In compliance with the
Canadian Privacy Legislation
some supporting forms
may have been removed from
this dissertation.

While these forms may be included
in the document page count,
their removal does not represent
any loss of content from the dissertation.

Statistical Approaches for Milk Composition Determination Using Combined Near Infrared, Raman, Conductivity, and Refractive Index Measurements

Kalumin Amila de Silva

Department of Chemistry
McGill University
Montréal, Québec
Canada

November 2002

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master's of Science

Your file Votre référence
ISBN: 0-612-88184-9
Our file Notre référence
ISBN: 0-612-88184-9

Abstract

Measurement of milk composition is a necessary step in production in the dairy industry. Determination of the major constituents of milk, fat, lactose and protein, provides important information in estimating animal health, economic value of milk, monitoring dairy herd management, and designation of milk for various dairy products. Current practices for routine milk composition determination employ commercial infrared systems. The use of SW-NIR and NIR FT-Raman spectra coupled with conductivity and refractive index could lead to more frequent and less costly analysis of fat, lactose and protein in milk.

The present study examines the potential of both SW-NIR absorbance spectrophotometry and NIR FT-Raman spectrophotometry to develop a model to estimate fat, lactose, and protein in whole milk of cows. To accomplish this, 79 milk standards, spanning the range of composition seen in practice, were obtained. Acquisition of NIR spectra over the wavelength range of 700 nm to 1018 nm was conducted. Between 0 and 3700 cm^{-1}, NIR FT-Raman spectrophotometric measurements of the milk samples were made using a 1064 nm Nd: YAG laser source. Conductivity and refractive index measurements were also obtained for the milk standards.

A partial least squares calibration with leave-N-out cross validation was made using spectra with conductivity and refractive index to estimate fat, lactose and protein contents. Calibrations were developed using 75% of the milk standards. Models were further validated using an independent test set comprised of the remaining 25% of the

data that had been excluded from the calibration. A second calibration was conducted using a genetic algorithm approach.

Increased accuracy was observed between estimated and reference concentrations when SW-NIR spectra with conductivity and refractive index were used as compared to using spectra alone. This is evidenced by standard errors for fat, lactose, and protein calibration being 0.59, 0.04, and 0.36 g/100g respectively. Accuracy achieved using Raman spectra was better than the SW-NIR calibration for fat and protein as indicated by standard errors for fat, lactose, and protein calibration being 0.21, 0.05, and 0.30 g/100g respectively. The genetic algorithm technique was found to improve estimation of lactose in both cases compared to the PLS calibrations. These findings show promise and emphasize the need to develop calibrations using NIR and NIR FT-Raman spectrophotometry for milk composition determination.

Résumé

Mesurer la composition du lait est une étape de production nécessaire dans l'industrie laitière. La détermination des composantes principales du lait, c.-à-d. gras, lactose et protéines, donne des renseignements importants pour estimer la santé animale et la valeur économique du lait, surveiller les troupeaux laitiers pour les gérer, et désigner l'usage du lait pour différents produits laitiers. Les méthodes courantes d'analyse laitière emploient des spectromètres infra-rouges commerciaux. L'emploi de spectres d'ondes courtes du proche infrarouge (SW-NIR) et NIR FT-Raman couplés avec des mesures de conductivité et d'indice de réfraction pourraient permettre des déterminations plus fréquentes et moins coûteuses du gras, du lactose et des protéines dans le lait.

Cette étude examine les spectrophotométries NIR et FT NIR-Raman en vue de développer un modèle pour estimer le gras, le lactose et les protéines dans le lait de vache entier. À ce but, 79 étalons de lait couvrant la gamme habituelle de compositions furent obtenus. Des spectres NIR furent mesurés entre 700 nm et 1018 nm. La spectrométrie FT NIR-Raman fut exécutée entre 0 et 3700 cm^{-1} sur les échantillons de lait en employant un laser Nd:YAG à 1064 nm comme source lumineuse. La conductivité et l'indice de réfraction des étalons de lait furent aussi enregistrés.

Un étalonnage par algorithme des moindres carrés avec validation croisée "leave-N-out" fut construit à partir des spectres et des données de conductivité et d'indice de réfraction pour estimer la teneur en gras, lactose et protéines. Les étalonnages furent développés à partir de 75% des étalons de lait. Les modèles furent ensuite validés au moyen d'un "test set" composé des 25% des données qui avaient été exclu de l'étalonnage. Un deuxième étalonnage fut construit à l'aide d'un algorithme génétique.

Un bon accord entre les concentrations estimées et les valeurs de référence fut observé quand les spectres SW-NIR furent utilisés en liaison avec la conductivité et l'indice de réfraction. Ceci se manifeste dans les erreurs associées à l'estimation du gras, du lactose et des protéines (respectivement 0.59, 0.04 et 0.36 g/100g). Au moyen de spectres Raman, les estimations de gras et de protéines furent améliorées. Les erreurs de calibration associées au gras, lactose et protéines sont de 0.21, 0.05 et 0.30 g/100g respectivement. L'algorithme génétique améliora l'estimation du lactose dans les deux cas. Ces conclusions prometteuses soulignent la nécessité de développer des étalonnages exploitant les spectrophotométries NIR et NIR-Raman pour déterminer la composition du lait.

Table of Contents

Abstract..... i
Résumé..... ii
List of Tables..... vi
List of Figures..... viii
Acknowledgements..... xii

Chapter 1 Introduction..... 1
1.1 Research Objectives..... 1
1.2 Overview of Milk Composition..... 2
1.3 Methods for Determining Milk Composition..... 6
1.4 Short-Wave Near-Infrared Spectrophotometry of Milk..... 12
1.5 Refractive Index of Milk..... 16
1.6 Electrical Conductivity of Milk..... 16
1.7 NIR FT-Raman Spectrophotometry of Milk..... 18
1.8 Overview of Research..... 19

Chapter 2 Experimental Methods..... 20
2.1 Milk Standards..... 20
2.2 Near Infrared Spectrophotometry of Milk Standards..... 21
2.3 Conductivity Analysis of Milk Standards..... 22
2.4 Refractive Index Analysis of Milk Standards..... 23
2.5 Raman Spectroscopy of Milk Standards..... 23
2.6 Pre-processing of Data for PLS..... 24
2.7 Partial Least Squares Analysis..... 25
2.8 Genetic Algorithm Approach for Analysis of Data..... 30

Chapter 3 Estimation Of Milk Composition Using SW-NIR Spectrophotometry 36

3.1 Near Infrared Spectra of Milk 36

3.2 Calibration of Fat, Lactose, and Protein in Milk by Short-Wave Near Infrared Spectrophotometry Using PLS with Leave N Out Cross Validation 38

3.3 Conductivity and Refractive Index of Milk 43

3.4 Calibration of Fat, Lactose, and Protein in Milk using Short-Wave Near Infrared Spectrophotometry With the Addition of Conductivity and Refractive Index Using PLS with Leave N Out Cross Validation 46

3.5 Investigation of inter-sample set variation using PLS analysis of SW-NIR measurements with leave-one-set-out cross validation 52

3.6 GA approach for fat, lactose, and protein estimation using SW-NIR, conductivity and refractive index 57

Chapter 4 Estimation Of Milk Composition Using NIR FT-Raman Spectrophotometry 66

4.1 NIR FT-Raman spectra of milk 66

4.2 Calibration of Fat, Lactose, and Protein in Milk by NIR FT-Raman Spectrophotometry Using PLS 67

4.3 Calibration of Fat, Lactose, and Protein in Milk by NIR FT-Raman Spectrophotometry With the Addition of Conductivity and Refractive Index Using PLS 74

4.4 Investigation of inter-sample set variation using PLS analysis of NIR FT-Raman measurements with leave-one-set-out cross validation 80

4.5 Estimation of fat, lactose, and protein using GA modelling of NIR FT-Raman spectra, conductivity and refractive index 84

Chapter 5 Discussion and Conclusions 94

References 102

List of Tables

Table 1.1. Approximate Distribution of Milk Constituents. 3

Table 1.3. Assignment of infrared absorption bands to food constituents. 13

Table 3.1. Estimation of fat, lactose and protein in milk using NIR spectra using PLS with leave-N-out cross validation 42

Table 3.2. Regression of reference constituent concentration on conductivity and refractive index 44

Table 3.3. Estimation of fat, lactose and protein in milk using SW-NIR spectra with the addition of conductivity and refractive index using PLS with leave-N-out cross validation 49

Table 3.4. Estimation of fat, lactose and protein using PLS analysis with leave-one-set-out cross validation of SW-NIR spectra and SW-NIR spectra with the inclusion of conductivity and refractive index 56

Table 3.5. Estimation of fat, lactose and protein using SW-NIR spectra, conductivity and refractive index with GA model. 63

Table 4.1. Estimation of fat, lactose and protein in milk using NIR FT-Raman spectra using PLS with leave-N-out cross validation 74

Table 4.2. Estimation of fat, lactose and protein in milk using NIR FT-Raman spectra with the addition of conductivity and refractive index using PLS with leave-N-out cross validation 78

Table 4.3. Estimation of fat, lactose and protein using PLS analysis with leave-one-set-out cross validation of NIR FT-Raman spectra and NIR FT-Raman spectra with the inclusion of conductivity and refractive index 84

Table 4.4. Estimation of fat, lactose and protein using NIR FT-Raman spectra, conductivity and refractive index with GA model. 92

Table 5.1. Summary of methods yielding most accurate estimation of constituent concentration using SW-NIR spectra 96

Table 5.2. Summary of methods yielding most accurate estimation of constituent concentration using NIR FT-Raman spectra 98

List of Figures

Figure 1.1. Conformational structures of β-lactose (a) and α-lactose (b) 5

Figure 3.1. Original NIR absorbance (a) spectra of 79 milk samples where each symbol represents a different sample set and smoothed, mean centered spectra for PLS (b) 37

Figure 3.2 Calibration coefficients for fat (a), lactose (b), and protein (c) estimation using SW-NIR spectra using PLS with leave-N-out cross validation..... 39

Figure 3.3 Estimation of fat, lactose, and protein in calibration set (a - c) and test set (d - f) using SW-NIR using PLS with leave-N-out cross validation..... 41

Figure 3.4. Conductivity and refractive index of milk samples 45

Figure 3.5 Calibration coefficients for fat (a), lactose (b), and protein (c) estimation using SW-NIR spectra with the addition of conductivity, (Cond.) and refractive index (Refr.) using PLS with leave-N-out cross validation..... 48

Figure 3.6 Estimation of fat, lactose, and protein in calibration set (a - c) and test set (d - f) using SW-NIR with the addition of conductivity and refractive index using PLS with leave-N-out cross validation 50

Figure 3.7 Calibration coefficients for fat (a), lactose (b), and protein (c) estimation using on SW-NIR spectra using PLS with leave-one-set-out cross validation 54

Figure 3.8 Calibration coefficients for fat (a), lactose (b), and protein (c) estimation using on SW-NIR spectra with the addition of conductivity (Cond.) and refractive index (Refr.), using PLS with leave-one-set-out cross validation 55

Figure 3.9 Standard Error at 800 generations for 1 to 15 wavelength model for fat (a), lactose (b), and protein (c) estimation using SW-NIR spectra, conductivity and refractive index. 58

Figure 3.10 Progression of standard error with increasing generations for optimal wavelength model for fat (a), lactose (b), and protein (c) estimation using SW-NIR spectra, conductivity and refractive index 60

Figure 3.11 Regression coefficients for GA selected wavelengths for fat (a), lactose (b), and protein (c) estimation using SW-NIR spectra, conductivity and refractive index plotted against average SW-NIR spectrum of milk samples 62

Figure 3.12 Estimation of fat, lactose, and protein in calibration set (a - c) and validation set (d - f) using SW-NIR, conductivity, and refractive index using GA calibration model. 64

Figure 4.1. Original NIR FT-Raman (a) spectra of 68 primary milk standards where each symbol represents a different sample set and smoothed, mean-centered spectra for PLS (b) 68

Figure 4.2 Calibration coefficients for fat (a), lactose (b), and protein (c) estimation using NIR FT-Raman spectra using PLS with leave-N-out cross validation 71

Figure 4.3 Estimation of fat, lactose, and protein in calibration set (a - c) and test set (d - f) using NIR FT-Raman spectra using PLS with leave-N-out cross validation 73

Figure 4.4 Calibration coefficients for fat (a), lactose (b), and protein (c) estimation using NIR FT-Raman spectra with the addition of conductivity (Cond.) and refractive index (Refr.) using PLS with leave-N-out cross validation. 76

Figure 4.5 Estimation of fat, lactose, and protein in calibration set (a - c) and test set (d - f) using NIR FT-Raman spectra with the addition of conductivity and refractive index using PLS with leave-N-out cross validation 79

Figure 4.6 Calibration coefficients for fat (a), lactose (b), and protein (c) estimation using on NIR FT-Raman spectra using PLS with leave-one-set-out cross validation 82

Figure 4.7 Calibration coefficients for fat (a), lactose (b), and protein (c) estimation using on NIR FT Raman spectra with the addition of conductivity (Cond.) and refractive index (Refr.), using PLS with leave-one-set-out cross validation 83

Figure 4.8 Standard Error at maximum generations for 1 to 10 wavenumber model for fat (a), lactose (b), and protein (c) estimation using NIR FT Raman spectra, conductivity and refractive index 86

Figure 4.9 Fitness with increasing generation for 3 (fat (a)), 6 (lactose (b)), and 3 (protein(c)) wavenumber models using NIR FT Raman spectra, conductivity and refractive index 88

Figure 4.10 Regression coefficients for GA selected wavelengths for fat (a), lactose (b), and protein (c) estimation using NIR FT-Raman spectra, conductivity and refractive index plotted against average NIR FT-Raman spectrum of milk samples 90

Figure 4.11 Estimation of fat, lactose, and protein in calibration set (a - c) and validation set (d-f) using NIR FT Raman spectra, conductivity, and refractive index using GA calibration model 93

Acknowledgements

I would like to thank my supervisor Dr. David H. Burns, for his guidance and encouragement. I have learned so much during my two years at McGill.

À mes filles: Holly Zulyniak, Kamilah Smith, Rashida Smith, and Debbie Mitra for their indispensable love and support. Thanks to Claudia Gributs for being a terrific labmate and mentor and an even better friend.

Finally thanks to my parents and two brothers. Your love is always cherished.

Chapter 1 Introduction

1.1 Research Objectives

Milk composition analysis is necessary for estimating animal health, determining economic value of milk, decision-making in dairy herd management, and designating milk for various dairy products. Current practices for routine milk composition determination employ commercial infrared systems that are costly and cannot be used for daily measurements. Use of NIR and/or NIR-Raman spectra coupled with conductivity and refractive index could lead to more frequent and less expensive analysis of fat, lactose and protein analysis in milk. In addition, these measurements are relatively easy to conduct and have the potential to be incorporated in automated milking instrumentation. This could lead to simultaneous acquisition of product and analysis. The objective of this research is to accurately estimate concentrations of the major constituents of milk, fat, lactose, and protein. Towards this goal, two approaches have been applied to quantify these milk constituents: 1) short-wave near-infrared (SW-NIR) spectrophotometry and 2) NIR-Raman spectrophotometry. Further accuracy was sought for by combining spectrophotometry with conductivity and index of refraction measurements. Multivariate statistical analysis of these measurements was used to develop calibrations for milk composition.

In this chapter, an outline of research objectives will first be covered, followed by a description of milk and its major constituents, fat, lactose, and protein. Standard reference methods for milk composition determination are then discussed. Finally, an

elaboration of tools used in this study, short-wave NIR spectrophotometry, refractive index, conductivity, and NIR FT-Raman measurements, are presented.

1.2 Overview of Milk Composition

All mammals nourish their offspring with milk, which is secreted by mammary glands. This nourishment is provided by the numerous constituents in this complex biological fluid. Composition of milk varies with each species, within each species, from animal to animal, geographically and temporally. Factors that have been found to influence milk composition are lactation stage, nutrition of the animal, seasonal and temperature variation, udder infection, and variations in milking procedure[1]. An overview of milk composition is presented in Table 1.1. The reported ranges were found to be typical for lowland breeds. Water is the largest constituent, comprising upwards of 85% by weight in bovine milk. All other constituents are emulsified, dispersed colloidally, or dissolved in the water of milk.

1.2.1. Lipids in Milk

Approximately 98% (by weight) of the lipids of milk are triglycerides while the rest are diacylglycerols (0.28 – 0.59%), free sterols (0.22 – 0.41%), phospholipids (0.2 – 1.0%), free fatty acids (0.10 – 0.44%), monoacylglycerols (0.016 – 0.038%), hydrocarbons (trace), sterol esters (trace)[1]. Almost all of the lipids in milk are in globule form with diameters ranging from 0.1 to 15 μm with a mean diameter of 3.5 μm in bovine milk[2]. These globules, which are suspended in the aqueous phase of milk and are protected by a surface membrane called the milk fat globule membrane (MFGM).

Table 1.1. Approximate distribution of milk constituents[3].

Constituent	Range in milk (% w/w)
Water	85.3 – 88.7
Lactose	3.8 – 5.3
Fat	2.5 – 5.5
Protein	2.3 – 4.4
casein (type of protein)	1.7 – 3.5
Mineral substances	0.57 – 0.83
Organic acids	0.12 – 0.21

The MFGM contains 60% of the phospholipid content and 85% of the cholesterol content in milk but phospholipids also exist as lipoprotein complexes[4]. Fat in milk contributes to its flavour, aroma, colour and texture. Triacylglycerols are an important source of energy and have been found to produce twice the amount of energy per gram of carbohydrates[4]. Milk fat is also a quick source of energy because of the short chain fatty acids, which are absent in vegetable oils. The short-chained fatty acids can be absorbed through the intestinal wall and do not have to be re-synthesized into glycerides. Another role milk fat plays in human nutrition is that it contains many fat-soluble vitamins such as vitamins A, D, E, and K. Fat content of milk has traditionally been a large factor in monetary value of milk especially when milk was used primarily for butter production. Protein level in milk now influences economic value due to its importance in the manufacturing of dairy products such as cheese. As an example, a differential payment plan has been established in the Netherlands to dairymen depending on fat and protein content in milk[5].

1.2.2. Lactose in Milk

The main carbohydrate in milk is lactose, a distinct and unique product formed by the mammary gland. Lactose concentration in bovine milk ranges from 3.8 to 5.3 g/100g. Small amounts of other carbohydrates are present in the form of monosaccharides, mainly glucose and galactose, and oligosaccharides. Glucose and galactose concentration ranges from 0.00002 to 0.00014 g/100g and 0.00000524 to 0.0000875 g/100g respectively[6]. Oligosaccharide content ranges from 0.0117 to 0.0136g/ 100g[6]. Lactose is a disaccharide that is composed of galactose linked to glucose by a glycosidic bond. One of the carbons in the glucose molecule of lactose is anomeric and unstable. It therefore mutarotates from the α to the β isomer and vice versa until an equilibrium is established between the two isomers as shown in Figure 1.1.

Lactose with soluble salts such as sodium, potassium and chloride ions maintains osmotic pressure in the mammary system. Fluctuations in lactose content are associated with changes in the amount of soluble salts. The presence of lactose in milk is advantageous nutritionally because as a disaccharide, it provides twice the amount of energy provided by a monosaccharide at a given osmotic pressure. Lactose is responsible for the low relative sweetness in milk. Presence of lactose in milk is a concern to many people due to two undesirable conditions: lactose intolerance and galactosaemia. Those who are lactose intolerant are unable to hydrolyze lactose sufficiently in the small intestine and the result is a large influx of water. This leads to symptoms such as nausea, cramps, bloating, gas, and diarrhea. Those who have galactosaemia are unable to metabolize galactose as a result of a hereditary deficiency of the necessary enzymes. The outcome is a buildup of galactitol in the lenses of eyes and subsequently cataracts.

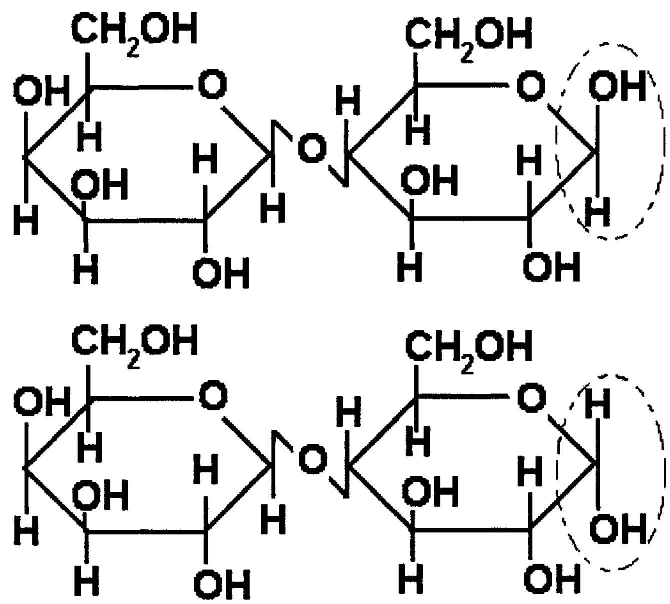

Figure 1.1. Conformational structures of β-lactose (a) and α-lactose (b)

1.2.3. Protein in Milk

The amount of protein in bovine milk changes during lactation and the changes reflect the two-fold function of milk proteins to young mammals. One function is to provide offspring with essential amino acids to develop muscle and other protein-containing tissue. The other function of protein is to supply biologically active proteins such as immunoglobins, vitamin-binding proteins, metal-binding proteins and protein hormones[3]. There are two main classes of proteins in milk, casein and whey protein. Casein refers to the fraction of protein in milk that precipitates out of solution when milk is acidified to pH 4.6 at 30 °C[3]. This accounts for nearly 80% of all protein in milk[3]. The remainder, which is soluble under the same conditions, is referred to as whey protein. Casein is stable at high temperatures and will not coagulate in milk when heated up to 100 °C at its natural pH[3]. Whey protein is more sensitive to heat and will be completely

denature if heated at 90 °C for 10 minutes[3]. Casein is a phosphoprotein, containing about 0.85% phosphorous whereas phosphorous is not a constituent of whey protein[3]. Presence of phosphorous (in the form of phosphate) is responsible for the ability of casein to bind to calcium, which increases its nutritional significance. Whey proteins are more rich in sulfur compared to casein and these sulfur-containing amino acids are partially responsible to some changes in heated milk such as cooked flavour and increased time for rennet coagulation. Rennet, found in the stomachs of cows, is used to coagulate milk for cheese manufacture. Casein is unique to the mammary gland while some of the whey proteins are blood-derived. Casein exists in milk as large colloidal aggregates (micelles) while whey protein is dispersed in solution. Milk protein has high nutritional value compared to other proteins because it contains all of the amino acids required by humans and the distribution pattern of the amino acids in milk resembles what is needed in humans[1]. Both casein and whey protein contain more of the following essential amino acids lysine, threonine, methionine, and isoleucine. Vegetable proteins, especially in cereals, are limited in these amino acids. It is for this reason that milk plays a large role in food interventions in developing countries where many children who have a diet heavily based in cereals, suffer from protein-energy malnutrition.

1.3 Methods for Determining Milk Composition

The International Dairy Federation (IDF), International Organization for Standardization (ISO), and Association of Official Analytical Chemists (AOAC International) cooperate in the establishment of methods for analysis of milk and milk

products. Inter-laboratory studies are conducted to establish and evaluate the performance of these methods. The methods have been designated as either reference or routine. This section covers official reference methods by the IDF, ISO, and AOAC International for the determination of fat, lactose, and protein in milk.

1.3.1. Fat Determination

Reference methods for fat determination in milk include Roese-Gottlieb method, Babcock Method and modified Mojonnier method. The modified Mojonnier method (IDF method 1D, 1996, ISO method 1211, 1999, AOAC International Method 989.05) was used in this study[7]. In the Mojonnier method, the milk sample is warmed to 38.0 +/- 0.1 °C to melt the fat. The homogenized sample is then weighed into an extraction flask. Ethanol is added to the sample followed by ammonium hydroxide to neutralize acid and dissolve casein. Phenolphthalein indicator is also added to sharpen the appearance of the boundary between the organic and aqueous layers. Nonpolar solvents ethyl and petroleum ether are used as the extracting agents. The extraction is repeated at least twice and the combined ether phases, which include the fat constituent of the milk, are evaporated at ≤100 °C. The dried extracted phase is subjected to 70 °C under pressure in a vacuum oven to reach a constant weight. Reproducibility for this technique, defined as standard deviation of inter-laboratory data, was found to be 0.020 g/100g[8].

1.3.2. Lactose Determination

The IDF, ISO, and AOAC International have been unable to reach a consensus upon a reference method for lactose determination. Lactose can be determined using a number of techniques such as gravimetrically (AOAC International method 930.28), enzymatically (AOAC International method 984.15), polarimetrically (AOAC

International method 896.01), and titrimetrically with chloramine T and potassium iodide (IDF method 28A, 1974)[9]. Likewise, lactose can also be determined by HPLC[8]. This was the method used in this study. For HPLC analysis, a milk sample is combined with 0.9 N sulfuric acid to form a precipitate, consisting mainly of protein and lactose. This precipitate is diluted with laboratory grade water and filtered. The filtrate, typically a clear and colourless liquid, is analyzed by HPLC using a mixture of acetonitrile and water as the mobile phase. As standards, α-lactose and β-lactose are employed. Using the HPLC method, precision has been found to be 0.06 g/100g[8].

1.3.3. Protein Determination

The reference method (IDF method 20B, 1993, ISO draft international standard method 8968-5, AOAC International method 991.22) for protein determination in milk is based on the Kjeldahl principle[8]. This method was used in the work presented here. In this method, protein is precipitated from the milk using trichloroacetic acid in a Kjeldahl flask. The nonprotein nitrogen constituents (such as urea) are removed by filtration. Potassium sulfate, a boiling point elevator, sulfuric acid for digestion and copper sulfate, a catalyst, are combined with the filtrate. The resulting mixture is digested in a Kjeldahl flask, which releases nitrogen from the protein and the nitrogen is retained as an ammonium salt. Following digestion, concentrated sodium hydroxide is added to the acid digestion mixture to release ammonia. The ammonia is distilled and collected in a boric acid solution to be titrated with hydrochloric acid. The necessary calculations are:

$$\%\text{Nitrogen} = [1.4*(Vs - Vb)*N]/W \quad [9] \qquad (1.1)$$

where Vs and Vb is the volume of hydrochloric acid used for the sample and blank in ml respectively, N represents the normality of the hydrochloric acid and W is the weight of

the original milk sample in g. The factor 1.4 is used because 1 ml of 0.1 N hydrochloric acid is equivalent to releasing 0.0014 g of nitrogen. The percent content of protein in the original milk sample is then calculated by:

$$\% \text{ protein} = \% \text{ Nitrogen} * 6.38 \quad [8] \qquad (1.2)$$

Here, the factor 6.38 is used because in dairy products, one part nitrogen is equivalent to 6.38 parts protein. This factor reflects the sample matrix. Reproducibility of this technique based on inter-laboratory studies is 0.021 g/100g[10].

1.3.4. Current Practices for Routine Measurement of Fat, Lactose, and Protein

The outlined referenced methods are often both time-consuming and destructive to the sample. All of these methods involve wet chemistry. For dairy herd management rapid and accurate measurements of fat, lactose and protein are necessary.

