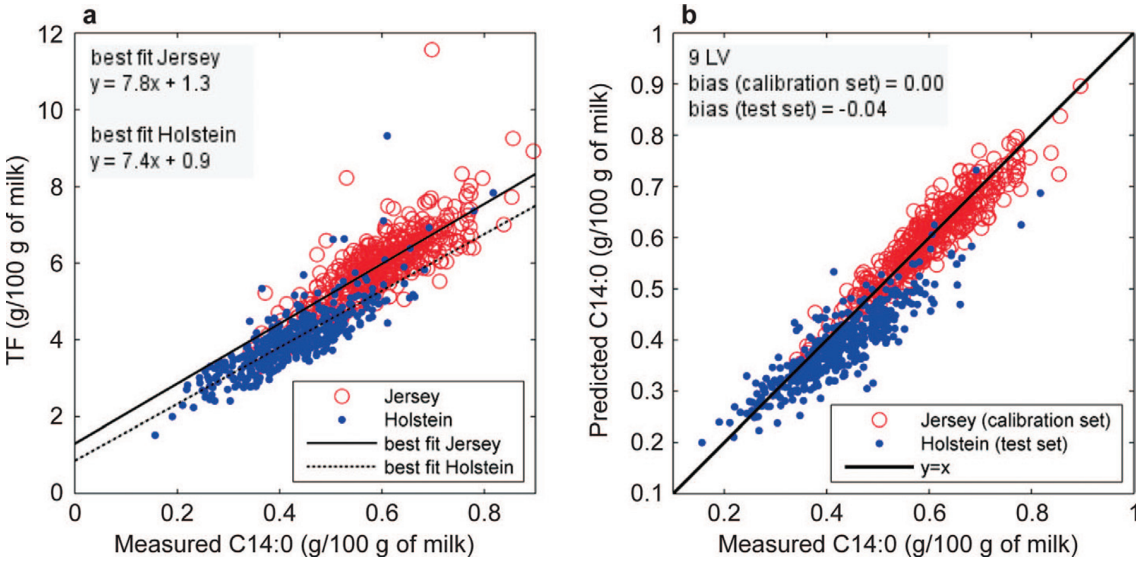


Figure 5. (a) Relationship between measured C14:0 and total fat content (TF) in raw milk samples from Jersey and Holstein cows; (b) measured versus predicted values of C14:0; C14:0 was predicted by partial least squares (PLS) regression applied to Fourier transform infrared measurements. The PLS model was calibrated on Jersey and tested with Holstein samples. LV = latent variable. Color version available in the online PDF.

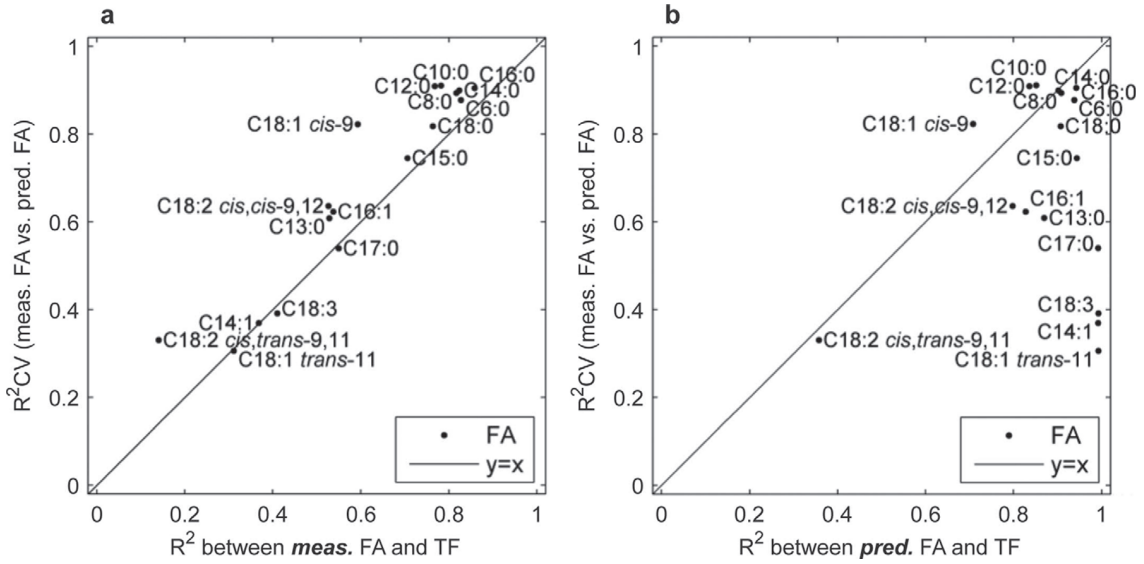


Figure 6. (a) Relationship between coefficient of determination from cross-validated PLS models (R^2CV) and coefficient of determination (R^2) between measured (meas.) FA concentrations and total fat content (TF); (b) relationship between R^2CV and R^2 between predicted (pred.) FA concentrations and TF.

ratio of the TF × Breed dependent part and the total percentage of explained variation for each individual FA will define α (from Figure 1). The ratio between the part dependent on the TF × Breed interaction and the part independent of the TF × Breed interaction is influenced by the nature of the sample set, as well as by the number of LV used in the particular model. Hence, results outlined in Figure 7 are not universal and the projection method should be viewed as a model diagnostic tool.

To further explore the relationship between individual FA and FT-IR absorption, R^2 values were calculated between the absorption intensities of each wavenumber of the raw FT-IR spectra and measured values of the individual FA (g of FA/100 g of milk). The R^2 values are presented in Figure 8 as a heat map. Figure 8 shows that individual FA primarily correlate with the regions associated with TF. None of the individual FA seem to have a unique relationship with the FT-IR spectra. Regions in the FT-IR measurements showing high correlation with individual FA correlate better with TF.

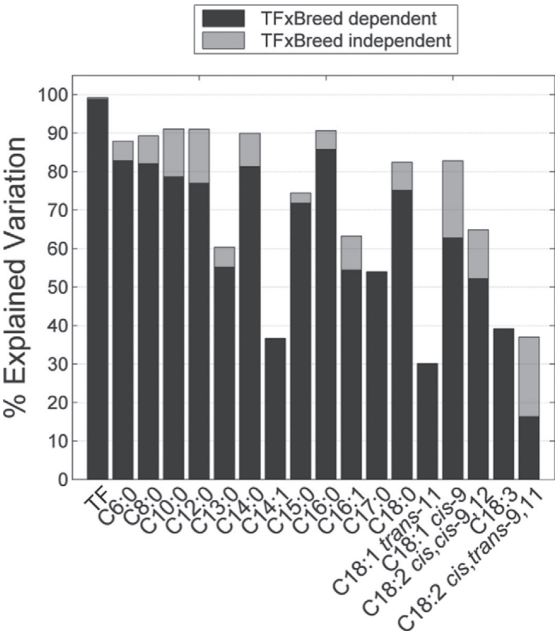


Figure 7. Prediction of FA by partial least squares regression applied to Fourier transform infrared measurements. The percent of explained variation in FA predictions was divided into that part related to the interaction of total fat content (TF) and breed (TF × Breed dependent; black) and a part independent of the interaction (TF × Breed independent; gray). The ratio between the TF × Breed dependent part and the total percent explained variation of each FA corresponds to α (from Figure 1).

This further documents that individual FA do not have unique absorption patterns in FT-IR measurements obtained from milk.

Spiked Milk Samples

The results of the spiking experiment are shown in Figure 9. The 3 plots show the RMSECV versus the number of LV for each of the 3 FA (C14:0, C16:0, and C18:1 *cis*-9). The RMSECV at 0 LV is the error obtained when all samples are predicted as the average value. Each monoglyceride was added to the skim milk in concentrations from 0 to 3 g/100 mL of skim milk. This corresponds approximately to the concentration range of C16:0 and C18:1 *cis*-9 in the raw milk samples (Table 1). Hence, for spiked and raw milk samples, the cross-validation error of C16:0 and C18:1 *cis*-9 is (to some extent) comparable. The FA C14:0 was found in smaller quantities in the raw milk samples compared with the spiked milk samples (Table 1); hence, the cross-validation error of C14:0 is expected to be smaller in raw milk samples than in spiked milk samples.

