



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

C. E. Eskildsen,<sup>\*1</sup> M. A. Rasmussen,<sup>\*</sup> S. B. Engelsen,<sup>\*</sup> L. B. Larsen,<sup>†</sup> N. A. Poulsen,<sup>†</sup> and T. Skov<sup>\*</sup>

<sup>\*</sup>Department of Food Science, University of Copenhagen, DK-1958 Frederiksberg, Denmark

<sup>†</sup>Department of Food Science, Aarhus University, DK-8830 Tjele, Denmark

### ABSTRACT

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

individual FA. The 2 breeds had different covariance structures between individual FA and total fat content. The results showed that the validation samples yielded biased predictions. This may limit the usefulness of FT-IR-based predictions of individual FA in milk recording as indirect covariance structures in the calibration set must be valid for future samples. Otherwise, future samples will show incorrect predictions.

**Key words:** bovine, fatty acid, infrared spectroscopy, milk, quantification

### INTRODUCTION

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

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

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

Received May 7, 2014.

Accepted August 14, 2014.

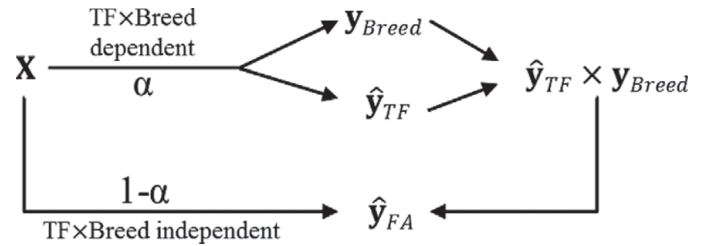
<sup>1</sup>Corresponding author: [carle@food.ku.dk](mailto:carle@food.ku.dk)

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

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

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

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



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

rect correlations between FA and the TF  $\times$  Breed interaction. In Figure 1,  $\alpha$  defines how dependent the FA prediction is on the TF  $\times$  Breed interaction. The FA prediction is more dependent on the TF  $\times$  Breed interaction as  $\alpha$  increases toward 1.