Commercial instruments such as the Milko-Scan (Foss Electric, Denmark) have been developed to analyze milk by infrared spectrophotometry specifically for semi-routine milk analysis. These are typically filter-based instruments that measure the absorbance at a specific wavelength in the mid-infrared region found to correlate with constituent quantity.

Infrared analysis in filter instruments consists of a single beam infrared system with one cuvette and no mirrors. It is equipped with an infrared light source that passes through filters to only allow light at the desired frequency to pass through a cuvette containing the sample and finally to the detector. Samples are homogenized prior to analysis.

The use of this type of instrument for fat, lactose, and protein analysis in milk has been deemed a standard method (IDF method 141B, 1996, ISO standard method 9622,

1999, AOAC International method 972.16)[9]. The wavelengths employed by the spectrometer are summarized in Table 1.2.

Table 1.2. Wavelengths in mid infrared spectral region used to determine fat, lactose and protein in milk[9].

Constituent	Wavelength (μm)	Functional Group
Fat	3.480	CH groups in fatty acid chains
	5.723	Carbonyl groups in ester linkages of glyceride
Protein	6.465	Secondary amide groups of peptide bonds
Lactose	9.610	Hydroxyl groups

This instrument must be calibrated regularly with standards that have been analyzed by the prior mentioned chemical reference methods. Accuracy of filter instruments is affected by changes in concentration of some interfering compounds not measured by the instrument and by the fluctuation in the composition of measured constituents[11]. For example, the accuracy of fat determination is influenced by the average molecular weight of the fatty acids and proportion of unsaturated fatty acids[11]. Likewise, protein determination is influenced by the proportion of non-protein nitrogen constituents, citrate, free fatty acids and phosphorous[11]. In addition, turn-around time for laboratories equipped with MilkoScan instruments to return milk constituent concentration information can be up to two weeks and in this case, MilkoScan instruments would not aid in daily management of dairy farms[12].

According to the International Dairy Federation, acceptable accuracy in a robust calibration is defined by the magnitude of the standard error of calibration (SEC). This is the standard deviation between reference constituent concentration and constituent concentration determined by the method being evaluated:

$$\mathrm{SEC} = \sqrt{\frac{\sum_{i=1}^{n}(c_i - \hat{c}_i)^2}{n-1}} \tag{1.3}$$

where c is the constituent concentration provided by the reference method, $\hat{c}$ is the constituent concentration provided by the method being tested, and n is the number of samples in the calibration set.

The standards set by the International Dairy Federation is that SEC should be no greater than 0.07 g per 100 g of milk for herd milk samples and 0.10 g per 100 g of milk for individual milk samples. A study by Lefier et al. compared the accuracy of fat, lactose, and protein determination determined by a conventional filter-based infrared milk analyzer, the MilkoScan to that determined by chemical reference methods[11]. Lefier et al. reported that an SEC less than 0.07 g/100g could be achieved using MilkoScan when performing calibrations in 6 trials, where each trial involved the analysis of 11 reconstituted milks made from raw milk constituents[11]. However, when a single calibration was made using all 66 milk samples, collected over six months, SEC results were: 0.130 g/100 g for fat, 0.121 g/100 g for protein, and 0.083 g/100 g for lactose[11]. In the research presented here, calibrations were also conducted using a set of reconstituted samples collected over a year. Because MilkoScan is the current accepted method for routine milk analysis and due to the similarity in calibration methodology, results of this research will be compared to results found by Lefier et al.

1.4 Short-Wave Near-Infrared Spectrophotometry of Milk

Near-infrared spectrophotometry is an attractive option for the dairy industry for a number of reasons. Acquiring near-infrared spectra is fast and non-destructive, requiring no sample pre-treatment. Furthermore, it is a multi-purpose technique because each spectrum contains information about a multitude of milk constituents. It can provide quantitative assessment of milk composition in real-time by monitoring milk spectra during milking and through the use of fibre optics, remote acquisition is possible. These features could allow for daily measurement of milk composition for dairy management and bio-monitoring.

The near-infrared region of the electromagnetic spectrum lies in the wavelength range from 770 nm to 2500 nm. Low energy electronic transitions, overtones, and combinations of hydrogen vibrations in C-H, N-H, and O-H groups can be observed in this region, indicating the presence of functional groups in the sample that can be quantified. Assignment of some near-infrared wavelengths to food constituents is presented in Table 1.3.

The challenge of near-infrared spectrophotometry for milk analysis is that spectra of many individual constituents have broad regions of overlapping bands and therefore, the spectra consist of wide absorption bands that appear difficult to interpret and quantify[14].

Table 1.3. Assignment of infrared absorption bands to food constituents.

Wavelength (nm)	Food Constituent	Bond Vibration
910	Protein	3rd overtone C-H [13]
928	Lipid	3rd overtone C-H [13]
990	Carbohydrate	2nd overtone O-H [13]
1200	Lipid	2nd overtone C-H [14]
1440	Carbohydrate	1st overtone O-H [14]
1730	Lipid	1st overtone C-H [14]
1780	Lipid	1st overtone C-H [14]
1980	Protein	Combination N-H [14]
2080	Carbohydrate	Stretching+deformation O-H [14]
2180	Protein	Combination C=O, N-H [14]
2320	Lipid	Combination C-H [14]
2350	Lipid	2nd overtone C-H [14]

Further complication arises in this region from the water absorption, which is very large compared to the absorption of fat, protein and lactose[15]. Fat globules and protein micelles of milk cause spectral deformations by scattering light, resulting in an increase in absorbance from the increase in optical pathlength. These are major reasons why this spectral region has been widely ignored for milk composition analysis until the advances of computers and chemometrics.

Through application of statistical analysis, near-infrared spectrophotometry has been used in the determination of fat, lactose, and protein in milk and other dairy products. Much of the published work on milk constituent determination using near

infrared spectrophotometry has focused on analysis between 1100 and 2400 nm[12, 15, 16, 17, 18, 19]. Of these, the most accurate results were found using the following techniques. Sato et al. were able to achieve an SEC of 0.0901 g/100g for fat estimation in 50 homogenized milk samples analyzed in transflectance mode in the region of 1100 to 2498 nm using a 0.25 mm pathlength[18]. Transflectance methods involve passing light through sample and reflecting the light from the bottom of the sample holder to a detector. Stepwise multiple linear regression analysis between reference concentrations and near-infrared data was used to choose wavelengths for the calibration. The samples used in this study were from Holstein cows but no information is provided about the number of cows or the length of time over which the samples were collected. Tsenkova et al. were able to obtain SEC of 0.066 g/100g for lactose calibration by obtaining near infrared transmittance spectra in the wavelength range from 1100 to 2400 nm of 84 samples from one cow using a 1 mm path length[16]. Partial least squares regression was used to form the calibration, however, an independent sample set was not used to validate the model. It should also be pointed out that more than one cow is necessary to acquire samples to achieve a robust calibration because milk is such a complex biological fluid. The most accurate protein estimation using near infrared spectrophotometry was found by Laporte et al. [17]. This study analyzed 96 homogenized and unhomogenized milk samples by transmittance spectrometry from 1100 to 2500 nm using a 0.5 mm pathlength[17]. Partial least squares was used to perform the calibration yielding an SEC of 0.04 g/100g for protein calibration.

Very few studies have examined the use of short-wave near infrared (SW-NIR) spectrophotometry (700 nm to 1000 nm) for milk analysis[12, 20, 21]. Increasingly, dairy

farming is moving toward automated milking. This move presents an opportunity for on-line instrumentation for routine milk analysis. With the advent of inexpensive silicon sensors in the 700 to 1100 nm range on the market, coupled with fibre optic probes, on-line instrumentation could be developed if an accurate calibration is possible. Of the mentioned references, the best results for fat, lactose and protein estimation using SW-NIR are as follows. Šašić et al. were able to obtain an SEC of 0.102 g/100g for fat calibration using transmission spectra of 40 homogenized milk samples from 800 to 1100nm collected using a 1.0 mm pathlength[20]. Partial least squares analysis with leave one out cross validation was used to develop the calibration model. Further validation was conducted using an independent set of 60 samples that yielded a standard error of 0.119 g/100g [20]. Limitations of this study are found in sample variation because the samples were acquired over 2 days and in the use of the Milkoscan instrument as a reference method. For a robust calibration, temporal variation is necessary in the sample set and Milkoscan is not a primary reference method. In addition, it was not mentioned how many cows the samples were collected from. Tsenkova et al. have obtained results for lactose and protein estimation using SW-NIR with the lowest error, comparatively[12]. For lactose calibration, SEC was found to be 0.084 g/100g using 4.0mm pathlength transmittance spectra, over 700 to 1100 nm, of 258 unhomogenized milk samples collected from 3 cows over six. months[12]. The calibration was performed using partial least squares with leave-N-out cross validation. For protein calibration, SEC was found to be 0.082 g/100g using the same spectra and regression method[12]. Limitations of this model are the lack of independent validation set, small number of animals samples were collected from, and use of MilkoScan as a reference method.

1.5 Refractive Index of Milk

Protein, lactose, and mineral salts contribute to refractive index of milk[22]. The index of refraction of a liquid is represented by, n, and is defined as the ratio of the velocity of light in a vacuum to the velocity of light in the liquid. Refractive index is a function of wavelength and temperature and in milk, this property is normally in the range of 1.3440 – 1.3485 using the sodium D line at 20°C [3]. A linear relationship has been found between the solids content of milk and its refractive index but in spite of this, the estimation of percent solids in milk using refractometry is challenging since the contribution of each milk constituent differs and is additive[3]. Due to the high degree of opacity of milk, refractive index is difficult to measure but the most satisfactory measurements have been made on an Abbe refractometer where only a thin layer of sample is used[1]. In this research, an Abbe refractometer with a white light sources was used. In an Abbe refractometer, the prisms used are Amici prisms, which are a composite of two different kinds of glass. This produces a large amount of dispersion without angular deviation of light. Therefore, it is possible to use a white light source instead of the usual sodium D line because the prisms compensate for the dispersion of the sample[23]. It has been found that a linear differentiating refractometer can measure refractive index of milk more accurately than the Abbe refractometer[24].

1.6 Electrical Conductivity of Milk

Electrical conductivity of milk has been used to detect udder infection at the subclinical level, including mastitis since elevated levels of sodium, potassium, and

chloride ions cause an increase in conductivity[25]. Conductance is the reciprocal of the resistance measured between opposing faces of 1 cm cube of the liquid of interest. Units for conductance is $\Omega^{-1}cm^{-1}$ but the SI unit for Ω^{-1} is siemens (S) and therefore, conductance is typically expressed in S/cm. Instrumentation has been developed to convert conductance to conductivity, a property that can be used to compare results from different experiments. Conductivity is the conductance multiplied by the cell constant. The cell constant, a function of the physical characteristics of the measuring cell, specifically refers to the distance between the two measuring electrodes divided by the cross sectional area of the electrodes. Recently manufactured conductivity meters automatically multiply the measured conductance by the cell constant unique to the measuring probe so that the output is in Siemens directly.

One study found the mean conductivity in uninfected bovine milk to be 4916 ± 506 μS based on a sample population of 92 cows; however the temperature at which the measurements were made was not listed [26]. Quantitative conductivity can be measured to ± 1 to 2 % when the resistance between electrodes is within 1000 Ω and the temperature is maintained to within ± 0.1°C [27]. Keeping a constant temperature is important in conductivity analysis since, as the temperature of the sample increases, there is a corresponding increase in the dissociation of electrolytes and a decreasing viscosity[1]. Likewise, fat globules reduce the conductivity measured of milk because they occupy volume and impede ionic movement but homogenization has not been found to influence the conductivity[1].

1.7 NIR FT-Raman Spectrophotometry of Milk

NIR FT-Raman techniques demonstrate potential in food analysis because unlike conventional Raman spectrophotometry with excitation in the visible range, fluorescence typical of major food constituents is dramatically reduced[28]. Also, NIR FT-Raman spectrophotometry can be used for remote sampling where laser and scattering wavelengths can be transmitted efficiently through fibre optics. In general, Raman spectrophotometry can be an alternative to infrared spectrophotometry because water produces only a weak signal and thus aqueous samples can be analyzed without major interference of water peaks.

In Raman spectrophotometry, light scattered by molecules at wavelengths different from the incident radiation, is monitored. Raman transition occurs when a photon excites an atom to a virtual state and then quickly relaxes to an eigenstate by releasing a photon. This is different from fluorescence because Raman scattering does not involve transfer of electron population to the intermediate state. The scattered light is dispersed and separated according to wavelength to produce a spectrum that corresponds to orbital energy levels.

Raman spectrophotometry has been used qualitatively to study the effect of various conditions on dairy products. For example, Raman spectroscopy was used to investigate the effect of freezing upon casein in ewe's milk[29]. A recent study looked at the Raman spectra of butter, potassium caseinate and alpha lactose for indication of band positions for fat, protein and carbohydrates and milk[30]. Casein was precipitated from the milk and then analyzed in the 1800 to 1350 cm^{-1} region using an argon ion laser[29].

Fehrmann et al. were able to achieve SEC of 0.32 g/100g for fat estimation using Raman spectrophotometry compared to the chemical reference method [31].

1.8 Overview of Research

The next chapter details the experimental methods implemented in this study for SW-NIR, conductivity, refractive index, and NIR FT-Raman measurements of milk. The statistical approaches used to develop the calibrations are then described. In Chapter 3, SW-NIR spectrophotometric calibrations and subsequent SW-NIR, conductivity and refractive index calibrations are developed using PLS analysis. The final section of Chapter 3 describes the SW-NIR spectral, conductivity and refractive index calibrations constructed using the genetic algorithm approach. Similarly, Chapter 4 examines PLS analysis of NIR FT-Raman spectra and then NIR FT-Raman spectra with conductivity and refractive index for milk composition estimation. As a comparison, the GA method for constructing calibrations using NIR FT-Raman, conductivity and refractive index is then investigated. In the last chapter, conclusions are presented along with recommendations for future work.

Chapter 2 Experimental Methods

2.1 Milk Standards

Milk Standards were obtained from Programme d'analyse des troupeaux laitiers du Quebec (PATLQ, Quebec). These standards are sold commercially for the intended use of infrared spectrometer calibration in order to quantify raw milk composition. Standards are prepared using a physical fractionation of milk to ensure consistency and meet the estimated shelf life needs of industry. Methodology for sample preparation is in accordance with that defined by the International Dairy Federation[32]. In this method, a bulk milk sample is made by pooling at least 60 herd milks. From this bulk milk sample, 5 fractions are obtained. The first fraction is the whole bulk milk sample itself. Bulk milk pasteurized and then stored for 12 hours at 4 °C for natural creaming. Cream is skimmed off to yield a cream fraction and a skim milk fraction. The skim milk fraction is centrifuged to reduce its fat content. The skim milk fraction is then filtered using an ultrafiltration module with a 10000 Daltons separation membrane. Ultrafiltration technology refers to the separation of components based on solvated size and structure using pressure and a semi-permeable membrane without the use of heat. The ultrafiltration step yields the fourth and fifth fractions. The fraction that moves through the membrane is called the "ultrafiltrate" and the fraction that doesn't is the "retentate". Of the original skim milk, the retentate contains approximately 100% fat, 100% protein, 50% lactose, and 50% minerals. To make the milk samples, the skim milk, cream, permeate and retentate fractions are combined. The standards were pasteurized and

partially homogenized (65°C at 4000 psi). All standards contained Brotab, a commercial preservative specifically made for milk used in analytical testing laboratories. Brotab is composed of 30% 2-Bromo-2-Nitropropane-1,3-Diol (Bronopol) and 1.4% Pimaricin. Both of these ingredients are antimycotic preservatives that prevent yeast and mold. The milk standards were stored at 4 °C in the dark and were reported, by the manufacturer, to be stable in composition for a month from the production date. Reference values for fat, protein, and lactose were provided by the supplier citing the following methods: Kjeldahl for protein, HPLC for lactose, and Mojonnier for fat. The experimental design included milk with a range, by weight, of protein content 2.4 to 5.2%, lactose content 4.30% to 4.65 %, and fat content 0% to 6%.

2.2 Near Infrared Spectrophotometry of Milk Standards

The NIR transmittance system used in the analysis of the milk standards was composed of several components. The light source was a current regulated 250 W Quartz Tungsten Halogen lamp (Oriel, Stratford, CT). Non-contact measurements were made via two 3 mm diameter fiber optics for illumination and collection. Through the use of optics, a collimated source was possible by placing the fiber optic cables at the focal length of the lenses. Milk samples were contained in a glass cuvette with 10 mm pathlength (Cole-Parmer, Quebec). The cuvette was stationary in a thermally controlled copper cell holder. This cell was warmed with a water bath circulator (Neslab, NH) so that the temperature of the milk at the time of analysis was 40 °C. The cell was fixed atop a magnetic stirrer with a stir bar in the cuvette to ensure thorough mixing of the

sample contents. The collection fiber was connected to a spectrograph (100S, American Holographic, MA), equipped with a concave grating (model #446.34/L, American Holographic, MA) with 20 nm/ mm dispersion and wavelength range of 360 to 1095 nm. The wavelength range of analysis was adjusted to 510 to 1020 nm using a micrometer connected to the grating. Width of the entrance slit to the detector was set to 100 μm. The spectrograph had a 512 linear diode array detector (C4070, Hamamatsu Corp., NJ) with a pixel size of 2.5 mm x 25 μm. Each diode sampled 0.97 nm. A 16-bit 100 kHz data acquisition board (AT-MIO-16X, National Instruments, TX) sampled the spectra from the diode array. Software in the C language served as an interface from the diode array and data acquisition board. Data was collected and stored in a 486-66 MHz PC. Integration time for each scan was set to 0.083 seconds and 30 scans were averaged for each resulting spectrum. Wavelengths of the spectrophotometer were calibrated using a didymium filter prior to each day of milk analysis. Absorbance spectra of the didymium filter were found to be reproducible and did not drift over the entire year. Absorbance was calculated using the negative log of the ratio of milk to air spectra.

2.3 Conductivity Analysis of Milk Standards

A TDS/conductivity meter (Oakton Instruments, IL) was used to perform all conductivity measurements. The meter was calibrated with a potassium chloride standard solution (Fisher Scientific, ON) prior to milk analysis. Measurements were made in the automatic temperature compensation mode so that the meter calculated conductivity

values referenced to 25 °C. All measurements were made with constant stirring of the milk to ensure thorough mixing of sample contents.

2.4 Refractive Index Analysis of Milk Standards

Refractive Index measurements were made using an Abbe refractometer (Bausch&Lomb, NY) illuminated by a white light source. A temperature-controlled water bath was connected to the refractometer so that the prisms of the refractometer and milk samples were at 20.0 °C.

2.5 Raman Spectroscopy of Milk Standards

The milk standards were analyzed in flint glass test tubes (6 mm ID, 10 mm OD) using a NIR Fourier Transform Raman system (Bruker IFS-88). A 1064nm Nd:YAG laser was used as the excitation source. The air-cooled, diode-pumped laser had a maximum output power of 350 mW. A lens with focal length 150 mm was used to focus the laser beam to 100μm and 180° scattering arrangement was implemented. In the 180° arrangement, the laser beam hits the sample at the same side which emits the scattered radiation. The detector element, consisting of a Ge diode, and pre-amplifier were cooled with liquid nitrogen before analysis. Software provided by Bruker was used to record the intensities and frequencies of the Stokes lines. Spectral acquisition was set from 0 to 3700 cm^{-1} with a resolution 2 cm^{-1}. This frequency range was selected since bands corresponding to fat, lactose, and protein in milk have been identified previously using Fourier transform infrared spectrophotometry[33]. Integration time of one scan was 2

seconds but each spectrum was an average of 200 scans. Therefore spectral acquisition time was approximately 6.7 minutes. Samples were analyzed at ambient temperatures (22 - 25 °C).

2.6 Pre-processing of Data for PLS

Much of the published work in this field presents NIR spectra that have been referenced to a scattering standard such as a ceramic plate. As mentioned, the spectra presented here were referenced to air leading to absorbance spectra quite different in appearance to those published. Air is a relatively weak scattering medium compared to a ceramic plate. Scattering intensity is a function of wavelength and therefore, lower attenuation of is found at shorter wavelengths. Rayleigh scattering is the type of scattering obtained using a ceramic plate. For a comparison to published work, a calibration transfer was made on all of the NIR spectra. This was done by taking the difference between the average spectrum of milk samples and the average spectrum of milk samples from a collaboration with the Department of Environmental Information and Bio-production Engineering, Kobe University, Kobe, Japan[12]. This was then added to each spectrum in the data set. Calibration transfer does not affect the results because the mean spectrum is subtracted from the data set when mean-centering prior to any of the analyses used in this study.

Only the 700 nm to 1019 nm range of NIR spectra was used in this study. Boxcar smoothing was done on the NIR spectral data set over a 3-point window (3 nm) and over a 21 point (42 cm^{-1}) window on the Raman spectral data set to reduce the

random noise. Spectra were mean-centered by calculating the mean spectrum from the spectral data set and then subtracting this average spectrum from each spectrum in the data set.

Conductivity data of the milk samples was autoscaled. This was accomplished by first calculating mean and standard deviation of the conductivity of the milk samples. Mean conductivity of the milk samples was subtracted from the conductivity of each sample. Following this subtraction, conductivity was divided by the standard deviation of the conductivity for the sample set. Refractive index data was also autoscaled using the same procedure.

2.7 Partial Least Squares Analysis

Partial least squares (PLS) regression, a multivariate statistical analysis approach, was used to analyze the correlation of spectral changes with changes in milk composition. The PLS method involves condensing spectral variations to predominant factors. These factors are used to create a calibration that relates spectra to milk constituent concentration. In this study, each milk sample had fat, lactose, and protein concentrations determined by chemical reference methods. The vector of fat, lactose, or protein reference values can be represented by **Y**. The number of rows in Y is equal to the number of samples, m.

The matrix of dependent variables is represented by **S** with m rows and the number of columns equal to the number of wavelengths or wavenumbers, n. For example, if **S** represents the Raman spectra, conductivity and refractive index, **S** would consist of 68 rows by 1921 columns because the spectra were collected over 1919

wavelengths/wavenumbers plus 2 additional wavelengths/wavenumbers for conductivity and refractive index data.

The independent and dependent variables are related to each other by the expression:

$$\mathbf{Y} = \mathbf{S} * \mathbf{P} \qquad (2.1)$$

dimensions (m x 1) (m x n) (n x 1)

where **P** is a column of calibration coefficients, each corresponding to each wavelength/wavenumber of **S**. PLS is used to determine **P**.

A subset of the data is designated as the calibration set, consisting of spectral absorbances **S** and milk constituent concentrations **Y**. By multiplying each side of equation 2.1 by the inverse of **S**, **P** is solved for:

$$\mathbf{P} = \mathbf{S}^{-1} * \mathbf{Y} \qquad (2.2)$$

When the inverse of a matrix is multiplied by that matrix, the result is the identity matrix, **I**, or 1. Calculating the inverse of a matrix is a complex undertaking and PLS can be used to approximate the inverse. This is done by decomposing **S** into the matrices **T** and **L**. The loading matrix, **L**, defines a new spectral coordinate system and **T** is the scores matrix, which defines intensities in the new coordinate system. Therefore **S** is defined in the new coordinate system as:

$$\mathbf{S} = \mathbf{T} * \mathbf{B} + \mathbf{E}_s \qquad (2.3)$$

Dimensions of **L** are *f* by m, where the n spectral wavelengths are now represented by *f* basis vectors. Dimensions of **T** are n by *f*, where *f* represents intensity in the new coordinate system. The residual portion of the data that could not be correlated is

represented as $\mathbf{E_s}$. Thus the prominent features of the spectra have been described using f optimal factors.

Independent variable **Y** must also be described in the coordinate system. This is done by relating **T** to the constituent concentration using vector **V** in the following manner:

$$\mathbf{Y} = \mathbf{T} * \mathbf{V} + \mathbf{E_c} \tag{2.4}$$

where **V**, with dimensions h by 1, is specific to the independent variable and $\mathbf{E_c}$ is the residual data that could not be described in the new coordinate system.

Matrices **T**, **L**, and **V** are found by examining the covariance between the dependent and independent variables. The PLS algorithm consists of a series of iterations where each iteration is designated by h. Each iteration uses a linear least squares analysis between the spectra and concentration.

The first factor $\mathbf{W}_h$ is found by the regression of **S** onto **Y**:

$$\mathbf{W}_h = \mathbf{S}' * \mathbf{Y} / (\mathbf{Y}' * \mathbf{Y}) \tag{2.5a}$$

The first score vector $\mathbf{T}_h$ is found by the regression of **S** on $\mathbf{W}_h$:

$$\mathbf{T}_h = \mathbf{S}' * \mathbf{W}_h / \mathbf{W}_h' * \mathbf{W}_h \tag{2.5b}$$

Similarly $\mathbf{V}_h$, the scalar score vector is found by the regression of $\mathbf{T}_h$ on **Y**:

$$\mathbf{V}_h = \mathbf{T}_h' * \mathbf{Y} / \mathbf{T}_h' * \mathbf{T}_h \tag{2.5c}$$

The loading vector $\mathbf{L}_h$ is found by regression of **S** on $\mathbf{T}_h$:

$$\mathbf{L}_h = \mathbf{S}' * \mathbf{T}_h / \mathbf{T}_h' * \mathbf{T}_h \tag{2.5d}$$

By multiplying $\mathbf{T}_h$ by $\mathbf{L}_h$, the h^{th} order approximation to **S** is obtained. For example, during the first iteration, $h = 1$ the first factor $\mathbf{W}_h$ represents the spectral

contributions of pure components to component concentration. To calculate the rest of the factors, the first approximations (PLS estimates) to **S** and **Y** must be subtracted from **S** and **Y** respectively to calculate the residuals:

$$\mathbf{E_s} = \mathbf{S} - \mathbf{T}_h * \mathbf{L}_h \tag{2.5e}$$

$$\mathbf{E_c} = \mathbf{Y} - \mathbf{V}_h * \mathbf{T}_h \tag{2.5f}$$

Matrices **S** and **Y** are then re-defined as $\mathbf{E_s}$ and $\mathbf{E_c}$ respectively and calculations 2.5a through 2.5f are repeated to calculate the next factor. This is repeated until h factors have been computed.

The vector of calibration coefficients, **P**, using h^* factors is then found by:

$$\mathbf{P}_{fh^*} = \mathbf{W} * (\mathbf{L}^t * \mathbf{W})^{-1} * \mathbf{V}^t \tag{2.6}$$

By substituting $\mathbf{P}_{fh^*}$ in equation 2.1, an estimate of the constituent concentration, **Y**, is obtained. As described, there are h factors. Early factors capture a greater amount of the variance as compared to later factors. In fact, later factors might weaken the calibration model to estimate independent samples as the later factors are modeling interferences and noise. The optimal number of factors, h^*, is the minimum number of statistically significant factors leading to a model that neither under-estimates nor over-estimates.

Predicted Residual Error Sum of Squares (PRESS) was used to determine the optimum number of factors to form the calibration model. The PRESS was calculated between the PLS estimates and the known concentrations for all values of h, where:

$$\text{PRESS} = \Sigma (\mathbf{Y}_i - \mathbf{S}_i * \mathbf{P}_h)^2 \tag{2.7}$$

An F-test at 95% significance on the ratios of PRESS at adjacent values of h was used to determine the minimum number of factors with an associated PRESS that was statistically the same as the absolute minimum PRESS of all factors[34].