In theory, it should be possible to model the FT-IR measurements using 2 LV, as the experimental design (Figure 2) has closure and thus results in a rank 2 system. It may be argued to include 1 additional LV to model variation introduced by different heating times during sample preparation. However, even though C14:0, C16:0, and C18:1 *cis*-9 are modeled with 5 to 6 LV (imposing the risk of over-fitting), the RMSECV for the spiked milk samples (Figure 9) are notably greater than those reported in the literature (De Marchi et al., 2011; Soyeurt et al., 2011; Ferrand-Calmels et al., 2014) and observed in the current study for raw milk samples (Table 1). This indicates that good predictions of these FA in raw milk samples are most likely due to indirect correlations, which are not present in the spiked samples.

For C14:0, increased values of RMSECV were found when going from 2 to 3 LV; for C16:0, increased RMSECV were found going from 1 to 2 LV; and for C18:1 *cis*-9, increased RMSECV values were observed going from 1 to 3 LV (Figure 9). Different preprocessing did not remove the increase in RMSECV with increasing numbers of LV (data not shown). An increase in RMSECV with an increasing number of LV indicates that finding systematic variation in the spectra that correlates with individual FA is difficult. The behavior of the RMSECV may be due to the fact that the information in the FT-IR spectra specific to the individual FA is minor and accounts for a limited amount of the variation in the spectra. A few LV may then be needed to model (remove) the major systematic spectral variation before the information specific to the FA can be captured.

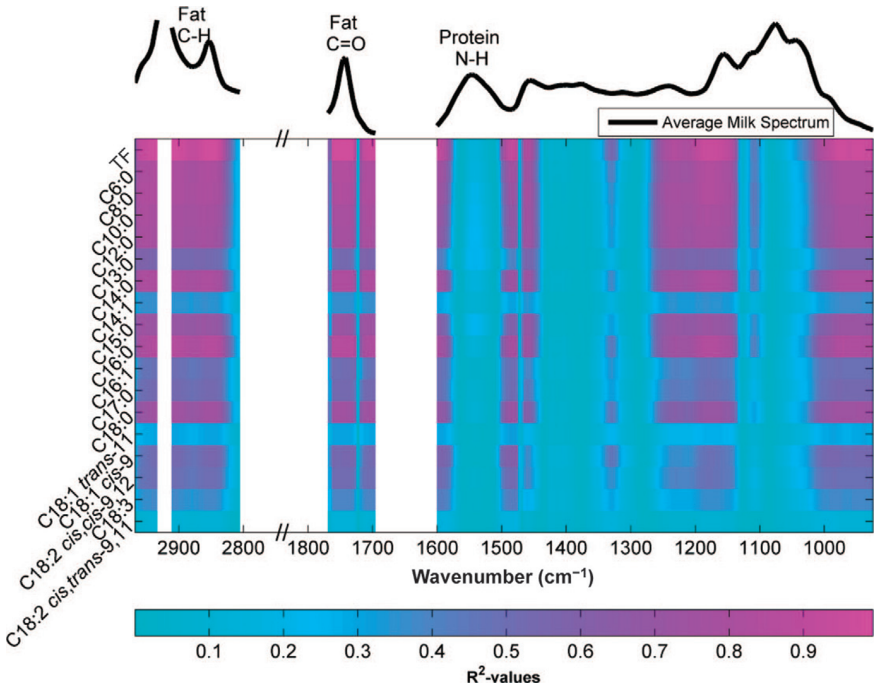


Figure 8. Coefficient of determination (R^2) between measured fatty acids (g/100 g of milk) and raw Fourier transform infrared measurements. TF = total fat content. Color version available in the online PDF.

However, as FA were spiked to a similar background, spectral variation that could hide information specific to the FA must be minimal and it seems more plausible that the increase in RMSECV is because no information specific to the individual FA is present in the FT-IR spectra. The lower RMSECV at 5 to 6 LV may be the consequence of over-fitting due to the combination of a limited number of samples (17) and the relatively

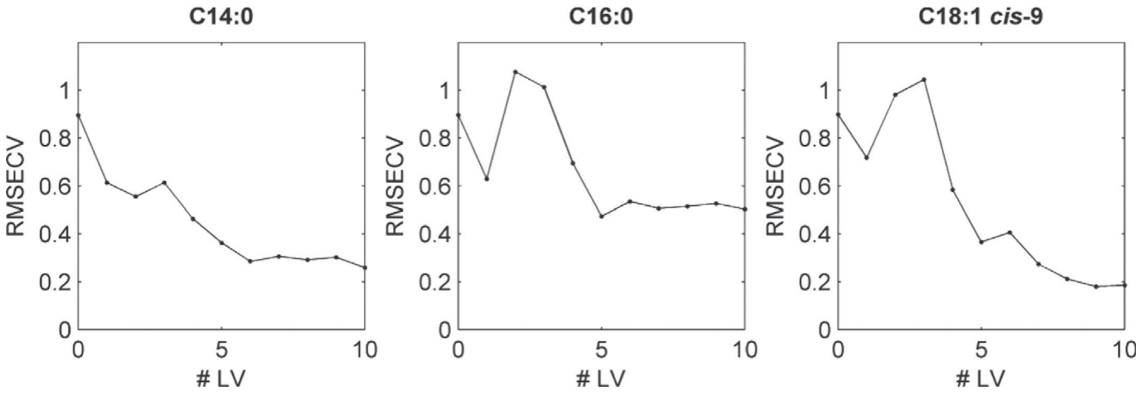


Figure 9. Prediction of monoacid triglycerides (C14:0, C16:0, and C18:1 *cis*-9) in spiked milk samples by partial least squares (PLS) regression applied to Fourier transform infrared measurements. Root mean squared error of cross-validation (RMSECV) from PLS models as a function of number of latent variables (LV) included in the model.

high number of LV. Furthermore, some indirect correlations between FA and the FT-IR measurements may still occur in the spiked milk samples. As the 3 FA do not have the same molecular weight, the FA are not spiked in equal amounts (i.e., moles of FA per 100 mL of milk). Hence, C14:0 would in general be present at a higher molarity than either C16:0 or C18:1 *cis*-9. Therefore, FT-IR absorption associated with carbonyl and the glycerol backbone would be affected more by C14:0 than by the other FA. Even though the errors of Figure 9 were notably greater than those obtained from raw milk samples, the errors of Figure 9 may still be overoptimistic, because indirect correlations may occur due to the simple experimental setup.

CONCLUSIONS

Concentrations of individual FA were predicted by applying PLS to FT-IR measurements of raw milk samples. Spectral variation associated with the interaction of TF and breed was important in predicting the FA concentrations of the samples available in this study. Using a spiking experiment, we found that FT-IR absorption signals due to specific FA (C14:0, C16:0, and C18:1 *cis*-9) cannot be detected when FA are mixed in a matrix of skim milk. Similarly, FA calibrations made on one breed and applied to another breed exhibited biased prediction results. Thus, predictions of individual FA by FT-IR measurements in milk are indirect and are based primarily on covariation between the FA and TF. The FA calibration models are not based on causal relationships but indirect correlations. Recommendations on implementing FT-IR–based FA models in milk recording schemes, for example, must account for the universal validity of these indirect correlations. Recommendations cannot be done solely on the basis of PLS model performance parameters such as R^2 and RMSECV. In contrast to previous studies, we suggest that indirect FA models may not be useful in milk recording systems or breeding programs because the FA models are providing information related to total fat rather than to individual FA. The indirect correlations responsible for successful FA predictions in a sample set may not be valid for samples of a different nature (different FA profiles), meaning that developed models may result in incorrect predictions of FA in future samples.