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

## MATERIALS AND METHODS

### Raw Milk Samples

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

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

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

### Spiked Milk Samples

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

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

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

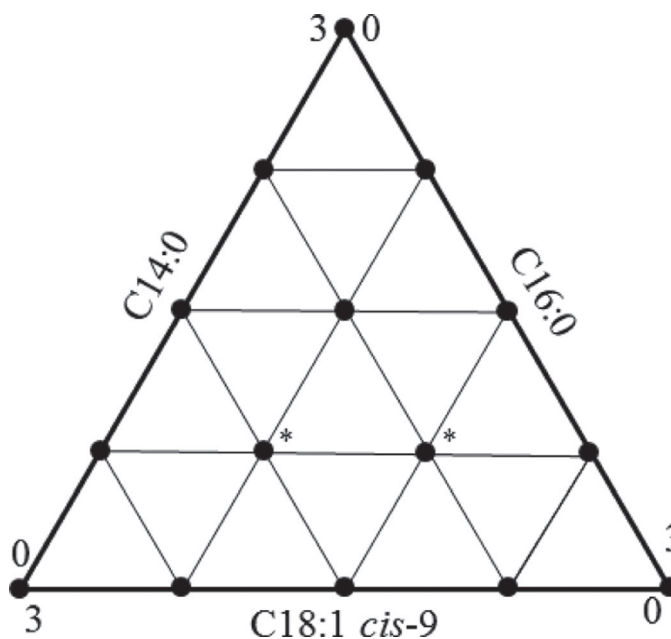
### Data Analysis

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

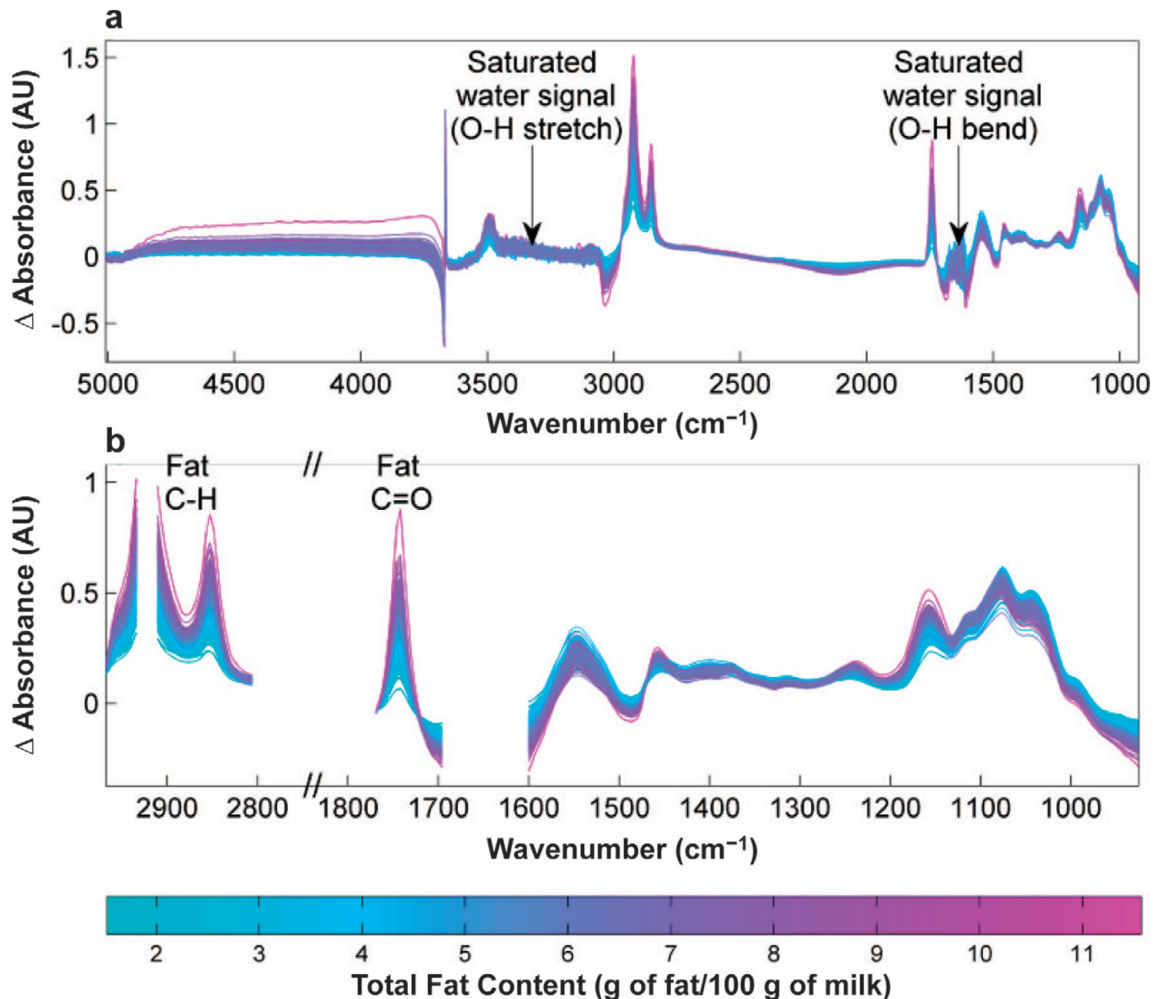
The FT-IR spectra were obtained in transmittance mode in the range from 5,008 to 925  $\text{cm}^{-1}$ , with a total

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

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



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



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

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

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

The PLS models calibrated using spiked milk samples were validated by leave-one-out cross-validation.