The PLS regression analysis was conducted using pre-processed spectra and then pre-processed spectra combined with auto-scaled refractive index and conductivity data with Matlab (The MathWorks Inc., MA, Version 5.3 Release 11) software developed in our laboratory.

The PLS algorithm was first computed using the leave-N-out cross validation method. The calibration model was constructed using 75% of the data. This data set, called the "calibration set", included the samples that were at the extreme concentrations of fat, lactose, and protein. In the leave-N-out method, N spectra are left out of the data set for a set number of iterations. For each iteration, the calibration model is constructed using the samples in the calibration set with the exception of the N samples that are kept aside. The PRESS is then calculated using the N samples and the calibration model. At the end of the iterations, the cumulative PRESS is calculated and the 95% F-test is used to determine the optimal number of factors. The calibration vector is then developed using PLS analysis of the entire calibration set at the optimal number of factors. Using this calibration model, SEC and R^2 are calculated for the calibration data set. The test set, which consisted of the 25% of the data that were set aside and not used at all to compute the calibration, is used to calculate the SEP and R^2. In the leave-N-out cross validation method, PLS was configured to leave 10 samples out for five iterations.

As mentioned, the milk samples were collected in 6 monthly sets. To investigate the influence of variation between milk sample sets upon the calibration, the regressions

were determined using another mode of cross-validation, leave-one-set-out cross validation. In this configuration, one set of samples is excluded from the calibration and the calibration is determined using the remainder of the samples. The calibration model is then used to estimate the set that has been excluded. This is repeated until each set has been left out once, resulting in 6 calibration vectors. Residual errors between estimated and reference values are combined and standard error is computed. In this case, the standard error calculated was called the standard error of cross validation (SECV) because the samples used to form the calibration are also used in the estimation.

2.8 Genetic Algorithm Approach for Analysis of Data

Genetic algorithms (GA) can be described as methods for selecting variables most correlated to a component of interest using Darwinian selection theory to optimize the fit of a regression. In Darwinian natural selection theory, evolutionary change is the result of the production of vast genetic variation in each generation. The few individuals who survive give rise to the next generation due to a well-adapted combination of inheritable qualities. Mutation and recombination introduce variation into the population. Mutation is a random event where a gene is transformed and recombination occurs during a mating event when genes of two parents combine to produce new genes in an offspring. Introduction of variation may lead to an elevated or depressed fitness compared to that of the parents.

Wavelengths (SW-NIR) or wavenumbers (NIR FT-Raman) in the dependent variables were selected using the GA approach to estimate fat, lactose, and protein by

multiple linear regression (MLR). The objective was to obtain robust estimations of fat, lactose, and protein using SW-NIR or Raman measurements and physical properties (conductivity and refractive index) of milk.

Before outlining the specific design of the GA method used in this study, one must note the following. The NIR spectra span 160 wavelengths and the conductivity and refractive index data were added as two additional wavelength/wavenumbers to each spectrum making a total of 162 "wavelengths". The Raman spectra consisted of 1919 wavenumbers and with the physical property measurements, this results in a total of 1921 wavenumbers. For use in the GA method, spectra and corresponding reference concentrations were divided into calibration and validation subsets. Calibration sets consisted of 75% of the original data including the smallest and largest fat, lactose, and protein concentrations. The remaining 25% of the original data were designated as the validation set.

There are many ways to configure GA techniques[35, 36, 37]. The version of GA employed in this study is an alteration of that used by Jang[38]. In this implementation of GA, the objective was to determine the combination of wavelengths leading to the best estimation of milk constituent concentration according to the expression:

$$\mathbf{Y} = c_1 * \mathbf{S}_{\lambda a} + c_2 * \mathbf{S}_{\lambda b} + c_3 * \mathbf{S}_{\lambda c} + c_4 * \mathbf{S}_{\lambda d} + \ldots + c_m * \mathbf{S}_{\lambda n} + \mathrm{b} \qquad (2.8)$$

where **Y** is the milk constituent concentration, $\mathbf{S}_{\lambda n}$ is the absorbance at wavelength n, m is the number of wavelengths in the model, c are the coefficients determined by MLR and b is an offset (intercept) determined by MLR. The following is an outline of the sequence

of steps involved in the GA approach developed. A brief description of each step (with the exception of steps 5 and 6) follows the outline.

1. Define:
    a. Range of potential wavelength/wavenumbers (wavelengths, conductivity, refractive index).
    b. Number of wavelength/wavenumbers to be used in the model of estimation.
    c. Number of generations
    d. Mutation rate
2. Create a random initial population of individuals.
3. Evaluate standard error associated with each individual.
4. Create the next population consisting of the two most fit individuals of the preceding population and new individuals resulting from crossover in the preceding population.
5. Repeat steps 3 and 4 for the number of generations defined in step 1e.
6. Repeat 1 through 5 incrementing the number of wavelength/wavenumbers to be used in the model (defined in 1c) by 1.
7. Choose the optimal number of wavelength/wavenumbers.

1a. Defining the range of potential wavelength/wavenumbers.

The region of analysis of the NIR spectra of milk samples was from 700 to 1018 nm, which consisted of 322 wavelengths. Conductivity and refractive index were

included with the spectra, which resulted in a total of 324 potential wavelength/wavenumbers to choose from. Therefore the range of wavelength/wavenumbers was defined from 1 to 324. Each Raman spectrum was 1919 wavenumbers. Time consumption of GA was decreased by including only every 8^{th} wavenumber in the data set, which reduced the number of wavenumbers to 240. With conductivity and refractive index, the range of wavelength/wavenumbers for Raman analysis by GA was defined from 1 to 240.

1b. Define number of wavelength/wavenumbers to be used in the model of estimation.

The model was initially constructed using 2 wavelength/wavenumbers. This number was incremented in steps of one to a maximum of 15 wavelength/wavenumbers.

1c. Define number of generations.

Through trial and error, it was decided that a maximum of 750 generations was suitable.

1d. Define mutation rate.

The mutation rate was defined as 0.50. This parameter will be discussed further in step 4.

2. Create a random initial population of individuals.

Each individual in the population consisted of the number of wavelength/wavenumbers to be included in the model. The number of each wavelength/wavenumber was binary encoded. For example, the decimal equivalent of 1 0 1 0 0 1 0 1 is 165 but only 240 potential wavelength/wavenumbers were in the Raman spectra. Therefore, the decimal values were scaled to obtain the encoded wavelength using the following formula:

$$\lambda = \{\mathbf{decval} * [(\mathbf{maxr} - \mathbf{minr}) \div (2^{n}-1)\ ]\} + 1 \qquad (2.9)$$

where λ was the output wavelength within the specified range, **decval** was the decimal value, **maxr** was the maximum value of the range, **minr** was the minimum value of the range, and ***n*** was the length of the binary code. Therefore 1 0 1 0 0 1 0 1 corresponded to the 105th wavelength/wavenumber in the wavelength range.

3. Evaluate standard error associated with each individual.

Coefficients $\mathbf{c_1}$ to $\mathbf{c_m}$ of equation 2.8 were calculated for each individual using the spectral measurements in the calibration set at the selected wavelengths/wavenumbers. By comparing the GA estimates to the known concentrations in the calibration set, the SEC was determined. This was repeated using the validation set but in this case, the standard error was referred to as SECV. Fitness was maximized by using the GA technique where both SECV and SEC were to be minimized. Therefore, similar to work by Ding et al., the fitness of the individual was defined as 1/(SEC*SECV) [35].

4. Create the next population.

The two most fit individuals of the preceding population were automatically incorporated into the new population. The rest of the population was formed through crossover and recombination of the other individuals in the previous population. Crossover was conducted by randomly choosing two individuals in the preceding population and also randomly, choosing a bit, deemed the crossover point, in the parent strings. All bits after the crossover point were exchanged between the two individuals, creating two new individuals. This was repeated until the new population was complete. The new population, with the exception of the two previously most-fit individuals, was subject to mutation. This was desirable to avoid getting trapped into local optima.

Mutation rate, in this configuration, represented the percentage of bits in the population that reverted to their other binary form.

7. Choose the optimal number of wavelength/wavenumbers.

Steps 5 and 6 are skipped as they are self-explanatory. The product of SEC and SECV was plotted versus the number of wavelength/wavenumbers in the model. The optimal number of wavelength/wavenumbers selected was determined to be the minimum number that showed no statistical difference between its associated SEC*SECV and the absolute minimum SEC*SECV based on F-tests at the 95% confidence level.

The use of the GA method in selecting wavelengths for near infrared calibrations has been attempted in previous work by others and it was found that the GA approach leads to a reduction in prediction errors compared to PLS[39]. One of the differences between the GA method and PLS is that in PLS, the entire spectrum is used to generate a calibration vector whereas when using the GA technique, only the wavelengths that lead to the best fitness are used. In fact, the GA-MLR approach has been found in some instances to be superior to PLS because it reduces the influence of data not containing critical information on the calibration models[40]. One of the drawbacks of the GA method is that the user is faced with a large number of adjustable parameters that affect the outcome such as fitness function, mutation rate, crossover scheme, number of generations and population size[40]. However, extensive investigation has been done to determine optimum values for these parameters[40].

Chapter 3 Estimation Of Milk Composition Using SW-NIR Spectrophotometry

The potential utility of SW-NIR spectrophotometry for quantitative analysis of major constituents of milk could be an asset in evolving dairy herd management practices. It is a fast, non-destructive, inexpensive technique with the capacity for remote sensing via fibre optics. In this chapter, SW-NIR spectrophotometry of milk to determine composition is examined. First, calibrations were made using PLS analysis of spectra alone and then on the spectra combined with conductivity and refractive index measurements. As an alternative method, the GA approach was used to construct a calibration using SW-NIR spectra, conductivity and refractive index.

3.1 Near Infrared Spectra of Milk

To develop a calibration for milk composition determination using SW-NIR spectrophotometry, milk samples with reference values for fat, lactose, and protein content were obtained. The monthly sample sets were collected from July to December 2001. Each set consisted of either 12 or 16 samples but due to sample mishandling, one sample was rejected of the 80 samples. These 79 samples were analyzed as they were acquired on a monthly basis, using SW-NIR transmittance spectrophotometry. Absorbance spectra of the milk samples are presented in Figure 3.1a. The most striking feature of the spectra is the peak at 970 nm, which corresponds to 2^{nd} overtone O-H stretching of water. Baseline changes occur in NIR spectra of milk due mainly to light scattering by fat globules[20]. However in Figure 3.1a, offsets are observed

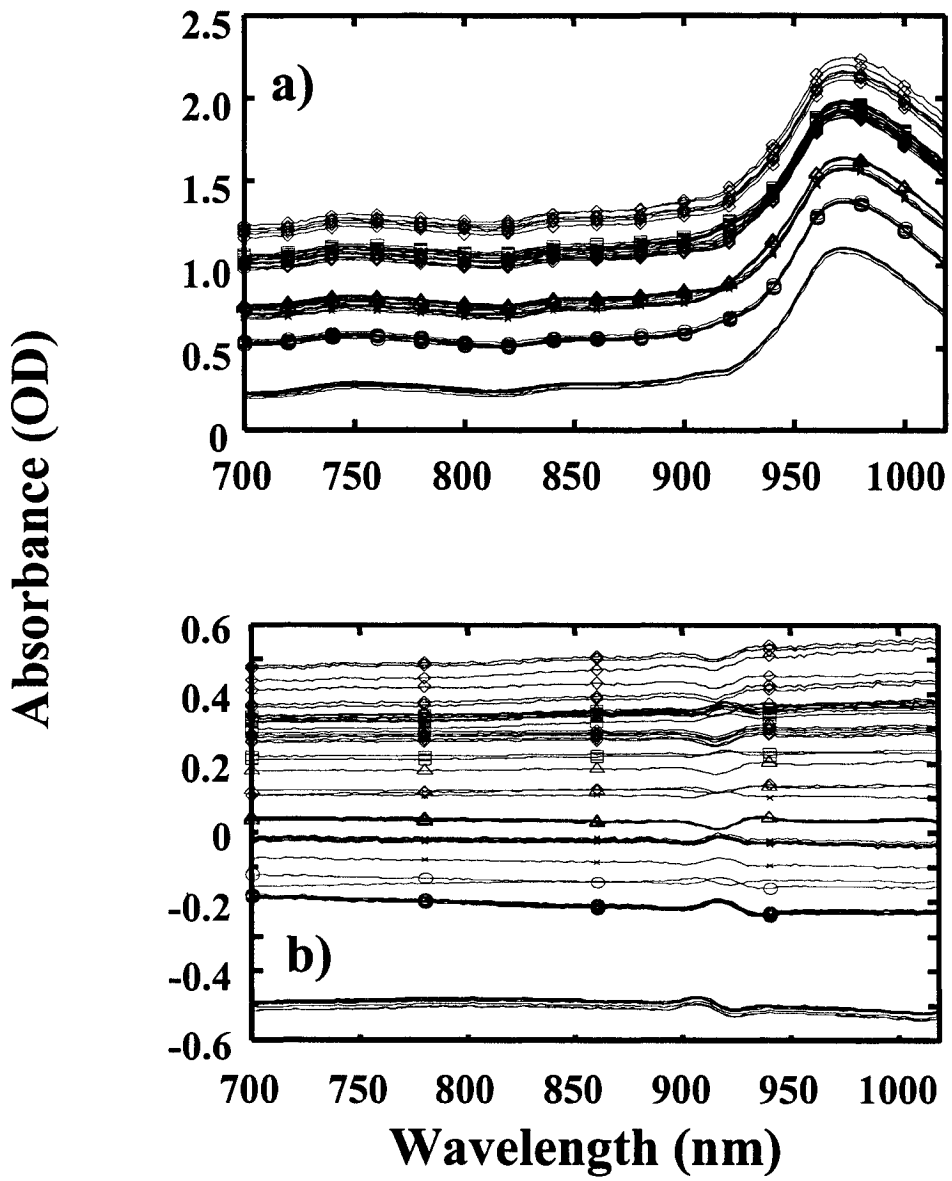

Figure 3.1. Original NIR absorbance (a) spectra of 79 milk samples where each symbol represents a different sample set and smoothed, mean centered spectra for PLS (b)

corresponding to sample sets. Analysis of this effect on calibration is given in Section 3.5. Baseline changes are also visible in smoothed and mean-centered spectra presented in Figure 3.1b. Though the baseline variation is still evident, these spectra are centered about zero absorbance units with subtle variations between 900 and 950 nm emphasized.

3.2 Calibration of Fat, Lactose, and Protein in Milk by Short-Wave Near Infrared Spectrophotometry Using PLS with Leave N Out Cross Validation

To assess the use SW-NIR spectrophotometry for milk constituent determination, calibrations were conducted using PLS analysis. To fulfill this, 75% of the samples were designated as the calibration set and subsequent models for fat, lactose, and protein were developed using PLS analysis of the spectra. In the configuration of PLS used, leave-N-out cross validation was implemented. Of the calibration set, 10 randomly selected samples were left out for 5 iterations. At each iteration, the calibration model was validated using the samples that were left out. The optimum number of factors used to construct the final calibration was found using the cumulative PRESS. An F-test at a 95% confidence level determined that fat, lactose, and protein calibrations required 7, 6, and 7 factors respectively as described in Chapter 2. Calibration vectors, each consisting of PLS regression coefficients from 700 to 1018 nm, are shown in Figure 3.2. Heavily weighted regions in the vectors are those that have large magnitude coefficients, both negative and positive. Regression coefficients give information relevant for the calibration of a constituent but weighting of certain wavelengths may not arise solely from this constituent. Milk constituents largely responsible for spectral variation

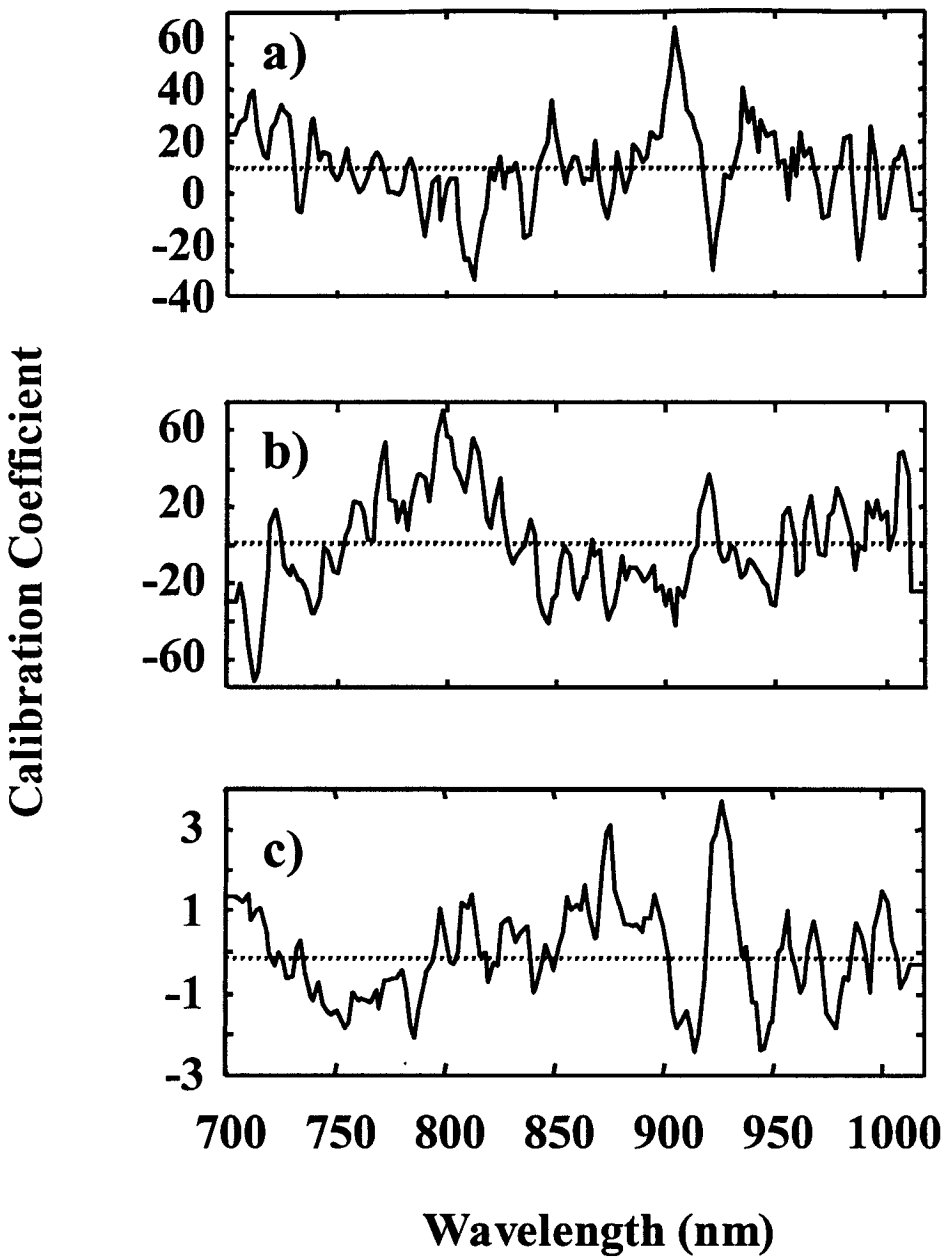

Figure 3.2 Calibration coefficients for fat (a), lactose (b), and protein (c) estimation using SW-NIR spectra using PLS with leave-N-out cross validation

influence the calibration of those constituents that have only minor contributions to spectra. Proposed assignment of heavily weighted regions in the calibration vector to functional groups was done using correlation charts and previously published results.

Triglycerides, the major constituent of milk fat, are composed of esters and C-H groups in fatty acid chains. The calibration vector for fat is shown in Figure 3.2a. The general trend of the calibration vector for fat consists of positive correlation between 750 and 800 nm, positive correlation at 920 nm, and negative correlation centered at 948 nm. All of these regions correspond to C-H stretching[13]. Also, there is negative correlation from 830 nm to 900 nm, which is not a region typical of C-H but instead corresponds to overtones of N-H stretching[13].

The carbohydrate, lactose, was shown in Figure 1.1. It has C-H groups, O-H groups and ether groups. These groups are accounted for in the calibration vector for lactose estimation, presented in Figure 3.2b. General trends in the calibration vector include positive correlation from 800 to 890 nm, intense negative correlation from 906 to 914 nm, and intense positive correlation from 922 to 926 nm, which all agree to C-H stretching[13]. Negative weighting from 750 to 790 nm and from 940 nm to 950 nm can be assigned to O-H stretching[13].

Protein is comprised of amino acids connected by peptide linkages. Functional groups in proteins are C-H, N-H, and C=O groups. The shape of the calibration vector for protein estimation in Figure 3.2c consists of negative weighting from 780 nm to 840nm, corresponding to N-H groups, positive correlation at 904 nm, assigned to C-H stretching, and negative correlation at 922 nm, corresponding to N-H stretching in protein[13].

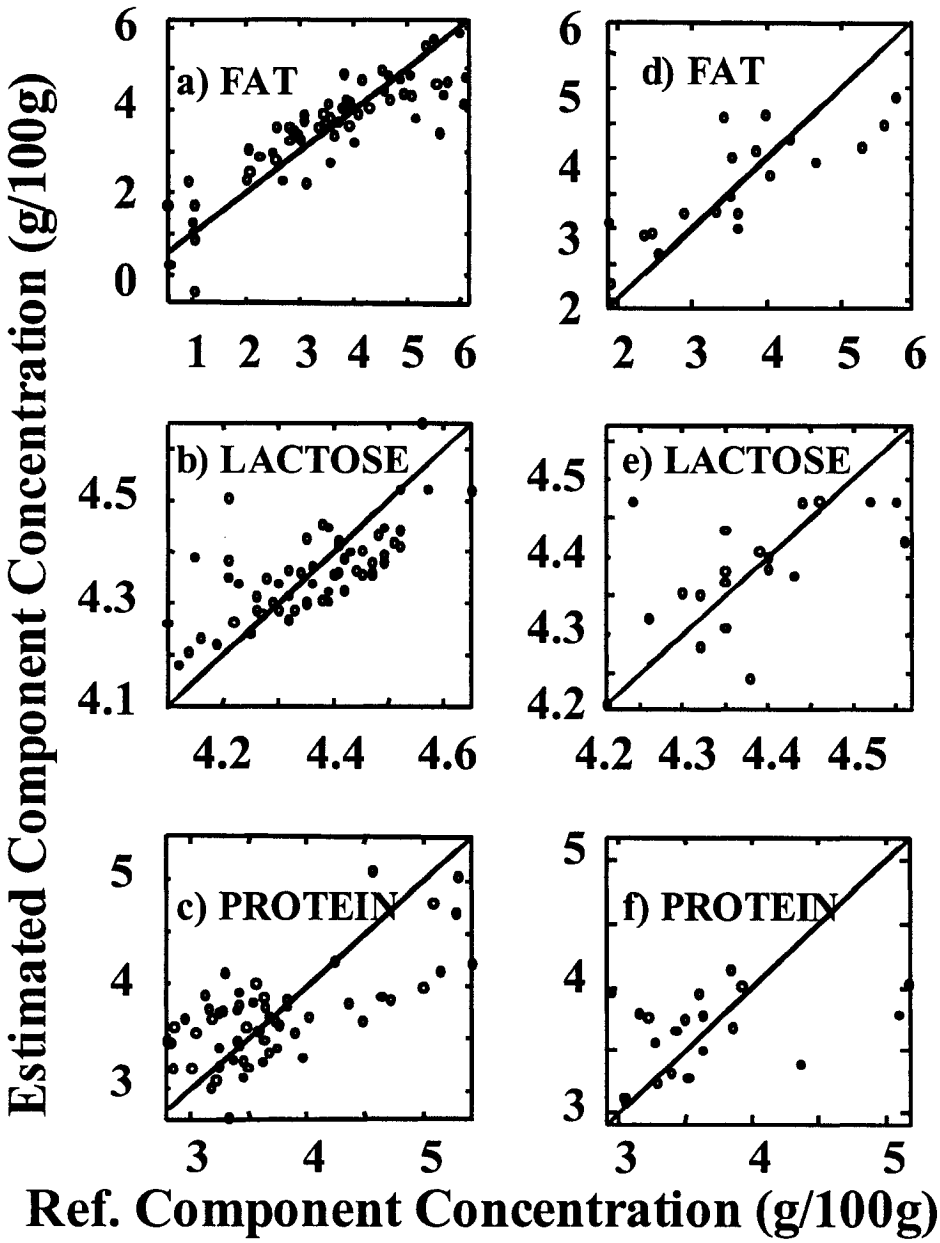

Figure 3.3 Estimation of fat, lactose, and protein in calibration set (a - c) and test set (d - f) using SW-NIR using PLS with leave-N-out cross validation

Utility of calibration vectors in Figure 3.2 was assessed by using them to determine fat, lactose, and protein in the calibration sample set. Estimation of milk constituents from PLS with leave-N-out cross validation is shown in Figure 3.3(a to c). The line of identity represents an ideal estimation. The models were further validated by using them to estimate milk constituent concentration in the independent validation set. These estimations are presented in Figure 3.3 (d - f). Measures of accuracy were indicated by R^2 and standard error (SEC for calibration set and SEP for validation set) between estimates and line of identity, which are listed in Table 3.1.

Table 3.1. Estimation of fat, lactose and protein in milk using NIR spectra using PLS with leave-N-out cross validation

	Calibration Set		**Validation Set**	
Constituent	**R^2**	**SEC (g/ 100 g)**	**R^2**	**SEP (g/ 100 g)**
Fat	0.76	0.74	0.66	0.68
Lactose	0.49	0.09	0.35	0.08
Protein	0.46	0.49	0.14	0.58

These results show that the models do not satisfactorily estimate milk composition. Although the general trends in the calibration vectors were accounted for, the vectors were quite detailed. It is likely that the fine details of the calibration vector indicate the modeling of noise. This may have enhanced the estimation of some samples and hindered that of others. For example, in Figure 3.3a, estimation of milk samples with a fat content greater than 5 g/100g and less than 2 g/100g have reduced accuracy

compared to the standards with moderate fat content. Estimation of the samples in the validation set deviate further from the reference values compared to estimation of calibration set. This implies that the calibrations have described specifically the samples in the calibration set but not an overall picture of milk composition that could be applied to samples outside of the calibration set.

3.3 Conductivity and Refractive Index of Milk

Conductivity and refractive index are physical properties of milk related to composition. Conductivity has been used to indicate mastitis and monitor concentration and composition of solids during dairy processing[1]. Refractive index has also been investigated as a means to determine total solids and added water in milk[1]. In order to improve accuracy in estimation of milk composition by SW-NIR spectrophotometry, conductivity and refractive index were included in the calibration.

Conductivity in the milk samples ranged from 4.05 to 4.85 mS with an average conductivity of 4.39 mS and standard deviation of 0.15 mS. Refractive index measurements of the samples ranged from 1.3454 to 1.3570 with an average of 1.3528 and standard deviation of 0.0027.