ACKNOWLEDGMENTS

The Danish Council for Strategic Research is acknowledged for generous financial support to the project entitled "Exploration of high-throughput FTIR-based solutions as an effective tool for correlating economically important milk quality characteristics with genetic markers" under the inSPIRe (Danish Industry–Science Partnership for Innovation and Research in Food Science) consortium (Copenhagen, Denmark). Furthermore, Arla Foods amba (Viby J, Denmark), Foss Analytical A/S (Hillerød, Denmark), Danish Cattle Federation (Aarhus, Denmark), as well as the Milk Levy Fund (Denmark) are acknowledged for financial support.

REFERENCES

Chapman, D. 1965. Infrared spectroscopy of lipids. J. Am. Oil Chem. Soc. 42:353–371.

Couvreur, S., C. Hurtaud, C. Lopez, L. Delaby, and J. L. Peyraud. 2006. The linear relationship between the proportion of fresh grass in the cow diet, milk fatty acid composition, and butter properties. J. Dairy Sci. 89:1956–1969.

De Marchi, M., M. Penasa, A. Cecchinato, M. Mele, P. Secchiari, and G. Bittante. 2011. Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of Brown Swiss bovine milk. Animal 5:1653–1658.

De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa. 2014. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. J. Dairy Sci. 97:1171–1186.

Ferrand, M., B. Huquet, S. Barbey, F. Barillet, F. Faucon, H. Larroque, O. Leray, J. M. Trommenschlager, and M. Brochard. 2011. Determination of fatty acid profile in cow's milk using mid-infrared spectrometry: Interest of applying a variable selection by genetic algorithms before a PLS regression. Chemom. Intell. Lab. Syst. 106:183–189.

Ferrand-Calmels, M., I. Palhiere, M. Brochard, O. Leray, J. M. Astruc, M. R. Aurel, S. Barbey, F. Bouvier, P. Brunschwig, H. Caillatt, M. Douguet, F. Faucon-Lahalle, M. Gele, G. Thomas, J. M. Trommenschlager, and H. Larroque. 2014. Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. J. Dairy Sci. 97:17–35.

German, J. B., R. A. Gibson, R. M. Krauss, P. Nestel, B. Lamarche, W. A. van Staveren, J. M. Steijns, L. C. P. G. M. de Groot, A. L. Lock, and F. Destaillats. 2009. A reappraisal of the impact of dairy foods and milk fat on cardiovascular disease risk. Eur. J. Nutr. 48:191–203.

Jensen, R. G. 2002. The composition of bovine milk lipids: January 1995 to December 2000. J. Dairy Sci. 85:295–350.

Kaylegian, K. E., J. M. Lynch, J. R. Fleming, and D. M. Barbano. 2009. Influence of fatty acid chain length and unsaturation on mid-infrared milk analysis. J. Dairy Sci. 92:2485–2501.

Luinge, H., E. Hop, E. Lutz, J. Vanhemert, and E. Dejong. 1993. Determination of the fat, protein and lactose content of milk using Fourier-transform infrared spectrometry. Anal. Chim. Acta 284:419–433.

Maurice-Van Eijndhoven, M. H. T., H. Soyeurt, F. Dehareng, and M. P. L. Calus. 2013. Validation of fatty acid predictions in milk using mid-infrared spectrometry across cattle breeds. Animal 7:348–354.

Montgomery, D. C. 2009. Response surface methodology. Pages 417–485 in Design and Analysis of Experiments. 7th ed. John Wiley & Sons Inc., Hoboken, NJ.

Næs, T., I. Mage, and V. H. Segtnan. 2011. Incorporating interactions in multi-block sequential and orthogonalised partial least squares regression. J. Chemometr. 25:601–609.

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

Poulsen, N. A., F. Gustavsson, M. Glantz, M. Paulsson, L. B. Larsen, and M. K. Larsen. 2012. The influence of feed and herd on fatty acid composition in 3 dairy breeds (Danish Holstein, Danish Jersey, and Swedish Red). J. Dairy Sci. 95:6362–6371.

Roger, J., F. Chauchard, and V. Bellon-Maurel. 2003. EPO-PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits. Chemom. Intell. Lab. Syst. 66:191–204.

Rutten, M. J. M., H. Bovenhuis, K. A. Hettinga, H. J. F. van Valenberg, and J. A. M. van Arendonk. 2009. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. J. Dairy Sci. 92:6202–6209.

Skov, T., D. Ballabio, and R. Bro. 2008. Multiblock variance partitioning: A new approach for comparing variation in multiple data blocks. Anal. Chim. Acta 615:18–29.

Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. J. Dairy Sci. 89:3690–3695.

Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. P. Berry, M. Coffey, and P. Dardenne. 2011. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. J. Dairy Sci. 94:1657–1667.

Soyeurt, H., F. Dehareng, P. Mayeres, C. Bertozzi, and N. Gengler. 2008. Variation of Δ^9-desaturase activity in dairy cattle. J. Dairy Sci. 91:3211–3224.

Strang, G. 2006. Orthogonality. Pages 141–200 in Linear Algebra and its Applications. 4th ed. Thomson Learning Inc., Toronto, ON, Canada.

APPENDIX: Cross-Validation of the Projection Procedure

The FT-IR measurements ($\mathbf{X}$), the FA ($\mathbf{y}_{FA}$), and the matrix $\mathbf{V}$ (Equation [4]) are split into calibration sets ($\mathbf{X}_{Cal}$, $\mathbf{y}_{Cal,FA}$, $\mathbf{V}_{Cal}$) and validation sets ($\mathbf{X}_{Valid}$, $\mathbf{y}_{Valid,FA}$, $\mathbf{V}_{Valid}$); $\mathbf{X}_{Cal}$ and $\mathbf{y}_{Cal,FA}$ are centered in Equations [8] and [9], respectively:

$$\mathbf{X}_{Cal,mncn} = \mathbf{X}_{Cal} - \mathbf{1} \cdot \bar{\mathbf{x}}_{Cal}, \quad [8]$$

where $\mathbf{X}_{Cal,mncn}$ is the mean centered calibration spectra, $\mathbf{1}$ is a vector of ones, and $\bar{\mathbf{x}}_{Cal}$ is a row vector containing the average value of each of the columns in $\mathbf{X}_{Cal}$; and

$$\mathbf{y}_{Cal,FA,mncn} = \mathbf{y}_{Cal,FA} - \mathbf{1} \cdot \bar{y}_{Cal,FA}, \quad [9]$$

where $\mathbf{y}_{Cal,FA,mncn}$ is the mean centered FA calibration set, $\bar{y}_{Cal,FA}$ is the average value of $\mathbf{y}_{Cal,FA}$.

Likewise, $\mathbf{X}_{Valid}$ and $\mathbf{y}_{Valid,FA}$ are centered by $\bar{\mathbf{x}}_{Cal}$ and $\bar{y}_{Cal,FA}$ in Equations [10] and [11], respectively:

$$\mathbf{X}_{Valid,mncn} = \mathbf{X}_{Valid} - \mathbf{1} \cdot \bar{\mathbf{x}}_{Cal}, \quad [10]$$

where $\mathbf{X}_{Valid,mncn}$ is the centered validation spectra, and

$$\mathbf{y}_{Valid,FA,mncn} = \mathbf{y}_{Valid,FA} - \mathbf{1} \cdot \bar{y}_{Cal,FA}, \quad [11]$$

where $\mathbf{y}_{Valid,FA,mncn}$ is the centered FA validation set.

The multiple linear regression (MLR) model between $\mathbf{X}_{Cal,mncn}$ and $\mathbf{V}_{Cal}$ is found in Equation [12]:

$$\mathbf{X}_{Cal,mncn} = \mathbf{V}_{Cal} \cdot \mathbf{D} + \mathbf{E}, \quad [12]$$

where $\mathbf{D}$ is the regression coefficients and $\mathbf{E}$ is the residuals.