### Predicting FA—Contribution from TF and Breed

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

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

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

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

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

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

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

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

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

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

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

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

## RESULTS AND DISCUSSION

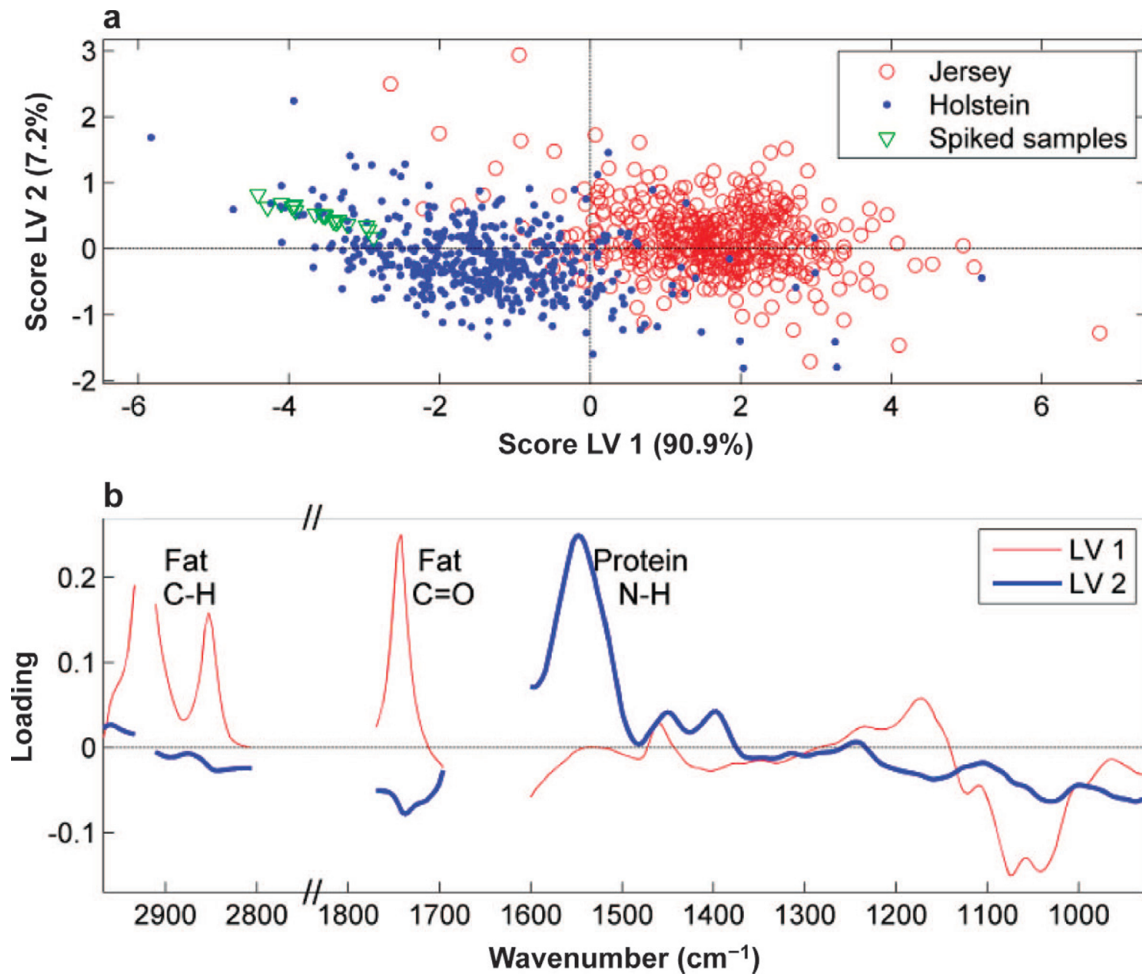
The FT-IR measurements obtained on raw milk samples are presented in Figure 3. The absorption band located just above 3,000  $\text{cm}^{-1}$  (Figure 3a) remains from the olefinic, =C-H stretch and should correlate to the degree of unsaturated fat in the milk samples. However, in this study, it was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

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

### Raw Milk Samples

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

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



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

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

will then result in increased predictions of an individual FA, even though other FA are responsible for the increase in TF. Hence, if the FA-to-TF ratio found in the calibration set is not conserved in future samples, future samples will show incorrect and biased predictions.