To evaluate the use of conductivity or refractive index without spectral input for milk composition determination, linear regression was carried out between each physical property and reference fat, lactose, and protein concentrations of the milk. The capability of the physical properties to estimate milk constituents was judged by R^2 and standard error between the actual constituent concentrations and the regression values. Results of

the linear regression are presented in Table 3.2. From the results in Table 3.2, it is apparent that the two physical properties are not suitable for accurately estimating milk composition. Of all constituents, lactose demonstrates the strongest correlation to the physical properties, particularly conductivity. This is expected because lactose concentration changes have been found to relate to changes in sodium, potassium, and chloride ions in milk as a means of maintaining osmotic pressure in the mammary system[6]. The two properties show some relation to each other inversely, as presented in Figure 3.4. Therefore, it is expected that where one constituent shows a positive relationship with conductivity, refractive index should be negatively related to the same constituent.

Table 3.2. Regression of reference constituent concentration on conductivity and refractive index

	Conductivity		Refractive Index	
Constituent	R^2	Standard error	R^2	Standard error
Fat	0.50	1.50	0.66	1.42
Lactose	0.76	0.01	0.65	0.12
Protein	0.24	0.71	0.24	0.65

In the case of fat, lactose and protein, this trend was observed. Both fat and protein were found to have an inverse relationship to conductivity whereas lactose had the opposite trend. A positive relationship was observed between refractive index and fat and protein content and again, lactose shows the opposite trend. It is not surprising that fat and

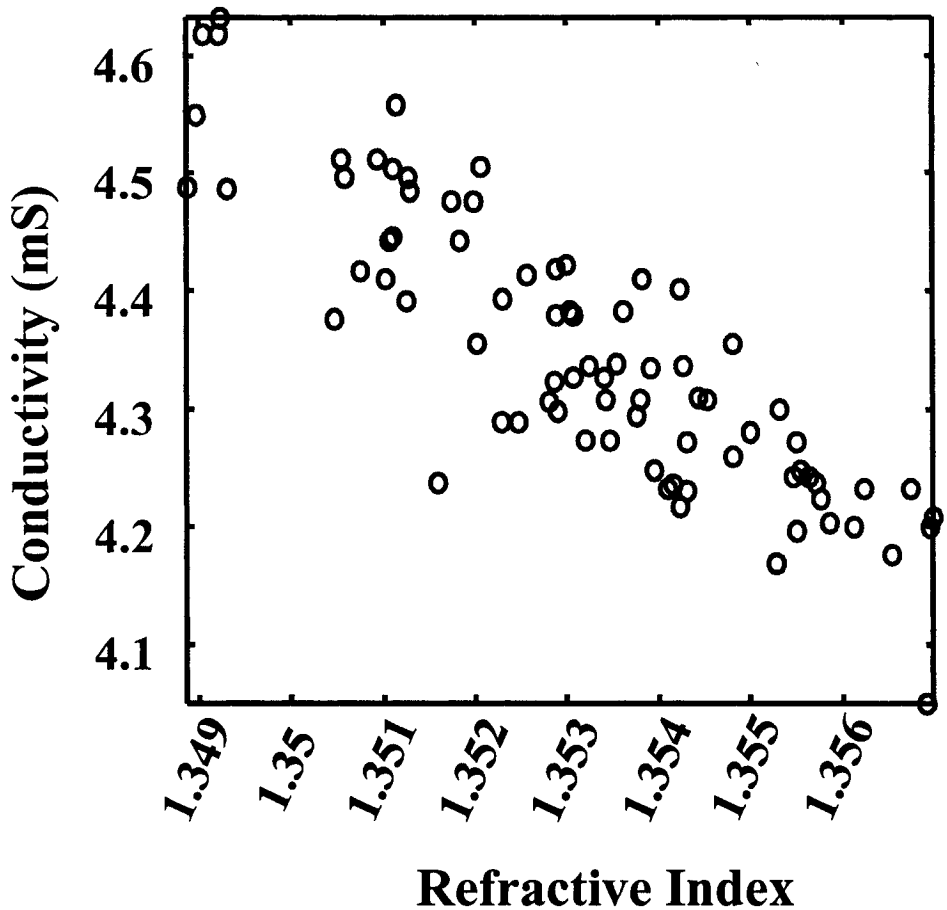

Figure 3.4. Conductivity and refractive index of milk samples

protein would have similar trends in refractive index because both exist colloidally in milk and contribute to the light-scattering properties of milk.

3.4 Calibration of Fat, Lactose, and Protein in Milk using Short-Wave Near Infrared Spectrophotometry With the Addition of Conductivity and Refractive Index Using PLS with Leave N Out Cross Validation

Calibration of fat, lactose and protein was not sufficient using SW-NIR spectra. Although the calibration vectors demonstrated expected trends, marked noise content interfered with the estimation. Conductivity and refractive index have been found to be related to milk composition. By including this added information with spectra for PLS analysis, construction of a more accurate calibration was attempted.

The optimum number of factors used to construct each calibration was found by PRESS using leave-N-out cross validation. At a 95% confidence level, an F-test was used to ascertain that 8, 3, and, 9 factors were significant for fat, lactose, and protein calibrations, respectively. On the entire calibration set, PLS regression was then conducted to develop models using the optimal number of factors. Resulting calibration vectors for fat, lactose and protein are shown in Figure 3.5 (a - c).

The calibration vector for fat estimation using SW-NIR spectra, conductivity, and refractive index is presented in Figure 3.5a. It has been split into two windows, one consisting of the spectral regression coefficients and the other for viewing regression coefficients corresponding to conductivity and refractive index. Spectral coefficients in Figure 3.5a showed the same weighting pattern observed in the calibration vector for fat, constructed using only SW-NIR spectra (Figure 3.2a). This included C-H stretching corresponding to positive correlation between 750 and 800 nm, at 920 nm, and negative

weighting near 948 nm[13]. In addition, negative weighting of the region from 830 nm to 910 nm, assigned to N-H stretching was observed[13]. Conductivity and refractive index were not strongly weighted according to the magnitude of their correlation coefficients. However, inclusion of these two properties influenced the calibration. Although a larger number of factors was used to construct the calibration, spectral regression coefficients were less noisy than those observed in the calibration using only SW-NIR spectra. This indicates that the addition of conductivity and refractive index accentuate variation in spectra and thus weighting of irrelevant information is reduced.

Calibration for lactose, presented in Figure 3.5b, required only 3 factors. Only the two physical properties were primarily used to describe lactose content in milk with minor contributions from the spectra. This theory is supported by the weak magnitudes of the spectral coefficients relative to the dominating conductivity and refractive index. In the spectral portion of the calibration vector, significant correlation is only visible near 900 nm, which was also observed in the spectra-only calibration vector for lactose, and from 970 nm to 1018 nm. Weighting near 900 nm agrees to C-H stretching and the second region corresponds to O-H stretching[41].

Like fat, calibration for protein using SW-NIR spectra, conductivity and refractive index used more factors compared to the calibration constructed without the physical properties. The calibration vector for protein, presented in Figure 3.5c, was noisier than the spectra-only calibration. Neither conductivity nor refractive index were weighted significantly relative to the magnitudes of the spectral calibration coefficients. This indicated that the increased number of factors reflects description of noise in the spectra.

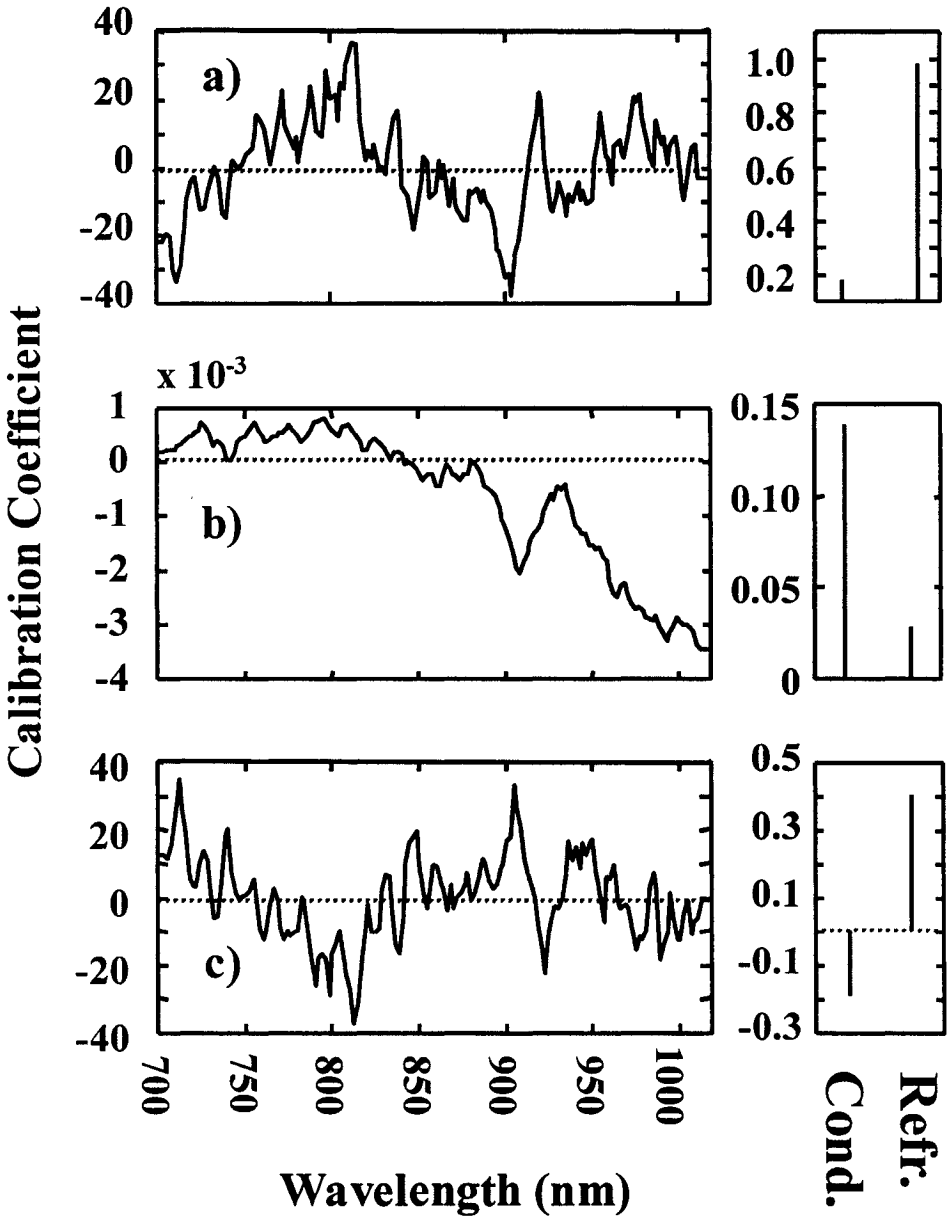

Figure 3.5 Calibration coefficients for fat (a), lactose (b), and protein (c) estimation using SW-NIR spectra with the addition of conductivity, (Cond.) and refractive index (Refr.) using PLS with leave-N-out cross validation

Overall pattern of spectral regression coefficients was equivalent to the calibration without conductivity and refractive index. In general, N-H stretching was apparent in the negative weighting between 780 nm and 840 nm and at 920 nm[41]. Correlation at 904 nm corresponded to C-H stretching[41].

Once the calibration vectors had been constructed, they were used to estimate fat, lactose, and protein content in the calibration set milk samples, shown in Figure 3.6 (a-c). Validation set estimates are presented in Figure 3.6 (d-f). There was a striking improvement in accuracy as compared to the estimation found using spectra alone. Table 3.3 lists R^2 values and standard errors found between the estimated and reference milk constituent concentrations. Including conductivity and refractive index improved estimation of fat, lactose, and protein by 27%, 55%, and 27% respectively, based on decreases in SEC, compared to spectra-only calibrations. Improvements in SEP for fat, lactose, and protein were 29%, 50%, and 52 % respectively.

Table 3.3. Estimation of fat, lactose and protein in milk using SW-NIR spectra with the addition of conductivity and refractive index using PLS with leave-N-out cross validation

	Calibration Set		**Validation Set**	
Constituent	**R^2**	**SEC (g/ 100 g)**	**R^2**	**SEP (g/ 100 g)**
Fat	0.84	0.59	0.83	0.48
Lactose	0.81	0.05	0.82	0.04
Protein	0.71	0.36	0.80	0.28

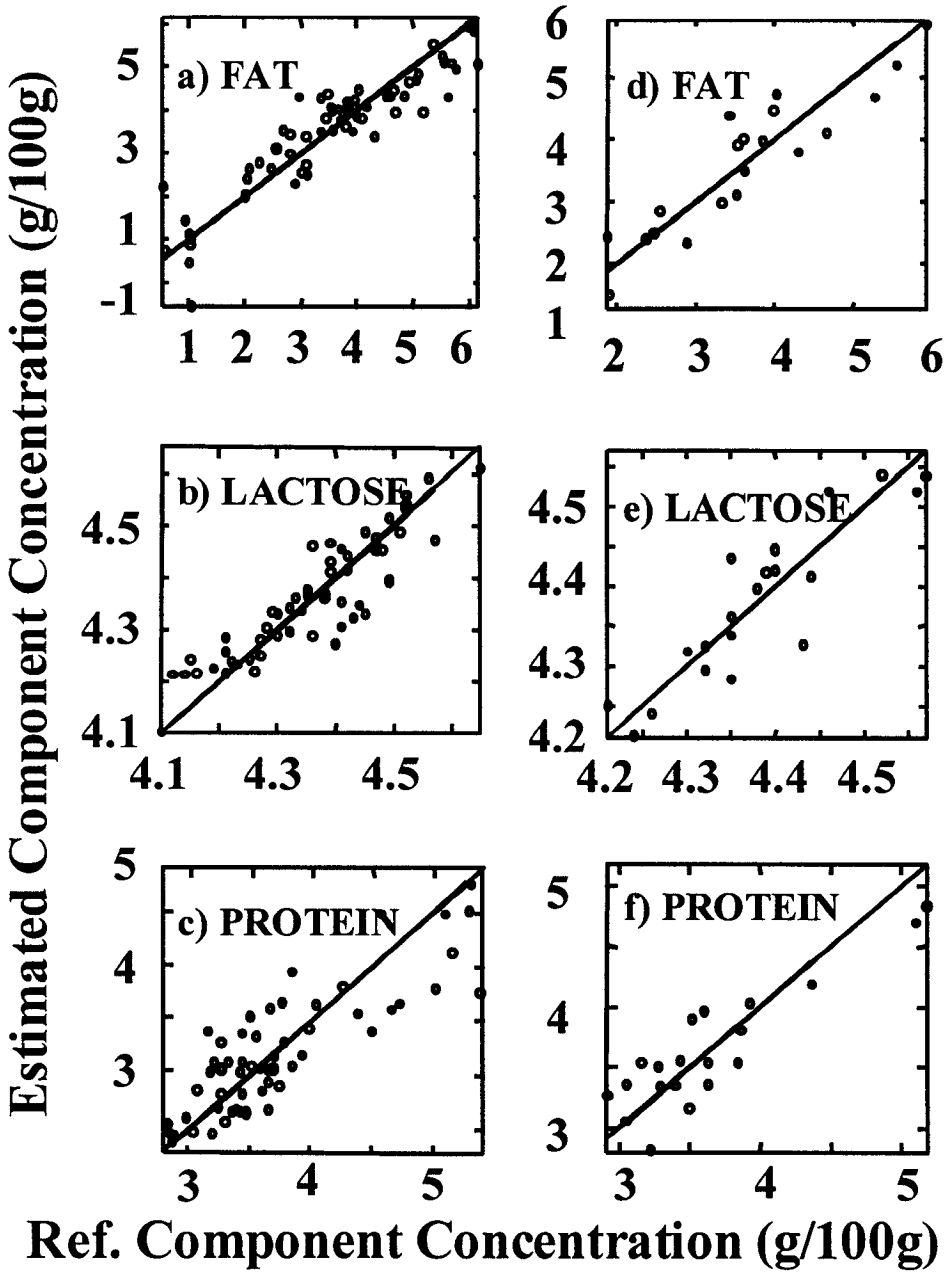

Figure 3.6 Estimation of fat, lactose, and protein in calibration set (a - c) and test set (d - f) using SW-NIR with the addition of conductivity and refractive index using PLS with leave-N-out cross validation

In the case of fat, the calibration required more factors when including the additional parameters and the calibration vector appeared less noisy. Judging by the improvement in accuracy using this calibration vector, it is likely that inclusion of conductivity and refractive index accentuated relevant changes in the spectra. Influence of conductivity and refractive index on the calibration for lactose is somewhat different. In this case, conductivity and refractive index directly aided the calibration and weighting of the majority of the spectral wavelengths was diminished. Nevertheless, spectral contributions, although minor were significant. In Table 3.2, linear regression of conductivity on lactose content only resulted in an R^2 of 0.76 and for refractive index, only 0.65 but as shown in Table 3.3, using the combination of spectra, conductivity and refractive index for lactose estimation exceeds those results. Although protein calibration using spectra with the two physical properties shows an improvement in accuracy over that achieved using only spectra, it is suspected that a significant amount of noise is modeled in the calibration. One extra factor was needed to construct the calibration using spectra, conductivity and refractive index but the general trend of the calibration vector remained unchanged. It was, however, noisier, which implied that the extra factor described noise.

Also of note is the under-estimation of high (> 5 g/ 100 g) protein milk samples as shown in Figure 3.5 c) and f). It was observed that high protein standards coagulated after 1 week when the reported shelf life of the standards at 4 °C was 4 weeks. When the manufacturer was informed of this, it was advised that the standards be analyzed within a week of shipment, because the high protein samples would denature completely after one week. The manufacturer also informed us that high protein standards had been prepared

in different vessels compared to the other standards but they were unable to link the source of the problem to this. Slow denaturing of the protein in these standards reduced the protein concentration. Because of this, reference concentrations for high protein standards were not representative at the time of SW-NIR analysis. The calibration was repeated with the omission of the higher protein samples but the calibration was found to be less accurate. It is possible that the high protein standards were necessary in the calibration due to the information provided even if the concentration was non-linearly represented.

3.5 Investigation of inter-sample set variation using PLS analysis of SW-NIR measurements with leave-one-set-out cross validation

In Figure 3.1, in addition to baseline offsets due to scattering, offsets between sample sets were observed. Influence of variation between sample sets was examined by re-constructing the calibrations using PLS with leave-one-set-out cross validation. In this cross validation mode, the calibration was done 6 times leaving out one sample set each time and estimating the samples in the set that were left out with the calibration models developed using the samples of the remaining sets. In the absence of inter-sample set variation, the calibration vectors should overlap entirely.

For fat, lactose and protein calibration using SW-NIR spectra using PLS using leave-one-set-out cross validation, it was found that 7, 2, and 3 factors respectively were necessary at the 95 % confidence level. The calibration vectors are presented in Figure 3.7. Baseline offsets were apparent in the calibration vectors in Figure 3.7 between each

iteration where a sample set was removed from the data. All six calibrations resulted in similar spectral weighting but there is a magnitude difference between vectors. No obvious trend was observed, meaning, one wavelength may be weighted less in one calibration vector compared to the other calibrations, but another wavelength would be weighted more in the same calibration vector. In the calibration vectors for lactose estimation, not only does the baseline vary but also the general trend of the calibration vectors differs between 900 nm and 950 nm. This is of particular interest because this wavelength region was significantly weighted in the calibration obtained using PLS with leave-N-out cross validation for lactose estimation.

Calibration was repeated with the same type of cross-validation using SW-NIR spectra with the addition of conductivity and refractive index. For fat, lactose and protein, the calibrations were found to require only 1 factor at the 95 % confidence level. Regression coefficient vectors of these calibrations are presented in Figure 3.8. The same variation was apparent in the spectral calibration coefficients. In addition, regression coefficients corresponding to both conductivity and refractive index differ in magnitude between calibrations.

Accuracy in the estimation by leave-one-set-out cross validation approach is much worse than that found using leave-N-out cross validation. Correlation between cumulative estimates and reference values for fat, lactose, and protein content using the leave-one-set-out cross validation approach is listed in Table 3.4.

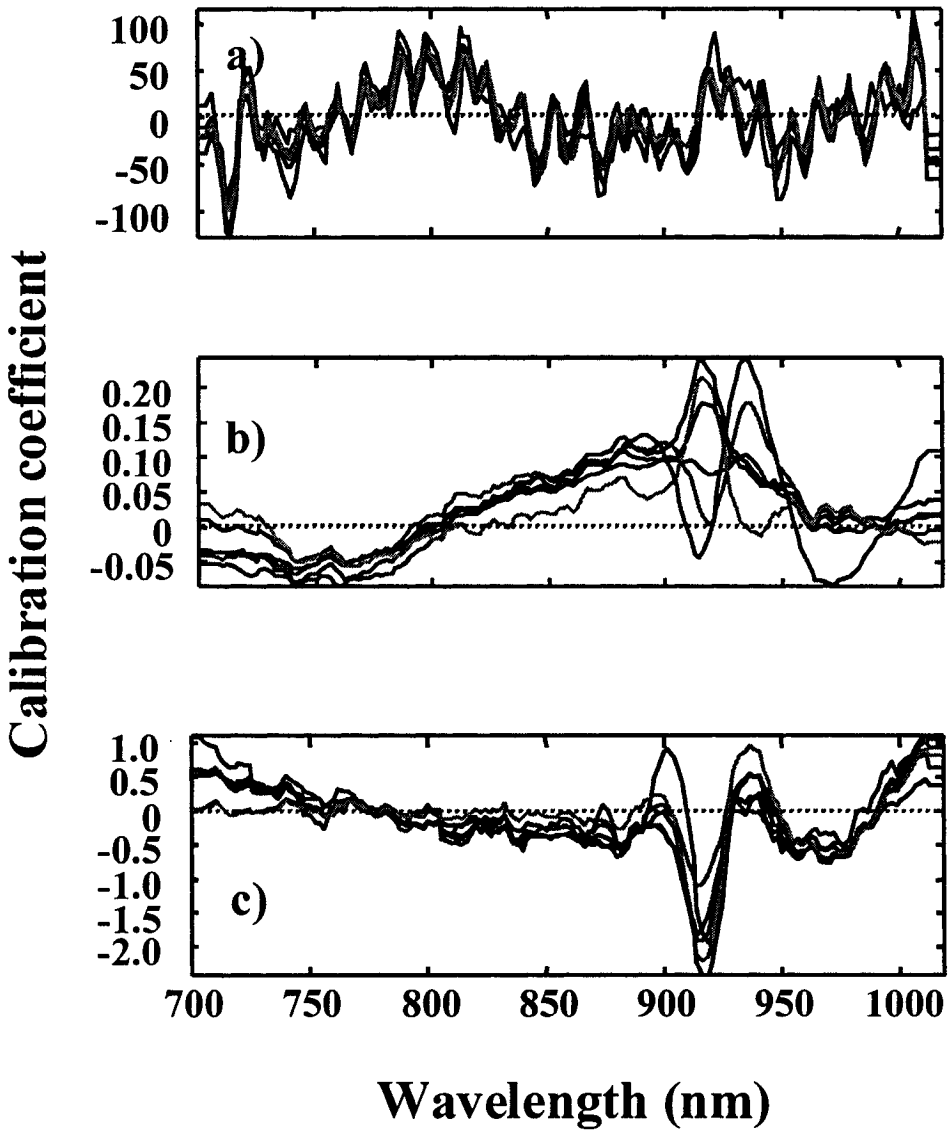

Figure 3.7 Calibration coefficients for fat (a), lactose (b), and protein (c) estimation using on SW-NIR spectra using PLS with leave-one-set-out cross validation

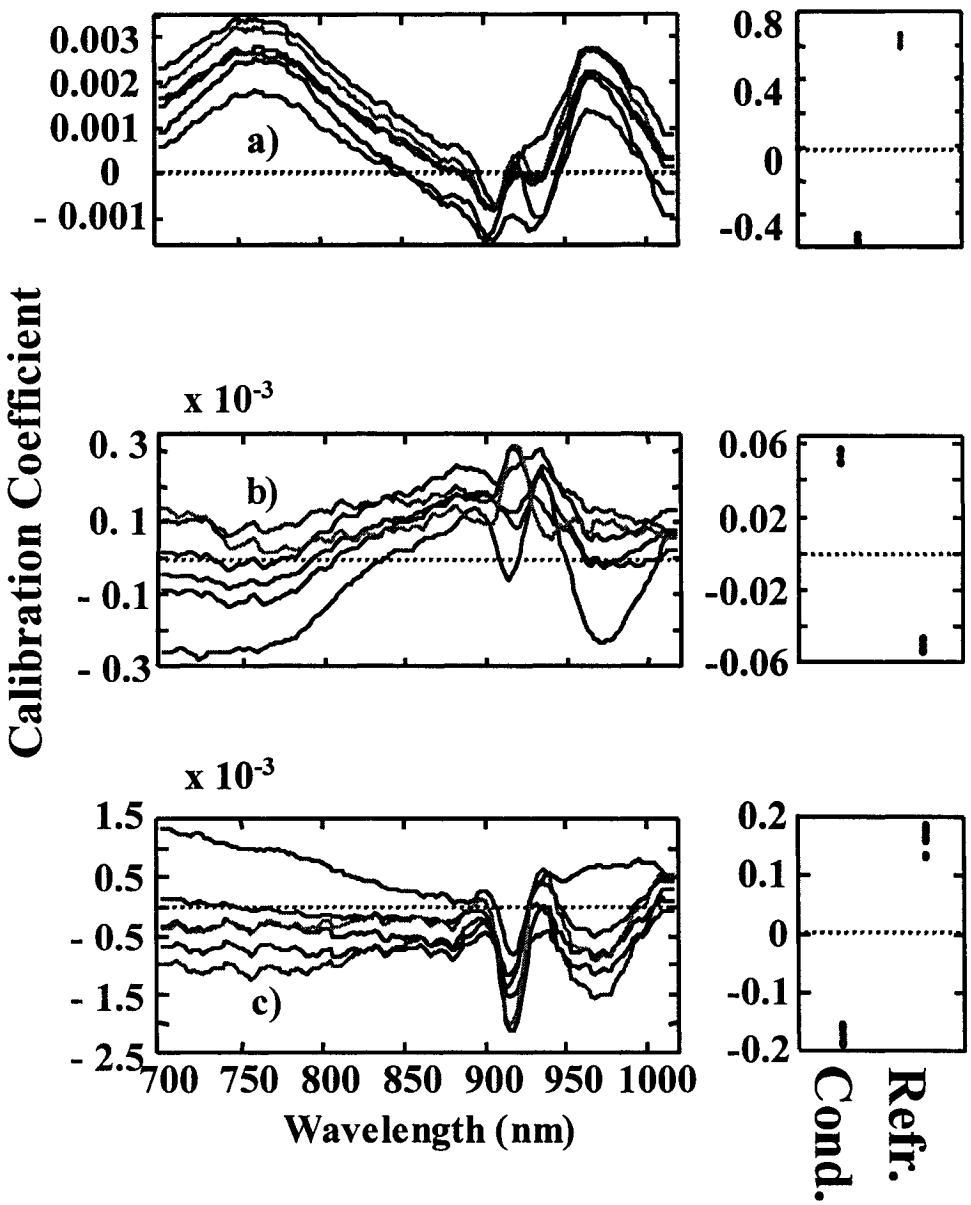

Figure 3.8 Calibration coefficients for fat (a), lactose (b), and protein (c) estimation using on SW-NIR spectra with the addition of conductivity (Cond.) and refractive index (Refr.), using PLS with leave-one-set-out cross validation

Table 3.4. Estimation of fat, lactose and protein using PLS analysis with leave-one-set-out cross validation of SW-NIR spectra and SW-NIR spectra with the inclusion of conductivity and refractive index

	Estimation using NIR spectra		Estimation using NIR spectra with conductivity and refractive index	
Constituent	R^2	SECV (g/ 100 g)	R^2	SECV (g/100 g)
Fat	0.49	0.91	0.60	0.90
Lactose	0.00	0.13	0.76	0.01
Protein	0.00	0.71	0.24	0.70

In comparison with the SEC results found using leave-N-out cross validation, SECV determined using leave-one-set-out cross validation increases by 23%, 44%, and 45% for fat, lactose, and protein estimation, respectively, when using only SW-NIR spectra. Similarly, in the calibrations using SW-NIR, conductivity, and refractive index, SECV increases by 32% and 21% for fat and protein when implementing leave-one-set-out compared to leave-N-out cross validation. However in the case of lactose, SECV decreases by 88% compared to SEC achieved using leave-N-out cross validation using spectra and the two physical properties. The R^2 and SECV calculated using leave-one-set-out cross validation for lactose estimation is the same as that achieved using the linear regression of conductivity on the reference values as listed in Table 3.2. This implies that the variation in spectra between sample sets is so significant that lactose is estimated almost entirely by conductivity. In general, the diminished accuracy in fat, lactose and protein estimation using this type of cross validation demonstrates that variation between

sample sets is detrimental to the calibration. In order to develop a robust and accurate calibration, description of this inter-sample set variation is necessary.