The solution to $\mathbf{V}_{Cal} \cdot \mathbf{D}$ will obviously be a linear combination of $\mathbf{V}_{Cal}$ (i.e., the TF × Breed interaction). The regression coefficients $\mathbf{D}$ are estimated from Equation [13]:

$$\mathbf{D} = (\mathbf{V}_{Cal}^T \cdot \mathbf{V}_{Cal})^{-1} \mathbf{V}_{Cal}^T \cdot \mathbf{X}_{Cal,mncn}, \quad [13]$$

where $\mathbf{V}_{Cal}^T$ is the transpose of $\mathbf{V}_{Cal}$.

The part of $\mathbf{X}_{Valid,mncn}$ being a linear combination of the TF × Breed interaction is estimated in Equation [14]:

$$\mathbf{X}_{Valid,TF\times Breed} = \mathbf{V}_{Valid} \cdot \mathbf{D}, \quad [14]$$

and the part of $\mathbf{X}_{Valid,mncn}$ being independent of the TF × Breed interaction is estimated in Equation [15]:

$$\mathbf{X}_{Valid,O} = \mathbf{X}_{Valid,mncn} - \mathbf{X}_{Valid,TF\times Breed}. \quad [15]$$

A PLS model is fitted between $\mathbf{X}_{Cal,mncn}$ and $\mathbf{y}_{Cal,FA,mncn}$, and the regression coefficients $\mathbf{b}$ are obtained. The part of $\hat{\mathbf{y}}_{Valid,FA,mncn}$ being a linear combination of the TF × Breed interaction is estimated in Equation [16]:

$$\hat{\mathbf{y}}_{Valid,FA_{TF\times Breed}} = \mathbf{X}_{Valid,TF\times Breed} \cdot \mathbf{b}, \quad [16]$$

and the part of $\hat{\mathbf{y}}_{Valid,FA,mncn}$ being orthogonal to the TF × Breed interaction is estimated in Equation [17]:

$$\hat{\mathbf{y}}_{Valid,FA_O} = \mathbf{X}_{Valid,O} \cdot \mathbf{b}. \quad [17]$$

In the current study, the projection procedure was cross-validated by the venetian blind method with 10 data splits (i.e., each validation set is determined by selecting every 10th sample in the data set, starting at sample 1 through 10).

PAPER V

Quantification of bovine milk protein composition and coagulation properties using infrared spectroscopy and chemometrics: A result of collinearity among reference variables

C. E. Eskildsen, T. Skov, M. S. Hansen, L. B. Larsen and N. A. Poulsen

In review, Journal of Dairy Science, 2016.

Quantification of bovine milk protein composition and coagulation properties using infrared spectroscopy and chemometrics: A result of collinearity among reference variables

C. E. Eskildsen*[1], T. Skov*, M. S. Hansen†, L. B. Larsen‡ and N. A. Poulsen‡

*Department of Food Science, University of Copenhagen, DK-1958 Frederiksberg, Denmark

†Arla Strategic Innovation Centre, DK-8220 Brabrand, Denmark

‡Department of Food Science, Aarhus University, DK-8830 Tjele, Denmark

[1]Corresponding author:
Carl Emil Eskildsen
Rolighedsvej 26
DK-1958 Frederiksberg
Phone: + 45 50901306
E-mail: carle@food.ku.dk

ABSTRACT

Predicting protein fractions and coagulation properties in bovine milk from Fourier transform infrared (FT-IR) measurements is desirable. However, such predictions may rely on correlations with total protein content. The aim of this study is to show how correlations between total protein content, protein fractions and coagulation properties are responsible for successful predictions of protein fractions and rennet-induced coagulation properties in milk samples. This study comprised 832 bovine milk samples from two breeds (426 Holstein and 406 Jersey). Holstein samples were collected from 20 Danish dairy herds from October to December 2009; Jersey samples were collected from 22 Danish dairy herds from February to April 2010. All samples were from conventional herds and taken while cows were housed. The results showed that κ-CN, α_{S1}-CN, α_{S1}-CN 8P and curd firming rate could be predicted from FT-IR measurements of the milk samples (with coefficient of determination, R^2 between 0.66 and 0.71). However, the study further demonstrated that the successfulness of these FT-IR based predictions is based on indirect relationships with total protein content. Hence, the FT-IR predictions rely on covariance structures with total protein content rather than absorption bands directly associated with the protein fractions and coagulation properties. If covariance structures between the protein fractions, coagulation properties and total protein content, used to calibrate Partial Least Squares (PLS) models, are not conserved in future samples, these samples will show incorrect predictions of the protein fractions and coagulation properties. This was demonstrated using samples of one breed to calibrate and samples of the other breed to validate PLS models for β-CN. The 2 breeds have different covariance structures between β-CN and total protein content. The results showed that the validation samples yielded incorrect predictions. This may limit the usefulness of FT-IR-based predictions of protein fractions in milk recording, as indirect covariance structures in the calibration set must be valid for future samples. Otherwise, future samples will show incorrect predictions.

Key Words: bovine, coagulation properties, infrared spectroscopy, milk, protein composition

INTRODUCTION

In recent years, Fourier transform infrared (**FT-IR**) spectroscopy has been suggested as a method with the ability to outcompete chromatography based analyses for detailed milk analyses (Soyeurt et al., 2006; Rutten et al., 2011; De Marchi et al., 2014). Especially the high-throughput measurements of FT-IR analyses is an advantage over the more time-consuming chromatography based analyses. If FT-IR based predictions do not compromise quality of the estimated detailed milk composition this would open possibilities for using FT-IR as a phenotyping tool in relation to e.g. breeding programs (Bovenhuis et al., 2013). Particularly the major fatty acids have been proposed to be routinely estimated from FT-IR measurements (Soyeurt et al., 2006; Rutten et al., 2009) whereas protein fractions in general have been predicted with less accuracy from FT-IR measurements (Bonfatti et al., 2011; Rutten et al., 2011).

Combined, the four caseins (α_{S1}-CN, α_{S2}-CN, β-CN and κ-CN) and two whey proteins α-lactalbumin and β-lactoglobulin (α-LA and β-LG) comprise approximately 90% of the bovine protein fraction (wt/wt; Walstra, 1999). Heterogeneity in the major milk proteins is related to underlying genetic polymorphism (Heck et al., 2009), as well as affected by e.g. days in milking and parity (Poulsen et al., 2016). In addition, post translational modification of the proteins adds to their isoform complexity and is shown to affect technological properties of the milk including milk coagulation (Frederiksen et al., 2011; Jensen et al., 2012). In particular, the relative fraction of κ-CN to total protein is important for milk coagulation together with the level of glycosylated κ-CN (Jensen et al., 2012; Bonfatti et al., 2014). Furthermore, bovine milk proteins are sold as ingredients based on their nutritional or functional properties and reliable high-throughput estimation of these are therefore warranted.

Several studies attempted using FT-IR measurements for predicting protein fractions or coagulation properties of bovine milk samples (Dal Zotto et al., 2008; Bonfatti et al., 2011; Rutten et al., 2011). Infrared radiation is absorbed by exciting fundamental vibrations of molecular bonds expressing a change in the dipole moment. For proteins in a milk matrix, the most pronounced absorption signal is related to the amide II band (primarily N-H bending vibration) located at approx. 1,540 cm^{-1} (Luinge et al., 1993). This band is originating from the peptide backbone. Specific milk proteins are differentiated by the composition of their amino acid residues. The different amino acids are present in more or less all the protein fractions, though in varying relative contents. However, the

FT-IR active groups in the amino acids are more or less the same (Walstra et al., 2006). Therefore, it might be difficult to obtain unique FT-IR signals for specific protein fractions in milk.

Eskildsen et al. (2014) argued that successful predictions of individual fatty acids in milk by FT-IR measurements were due to covariance between individual fatty acids and total fat content. Bonfatti et al. (2015) speculated that a similar covariance phenomenon between protein fractions and total protein (**TP**) content is responsible for successful predictions of protein fractions. The TP content is easily predicted from FT-IR measurements using primarily the amide II band (Luinge et al., 1993). Therefore, good correlations between TP and individual protein fractions might enable indirect predictions of protein fractions using the amide II. The indirect relationships between TP and protein fractions may change with factors like breed, feed, etc. Therefore, indirect relationships for calibrating e.g. partial least squares (**PLS**) regression models may compromise calibration robustness. If an indirect relationship is used to calibrate a regression model, the model will not be valid for future samples unless the indirect relationship is conserved in the new samples. If such an indirect relationship is not conserved, the model will provide incorrect predictions as shown by Eskildsen et al. (2014).