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

**Table 1.** Results from partial least squares models based on raw milk samples<sup>1</sup>

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

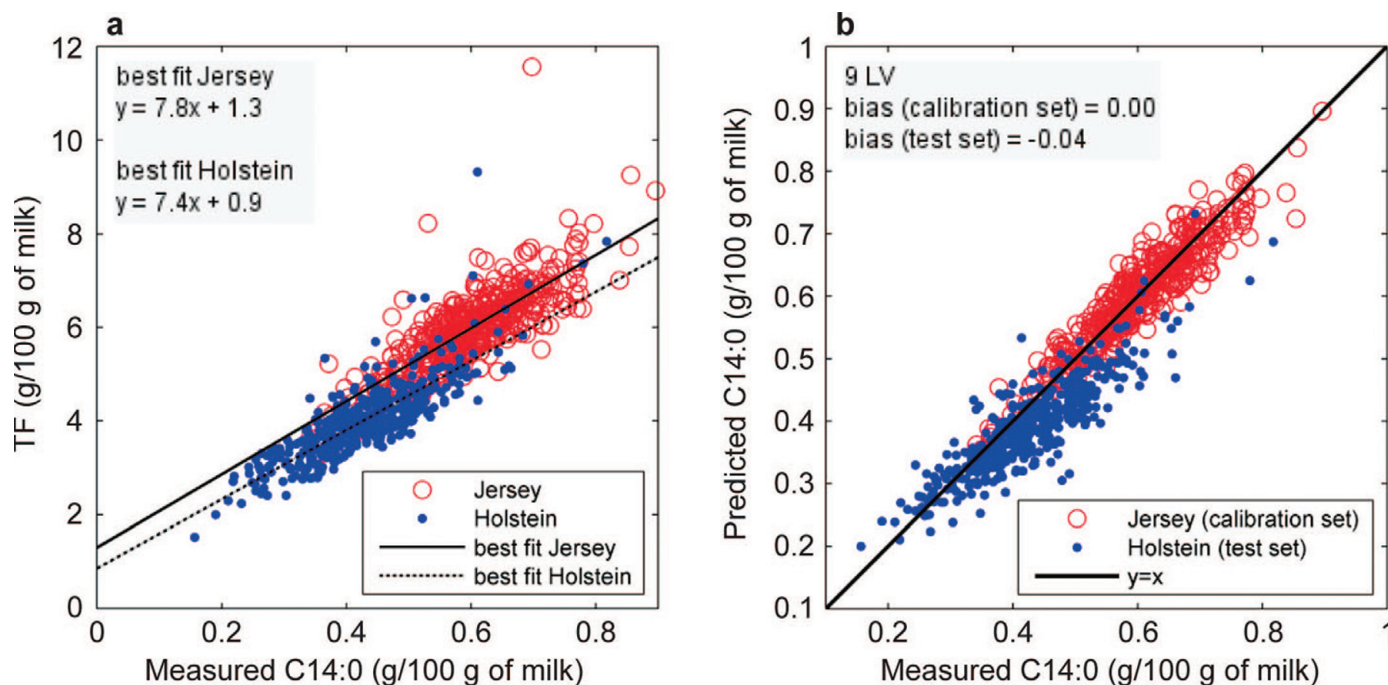
<sup>1</sup>LV = number of latent variables; R<sup>2</sup>CV = cross-validated coefficient of determination; RMSECV = root mean squared error of cross-validation; relative error = RMSECV/mean. Range, mean, SD, CV, and RMSECV are in units of grams of FA/100 g of milk.