3.6 GA approach for fat, lactose, and protein estimation using SW-NIR, conductivity and refractive index

Milk has a high degree of opacity and efficiently scatters light. This can be problematic in vibrational spectrometry for quantitative analysis because of reduced light intensity at the detector. One way of eliminating noise is by choosing spectral regions that encode information unique to the analyte of interest. The GA method is used for wavelength selection in order to quantify a constituent. This is an alternative approach to PLS, which uses the entire spectrum to estimate constituent concentration.

To compare PLS-constructed models for estimation of milk composition using SW-NIR spectra, conductivity and refractive index measurements, models were constructed using the GA approach. As in PLS, a subset of samples were designated as the calibration set, which contained 75% of the data including extreme concentrations and the remaining 25% defined the validation set. Estimation of concentrations of the calibration set and validation set yielded standard errors SEC and SECV respectively. The quantity to be maximized by the GA technique was the product of SEC and SECV.

The models were constructed using 1 to 15 wavelength(s). In order to determine the optimal number of wavelengths to be used in the model, the product of SEC and SECV was plotted against the number of wavelengths in the model at the maximum

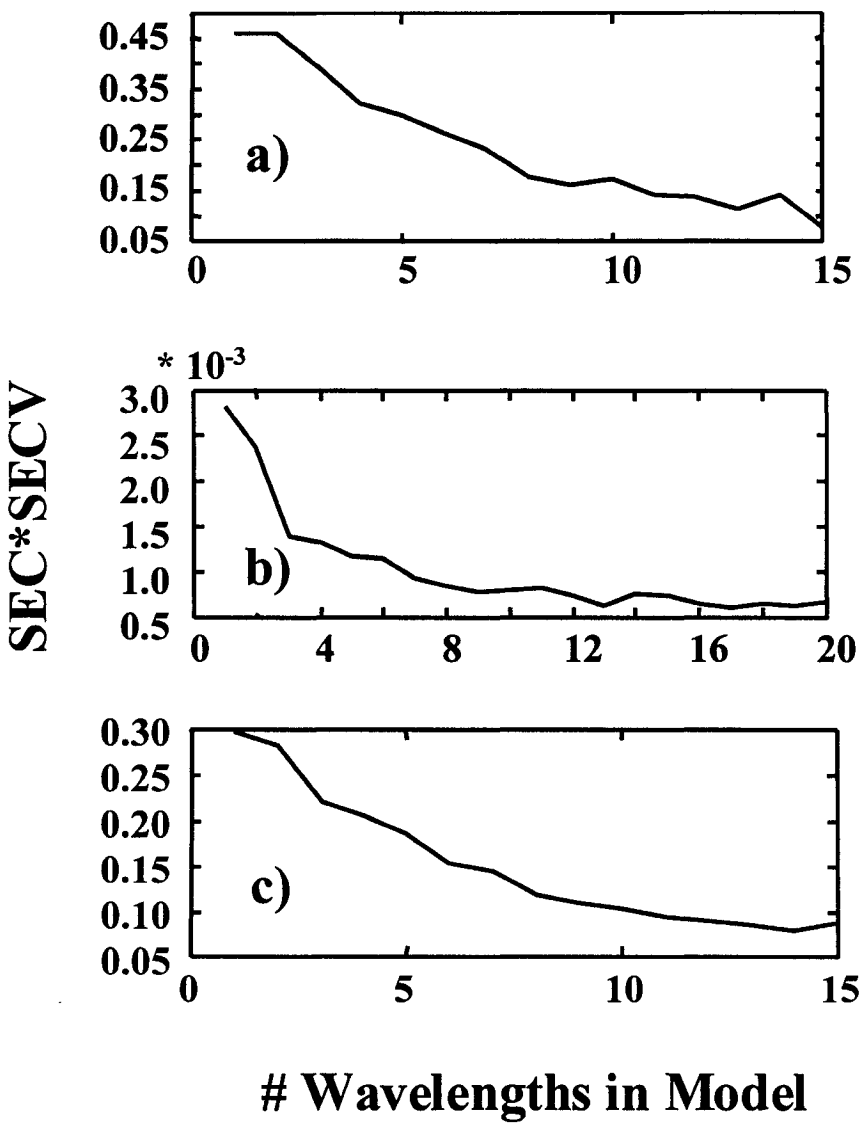

Figure 3.9 Standard Error at 800 generations for 1 to 15 wavelength model for fat (a), lactose (b), and protein (c) estimation using SW-NIR spectra, conductivity and refractive index.

number of generations, shown in Figure 3.9. An F-test was used to determine the optimal number of wavelengths at a 95% confidence level. For fat, lactose, and protein estimation, 9, 8, and 8 wavelengths were used in the models respectively. As described in the Section 2.8, the models were constructed with incrementing generation. A total of 150 generations was selected so that optimal wavelength selection could be reached. To ensure that the solution reaches a minimum, the product of the standard errors were plotted versus generation in Figure 3.10. This allowed observation of the evolution of the wavelength model and the progression of error. From these plots it appears as though the final models were stable.

The wavelengths selected by the GA method were then subjected to multiple linear regression using the calibration set to determine the numerical coefficients. The resulting models for fat, lactose and protein estimation are:

$$\begin{aligned}\text{Fat} = & -893.5 - 224.6*S_{724} - 870.9*S_{778} + 669.3*S_{780} + 915.4*S_{808} \\ & +154.0*S_{810}\, 792.3*S_{848} - 499.5*S_{860} + 649.3*S_{882} \\ & + 684.6*\text{refractive index}\end{aligned} \tag{3.1}$$

$$\begin{aligned}\text{Lactose} = & -0.6 - 33.9*S_{716} + 41.4*S_{760} - 7.1*S_{790} - 3.9*S_{846} - 25.5*S_{910} \\ & + 30.9*S_{924} - 3.3*S_{1010} + 0.8*\text{conductivity}\end{aligned} \tag{3.2}$$

$$\begin{aligned}\text{Protein} = & \ 59.3 + 559.4*S_{726} - 553.1*S_{780} - 260.7*S_{808} + 309.8*S_{848} + 142.2*S_{888} \\ & + 223.7*S_{912} - 409.4*S_{924} + 65.0*\text{refractive index}\end{aligned} \tag{3.3}$$

where S represents SW-NIR absorbance and the numbers in subscript are the wavelengths. Selected wavelengths for fat, lactose, and protein determination were

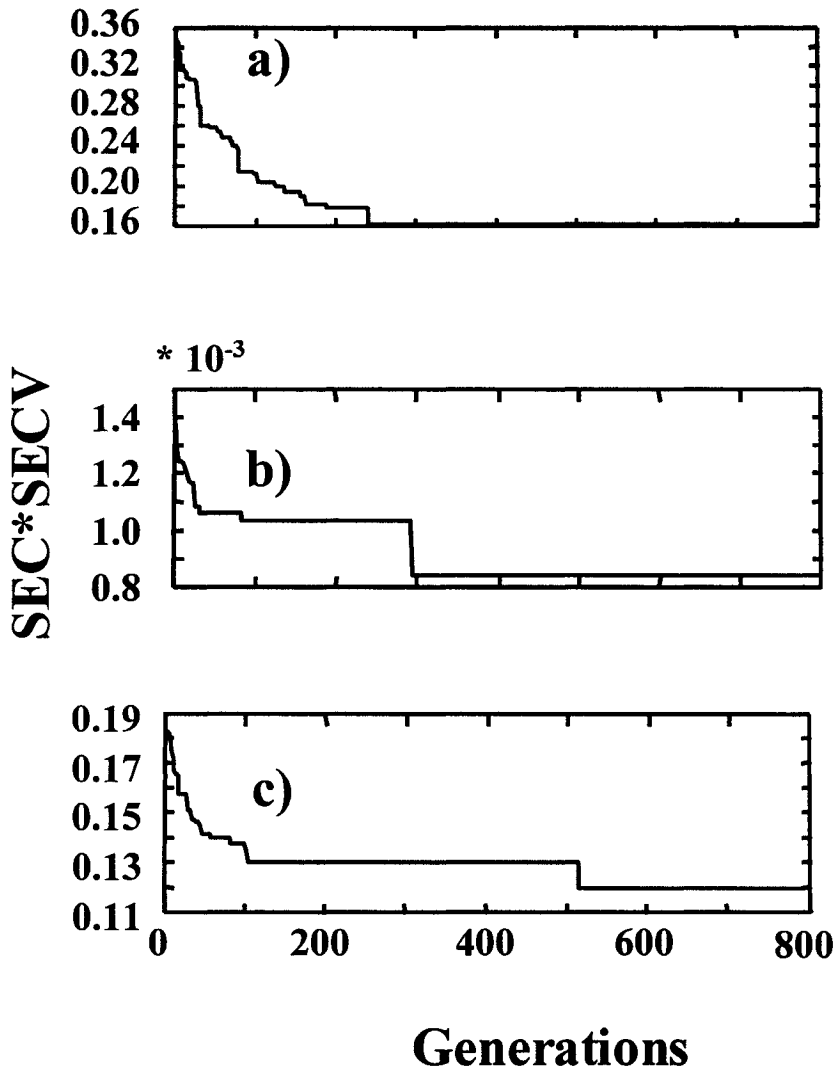

Figure 3.10 Progression of standard error with increasing generations for optimal wavelength model for fat (a), lactose (b), and protein (c) estimation using SW-NIR spectra, conductivity and refractive index

indicated on an average milk spectrum in Figure 3.11. The 2nd y-axis refers to the numerical coefficients corresponding to the wavelengths selected. The O-H stretching from water at 970 nm is largely avoided. Weighting of absorbances at 844, 861, and 879 nm correspond to C-H stretching. Selection of refractive index is due to contribution of fat globules to scattering properties of milk.

The selected wavelengths for lactose show that the weighting is heavier outside of the 800 nm to 900 nm region, contrary to the fat calibration. Absorbance at 716, 846, 910, and 924 nm agrees to C-H stretching and 716 nm, along with 1010 nm are assigned to O-H groups. Similar to the PLS results, conductivity was selected, which is reasonable because lactose concentration is correlated with the soluble salts in milk.

In the model for protein estimation, the selected wavelengths span most of the analyzed region, avoiding the broad water stretch at 970 nm. Two of the wavelengths selected for protein estimation can be assigned to N-H stretching, 780 and 808 nm and 828, 888, and 912 nm are correlated with C-H stretching. Refractive index was selected by the GA approach because like fat, protein exists in milk colloidally and influences the scattering properties of milk.

These models constructed using the GA method were used to estimate milk composition of the calibration sample set. These estimations are presented in Figure 3.12 (a - c). The selected wavelengths and coefficients determined using the calibration set were then used to estimate milk constituent concentrations in the validation set, which is shown in Figure 3.12 (d - f). Accuracy using the models is illustrated by the R^2 and standard error between the estimated and reference values. The results found for R^2,

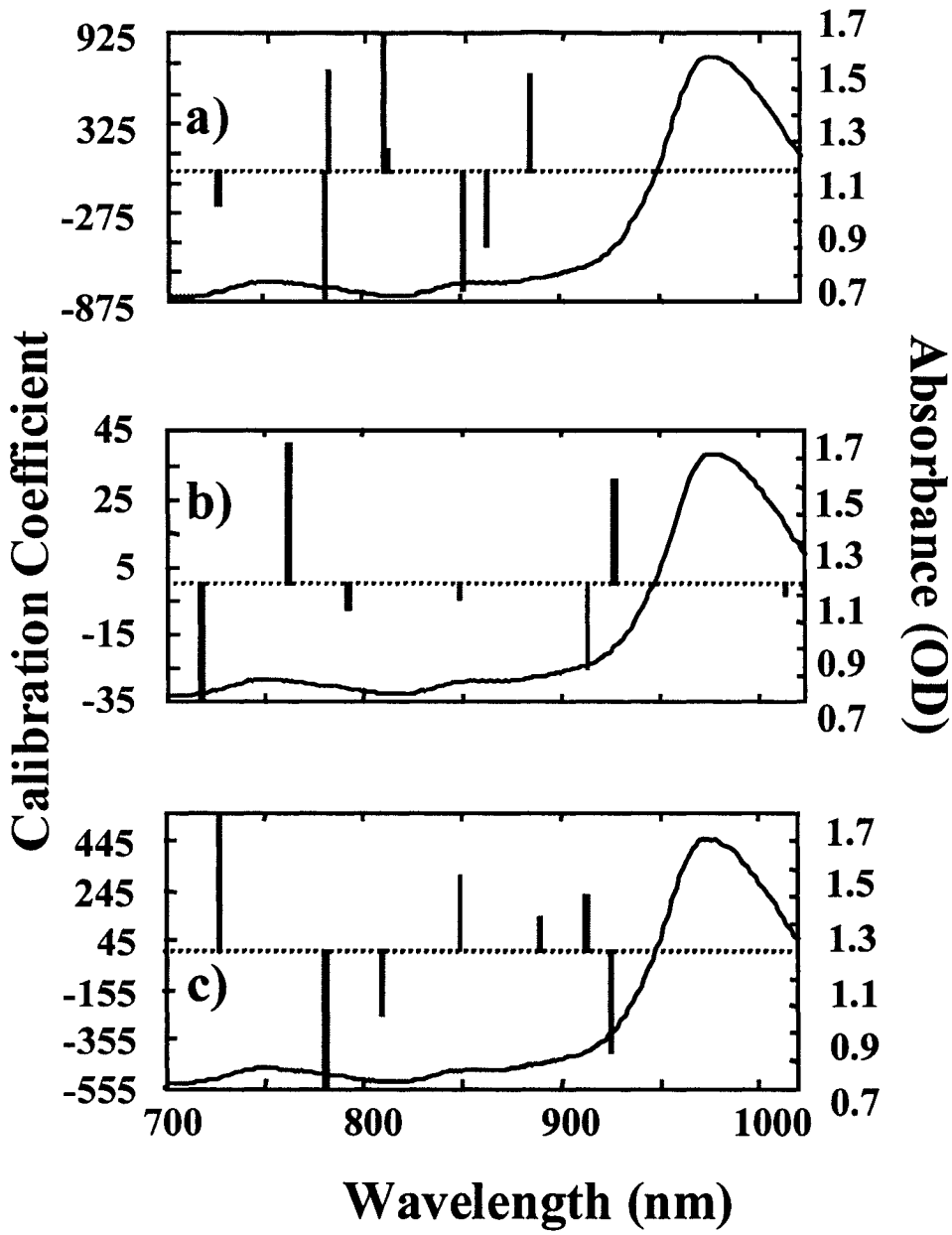

Figure 3.11 Regression coefficients for GA selected wavelengths for fat (a), lactose (b), and protein (c) estimation using SW-NIR spectra, conductivity and refractive index plotted against average SW-NIR spectrum of milk samples

SEC, and SECV using the GA method to estimate fat, lactose, and protein are listed in Table 3.5.

Table 3.5. Estimation of fat, lactose and protein using SW-NIR spectra, conductivity and refractive index with GA.

	Calibration Set		Test Set	
Constituent	**R^2**	**SEC (g/100g)**	**R^2**	**SECV (g/100g)**
Fat	0.83	0.63	0.95	0.26
Lactose	0.88	0.04	0.96	0.02
Protein	0.64	0.40	0.78	0.30

Estimation of fat in the calibration set shows a 7% decrease in SEC compared to the same property found using PLS with leave-N-out cross validation. The test set is even better estimated, showing a 56% reduction when comparing SECV to SEC obtained using PLS. Another notable improvement achieved using wavelength selection is that the model is better able to estimate extreme fat content. Estimation of lactose using the GA model shows an improvement in SEC and SECV by 20% and 60% respectively compared to SEC found using PLS. Not only do these results exceed those by PLS in this study, but they are better than any reported results found using SW-NIR spectrophotometry for the estimation of lactose in milk[12, 20]. For protein, the calculated SEC using PLS with leave-N-out cross validation is better than the result determined using the calibration set and the GA model by 10%. However, SECV obtained for the

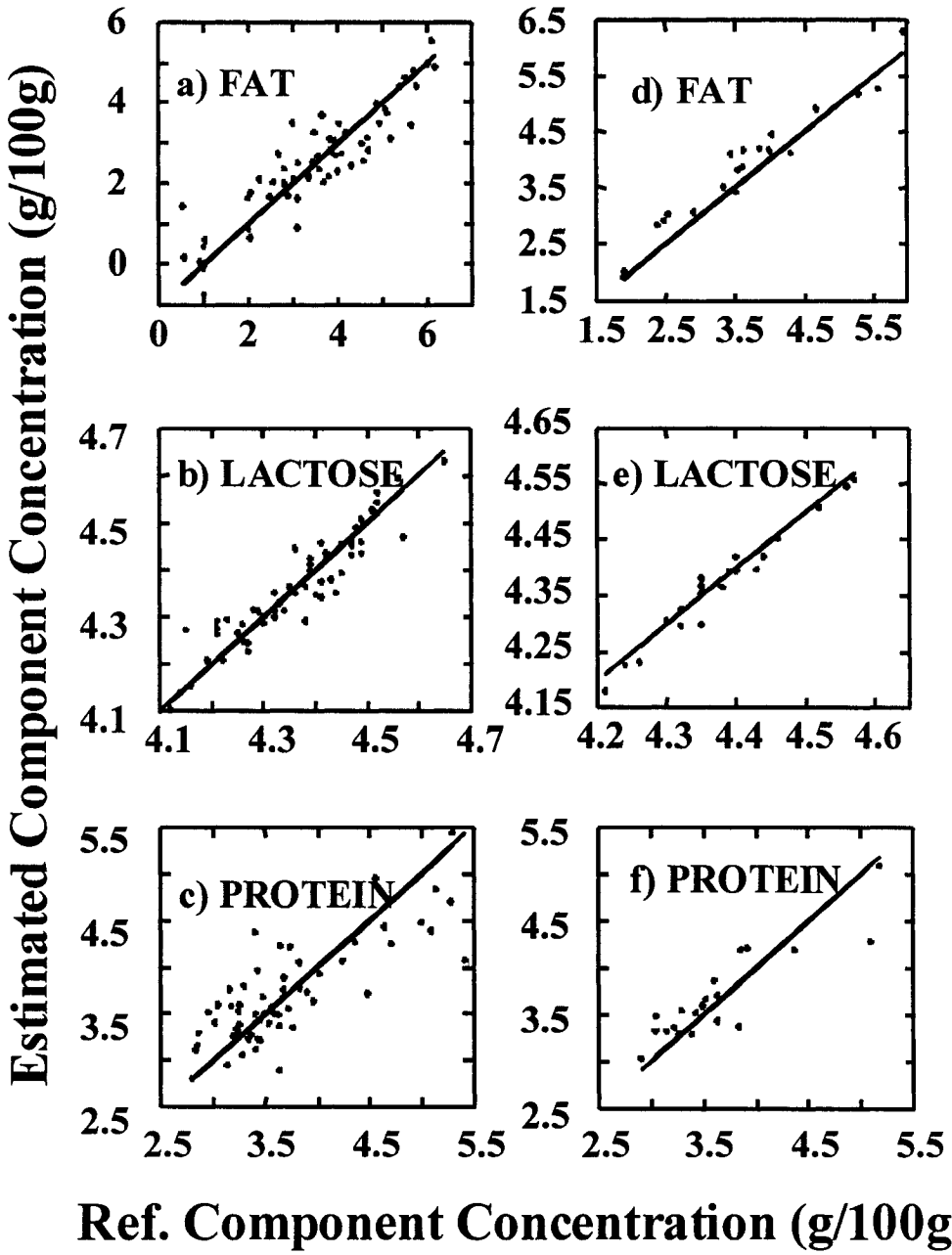

Figure 3.12 Estimation of fat, lactose, and protein in calibration set (a - c) and validation set (d - f) using SW-NIR, conductivity, and refractive index using GA calibration model.

validation set using the GA technique improves on the PLS result by 17%. Results for protein estimation from the GA approach were better than those obtained by Tsenkova et al. using PLS analysis of log(1/T) SW-NIR spectra acquired using a 10 mm pathlength[12]. However, the work by Tsenkova et al. is different from the research presented here because prepared primary milk standards were not used. Instead, 258 milk samples collected from just three cows analyzed by MilkoScan were employed and comparison to primary standards was not attempted[12].

An advantage of the GA method is that only wavelengths that lead to the smallest errors are used in the calibration, unlike in PLS where the entire spectrum is used. Therefore, although results were better for fat and protein using PLS, it should be emphasized that in the GA method no more than 10 wavelengths were used for estimation as opposed to 160. In spite of this, it is suspected that random correlation with some of the wavelengths was selected by the GA approach. A limitation of the present GA configuration is the assumption that the model for estimation took the form of a linear regression when it is unknown whether the estimation of milk composition is due to linear reactions. Finally, a drawback of the GA method was that it was more time consuming than PLS. A single calibration using PLS analysis of NIR spectra only took minutes but the same using GA took 4 hours. Once the calibration is obtained, however, subsequent estimations using the calibration take an equivalent amount of time compared to using the calibration developed using PLS regressions. In both cases, there is no need to recalibrate provided the sample set used to develop the calibration was large enough to capture sufficient variation.

Chapter 4 Estimation Of Milk Composition Using NIR FT-Raman Spectrophotometry

Like SW-NIR spectrophotometry, the use of NIR FT-Raman measurements for the quantification of milk constituents, like fat, lactose, and protein, could be beneficial to the dairy industry. These measurements are rapid without any sample pre-treatment and do not consume the samples themselves. Near Infrared FT-Raman spectrophotometry could be measured simultaneously as the samples are being collected using remote sensing through the application of fibre optic technology. The structure of this chapter follows that of Chapter 3. Fat, lactose, and protein calibrations were developed using PLS analysis of spectra alone and then on spectra with the inclusion of conductivity and refractive index. For comparison to calibration models found using PLS, the GA method was used to construct models using NIR FT-Raman spectra, conductivity and refractive index.

4.1 NIR FT-Raman spectra of milk

To investigate the use of NIR FT-Raman spectrophotometry for determining major milk constituents, prepared primary milk standards were obtained. These 68 samples were collected in 5 sets from August to December 2001. These samples were the same samples that were analyzed using SW-NIR excluding the July sample set. The samples were analyzed using NIR FT-Raman spectrophotometry at ambient temperatures. The spectra consisted of intensity of the scattered radiation versus the

Raman wavenumber, which is the difference between the frequency of the scattered radiation and laser frequency. Raw spectra of the milk samples are presented in Figure 4.1a. Unlike in the NIR data, offsets corresponding to sample sets were not observed in the Raman spectra. The intensity close to 0 cm^{-1} is representative of Rayleigh scattering, an elastic form of scattering where the scattered photons are of the same frequency as the excitation source. Smoothed and mean-centered spectra are presented in Figure 4.1b. Variations in the spectra appear at ~ 170, 350, 1050, 1250, 1450, 1950, 2950, and 3250 cm^{-1}. Bands at 170 cm^{-1} and 3250 cm^{-1} are indicative of the water constituent in milk. Although water has weak Raman scattering properties, it shows several distinct bands at 60, 170, and between 3200 and 3600 cm^{-1} [41]. Other bands are reflective of the organic constituent in milk. 350 cm^{-1}, 1050, 1250, and 2950 cm^{-1} correspond to chain expansion in alkanes, C-C stretching in alkane, CH_2 rocking, and CH stretching in fatty acids respectively [41]. As expected, the Raman spectra of milk appear less complex than NIR spectra because overtone and combination bands are less prominent and thus overlapping is less frequent[41].