McDermott et al. (2016) argued that when predicting protein fractions from FT-IR measurements, it is not only TP related information that is being used during prediction but the FT-IR spectra are providing additional information. However, if information being related with TP is just partly used for predicting a given protein fraction, this prediction will be indirect and a change in TP will affect the prediction of the protein fraction. Furthermore, McDermott et al. (2016) did not identify this additional information. It could as well be related to additional indirect relationships with other major milk components or be a consequence of overfitting.

This study investigated if FT-IR can be used to obtain reliable prediction models for the major milk proteins and coagulation properties and how PLS models for predicting these protein fractions as well as curd firming rate (**CFR**) and rennet coagulation time (**RCT**) depend on correlations between these traits and TP in raw milk samples from 2 Danish dairy breeds.

MATERIALS AND METHODS

Morning milk samples from 406 Jersey cows, collected from 22 Danish dairy herds from February to April 2010, and 426 Holstein cows, collected from 20 Danish dairy herds from October to

december 2009, were included in this study. All samples were from conventional herds and taken while cows were housed. The sample set used in this study is a subset of the sample set used in Eskildsen et al. (2014) for prediction of individual fatty acids. As compared with Eskildsen et al. (2014) some samples were excluded in this present study as the samples were not analyzed for content of protein fractions. As also pointed out by Eskildsen et al. (2014) the sampling strategy aimed at minimizing environmental variation but maximizing the genetic variation of the sample population (Poulsen et al., 2012). The milk samples were placed on ice during transport to the laboratory and analyzed the same day using MilkoScan FT2 (FOSS Analytical A/S, Hillerød, Denmark) and ReoRox4 rheometer (MediRox AB, Nyköping, Sweden). The FT-IR measurements were performed on the raw milk samples but the rheological analysis was performed on skimmed milk samples (Jensen et al., 2012). A part of each skim milk sample was stored at -20ºC for quantification of protein fractions by electrospray mass spectrometry coupled to liquid chromatography (**LC-ESI/MS**).

Protein fractions were quantified using LC-ESI/MS as described by Jensen et al. (2012). The relative reference values obtained from LC-ESI/MS were multiplied by TP in order to convert the protein fractions into quantities per unit milk (g of specific protein/100 g of milk) as quantities of protein fractions in units of TP are not comparable with FT-IR spectra obtained from liquid milk.

Rennet-induced coagulation of skimmed milk samples was determined by a ReoRox4 rheometer, as outlined in Poulsen et al. (2013) using addition of chymosin (ChyMax, Chr. Hansen, Hørsholm, Denmark). Milk coagulation properties for individual samples were described as RCT and CFR. The RCT was defined as the amount of time from chymosin addition to when the phase angle reached 45º ($\theta = 45º$). The CFR was calculated from consecutive points of the linear part of the gelation profile $[\Delta G'/\Delta t]_{lin}$. The descriptive statistics (including breed differences) for protein composition and milk coagulation properties of the samples are described by Poulsen et al. (2013) and Poulsen et al. (2016).

As also described by Eskildsen et al. (2014) FT-IR full-spectra were recorded on all samples using MilkoScan FT2. Each FT-IR measurement, was ratioed against a FT-IR water spectrum (as background). Spectra were obtained in triplicates from fresh whole milk samples and the average difference spectrum across the three replicates was used for further analysis. The TP content, total fat and total lactose content were determined using MilkoScan FT2 as also outlined by Eskildsen et al. (2014).

Data Analysis

The MATLAB version R2014a (8.3.0.532, MathWorks Inc., Natick, MA, USA) and PLS_Toolbox version 7.5 (Eigenvector Research Inc., Manson, WA, USA) was used for data analysis.

FT-IR spectra were obtained in transmittance and transformed into absorbance to obey Beer's law. Spectra were obtained in the range from 5,008 cm^{-1} to 925 cm^{-1}, with a total of 1,060 data points, for each sample. The FT-IR spectra were preprocessed by Savitzky-Golay 1st derivative (window size of 9 points and second-order polynomial) to remove offset differences between samples and then mean centered before PLS modeling. Both Holstein and Jersey samples were included (simultaneous) in fitting the PLS models. All PLS models were built with a mean centered univariate response variable.

Venetian blinds with 10 data splits (starting at sample 1 through 10, each validation set was determined by selecting every 10th sample in the data set) was used for validating PLS models calibrated on both Jersey and Holstein samples. Model parameters (cross-validated coefficient of determination, $\mathbf{R^2CV}$ and root mean squared error of cross-validation, **RMSECV**) are reported and were used for choosing number of Latent Variables.

In order to investigate whether indirect relationships are responsible for successful predictions of the protein fractions, 3 parameters are calculated and compared. The first parameter is the R^2CV from PLS models (measured vs. predicted). This parameter shows how good the individual milk proteins (or coagulation properties) are predicted. The second parameter is the coefficient of determination ($\mathbf{R^2}$) between the measured values of the individual proteins (or coagulation properties) and predicted TP. In the cases where the protein fractions are only predicted by the correlation to TP, the second parameter will equal the model performance (the first parameter). If the protein fractions are modeled only by the correlation with TP, then the protein fractions and TP will be modeled by the same wavenumbers of the FT-IR spectra (i.e. the same linear combination of the wavenumbers). This would cause the correlation between the predicted values and predicted TP to be 1. Hence, if the third parameter is 1 then the protein fractions are modelled only by the correlation with TP.

To highlight the consequences of indirect models, one additional PLS model predicting β-CN was calibrated on a subset of the Holstein samples. The subset included 16 Holstein herds (338 samples). The remaining 4 Holstein herds (88 samples) were used as one test set and all the Jersey samples were used as another test set. The 4 Holstein herds were randomly selected. The data were

preprocessed similar to the multi-breed PLS models. The root mean squared error of calibration (**RMSEC**) for the Holstein calibration samples was compared with the root mean squared error of prediction (**RMSEP**) for the Holstein test set samples and the Jersey samples, respectively.

RESULTS AND DISCUSSION

The FT-IR absorption spectra of the milk samples are presented in Figure 1a. The spectra are colored by TP. The region from 2,968 cm^{-1} to 5,008 cm^{-1} and from 1,692 cm^{-1} to 1,604 cm^{-1} was not included for modeling due to low signal-to-noise ratio. No peaks are found in the region from 1,773 cm^{-1} to 2,802 cm^{-1} and therefore this region was also not included for modeling. Furthermore, due to possible saturation of the single beam spectra, a narrow region from 2,915 cm^{-1} to 2,930 cm^{-1} was not included for modeling. Figure 1b shows the spectral parts included for modeling. The spectra in Figure 1b are also colored by TP.

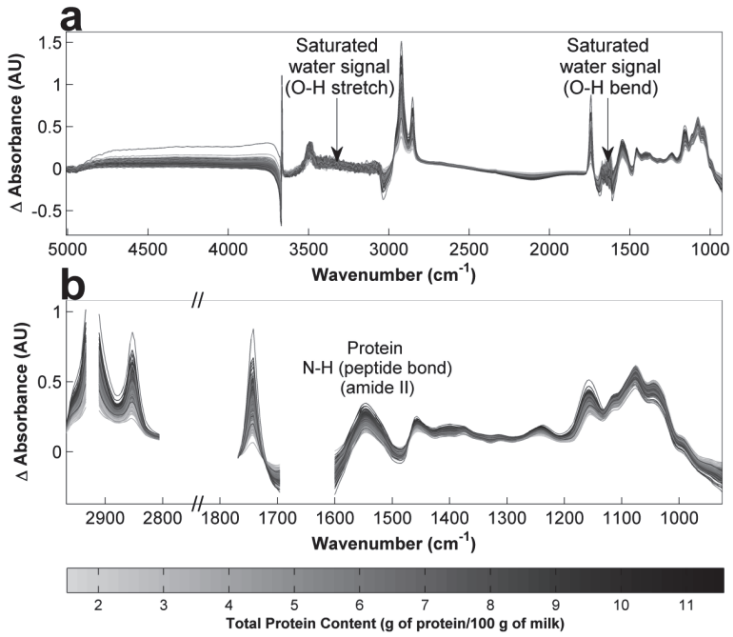


Figure 1. (a) Fourier transform infrared (FT-IR) difference spectra (water spectrum subtracted from each milk measurement); (b) parts of the FT-IR difference spectra included for PLS modeling. Spectra are colored (shaded) by total protein content. AU = absorbance units. Color version available in the online PDF. The figure is modified from Eskildsen et al. (2014).