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

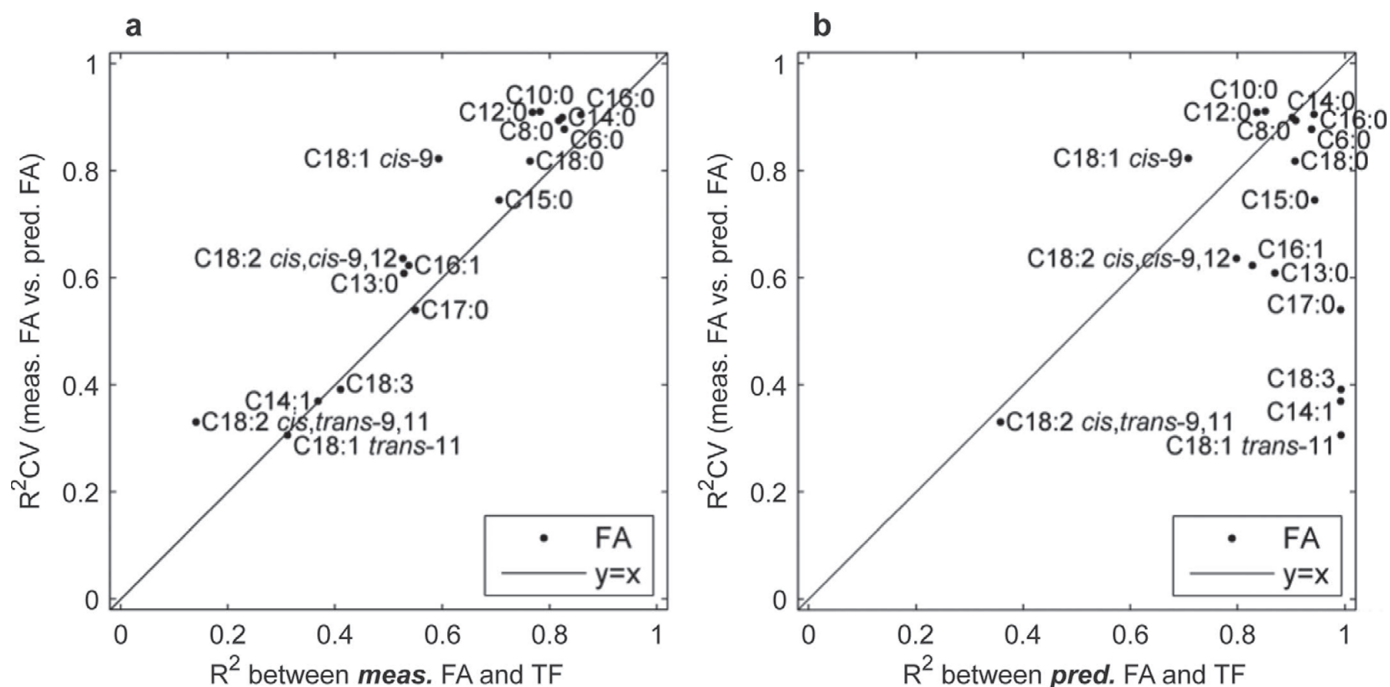
In a study based on 49 samples (using approximately 10 LV per PLS model), Soyeurt et al. (2006) found that the correlation coefficients (*r*) obtained from PLS models (measured vs. predicted FA) were greater than those between the measured FA and TF. Based on these findings, Soyeurt et al. (2006) suggested that predicted concentrations of FA were based on infrared absorption bands specific to individual FA rather than to TF. As TF (estimated from MilkoScan) is already a linear combination of the FT-IR measurements, correlations obtained from PLS models (measured vs. predicted FA) will always be greater than or equal to the correlation between measured FA and TF. If correlations from PLS models are greater, this indicates that information additional to TF is used for predicting the FA. However, this additional information is not necessarily the major contributor to the FA predictions. Furthermore, the additional information cannot be assigned as

absorption directly associated with FA without further investigations. In the current study, R<sup>2</sup> from PLS models (measured vs. predicted FA) were compared with R<sup>2</sup> between measured FA and TF (Figure 6a). Results similar to those obtained by Soyeurt et al. (2006) were found. However, when comparing R<sup>2</sup> from PLS models with R<sup>2</sup> between predicted FA and TF (Figure 6b), we found that the degree of correlation between predicted FA and TF exceeded model performance for most of the FA. In particular, the increased R<sup>2</sup> values between most FA and TF from Figure 6a to Figure 6b is an indication of the FA being primarily modeled (indirectly) by their relationship to TF.