4.2 Calibration of Fat, Lactose, and Protein in Milk by NIR FT-Raman Spectrophotometry Using PLS

To develop calibrations using Raman spectra for milk composition analysis, a subset of the milk samples, the calibration set, was used for PLS analysis. The calibration set consisted of 75% of the samples including extremes for fat, lactose, and protein concentrations. The remaining 25% of the samples were designated as the

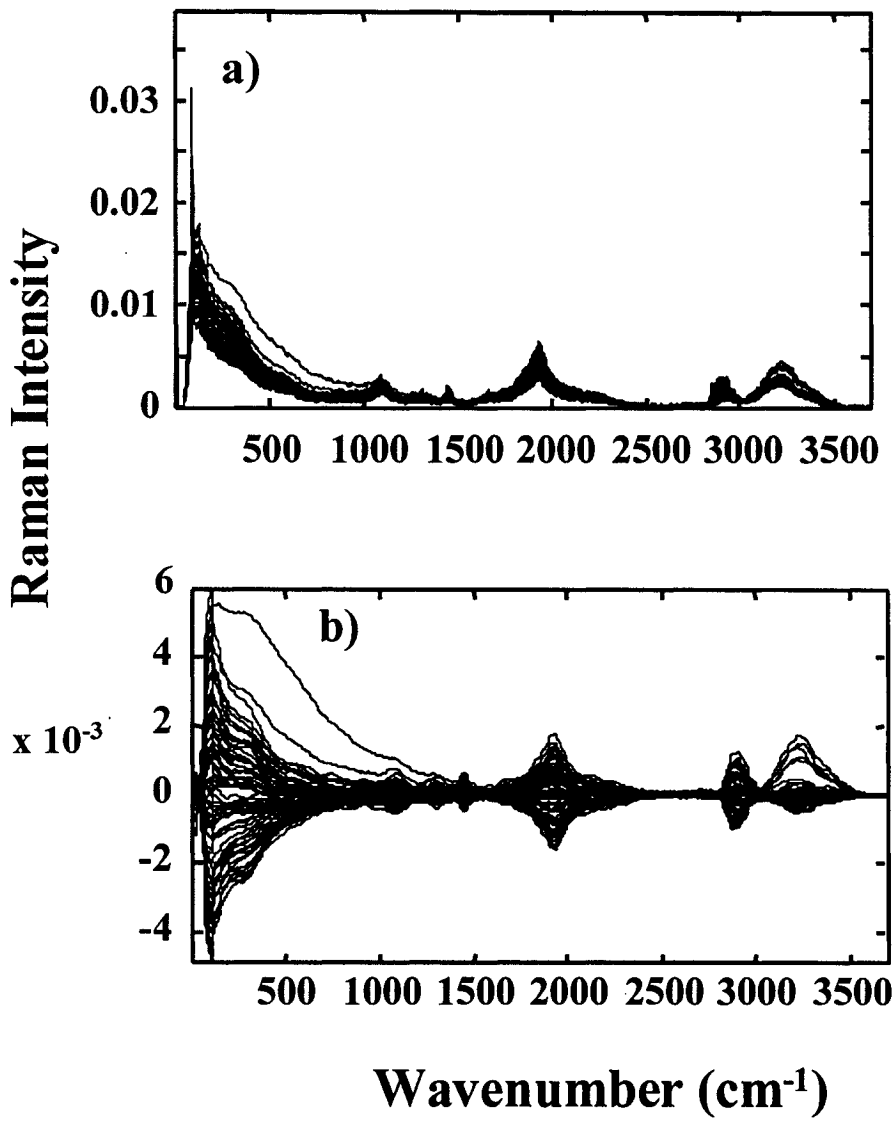

Figure 4.1. Original NIR FT-Raman (a) spectra of 68 primary milk standards where each symbol represents a different sample set and smoothed, mean-centered spectra for PLS (b)

independent validation set. Calibration models were constructed using PLS regression with leave-N-out cross validation. In this configuration, there were 5 iterations. At each iteration, 10 randomly selected samples were excluded from the calibration and the 10 samples that were left out were used to validate the model. For each iteration, there was a PRESS value for each factor. A cumulative (from all 5 iterations) PRESS plot was used to determine the optimum number of factors to be used in the calibration model. An F-test at a 95% confidence level determined that 5 factors were necessary for fat, lactose, and protein calibrations. The calibration vectors, which are comprised of PLS regression coefficients corresponding to spectral wavelengths, for each milk constituent are presented in Figure 4.2. These regression coefficients are relevant to the calibration of each species but may not have arisen from the species alone. Assignment of heavily weighted bands in each calibration vector to functional groups were made using correlation tables and published work.

All fat consists of esterified fatty acids. Fatty acids are composed of hydrocarbon chains with a terminal carboxyl group. The calibration vector for fat estimation is in Figure 4.2a. Low frequency bands 48 and 125 cm^{-1} are due to elastic scattering of fat globules. Another possibility is that these bands indicate water-lipid interactions because stretches associated with hydrogen bonding occurs between 200 and 50 cm^{-1} [41]. The band at 1447 cm^{-1} is due to CH_2 scissoring. This type of vibration has been identified at 445 cm^{-1} to quantify un-saturation of fat[33]. Weighting of 2874 cm^{-1} falls in the expected region of symmetric stretching of acyclic -CH_2- groups[42]. Furthermore, absorbance at this wavenumber is used in commercial infrared milk analyzers for the estimation of milk fat content. Intense weighting of 1650 cm^{-1} was assigned to C=C stretching modes of

lipids. This band was observed in FT-Raman spectra of butter[43]. Peaks at 3171 and 3474 cm^{-1} correspond to O-H stretching and may represent a water-fat interaction. This type of band was also visible in the SW-NIR calibration vector for fat estimation.

Lactose, a sugar, is a ring structure that is composed of C-H, O-H, and C-O-C groups. The calibration vector for lactose estimation is shown in Figure 4.2b. Hydrogen bonding to lactose is expected between 200 and 50 cm^{-1} [41]. The bands at 66, 112, and 350 cm^{-1} can be assigned to this phenomenon and/or elastic scattering. Primary and secondary alcohol C-O vibrations are visible in the weighting of 350 cm^{-1}, as this is expected between 330 and 390 cm^{-1} [42]. Weighting of 2860 cm^{-1} is an expression of CH_2 stretching, although it is unclear what specifically the vibration is ascribed to because asymmetric and symmetric stretching of CH_2 in $-CH_2O-$ and CH_2OH occur in this frequency region[42]. Correlation at 2974 cm^{-1} probably corresponds to C-H stretching because it has been found that this type of stretching occurs at 2982 cm^{-1} in another sugar, sucrose[33]. The band at 3294 cm^{-1} occurs in the region where O-H stretching in carbohydrates has been tabulated[42].

Amino acids are the building blocks of proteins and consist of structures with an amino group (NH_3^+) at one end and a carboxyl group (COO^-) at the other end. Peptide bonds (-CO-NH-) fuse the amino group and carboxyl group of different amino acids to form proteins. Spectral regression coefficients for the calibration of protein are shown in Figure 4.2c. The peak at 1001 corresponds to CNC symmetrical stretching in protein[42]. Weighting of 1265 cm^{-1} was assigned to amide III modes of protein, which consists of CN stretching, NH bending, C=O stretching, and O=C-N bending[42]. In addition, 1641cm^{-1} corresponds to amide I modes of protein, consisting of C=O stretching, C-N

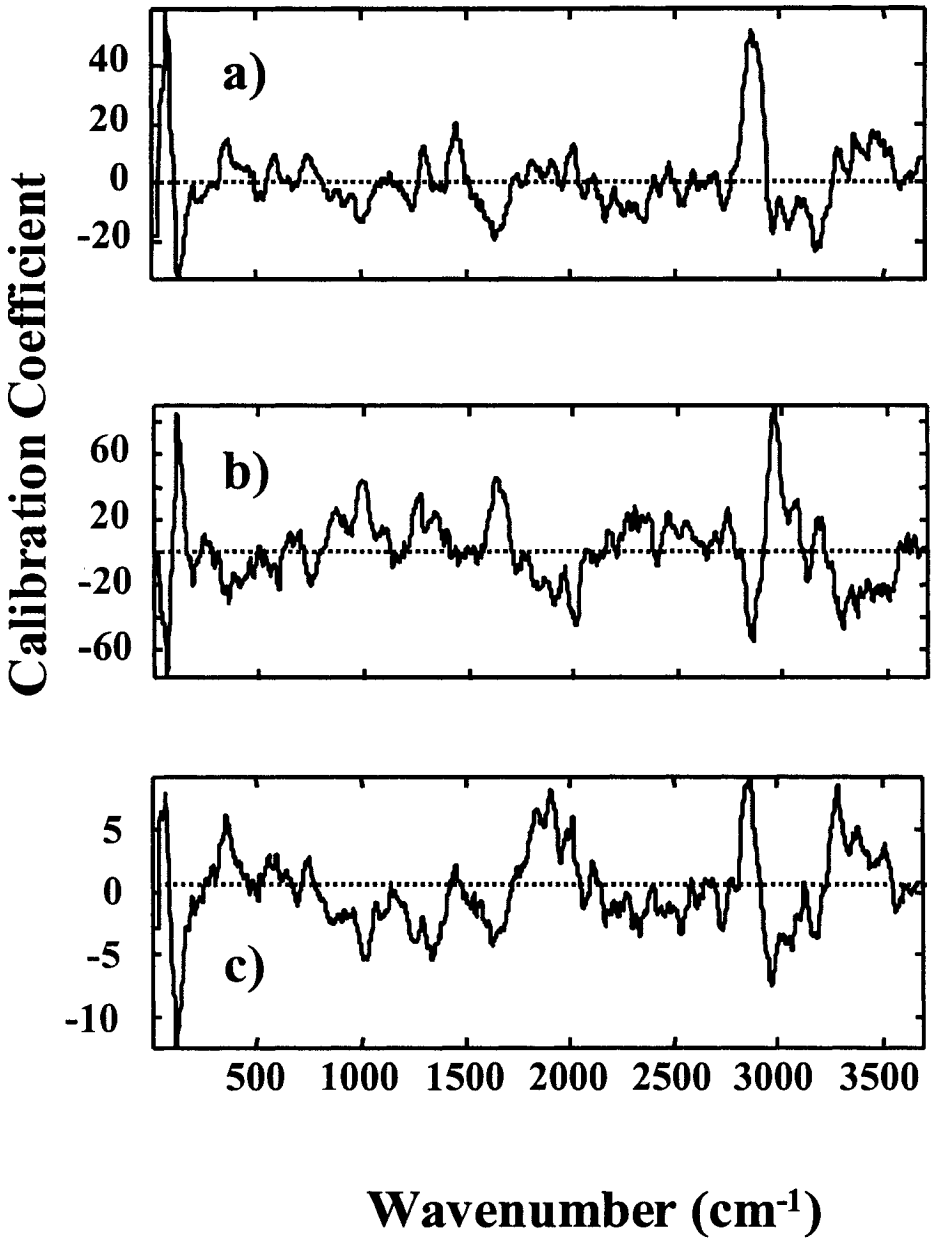

Figure 4.2 Calibration coefficients for fat (a), lactose (b), and protein (c) estimation using NIR FT-Raman spectra using PLS with leave-N-out cross validation

stretching, and NH bending vibrations[42]. Bands at 2864 and 2966 cm^{-1} were assigned to C-H stretching in proteins[33]. Weighting of 3163 cm^{-1} agrees to N-H stretching associated with intramolecular hydrogen bonding [42].

The calibration vectors obtained using PLS with leave-N-out cross validation were used to estimate fat, lactose, and protein in the standards of the calibration set. This was done by multiplying the spectral regression coefficients in the calibration vector by the NIR FT-Raman spectra of the samples. For validation of the models, estimation of the independent test set was also conducted. Figure 4.3 (a - c) illustrates the estimation of the calibration set and (d - f) the validation set where the line of identity represents an ideal estimation. Accuracy achieved using the models were determined by R^2 and the standard error (SEC for calibration set and SEP for validation set) between the estimated concentrations and concentrations determined by reference methods. These quantities are listed in Table 4.1. Results found using Raman spectra show a vast improvement over results obtained using SW-NIR spectra with and without conductivity and refractive index using PLS with leave-N-out cross validation. The results are also better than R^2 and standard errors found using the GA model of the combined SW-NIR, conductivity and refractive index with the exception of lactose. Results for the calibration set were the same between these two calibrations but there was a discrepancy between the validation set data. Overall, these results demonstrate that NIR FT-Raman spectrophotometry is a more accurate method for determining fat, lactose, and protein in milk.

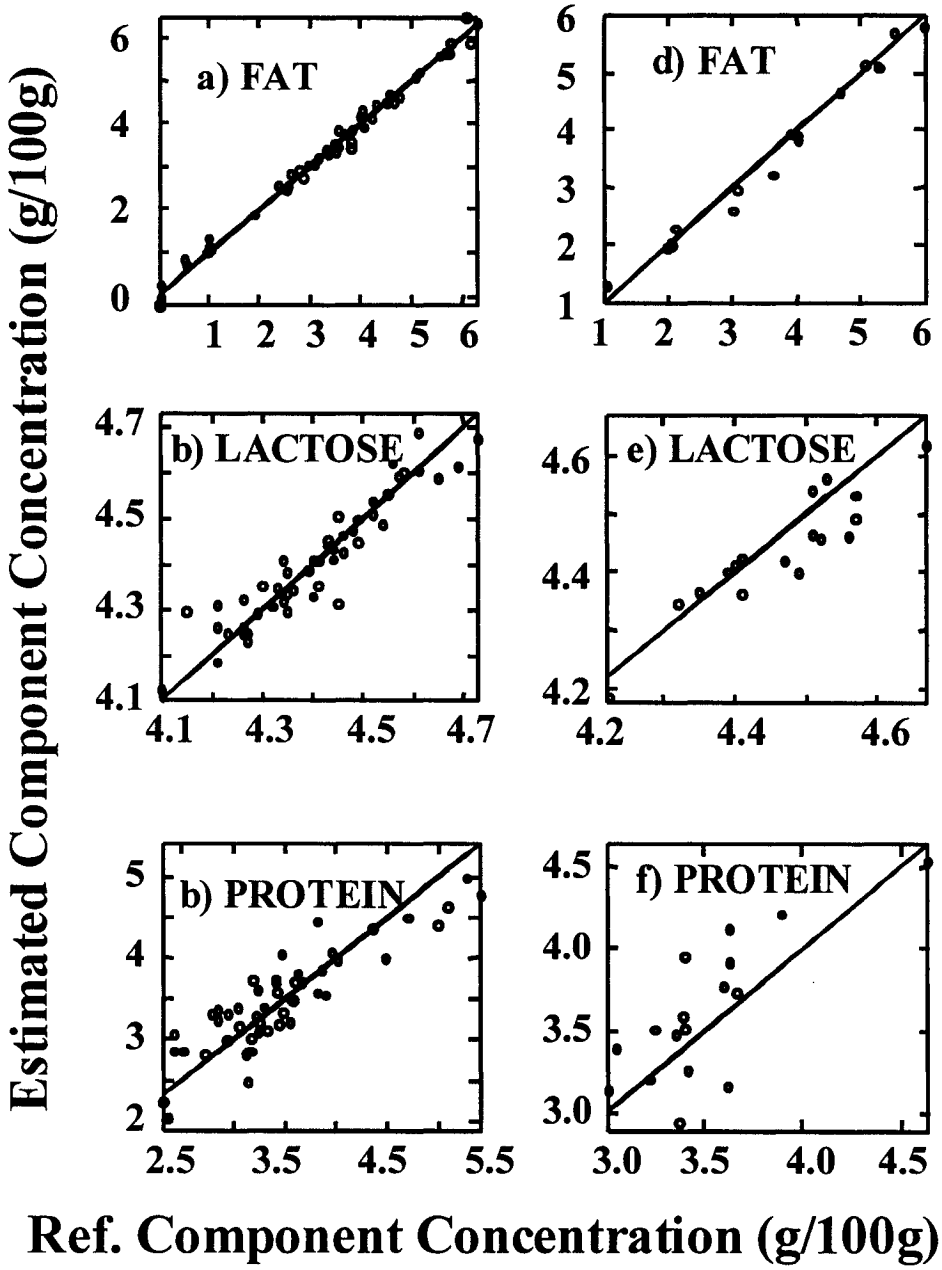

Figure 4.3 Estimation of fat, lactose, and protein in calibration set (a - c) and test set (d - f) using NIR FT-Raman spectra using PLS with leave-N-out cross validation

Table 4.1. Estimation of fat, lactose and protein in milk using NIR FT-Raman spectra using PLS with leave-N-out cross validation

	Calibration Set		Validation Set	
Constituent	**R^2**	**SEC (g/ 100 g)**	**R^2**	**SEP (g/ 100 g)**
Fat	0.99	0.17	0.98	0.16
Lactose	0.88	0.05	0.84	0.04
Protein	0.77	0.35	0.58	0.29

4.3 Calibration of Fat, Lactose, and Protein in Milk by NIR FT-Raman Spectrophotometry With the Addition of Conductivity and Refractive Index Using PLS

Estimation of milk constituents using Raman spectra were found to have better accuracy than obtained using NIR spectra. In the NIR calibration, accuracy was improved through the inclusion of conductivity and refractive index. A similar improvement was sought by adding the two physical properties to the Raman spectra.

Using PLS analysis with leave-N-out cross validation, the PRESS was calculated. An F-test significance comparison was made on the PRESS values to determine the minimum number of statistically significant factors at a 95% confidence interval. Using this method, 6, 5, and 6 factors were used to construct calibrations for fat, lactose, and protein respectively. The full calibration sample set was then used to compute the calibration vectors, shown in Figure 4.4, using PLS.

In Figure 4.4a, the calibration vector for fat is heavily weighted at 66, 1447, 2874, and 3171 cm^{-1}. All of these bands were present in the calibration vector obtained using PLS analysis of Raman spectra without the physical properties, as shown in Figure. 4.2a. Assignment of these bands was discussed in Section 4.2. In addition to these bands, the band at 1304 cm^{-1} was intensely correlated when conductivity and refractive index were included in the calibration. This peak corresponds to CH_2 twisting modes and is a prominent feature in Raman spectra of butter[43]. Neither conductivity nor refractive index was heavily weighted but their presence was found to influence the calibration. In general, when the number of factors used in a calibration is increased, the resulting calibration vector is noisier due to the modeling of extraneous information. However, in this case, an increased number of factors used in the calibration showed a decrease in noise compared to the calibration vector in Figure 4.2a. This implies that conductivity and refractive index were used to emphasize correlation of specific Raman wavenumbers and diminish the weighting of irrelevant spectral information

Like fat, the calibration vector for lactose estimation, shown in Figure 4.4b, contained features that were present in the Raman spectra-only derived calibration vector. These peaks were at 66, 112, 350, and 2860 cm^{-1}. Assignment of these Raman wavenumbers was discussed in Section 4.2. Relatively intense weighting was also observed at 1451 and 3204 cm^{-1}. The band at 1451 cm^{-1} is present in the infrared absorbance spectrum of lactose and has been ascribed to CH_2 deformation vibration of the primary alcohol $-CH_2$-OH [42]. Strong negative correlation at 3204 cm^{-1}, corresponds to O-H stretching in water, implying hydrogen bonding with lactose. Conductivity and

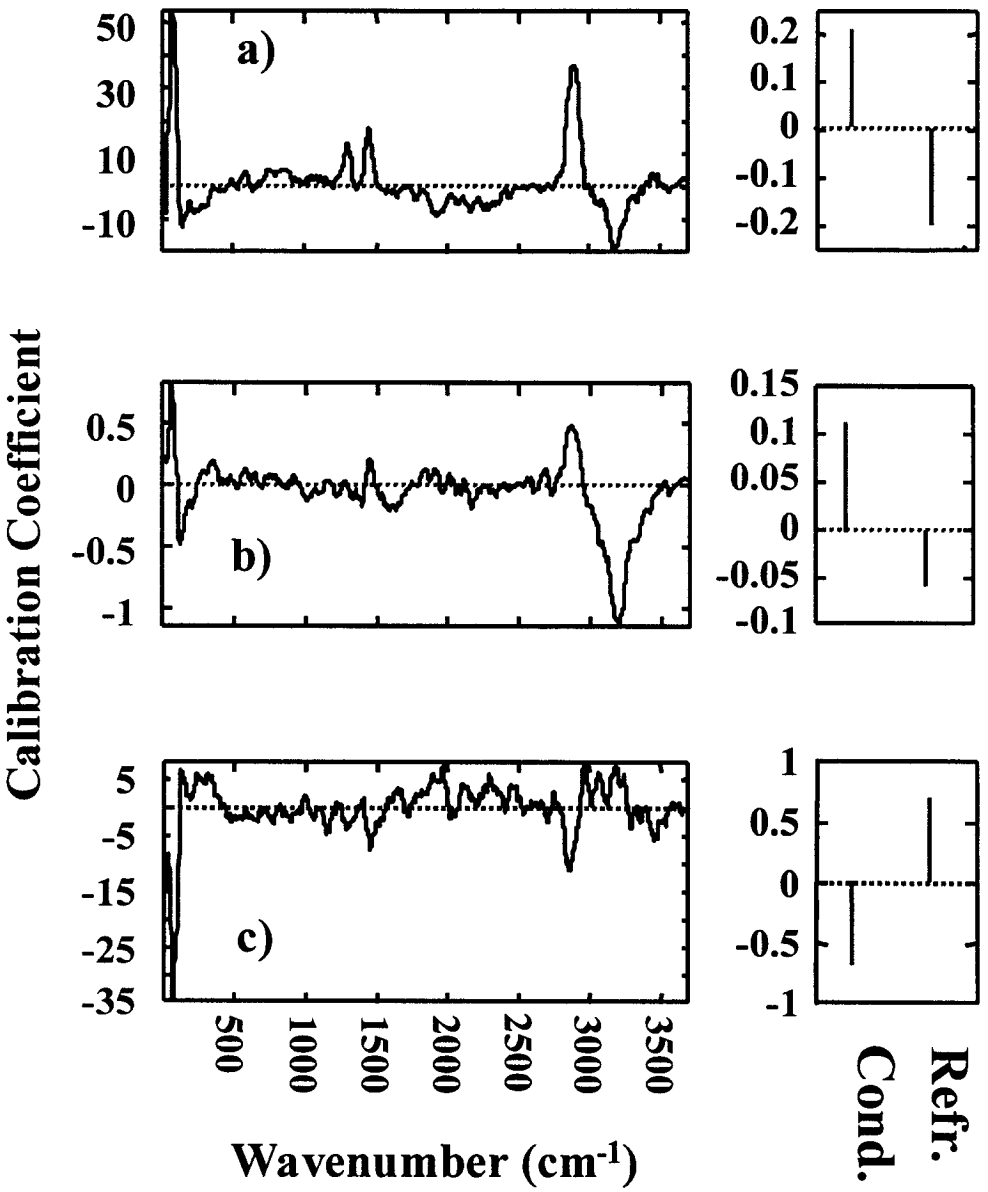

Figure 4.4 Calibration coefficients for fat (a), lactose (b), and protein (c) estimation using NIR FT-Raman spectra with the addition of conductivity (Cond.) and refractive index (Refr.) using PLS with leave-N-out cross validation.

refractive index were not heavily weighted in the calibration. However their addition is necessary in the weighting of 1445 and 3204 cm^{-1}. Also, in comparison to the spectra-only calibration of lactose (Figure 4.2b), bands at 350 and 1908 cm^{-1} are weighted much less when conductivity and refractive index are included.

Very few features in the calibration vector for protein, shown in Figure 4.4c, are weighted distinctly and heavily. Standout bands are at 62, between 100 and 400, 2864, and from 2966 to 3300 cm^{-1}. These regions were also found to be strongly correlated in the calibration vector constructed using Raman spectra alone, in Figure 4.2c. Discussion of assignment of these bands is in Section 4.2. Calibration coefficients corresponding to conductivity and refractive index are almost equal in magnitude but opposite in sign. This canceling out effect implies that refractive index and conductivity may not be directly significant in the calibration. However, the inclusion of these factors greatly influences the weighting of the spectral features in the calibration. This is shown by comparing the calibration vectors constructed using Raman spectra alone to Raman spectra with the physical properties. The same number of factors was used in both calibrations but on the calibration vector based on Raman and the additional information, the relevant frequencies are enhanced and the weighting of rest of the spectral frequencies has been drastically reduced.

All three calibration vectors were then used to estimate fat, lactose, and protein in the calibration set. Estimates of fat, lactose, and protein were plotted against concentrations found using the reference methods, shown in Figure 4.5 (a - c). Standard deviation (SEC) of the residual error and R^2 between estimated and reference values was computed. These results are listed in the first columns of Table 4.2. Validation of the

model was conducted by using the calibration models to estimate milk composition in an independent test set. Estimated concentration was plotted against reference concentration of the test set, shown in Figure 4.5 (d - f). Accuracy attained by the model for estimation in the validation set was determined by R^2 and the SEP, listed in the last 2 columns of Table 4.2.

Table 4.2. Estimation of fat, lactose and protein in milk using NIR FT-Raman spectra with the addition of conductivity and refractive index using PLS with leave-N-out cross validation

	Calibration Set		**Validation Set**	
Constituent	**R^2**	**SEC (g/ 100 g)**	**R^2**	**SEP (g/ 100 g)**
Fat	0.99	0.21	0.99	0.16
Lactose	0.82	0.06	0.77	0.05
Protein	0.81	0.30	0.83	0.21

Results for fat estimation in the validation set were the same using Raman spectra with or without the additional properties. However, in the calibration samples, R^2 is unchanged but the SEP increases by 24% when using Raman spectra with conductivity and refractive index. It was mentioned that the calibration vector used in this case was less noisy than that produced using only NIR FT-Raman spectra. Therefore, although there was an increase in SEP, this calibration may be more robust because less erroneous noise is modeled. Conductivity and refractive index were not found to enhance accuracy in estimating lactose using Raman spectra. In comparison to the spectra only calibration,

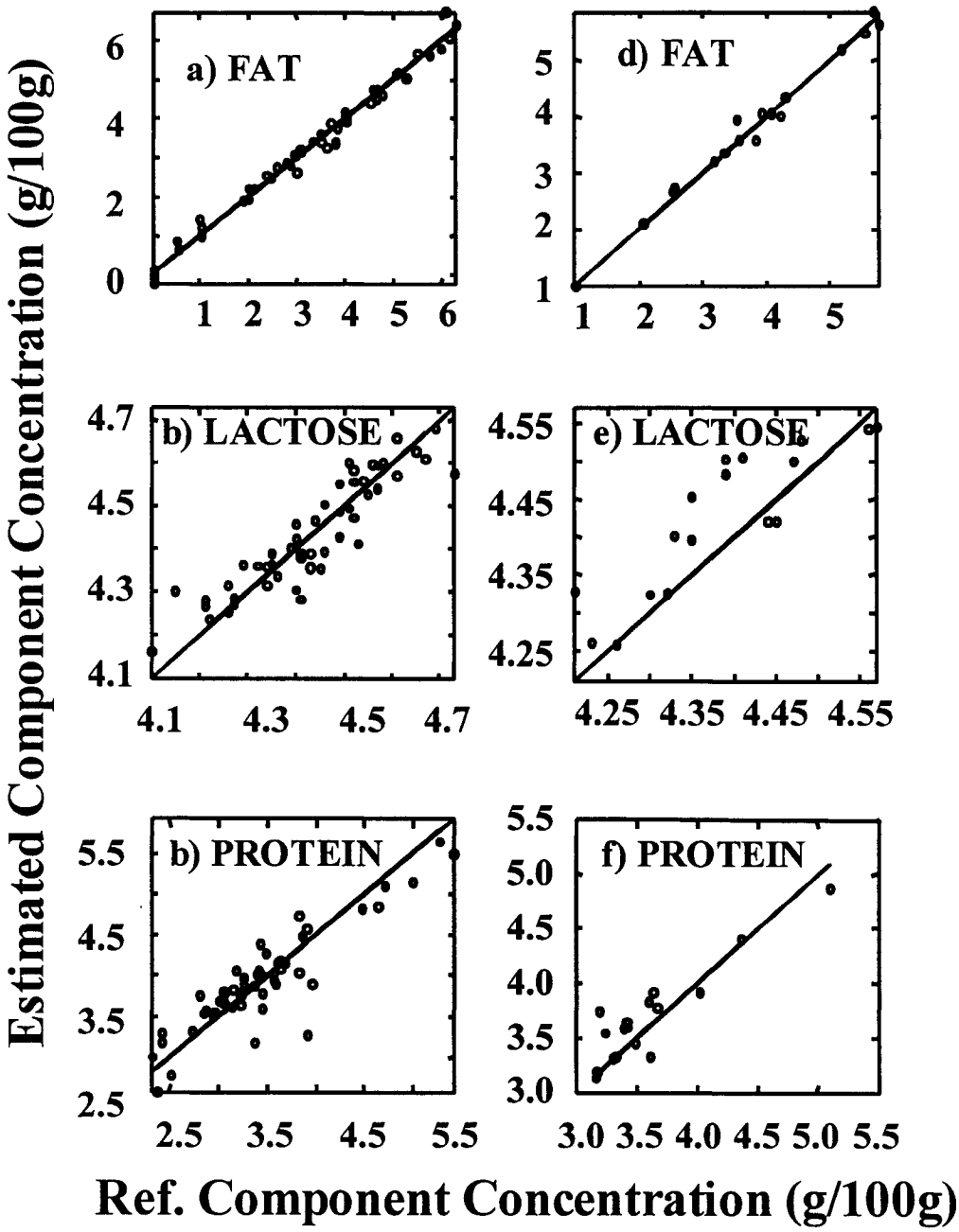

Figure 4.5 Estimation of fat, lactose, and protein in calibration set (a - c) and test set (d - f) using NIR FT-Raman spectra with the addition of conductivity and refractive index using PLS with leave-N-out cross validation

SEC increases by 20% and SEP increases by 25%. As in the fat calibration, the model constructed with all 3 measurements is likely to be more robust because the same number of factors produced less noise in the calibration vector. The two physical properties influenced the calibration model for protein estimation the most. With their inclusion, the SEC and SEP improved by 51% and 28% respectively. Standard error found using the calibration sample set is typically less than that of the validation set. This is due to the noise in the validation samples being independent of the information in the calibration samples for which the model was developed. Another notable improvement is that the estimation of higher protein content milk samples was shown to be more accurate using NIR FT-Raman spectra with the two physical properties compared to the estimation using SW-NIR spectra.