Offset differences between spectra may affect the PLS models. In order to remove offset differences the spectra were preprocessed by Savitzky-Golay 1st derivative prior modeling, also done in Eskildsen et al. (2014).

The spectral data were modeled by principal component analysis (results not shown). The spectral sample set used in this study only differ slightly from the spectral sample set used by Eskildsen et al. (2014). Therefore, the results from principal component analysis are also very similar. Eskildsen et al. (2014) found that Holstein and Jersey samples were to a large extent separated by spectral variation associated with total fat and TP, where the Jersey samples are having higher content of both total fat and TP.

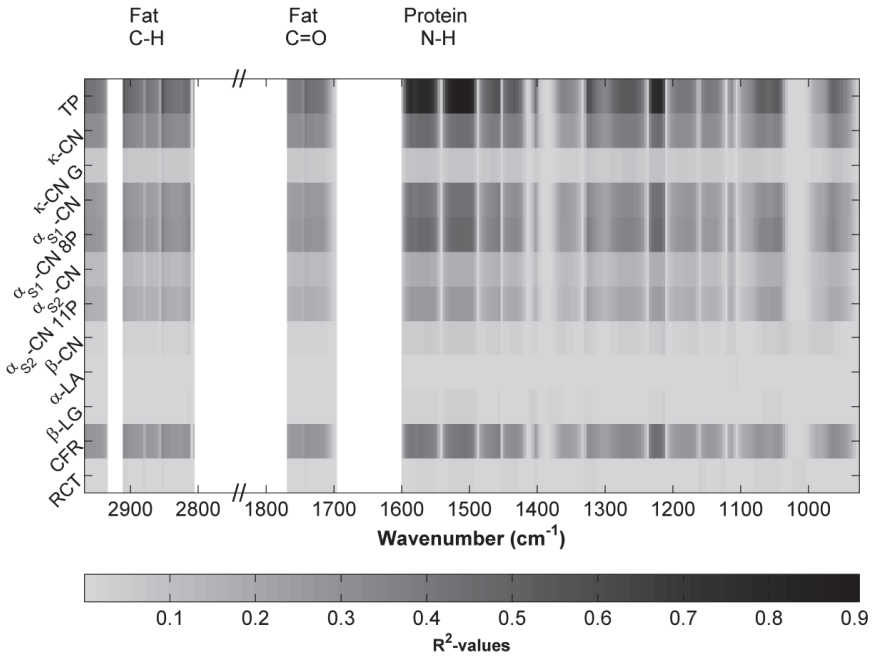


Figure 2. Coefficient of determination (R^2) between measured specific proteins (g/100 g of milk) and Fourier transform infrared measurements preprocessed by Savitzky-Golay first derivative. TP = total protein content. RCT = rennet coagulation time, CFR = curd firming rate. Color version available in the online PDF.

To explore how individual fatty acids related with the FT-IR spectra, Eskildsen et al. (2014) calculated R^2 values between the absorption intensities of each wavenumber and concentrations of individual fatty acids. Eskildsen et al. (2014) found that the fatty acids showed a similar correlation pattern as total fat content with FT-IR absorption intensities. Hence, the fatty acids primarily correlated with the spectral regions associated with symmetric and anti-symmetric stretching vibrations of methylene groups and the stretching vibration of carbonyl groups. In a similar fashion, R^2-values were calculated between absorption intensities of each wavenumber of the preprocessed (Savitzky-Golay 1st derivative) FT-IR spectra and the measured values of the individual protein fractions, CFR and RCT (Figure 2). Figure 2 shows a similar trend for TP and protein fractions as Eskildsen et al. (2014) found for total fat content and fatty acids. The individual protein fractions (like κ-CN and α_{S1}-CN 8P) and CFR have more or less the same correlation pattern as TP, which is having high correlation with the expected area between 1,600 cm^{-1} and 1,500 cm^{-1} related to the amide II band. This could very well indicate that the individual proteins and coagulation properties do not give rise to specific absorption bands in FT-IR and are predicted mainly be the amide II band.

Table 1 shows the results of the PLS models for protein fractions, CFR and RCT. Acceptable predictions were obtained for κ-CN, α_{S1}-CN, α_{S1}-CN 8P as well as CFR, which all yielded R^2CV-values round 0.7. The results for α_{S1}-CN and κ-CN are in agreement to what was reported by De Marchi et al. (2009b) and Bonfatti et al. (2011). In this study poor predictions were obtained for glycosylated κ-CN (**κ-CN G**), α_{S2}-CN, α_{S2}-CN 11P, β-CN, α-LA, β-LG, as well as RCT, which all have R^2CV-values from 0.06 to 0.47. The results for α_{S2}-CN and β-CN are in agreement with De Marchi et al. (2009b) and Bonfatti et al. (2011). However, our predictions for κ-CN G, β-CN, α-LA and β-LG are worse than what have been previously reported (De Marchi et al., 2009b; Bonfatti et al., 2011; Rutten et al., 2011). Moreover, in this study the prediction of RCT is remarkably worse compared to other studies (Dal Zotto et al., 2008; De Marchi et al., 2009a; De Marchi et al., 2013). It is worth noticing that some protein fractions and coagulation properties are well predicted in some studies and poorly predicted in other studies. As highlighted by De Marchi et al. (2014) different reference methods are used across studies. When calibration models are evaluated and compared based on R^2CV-values, differences in uncertainties among the reference methods will impact the R^2CV-values (DiFoggio, 1995; Faber and Kowalski, 1997; De Marchi et al., 2014). Furthermore, in this present study milk coagulation properties were determined on skimmed milk whereas this is not the case in other studies

(Dal Zotto et al., 2008; De Marchi et al., 2009a; De Marchi et al., 2013). This makes it difficult to directly compare calibration models across studies. Moreover, if protein fractions and coagulation properties are predicted by indirect relationships, then these indirect relationships will also impact evaluations of the calibration models. In some studies, certain indirect relationships may enable good predictions for specific traits, whereas these indirect relationships may not be present in other studies due to differences in e.g. sampling strategies and consequently the traits are showing poor predictions.

Table 1. Results from partial least squares models[1]

Protein fraction or coagulation trait	Range	Mean	SD	CV	LV	R^2CV	RMSECV	Relative error
κ-CN	0.11 – 0.40	0.25	0.06	0.25	2	0.71	0.03	0.13
κ-CN G	0.01 – 0.15	0.05	0.02	0.34	4	0.20	0.02	0.30
α_{S1}-CN	0.50 – 1.64	1.05	0.18	0.18	2	0.66	0.11	0.10
α_{S1}-CN 8P	0.29 – 1.40	0.80	0.16	0.20	2	0.71	0.09	0.11
α_{S2}-CN	0.09 – 0.47	0.20	0.06	0.30	2	0.36	0.05	0.24
α_{S2}-CN 11P	0.05 – 0.32	0.13	0.04	0.34	2	0.47	0.03	0.25
β-CN	0.57 – 1.81	1.25	0.16	0.13	7	0.25	0.14	0.11
α-LA	0.02 – 0.20	0.11	0.02	0.22	4	0.06	0.02	0.21
β-LG	0.09 – 0.51	0.27	0.06	0.24	9	0.34	0.05	0.19
CFR	0 – 44.85	14.80	8.42	0.57	2	0.66	4.90	0.33
RCT	5.53 – 30.35	13.29	2.33	0.18	4	0.12	2.19	0.17

[1]SD = Standard deviation; CV = Coefficient of variation; LV = number of latent variables; R^2CV = cross-validated coefficient of determination; RMSECV = root mean squared error of cross-validation; relative error = RMSECV/mean; CFR = curd firming rate; RCT = rennet coagulation time. Range, mean, SD, CV and RMSECV are in units of grams/100 grams of milk for protein fractions, units of Pa/min for CFR and units of min for RCT.