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



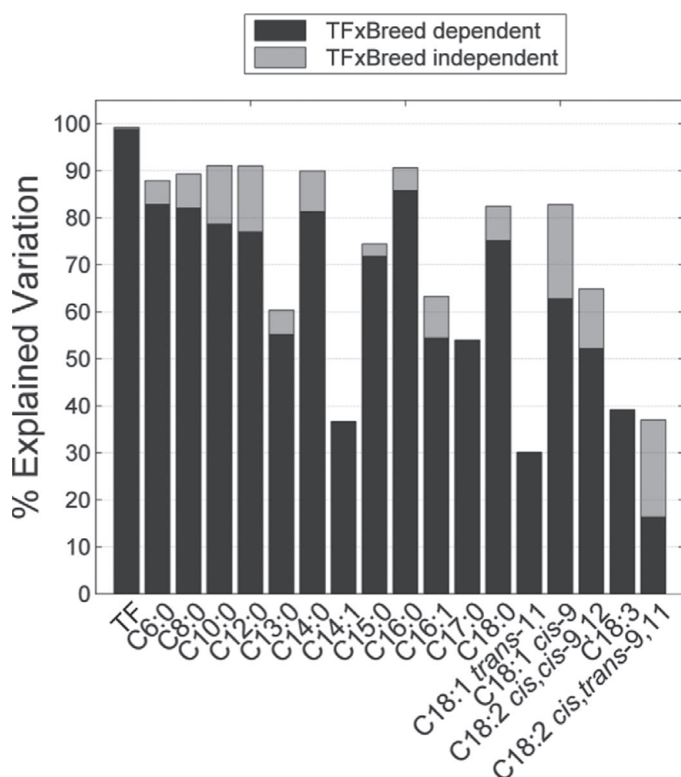
**Figure 5.** (a) Relationship between measured C14:0 and total fat content (TF) in raw milk samples from Jersey and Holstein cows; (b) measured versus predicted values of C14:0; C14:0 was predicted by partial least squares (PLS) regression applied to Fourier transform infrared measurements. The PLS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



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

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

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



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

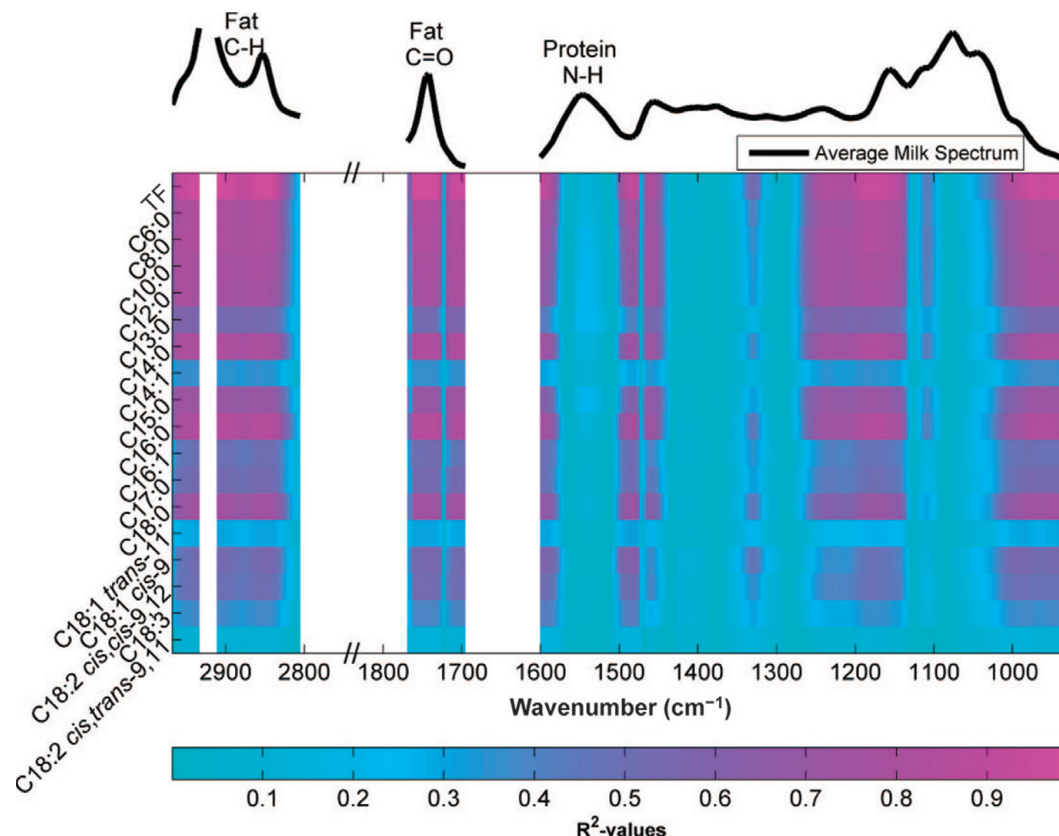
This further documents that individual FA do not have unique absorption patterns in FT-IR measurements obtained from milk.