4.4 Investigation of inter-sample set variation using PLS analysis of NIR FT-Raman measurements with leave-one-set-out cross validation

Using PLS analysis of SW-NIR spectra with leave-one-set-out cross validation, it was determined that the presence of variation corresponding to each sample set resulted in a diminished accuracy of calibration models. Although results found using Raman spectra showed a marked improvement over those obtained using SW-NIR spectrophotometry, the influence of between sample set variation was examined. This was accomplished by conducting PLS analysis of Raman spectral data with and without conductivity and refractive index with leave-one-set-out cross validation.

Using this method, fat, lactose and protein calibration based on only Raman spectra, required 5, 6, and 6 factors respectively at a 95 % confidence level. The spectral regression coefficients of these calibrations are presented in Figure 4.6. Similarly, fat, lactose and protein, the calibrations using Raman spectra with the physical properties were found to require 6, 1, and 4 factors respectively at a 95 % confidence level. These calibration vectors are presented in Figure 4.7.

In Figure 4.6, fluctuations in magnitude of the regression coefficients are apparent although not as pronounced as the SW-NIR calibrations using this type of cross validation. In Figure 4.7, where conductivity and refractive index are included, variation in calibration coefficient magnitude is apparent. However, unlike in the spectra-only calibration vectors where the variation occurred for the duration of frequency region examined, the variation was contained to certain wavelength regions. For example, in the calibration vector for lactose (Figure 4.7b), the variation was distinct at around 350 cm^{-1} while for protein (Figure 4.7d), variation in bands at 600 and 2800 cm^{-1} was observed. Calibration coefficients corresponding to conductivity and refractive index also change in magnitude with each calibration.

Overall estimates of fat, lactose, and protein were calculated using the combined validation results from each separate calibration. Accuracy found using leaving one-set-out cross validation was indicated by the SECV and R^2 between reference and estimated concentration. These results are listed in Table 4.3.

Using spectra alone, the SECV increased by 65, 40, and 14% for fat, lactose, and protein respectively when leave-one-set out cross validation is used as opposed to leave-N-out cross validation. Similarly, when spectra with conductivity and refractive index

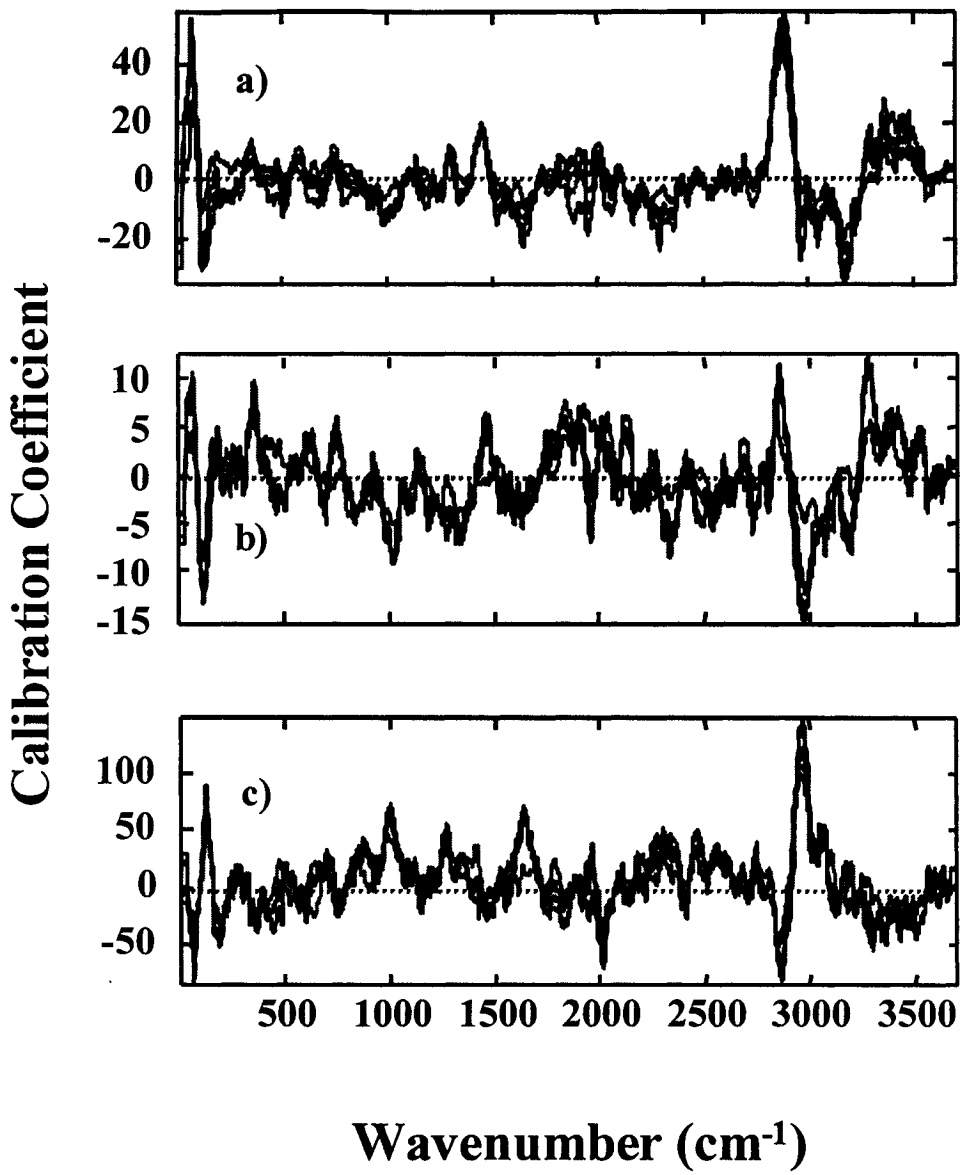

Figure 4.6 Calibration coefficients for fat (a), lactose (b), and protein (c) estimation using on NIR FT-Raman spectra using PLS with leave-one-set-out cross validation

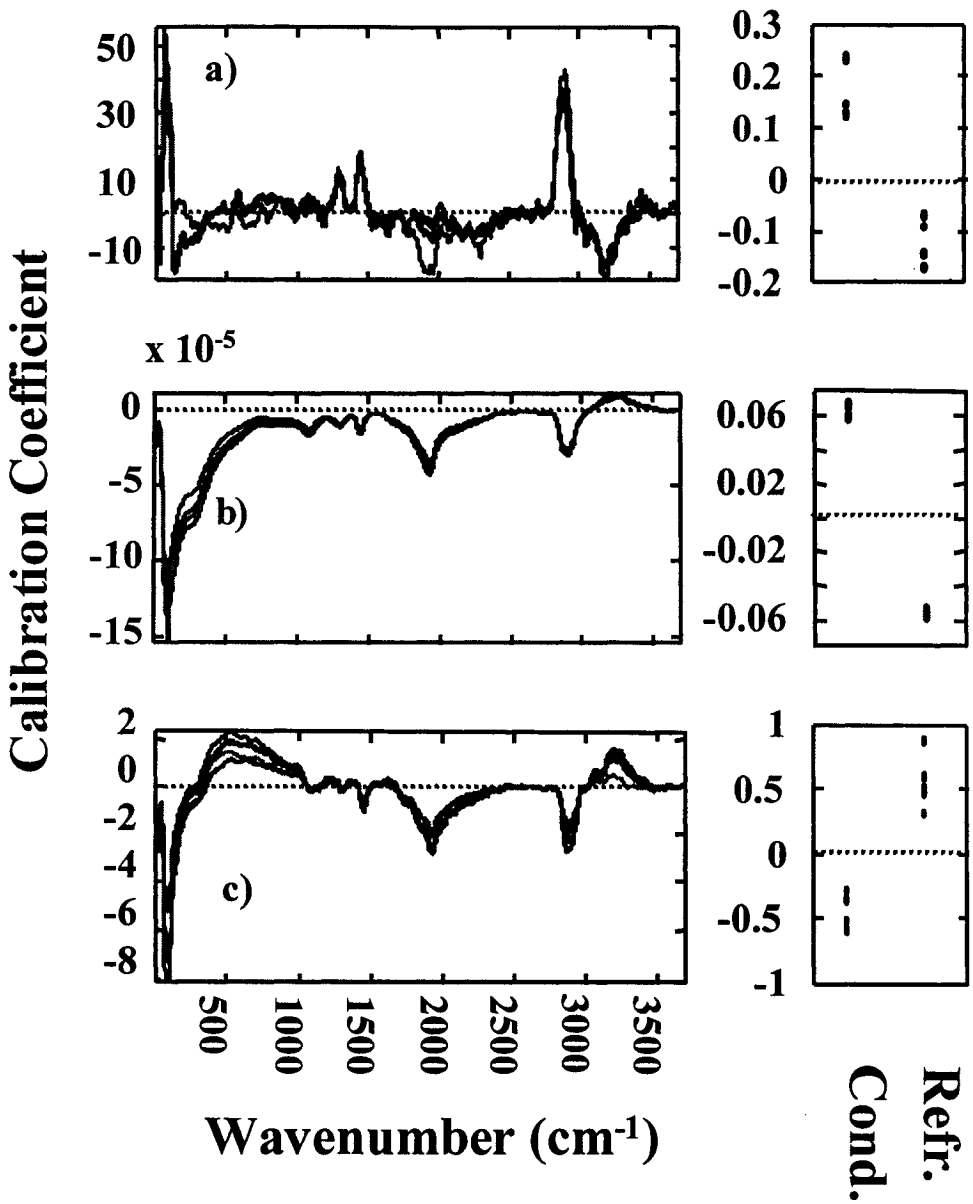

Figure 4.7 Calibration coefficients for fat (a), lactose (b), and protein (c) estimation using on NIR FT Raman spectra with the addition of conductivity (Cond.) and refractive index (Refr.), using PLS with leave-one-set-out cross validation

were used, the SECV increased by 33 and 43 % for fat and protein respectively when leave-one-set-out cross validation versus leave-N-out cross validation. For lactose estimation, the SECV does not change. However, R^2 decreases by 4% implying a reduction in accuracy. These results supports the need for inter sample set variation to be described in order to construct robust models for fat, lactose, and protein estimation in milk.

Table 4.3. Estimation of fat, lactose and protein using PLS analysis with leave-one-set-out cross validation of NIR FT-Raman spectra and NIR FT-Raman spectra with the inclusion of conductivity and refractive index

	Estimation using Raman spectra		**Estimation using Raman spectra with conductivity and refractive index**	
Constituent	R^2	SECV (g/ 100 g)	R^2	SECV (g/100 g)
Fat	0.98	0.28	0.97	0.28
Lactose	0.76	0.07	0.78	0.06
Protein	0.62	0.40	0.53	0.43

4.5 Estimation of fat, lactose, and protein using GA modelling of NIR FT-Raman spectra, conductivity and refractive index

Section 4.3 presented results for fat, lactose, and protein estimation using NIR FT-Raman spectra, conductivity, and refractive index using PLS regression, a multivariate statistical analysis. In PLS analysis, entire spectral measurements were used

to construct calibration models. In contrast to this method, the GA method is used to select a small subset of variables to build a multilinear regression model. This approach has been applied in the work of others to improve upon PLS models that were corrupted by the influence of spectral results not containing critical information [40, 44, 45]. With this work in mind, the GA method was used to estimate milk constituent concentration using NIR FT-Raman spectra with the two physical properties, as an alternative method to PLS analysis.

The same configuration of the GA method used in Section 3.6 was applied here where SW-NIR was replaced by Raman spectral measurements. The calibration set consisted of 75% of the samples including extreme concentrations and the remaining 25% made up the validation set. Numerical coefficients corresponding to Raman wavenumbers were computed by conducting a multiple linear regression using the calibration set. Estimation of concentration in the calibration set using this model yielded the SEC and in the validation set SECV. Combinations of Raman wavenumbers were selected based on maximum fitness. In this case, fitness was defined as the product of the SEC and SECV.

To determine the number of wavenumbers (consisting of Raman wavenumbers, conductivity and refractive index) to use in the calibration, models were constructed with incrementing wavenumbers, from 1 to 10. The optimal number of wavenumbers was found using an F-test to determine the model that resulted in the statistically minimum fitness at a 95% confidence level. Fitness was plotted against the number of wavelength/wavenumbers in the models for fat, lactose, and protein in Figure 4.8. For fat, lactose and protein estimation, it was determined that the optimal number of

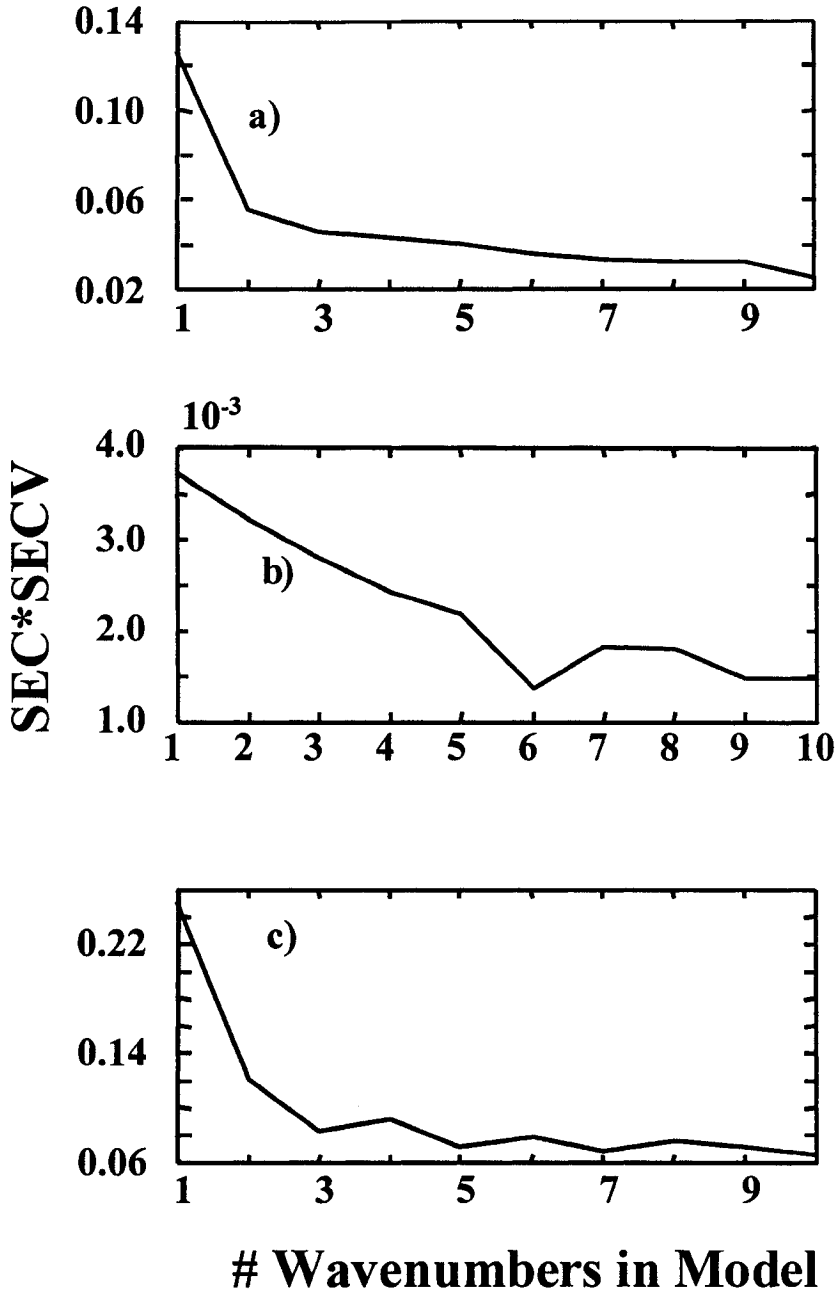

Figure 4.8 Standard Error at maximum generations for 1 to 10 wavenumber model for fat (a), lactose (b), and protein (c) estimation using NIR FT Raman spectra, conductivity and refractive index

wavelength/wavenumbers to be used in the models were 3, 6, and 5 respectively.

In the GA method, models are re-constructed with each generation. At each re-construction, recombination and mutation of individuals leads to a new population. Each individual in the population encodes a combination of wavelength/wavenumbers for fat, lactose, or protein estimation. The two individuals of the population yielding maximum fitness are carried over to the next population. At the 800^{th} generation, the wavelength/wavenumbers used in the model are found in the individual with the maximum fitness. It is desired that this final model is stable and the result of evolution through generations. As a measure of assurance in meeting this objective, fitness was plotted versus the number of generations for each model in Figure 4.9. This figure shows that by the 800^{th} generation, the final model is stable and does not appear to be caught in a local minimum of error. In fact the model was stable in less than 200 generations but 800 iterations was retained for the analysis to ensure convergence.

Multiple linear regression was used to determine numerical coefficients corresponding to the wavenumbers selected using the calibration sample set. The subsequent models for fat, lactose and protein estimation were found to be:

$$\text{Fat} = 1.2 + 3999*S_{2855} - 1725*S_{3163} + 3515*S_{3472} \quad (4.1)$$

$$\text{Lactose} = 2.1 - 808*S_{124} + 816*S_{139} - 317*S_{1219} + 345*S_{2855} - 95*S_{2978} + 0.6*\text{conductivity} \quad (4.2)$$

$$\text{Protein} = -415.4*S_{2839} + 2259.3*S_{2963} + 309.8*\text{refractive index} \quad (4.3)$$

where S represents NIR FT-Raman intensity at the GA selected subscript Raman wavenumbers.

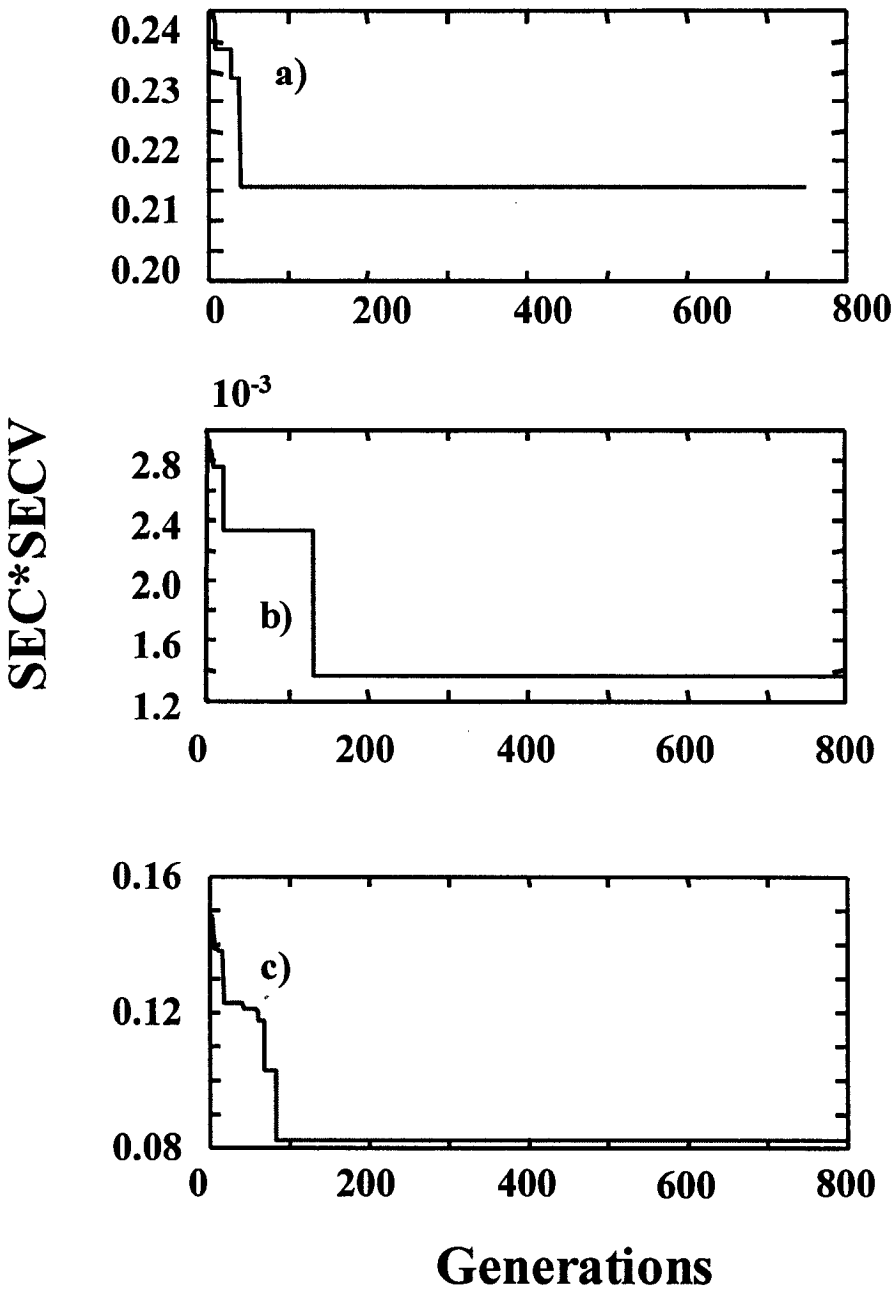

Figure 4.9 Fitness with increasing generation for 3 (fat (a)), 6 (lactose (b)), and 3 (protein(c)) wavelength/wavenumber models using NIR FT Raman spectra, conductivity and refractive index

Selected Raman wavenumbers for fat, lactose, and protein determination were indicated on an average milk spectrum in Figure 4.10. The 1st y-axis corresponds to the intensity of Raman scattering measured in the average milk spectrum. The 2nd y-axis refers to the numerical coefficients corresponding to the spectral frequencies selected.

For the fat calibration model, wavenumbers selected were all heavily weighted by PLS. In order to reduce the time consumption of the GA procedure, the Raman spectral data were reduced to 240 Raman points by only including every eighth wavenumber from the original data set consisting of 1919 wavenumbers. Of course, this resulted in a loss of resolution. In calibrations by PLS, 2874 cm^{-1} and 3474 cm^{-1} were heavily weighted. Although these wavenumbers weren't available in the condensed spectral data set, wavenumbers 2855 and 3472 cm^{-1}, close to these two frequencies, were selected by using the GA method . Raman signal at 2855 cm^{-1} corresponds to acyclic $-CH_2-$ symmetric stretching. This band is typical of fatty acids[33]. Selection of 3472 cm^{-1} corresponds to R-OH stretching due to water-fat interaction. The negatively weighted band at 3163 cm^{-1} corresponds to CH stretching in unsaturated hydrocarbons[41].

The Raman wavenumbers used in the model for lactose estimation are presented in Figure 4.11b. Selection of 1219 cm^{-1} by GA technique was assigned to C-O-C antisymmetric stretching [41]. Positive weighting of 2855 cm^{-1} agrees with the PLS regression, where 2860 cm^{-1} was one of the most heavily weighted Raman bands and corresponds to CH_2 stretching of lactose. Significance of 2978 cm^{-1} is due to CH stretching of carbohydrates[42]. The GA model selected both 124 cm^{-1} and 139 cm^{-1} , probably accounting for variation in scattering between samples as determined from

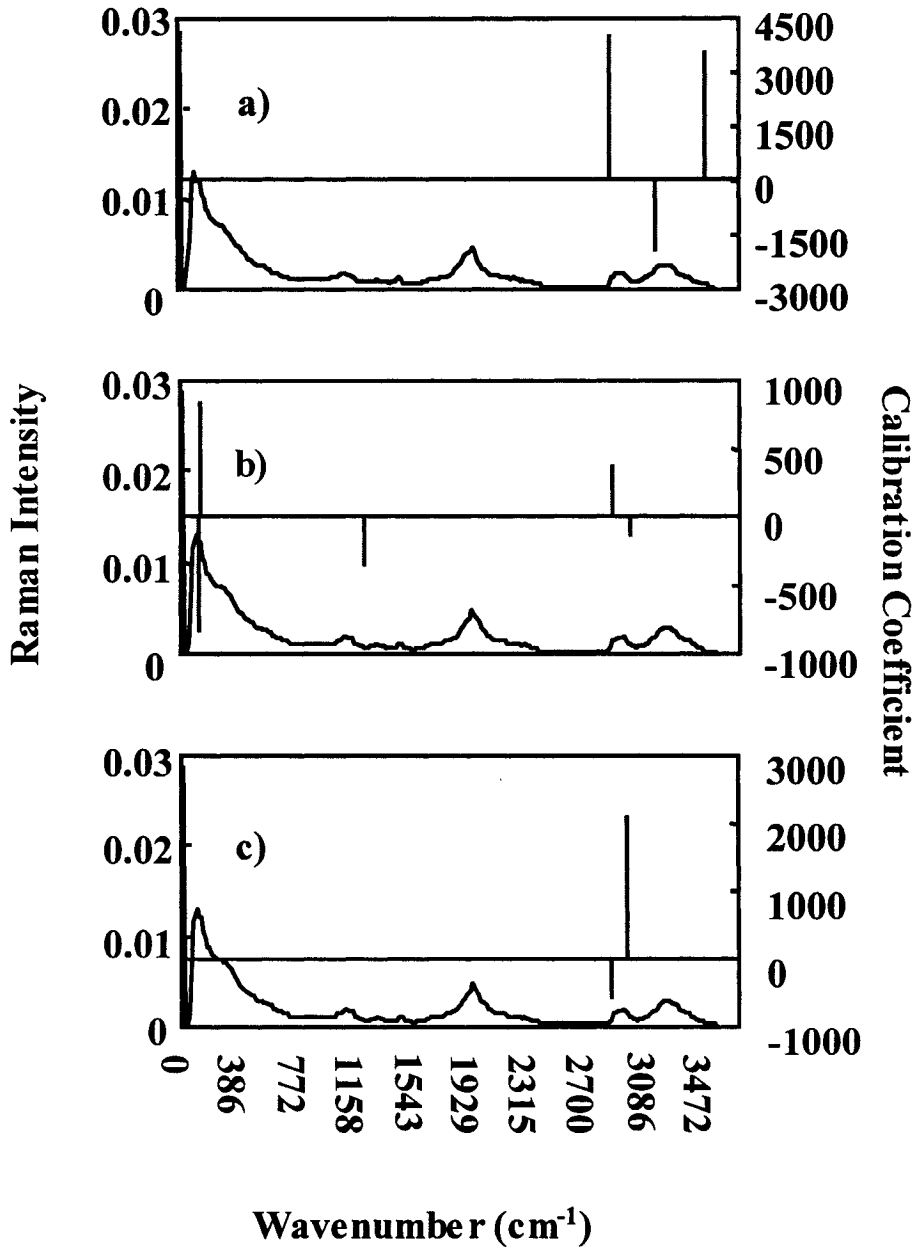

Figure 4.10 Regression coefficients for GA selected wavelengths for fat (a), lactose (b), and protein (c) estimation using NIR FT-Raman spectra, conductivity and refractive index plotted against average NIR FT-Raman spectrum of milk samples

changes in the Rayleigh scattering. Conductivity was selected in the estimation of lactose. Conductivity showed a strong correlation to lactose concentration as demonstrated in Table 3.2. It was also weighted in PLS and GA analyses of SW-NIR calibrations for lactose estimation. Only two Raman frequencies were necessary in the calibration model for protein. Both of the selected wavenumbers were significantly weighted in the PLS calibration.