In order to investigate these indirect relationships the three parameters, R^2CV (model performance), R^2 (measured traits vs predicted TP) and R^2 (predicted traits and predicted TP), were calculated and compared. Model performance (the first parameter) is plotted on the X-axis in Figure 3.

The further to the right the protein fractions are located in Figure 3, the better the protein fractions are predicted. On the Y-axis (Figure 3) the R^2 between the measured traits and predicted TP (the second parameter) is plotted with an open square and R^2 between predicted traits and predicted TP (the third parameter) is plotted with a filled triangle.

Figure 3 reveals that the protein fractions α_{S1}-CN 8P, κ-CN and α_{S1}-CN as well as CFR, which are all modeled well, are very dependent on the correlation with TP (note that R^2-values for α_{S1}-CN coincide with R^2-values for CFR and R^2-values for α_{S1}-CN 8P coincide with R^2-values for κ-CN coincide in Figure 3). Total fat and lactose are also modeled very well (Figure 3). However, these models are not dependent on the correlation with TP. This is shown by the R^2 between the measured values and TP (parameter 2) and the R^2 between the predicted values and TP (parameter 3) are almost the same. Hence, total fat and lactose must be predicted by parts of the FT-IR spectra, which are different from the parts used for predicting TP as also shown by (Luinge et al., 1993).

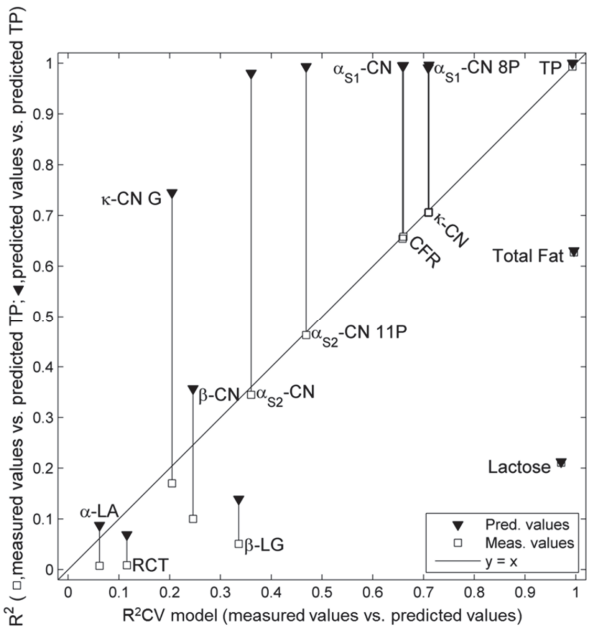


Figure 3. Relationship between cross-validated coefficient of determination (**R^2CV**) from cross-validated PLS models (X-axis) and coefficient of determination (**R^2**) between measured and predicted values of individual proteins, coagulation properties, lactose, total fat and predicted total protein (**TP**) content (Y-axis). RCT = rennet coagulation time, CFR = curd firming rate.

Figure 4 illustrates the problems in modeling individual proteins by the indirect relationship with TP. In Figure 4a, β-CN is plotted against TP for the 2 breeds (Holstein and Jersey). From Figure 4a it is clear that the relationship between β-CN and TP is different for the 2 breeds. There is a fairly good correlation between β-CN and TP for the Holstein samples, whereas this correlation is poor for the Jersey samples. Samples of the 4 randomly selected Holstein herds are highlighted with triangular symbols in Figure 4a. These 4 herds are kept aside together with the Jersey samples for testing and a PLS model is calibrated on the remaining Holstein samples belonging to 16 herds. Figure 4b shows the measured vs. predicted β-CN values for the calibration set and the two test sets. The RMSEC for the Holstein calibration set is 0.09 g β-CN pr. 100 g of milk. The RMSEP for the Holstein test set is also 0.09 g β-CN pr. 100 g of milk, whereas the RMSEP for the Jersey samples is 0.29 g β-CN pr. 100 g of milk, which is remarkably higher than the error for the Holstein test set. Hence, Figure 4 highlights the problems using models based on indirect correlations. The indirect relationship used to calibrate the PLS model on Holstein samples is not valid for Jersey samples. In order for a model based on indirect relationships to be valid, the indirect relationships used for calibrating the model must be conserved in the new data set.

Building universal models (using e.g. a high number of samples from multiple breeds in the calibration set) will not necessarily solve the problem of indirect models. Including a high amount of variation in the calibration set (different breeds, feeding systems, lactation stages, etc.) will most likely break the indirect relationship between protein fractions and TP. However, the consequence will presumably be a poor global model as the links between the FT-IR spectra and the protein fractions are lost. In Figure 4b the PLS model is calibrated on the Holstein samples only. There is a good relationship between β-CN and TP for the Holstein samples and consequently the Holstein samples are predicted well (Figure 4b). The PLS models presented in Table 1 are calibrated on both Jersey and Holstein samples. Figure 4a reveals that the overall correlation between TP and β-CN is poor if taking both breeds into account simultaneously. Consequently, the model for β-CN is poor when both breeds are included in the calibration set (Table 1). An option would then be to only construct indirect local models on e.g. single breeds. Here the correlations between protein fractions and TP are expected to be strong and thereby is the link between the FT-IR spectra and the protein fractions established. Nevertheless, such indirect relationships need to be validated. Furthermore, the indirect predictions of

protein fractions are of limited use as they (in principle) do not provide additional information (or very limited additional information) if compared to a model predicting TP.

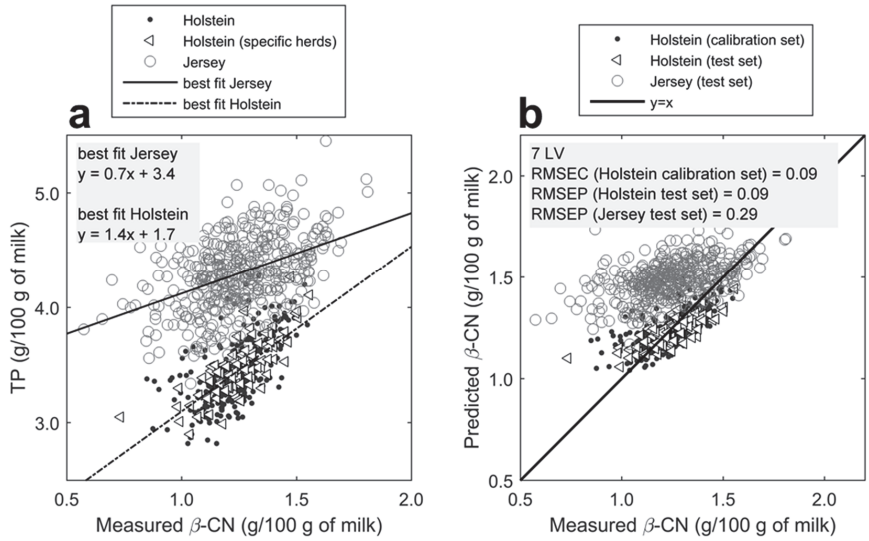


Figure 4. (a) Relationship between measured β-CN and total protein content (TP) in milk samples from individual Jersey and Holstein cows; (b) measured versus predicted values of β-CN; β-CN was predicted by Partial Least Squares (PLS) regression applied to Fourier transform infrared measurements. The PLS model is calibrated on a subset of Holstein sample and tested with the remaining Holstein samples and the Jersey samples. LV = latent variables, RMSEC = root mean squared error of calibration, RMSEP = root mean squared error of prediction. Color version available in the online PDF.