### Spiked Milk Samples

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

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

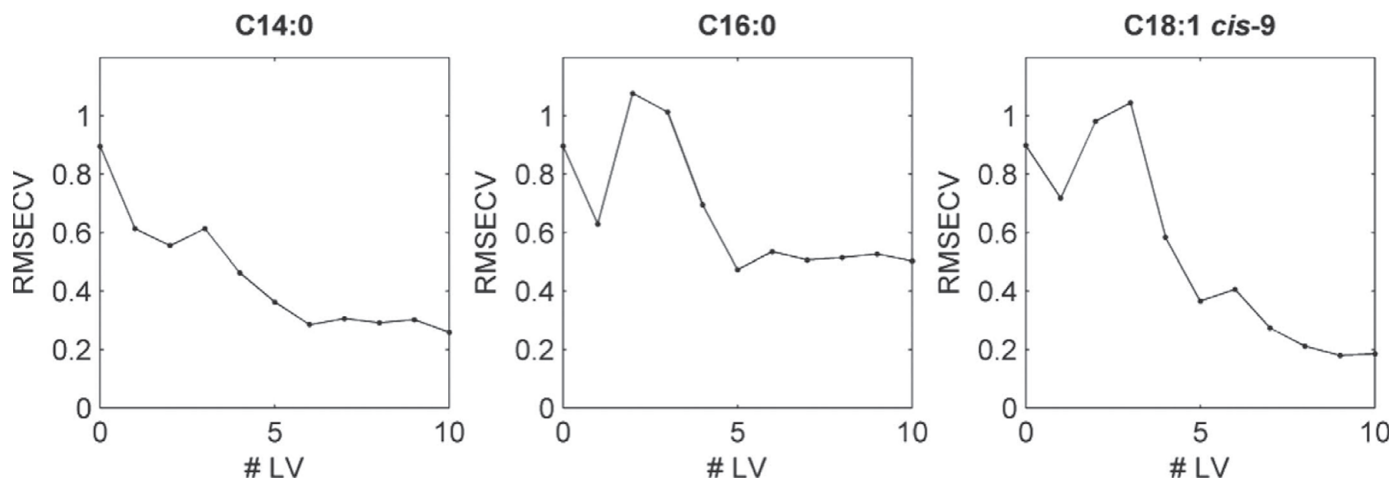
For C14:0, increased values of RMSECV were found when going from 2 to 3 LV; for C16:0, increased RMSECV were found going from 1 to 2 LV; and for C18:1 *cis*-9, increased RMSECV values were observed going from 1 to 3 LV (Figure 9). Different preprocessing did not remove the increase in RMSECV with increasing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



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

However, as FA were spiked to a similar background, spectral variation that could hide information specific to the FA must be minimal and it seems more plausible that the increase in RMSECV is because no informa-

tion specific to the individual FA is present in the FT-IR spectra. The lower RMSECV at 5 to 6 LV may be the consequence of over-fitting due to the combination of a limited number of samples (17) and the relatively