Negative weighting of 2839 cm^{-1} corresponds to O-H stretching in hydrogen bonded secondary amide groups and 2963 cm^{-1} corresponds to CH stretching common in protein structures. In addition to these two Raman regions, refractive index was selected to be used in the model. This was also weighted heavily in the PLS and GA models for protein estimation using SW-NIR spectra.

The calibration models derived using the GA method were then applied to quantify milk constituent concentration in the calibration and validation sets. Plots of estimated concentration versus reference concentration are presented in Figure 4.11. As measures of accuracy, the R^2 and standard errors were calculated between the estimated values and line of identity. These results are listed in Table 4.4.

Accuracy in estimation of fat in the calibration set was similar to that found using the PLS model. However, in the test set, standard error increases by 69% using the GA model versus the PLS model. This result favours the PLS model for fat estimation because higher accuracy was found using an independent validation set. However, the GA model is quite competent considering only 2 spectral frequencies along with conductivity were used in the model as opposed to all 1919 wavenumbers. Results found for lactose estimation using the GA model exceed those found using the PLS model in

which, Raman spectra, conductivity and refractive index were applied. In the estimation of protein, samples with a protein concentration >5 g/100g, deviate considerably from the line of identity. Lack of samples in this concentration range in the validation set was a factor in a relatively low SECV compared to SEC.

Table 4.4. Estimation of fat, lactose and protein using NIR FT-Raman spectra, conductivity and refractive index with GA model.

	Calibration Set		Test Set	
Constituent	R^2	SEC (g/100g)	R^2	SECV (g/100g)
Fat	0.99	0.21	0.97	0.27
Lactose	0.86	0.05	0.95	0.03
Protein	0.69	0.39	0.86	0.21

As discussed earlier, there was a problem with the higher protein samples, where coagulation causing the reported reference concentrations to be under-estimates at the time of spectral analysis may have been possible. Another possibility is that protein cannot be modeled efficiently using linear regression. Instead, another function might be better suited such as logarithmic. However it should be noted that these methods employed the same calibration samples used routinely for commercial milk composition analysis. The errors determined here would also be present in the commercial systems.

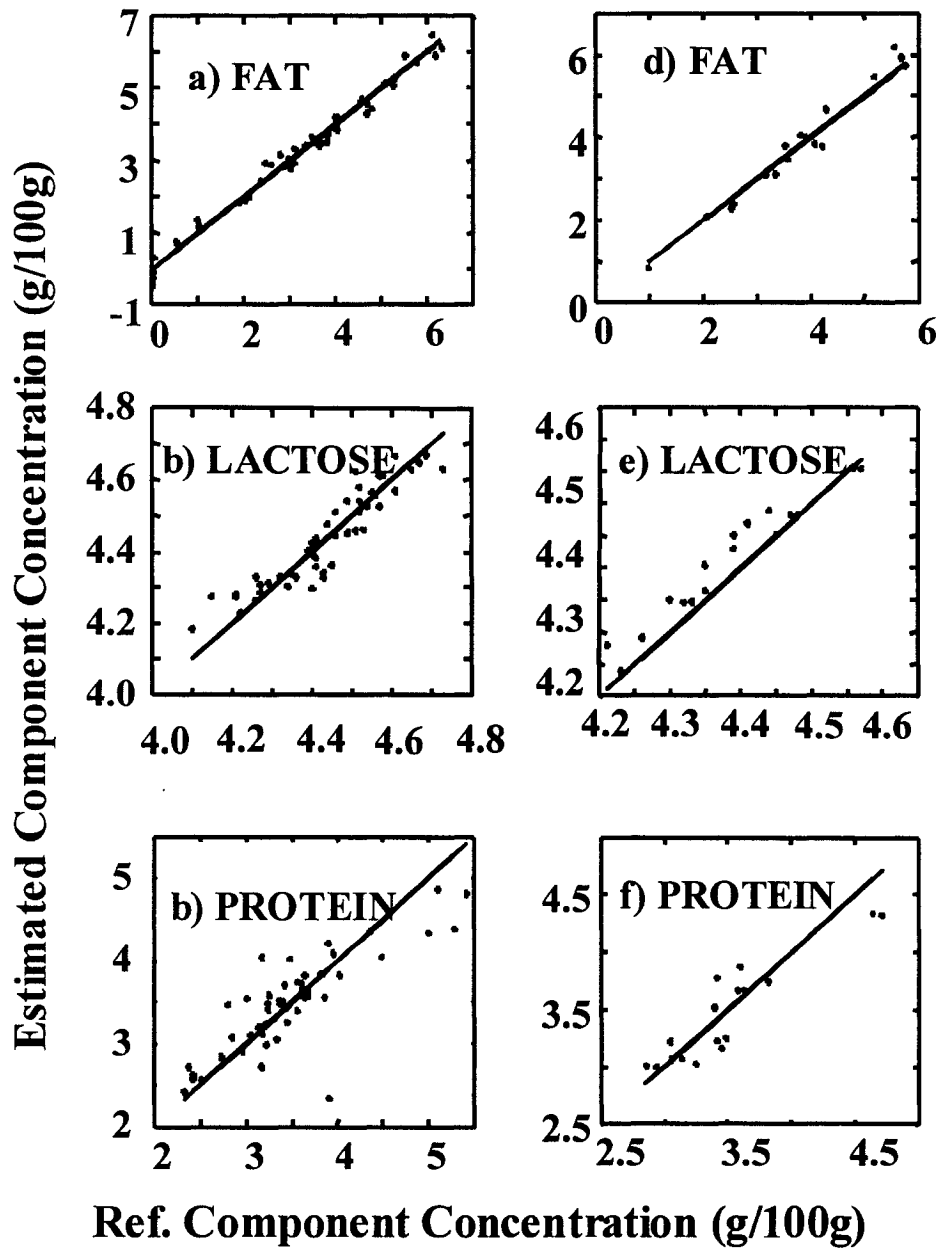

Figure 4.11 Estimation of fat, lactose, and protein in calibration set (a - c) and validation set (d-f) using NIR FT Raman spectra, conductivity, and refractive index using GA calibration model

Chapter 5 Discussion and Conclusions

This thesis has demonstrated the use of SW-NIR and NIR FT-Raman spectrophotometry with conductivity and refractive index in the determination of fat, lactose, and protein content in bovine milk. The initial stage of the project examined the use of PLS in calibrating SW-NIR spectra with and without conductivity and refractive index for fat, lactose, and protein determination. Next, the GA method was implemented to construct similar calibrations using spectra and the two physical properties. These steps were then repeated using NIR FT-Raman spectra.

Using SW-NIR spectra alone, the models determined using PLS analysis were found to be insufficient for milk composition estimation. Diminished accuracy was a result of noise modeled into the calibration vector. A vast improvement was found when incorporating conductivity and refractive index into the calibration. For fat, lactose, and protein estimations, SEC decreased by 27%, 55%, and 27% and SEP by 29%, 50%, and 52 %, respectively, with the inclusion of conductivity and refractive index compared to using only SW-NIR spectra. Using this method, good correlation between reference and estimated concentrations was achieved. This was indicated by R^2 values of 0.84, 0.81, and 0.71 and SEC results of 0.59 g/100g, 0.05 g/100g, and 0.36 g/100g in the calibration sample set (59 samples) for fat, lactose, and protein. Validation of these results was found using the models to estimate the test set. Here, R^2 was found to be 0.83, 0.82, and 0.80 and SEP 0.48 g/100g, 0.04 g/100g, and 0.28 g/100g for fat, lactose, and protein respectively. Because the Infrared based MilkoScan instrument is the method used for routine milk composition analysis, results of this study should be compared to the

accuracy found using this instrument. Of these results, the model for lactose estimation using SW-NIR spectra, conductivity and refractive index was found to give better results than those by Lefier et al., where an SEC of 0.083 g/100g was achieved based on 163 samples using a Milkoscan instrument.

When the GA method was used to select wavelengths to be used in multiple linear regression models, it was found that 9, 8, and 8 wavelengths were necessary to estimate fat, lactose, and protein respectively. Estimation of fat using PLS proved to be 6% more accurate than the GA model according to SEC values. The model for lactose estimation using GA, however, was more accurate than the PLS approach. This was indicated by an improvement in the SEC by 20%. Like fat, the PLS model was better for protein estimation compared to the GA model. Using the GA model, there was a 11% increase in SEC. In the GA calibration models for fat and protein, there were three mutual wavelengths. This represented an overlap in bands specific to each constituent. Table 5.1 presents the best results found for fat, lactose, and protein estimation using SW-NIR spectra. Accuracy found for protein and fat estimation was not as high as desired by the International Dairy Federation. Also, for fat and protein, the SEC values in Table 5.1 are 4.5 and 3 times greater than that reported by Lefier et al. respectively. In spite of this, results show promise for the technique. Definite correlation was observed when conductivity and refractive index were included with SW-NIR spectra.

Significant variation between sample sets was indicated by the baseline offsets in NIR spectra. This was further evidenced by variation in the calibration vectors obtained using PLS with leave-one-set-out cross validation. Description of this variation between

sample sets is necessary for a successful calibration as shown by the reduced accuracy observed in the results from this type of cross validation.

Table 5.1. Summary of methods yielding most accurate estimation of constituent concentration using SW-NIR spectra

Constituent	Measurement used in Calibration	Tools	R^2 (calibration)	SEC (g/ 100g)
Fat	SW-NIR spectra, conductivity, refractive index	PLS	0.84	0.59
Lactose	SW-NIR spectra, conductivity, refractive index	GA	0.88	0.04
Protein	SW-NIR spectra, conductivity, refractive index	PLS	0.71	0.36

Furthermore, 59 samples were used to acquire the calibration model and 20 samples were used to validate the calibration model. The partially homogenized milk samples were prepared using fractionation technology. Each sample set originates from a bulk milk sample consisting of 60 herd milks. In a study by Laporte et al., it was reported that increasing the number of samples to greater than 150 in a calibration leads to improvement in calibration and validation results[17]. In addition, it was found that a calibration consisting of both homogenized and unhomogenized milk samples leads to a more robust calibration with improvement in the determination of protein and casein due to the larger spectral variation[17].

The calibrations were repeated using NIR FT-Raman spectra in place of SW-NIR spectra. In contrast to SW-NIR results, NIR FT-Raman spectra proved to be an accurate calibration without conductivity and refractive index for fat and lactose estimation. Using spectra alone, the SEC was calculated to be 0.17g/100g, 0.05g/ 100g, and 0.35g/100g in the calibration set and the SEP was 0.16g/100g, 0.04g/100g, and 0.29g/100g in the validation set for fat, lactose, and protein estimation. Increase in accuracy was also supported by results for R^2. In the calibration set, R^2 was found to be 0.99, 0.88, and 0.77 and in the validation set, calculated R^2 was 0.98, 0.84, and 0.58 for fat, lactose, and protein respectively. Based on the SEC, these results are better than the best results found using SW-NIR for two of the milk constituents listed in Table 5.1. Calibrations were repeated where conductivity and refractive were included with the NIR FT-Raman spectra. For fat, lactose, and protein R^2 was found to be 0.99, 0.82, and 0.81 using the calibration set and 0.99, 0.77, and 0.83 using the validation set. Calculated standard error was 0.21, 0.06, and 0.30 g/100g in the calibration set and 0.16, 0.05, and 0.21 g/100g in the validation set for fat, lactose, and protein. In this case, there was a decrease in accuracy found for fat and lactose. However, it is believed that models in which the physical properties were included, were more robust than the calibrations based solely on Raman spectra. This is evidenced by diminished noise in the calibration vectors. Protein, on the other hand, shows a dramatic improvement when the physical properties are incorporated in the calibration. With their inclusion, the SEC and SEP improved by 51% and 28% respectively compared to the spectra-only calibration. In comparison to the best results found for protein estimation using SW-NIR spectra,

conductivity and refractive index, the result found using NIR FT-Raman spectra with the additional information is better by 17% (based on SEC) and 25% (based on SEP).

For comparison, the GA method was used to construct calibrations using NIR FT-Raman, conductivity and refractive index. Models used 3, 6, and 3 wavelength/wavenumbers for fat, lactose, and protein respectively. Lactose estimation using the GA model was better than that found using the PLS approach using NIR FT-Raman with conductivity and refractive index. However, accuracy in the estimation fat and protein decreased with the GA models. Table 5.2 lists the best results found using NIR FT-Raman spectra.

Table 5.2. Summary of methods yielding most accurate estimation of constituent concentration using NIR FT-Raman spectra

Constituent	Measurement used in Calibration	Tools	R^2 (calibration)	SEC (g/ 100g)
Fat	Raman spectra	PLS	0.99	0.17
Lactose	Raman spectra, conductivity, refractive index	GA	0.86	0.05
Protein	Raman spectra, conductivity, refractive index	PLS	0.81	0.30

In comparison to the accuracy found using the Milkoscan system with 163 samples in the work of Lefier et al., NIR FT-Raman results were much closer than SW-NIR results. Based on the information listed in Table 5.2, fat and protein yielded SEC values that were only 3 to 4 times greater than those reported by Lefier et al. under ideal

conditions. Like SW-NIR, accuracy found in the lactose estimation using NIR FT-Raman exceeds that found by Lefier et al. These results show promise for the use of NIR FT-Raman spectrophotometry in the quantification of fat, lactose and protein since the sample preparation step for both SW-NIR and Raman are significantly less stringent.

Using leave-one-set out cross validation, variation corresponding to each sample set was found. Without description of this variation in the calibration models, it was found that accuracy diminished.

Potential for SW-NIR and NIR FT-Raman spectrophotometry with the addition of conductivity and refractive index has been demonstrated for the estimation of fat, lactose, and protein in milk. All four of these properties have the potential to be measured on-line while milk is being collected. Difficulty in measurement is greatly diminished if hand-held conductivity meters and refractometers, presently available on the market presently, are employed. The combined cost of these items is relatively inexpensive compared to the cost of a commercial milk analyzer. Once a calibration has been developed that reaches standards set by the International Dairy Federation, mathematical treatment of the data could be simplistic. Ideally, all four parameters could be measured by a conglomerate device with built-in software that would compute and output fat, lactose, and protein content. This would allow for rapid results instead of the present scenario, which involves the collection of milk samples by dairymen, which are then sent to a lab specializing in analysis. Decreasing turnaround time would allow more frequent analysis as well as early detection of abnormalities in the milk that indicate illness.

Future work in this field should emphasize further development of the calibration. This may be accomplished using a larger sample set that includes un-homogenized raw

milk. The influence of another easily measurable parameter on the calibration such as pH or viscosity should be investigated as another descriptor which may improve accuracy. In terms of chemometrics, configuration of the GA method should be examined to construct the calibration with emphasis on the calibration set. Also, especially in the case of protein, it may be fruitful to use nonlinear forms of the calibration model.

Specifically in NIR FT-Raman spectral acquisition, several areas need investigation. One area is the influence of temperature on spectra. Unlike SW-NIR analysis, which was conducted at 40 °C, NIR FT-Raman spectra were acquired at room temperature. Maintaining temperature during analysis may result in a more robust calibration and may reduce inconsistency in inter-sample set variation. Furthermore, the long-term goal is to build a device that measures these properties as milk is being collected from the animal and it is thus important to recognize that the temperature will be closer to 40 °C. Fat melts at 40 °C, which changes its scattering properties and this most probably will affect resulting Raman spectra as it is a technique based on scattering. Another area for future work is assurance in maintenance of laser power. This could be achieved by employing a calibration standard that will not change over time such as a ceramic plate. Improvement in all of these areas may fine tune the calibration.

The cost of implementing an SW-NIR or NIR FT-Raman device for milk analysis based on the presented calibration models, would be much less than purchasing the current commercial Infrared based instruments, which cost approximately $750,000. The Raman instrument that was used in this study costs approximately $100,000. In addition the calibration identified wavelengths necessary for the estimation of milk components. A smaller-scale Raman instrument would be possible. This instrument would employ a

laser diode as the source, filters for wavelength isolation, and an avalanche photodiode in photon counting Geiger mode as the detector. Acquisition of these parts would cost approximately $10,000. A device at this price would be within reach of most dairy operations.

References

1. Wong, N.P. *Fundamentals of Dairy Chemistry*; 3rd ed; Jenness, R.; Keeney, M.; Marth, E.H., Eds.; Van Nostrand Reihold Co., New York, N.Y., 1988.

2. Fox, P.F.; McSweeney, P.L.H. *Dairy Chemistry and Biochemistry*; Blackie Academic and Professional: London, 1998.

3. Walstra, P., Geurts, TJ, Noomen, A., Jellema, A, van Boekel, MAJS. 1999. *Dairy Technology, Principles of Milk Properties and Processes*, Marcel Dekker Inc., New York, 4.

4. Varnam, A.H.; Sutherland, J.P. *Milk and Milk Products: Technology, Chemistry and Microbiology*; Chapman & Hall: London, 1994.

5. Burns, D.A.; Ciurczak, E.W., eds.; *Handbook of Near-Infrared Analysis*; 2nd ed.; Marcel Dekker, Inc.: New York, 2001, p499.

6. Jensen, R.G.; ed. *Handbook of Milk Composition.* Academic Press: San Diego, **1995**, 303-331.

7. Webber, G.; Lauwaars, M.; van Schaik, M. Inventory of IDF/ISO/AOAC International Adopted Methods of Analysis and Milk Products; 6th ed.; *Bulletin of the IDF.* **2000**, 350, 3-42.

8. Marshall, R.T. *Standard Methods for the Examination of Dairy Products*; 16th ed; American Public Health Association: Washington, D.C., 1993.

9. Association of Official Analytical Chemists. *Official Methods of Analysis of AOAC International*; 16th ed.; 5th revision; AOAC International: Arlington, **1999**, Vol 2., Chapt. 13.

10. International Dairy Federation. *Protein Definition*; IDF: Brussels, 1994, p42.

11. Lefier, D.; Grappin, R.; Pochet, S. Determination Of Fat, Protein, And Lactose In Raw Milk By Fourier Transform Infrared Spectroscopy And By Analysis With A Conventional Filter-Based Milk Analyzer; *Journal of AOAC International.* **1996**, 79, 711-717.

12. Tsenkova, R.; Atanassova, S.; Toyoda, K.; Ozaki, Y.; Itoh, K.; Fearn, T. Near-Infrared Spectroscopy For Dairy Management: Measurement Of Unhomogenized Milk Composition; *J. Dairy Sci.* **1999**, 82, 2344-2351.

13. Osborne, B.G.; Fearn, T. *Near Infrared Spectroscopy in Food Analysis*; Longman Scientific & Technical: New York, **1986**.

14. Giangiacomo, R.; Nzabonimpa, R. Approach To Near Infrared Spectroscopy; *Bulletin of the IDF.* **1994**, 298, 37-42.

15. Robert, P.; Bertrand, D.; Devaux, M.F. Multivariate Analysis Applied To Near-Infrared Spectra Of Milk; *Anal. Chem.* **1987**, 59, 2187-2191.

16. Tsenkova, R.; Atanassova, S.; Itoh, K.; Ozaki, Y.; Toyoda, K. Near Infrared Spectroscopy For Biomonitoring: Cow Milk Composition Measurement In A Spectral Region From 1,100 To 2,400 Nanometers; *J. Anim. Sci.* **2000**, 78, 515-522.

17. Laporte, M-F.; Paquin, P.; Near-Infrared Analysis Of Fat, Protein, And Casein In Cow's Milk; *J. Agric. Food Chem.* **1999**, 47, 2600-2605.

18. Sato, T.; Yoshino, M.; Furukawa, S.; Someya, Y.; Yano, N.; Uozumi, J.; Iwamoto, M. Analysis Of Milk Constituents By The Near Infrared Spectrophotometric Method; *Jpn. J. Zootech. Sci.*, **1987**, 58, 698-706.

19. Kamishikiryo-Yamashita, H.; Oritani, Y.; Takamura, H.; Matoba, T. Protein Content In Milk By Near-Infrared Spectroscopy; *Journal of Food Science.* **1994**, 59, 313-315.

20. Šašić, S.; Ozaki, Y. Short-wave Near-Infrared Spectroscopy Of Biological Fluids. 1. Quantitative Analysis Of Fat, Protein, And Lactose In Raw Milk By Partial Least-Squares Regression And Band Assignment; *Anal. Chem.* **2001**, 73, 64-71.

21. Chen, J.Y.; Iyo, C.; Kawano, S.; Terada, F. Development Of Calibration With Sample Cell Compensation For Determining The Fat Content Of Unhomogenised Raw Milk By A Simple Near Infrared Transmittance Method; *J. Near Infrared Spectrosc.* **1999**, 7, 265-273.

22. Rangappa, K.S. Contribution Of The Major Constituents To The Total Refraction Of Milk; *Biochimica et Biophysica Acta.* **1948**, 2,210-217.

23. Shoemaker, D.P.; Garland, C.W.; Nibler, J.W. *Experiments in Physical Chemistry*; 5th ed.; McGraw-Hill: New York,**1989.**

24. Jääskeläinen, A.J.; Peiponen, K.-E.; Räty, J.A. On Reflectometric Measurement Of A Refractive Index Of Milk; *J. Dairy Sci.* **2001**, 84, 38-43.

25. Fernando, R.S.; Spahr, S.L.; Jaster, E.H. Comparison Of Electrical Conductivity Of Milk With Other Indirect Methods For Detection Of Subclcinical Mastitis; *J.Dairy Sci.* **1985**, 68, 449-456.

26. Fernando, R.S.; Rindsig, R.B.; Spahr, S.L. Electrical Conductivity Of Milk For Detection Of Mastitis; *J. Dairy Sci.* **1982**, 65, 659-664.

27. Rubinson, J.F.; Rubinson, K.A.; *Contemporary Chemical Analysis*, Prentice Hall: Upper Saddle River, 1998, p505.

28. Li-Chan, E.C.Y. The Applications Of Raman Spectroscopy In Food Science; *Trends in Food Science and Technology.* **1996**, 7,361-370.

29. Fontecha, J.; Bellanato, J.; Juarez, M. Infrared And Raman Spectroscopic Study Of Casein In Cheese: Effect Of Freezing And Frozen Storage; *J. Dairy Sci.* **1993**, 76, 3303-3309.

30. Fehrmann, A.; Franz, M.; Hoffmann, A.; Rudzik, L.; Wüst, E. Dairy Product Analysis: Identification Of Microorganisms By Mid-Infrared Spectroscopy And Determination Of Constituents By Raman Spectroscopy; *Journal of AOAC International.* **1995**, 78, 1537-1542.

31. Fehrmann, A.; Franz, M.; Hoffmann, A.; Rudzik, L.; Wüst, E. Identification Of Micro-organisms Using Mid Infrared Spectroscopy And Quantitative Raman Spectroscopy In Dairies; *Journal of Molecular Structure.* **1995**, 348, 13-16.

32. International Dairy Federation. *Whole milk: Determination of milk fat, protein, and lactose content-Guide for the operation of Mid-Infrared Instruments;* IDF Standard 141 B; International Dairy Federation: Brussels, Belgium, 1996.

33. Chalmers, J.M.; Griffiths, P.R.; eds.; *Handbook of Vibrational Spectroscopy*; J. Wiley: New York, 2002,.Vol 5.

34. Haaland, D.M.; Thomas, E.V. Partial Least-Squares Methods For Spectral Analyses. 1. Relation To Other Quantitative Calibration Methods And The Extraction Of Qualitative Information; *Anal. Chem.* **1998**, 60, 1193-1202.

35. Ding, Q.; Small, G.W.; Arnold, M.A. Genetic Algorithm-Based Wavelength Selection For The Near-Infrared Determination Of Glucose In Biological Matrixes: Initialization Strategies And Effects Of Spectral Resolution; *Anal. Chem.* **1998**, 70, 4472-4479.

36. McShane, M.J.; Cote, G.L.; Spiegelman, C.H. Assessment Of Partial Least-Squares Calibration And Wavelength Selection For Complex Near-Infrared Spectra; *Applied Spectroscopy.* **1998**, 52, 878-884.

37. Gilbert, R.J.; Goodacre, R.; Woodward, A.M.; Kell, D.B. Genetic Programming: A Novel Method For The Quantitative Analysis Of Pyrolysis Mass Spectral Data; *Anal. Chem.* **1997**, 69, 4381-4389.

38. Jang, J.-S.R.; Sun, C.-T.; Mizutani, E.; eds. *Neuro-Fuzzy and Soft Computing. A Computational Approach to Learning and Machine Intelligence.* Prentice-Hall: New York, 1986, p175-180.

39. McShane, MJ., Cameron, BD, Cote, GL, Spiegelman, CH. Improving Complex Near-IR Calibrations Using A New Wavelength Selection Algorithm, *Applied Spectroscopy.* **1999**, 53, 1575-1581.

40. McShane, M.J.; Cameron, B.D.; Coté, G.L.; Motamedi, M.; Spiegelman, C.H.; A Novel Peak-Hopping Stepwise Feature Selection Method With Application To Raman Spectroscopy; *Analytica Chimica Acta.* **1999**, 388, 251-264.

41. Chalmers, J.M.; Griffiths, P.R.; eds.; *Handbook of Vibrational Spectroscopy*; J. Wiley: New York, 2002,.Vol 3.

42. Socrates, G. *Infrared and Raman characteristic group frequencies tables and charts*; 3rd edition; John Wiley and Sons: New York, 2001.

43. Ozaki, Y., Cho, R., Ikegaya, K., Muraishi, S., Kawauchi, K. Potential Of Near-Infrared Fourier Transform Raman Spectroscopy In Food Analysis, *Applied Spectroscopy*. **1992**, 46, 1503-1507.

44. Broadhurst, D.; Goodacre, R.; Jones, A.; Rowland, J.J.; Kell, D.B.; Genetic Algorithms As A Method For Variable Selection In Multiple Linear Regression And Partial Least Squares Regression With Applications To Pyrolysis Mass Spectrometry; *Analytica Chimica Acta*. **1997**, 348, 71-86.

45. Estienne, F.; Massart, D.L.; Zanier-Szydlowski, N.; Marteau, Ph.; Multivariate Calibration With Raman Spectroscopic Data: A Case Study; *Analytica Chimica Acta*. **2000**, 424, 185-201.