CONCLUSIONS

Concentrations of protein fractions as well as coagulation properties were predicted by applying PLS to FT-IR measurements of milk samples. This paper has illustrated that predictions of protein fractions and coagulation properties, for the samples available in this study, rely on indirect relationships with TP. This fact compromises robustness of the calibration models for protein fractions and coagulation properties. The calibration models will no longer be valid if these indirect relationships change. Hence, calibration models may not be valid for samples of a different nature (e.g. different breeds or lactation stages). Therefore, recommendations on applying FT-IR based estimates of protein fractions and coagulation properties for e.g. breeding purposes must account for these indirect relationships. It is simply not sufficient just evaluating model performance parameters like R^2CV and

RMSECV as done in previous studies. The indirect relationships must be fully understood and controlled. Furthermore, one should be aware that the indirect models are providing very limited additional information as compared with models on TP.

ACKNOWLEDGMENTS

The Danish Council for Strategic Research is acknowledged for generous financial support to the project entitled "Exploration of high-throughput FTIR-based solutions as an effective tool for correlating economically important milk quality characteristics with genetic markers" under the inSPIRe (Danish Industry-Science Partnership for Innovation and Research in Food Science) consortium (Copenhagen, Denmark). Furthermore, Arla Foods amba (Viby J, Denmark), FOSS Analytical A/S (Hillerød, Denmark), Danish Cattle Federation (Aarhus, Denmark), as well as the Milk Levy Fund (Denmark) are acknowledged for financial support.

REFERENCES

Bonfatti, V., A. Cecchinato and P. Carnier. 2015. Short communication: Predictive ability of fourier-transform mid-infrared spectroscopy to assess CSN genotypes and detailed protein composition of buffalo milk. J. Dairy Sci. 98:6583-6587.

Bonfatti, V., G. Chiarot and R. Carnier. 2014. Glycosylation of kappa-casein: Genetic and nongenetic variation and effects on rennet coagulation properties of milk. J. Dairy Sci. 97:1961-1969.

Bonfatti, V., G. Di Martino and P. Carnier. 2011. Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and contents of protein genetic variants of individual milk of simmental cows. J. Dairy Sci. 94:5776-5785.

Bovenhuis, H., M. Visker and A. Lundén. 2013. Selection for milk fat and milk protein composition. Advances in Animal Biosciences. 4:612-617.

Dal Zotto, R., M. De Marchi, A. Cecchinato, M. Penasa, M. Cassandro, P. Carnier, L. Gallo and G. Bittante. 2008. Reproducibility and repeatability of measures of milk coagulation properties and predictive ability of mid-infrared reflectance spectroscopy. J. Dairy Sci. 91:4103-4112.

De Marchi, M., C. C. Fagan, C. P. O'Donnell, A. Cecchinato, R. Dal Zotto, M. Cassandro, M. Penasa and G. Bittante. 2009a. Prediction of coagulation properties, titratable acidity, and pH of bovine milk using mid-infrared spectroscopy. J. Dairy Sci. 92:423-432.

De Marchi, M., V. Toffanin, M. Cassandro and M. Penasa. 2014. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. J. Dairy Sci. 97:1171-1186.

De Marchi, M., V. Toffanin, M. Cassandro and M. Penasa. 2013. Prediction of coagulating and noncoagulating milk samples using mid-infrared spectroscopy. J. Dairy Sci. 96:4707-4715.

De Marchi, M., V. Bonfatti, A. Cecchinato, G. Di Martino and P. Carnier. 2009b. Prediction of protein composition of individual cow milk using mid-infrared spectroscopy. Italian Journal of Animal Science. 8:399-401.

DiFoggio, R. 1995. Examination of some misconceptions about near-infrared analysis. Appl. Spectrosc. 49:67-75.

Eskildsen, C. E., M. A. Rasmussen, S. B. Engelsen, L. B. Larsen, N. A. Poulsen and T. Skov. 2014. Quantification of individual fatty acids in bovine milk by infrared spectroscopy and chemometrics: Understanding predictions of highly collinear reference variables. J. Dairy Sci. 97:7940-7951.

Faber, K. and B. R. Kowalski. 1997. Improved prediction error estimates for multivariate calibration by correcting for the measurement error in the reference values. Appl. Spectrosc. 51:660-665.

Frederiksen, P., K. Andersen, M. Hammershøj, H. Poulsen, J. Sørensen, M. Bakman, K. Qvist and L. Larsen. 2011. Composition and effect of blending of noncoagulating, poorly coagulating, and well-coagulating bovine milk from individual danish holstein cows. J. Dairy Sci. 94:4787-4799.

Heck, J. M. L., A. Schennink, H. J. F. van Valenberg, H. Bovenhuis, M. H. P. W. Visker, J. A. M. van Arendonk and A. C. M. van Hooijdonk. 2009. Effects of milk protein variants on the protein composition of bovine milk. J. Dairy Sci. 92:1192-1202.

Jensen, H. B., N. A. Poulsen, K. K. Andersen, M. Hammershøj, H. D. Poulsen and L. B. Larsen. 2012. Distinct composition of bovine milk from jersey and holstein-friesian cows with good, poor, or noncoagulation properties as reflected in protein genetic variants and isoforms. J. Dairy Sci. 95:6905-6917.

Luinge, H., E. Hop, E. Lutz, J. Vanhemert and E. Dejong. 1993. Determination of the fat, protein and lactose content of milk using fourier-transform infrared spectrometry. Anal. Chim. Acta. 284:419-433.

McDermott, A., G. Visentin, M. De Marchi, D. Berry, M. Fenelon, P. O'Connor, O. Kenny and S. McParland. 2016. Prediction of individual milk proteins including free amino acids in bovine milk

using mid-infrared spectroscopy and their correlations with milk processing characteristics. J. Dairy Sci.

Poulsen, N. A., H. P. Bertelsen, H. B. Jensen, F. Gustavsson, M. Glantz, H. L. Månsson, A. Andrén, M. Paulsson, C. Bendixen and A. J. Buitenhuis. 2013. The occurrence of noncoagulating milk and the association of bovine milk coagulation properties with genetic variants of the caseins in 3 scandinavian dairy breeds. J. Dairy Sci. 96:4830-4842.

Poulsen, N. A., F. Gustavsson, M. Glantz, M. Paulsson, L. B. Larsen and M. K. Larsen. 2012. The influence of feed and herd on fatty acid composition in 3 dairy breeds (danish holstein, danish jersey, and swedish red). J. Dairy Sci. 95:6362-6371.

Poulsen, N. A., H. B. Jensen, L. B. Larsen. 2016. Factors influencing degree of glycosylation and phosphorylation of caseins in individual cow milk samples. J. Dairy Sci. Accepted

Rutten, M. J. M., H. Bovenhuis, J. M. L. Heck and J. A. M. van Arendonk. 2011. Predicting bovine milk protein composition based on fourier transform infrared spectra. J. Dairy Sci. 94:5683-5690.

Rutten, M. J. M., H. Bovenhuis, K. A. Hettinga, H. J. F. van Valenberg and J. A. M. van Arendonk. 2009. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. J. Dairy Sci. 92:6202-6209.

Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. J. Dairy Sci. 89:3690-3695.

Walstra, P., J. T. M. Wouters and T. J. Geurts. 2006. Dairy Science and Technology. 2nd ed. CRC Press, Taylor & Francis Group, FL, USA.

Walstra, P. 1999. Casein sub-micelles: Do they exist? Int. Dairy J. 9:189-192.

DEPARTMENT OF FOOD SCIENCE
FACULTY OF SCIENCE · UNIVERSITY OF COPENHAGEN
PHD THESIS 2016 · ISBN 978-87-93476-06-6

CARL EMIL AAE ESKILDSEN

Prediction of milk quality parameters using vibrational spectroscopy and chemometrics – opportunities and challenges in milk phenotyping

SL grafik