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

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

## CONCLUSIONS

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

## ACKNOWLEDGMENTS

The Danish Council for Strategic Research is acknowledged for generous financial support to the project entitled "Exploration of high-throughput FTIR-based solutions as an effective tool for correlat-

ing economically important milk quality characteristics with genetic markers" under the inSPIRe (Danish Industry-Science Partnership for Innovation and Research in Food Science) consortium (Copenhagen, Denmark). Furthermore, Arla Foods a.m.b.a. (Viby J, Denmark), Foss Analytical A/S (Hillerød, Denmark), Danish Cattle Federation (Aarhus, Denmark), as well as the Milk Levy Fund (Denmark) are acknowledged for financial support.

## REFERENCES

- Chapman, D. 1965. Infrared spectroscopy of lipids. *J. Am. Oil Chem. Soc.* 42:353–371.
- Couvreux, S., C. Hurtaud, C. Lopez, L. Delaby, and J. L. Peyraud. 2006. The linear relationship between the proportion of fresh grass in the cow diet, milk fatty acid composition, and butter properties. *J. Dairy Sci.* 89:1956–1969.
- De Marchi, M., M. Penasa, A. Cecchinato, M. Mele, P. Secchiari, and G. Bittante. 2011. Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of Brown Swiss bovine milk. *Animal* 5:1653–1658.
- De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa. 2014. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. *J. Dairy Sci.* 97:1171–1186.
- Ferrand, M., B. Huquet, S. Barbey, F. Barillet, F. Faucon, H. Larroque, O. Leray, J. M. Trommenschlager, and M. Brochard. 2011. Determination of fatty acid profile in cow's milk using mid-infrared spectrometry: Interest of applying a variable selection by genetic algorithms before a PLS regression. *Chemom. Intell. Lab. Syst.* 106:183–189.
- Ferrand-Calmels, M., I. Palhiere, M. Brochard, O. Leray, J. M. Astruc, M. R. Aurel, S. Barbey, F. Bouvier, P. Brunschwig, H. Caillett, M. Douquet, F. Faucon-Lahalle, M. Gele, G. Thomas, J. M. Trommenschlager, and H. Larroque. 2014. Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. *J. Dairy Sci.* 97:17–35.
- German, J. B., R. A. Gibson, R. M. Krauss, P. Nestel, B. Lamarche, W. A. van Staveren, J. M. Steijns, L. C. P. G. M. de Groot, A. L. Lock, and F. Destailats. 2009. A reappraisal of the impact of dairy foods and milk fat on cardiovascular disease risk. *Eur. J. Nutr.* 48:191–203.
- Jensen, R. G. 2002. The composition of bovine milk lipids: January 1995 to December 2000. *J. Dairy Sci.* 85:295–350.
- Kaylegian, K. E., J. M. Lynch, J. R. Fleming, and D. M. Barbano. 2009. Influence of fatty acid chain length and unsaturation on mid-infrared milk analysis. *J. Dairy Sci.* 92:2485–2501.
- Luinge, H., E. Hop, E. Lutz, J. Vanhemert, and E. Dejong. 1993. Determination of the fat, protein and lactose content of milk using Fourier-transform infrared spectrometry. *Anal. Chim. Acta* 284:419–433.
- Maurice-Van Eijndhoven, M. H. T., H. Soyeurt, F. Dehareng, and M. P. L. Calus. 2013. Validation of fatty acid predictions in milk using mid-infrared spectrometry across cattle breeds. *Animal* 7:348–354.
- Montgomery, D. C. 2009. Response surface methodology. Pages 417–485 in *Design and Analysis of Experiments*. 7th ed. John Wiley & Sons Inc., Hoboken, NJ.
- Næs, T., I. Mage, and V. H. Segtnan. 2011. Incorporating interactions in multi-block sequential and orthogonalised partial least squares regression. *J. Chemometr.* 25:601–609.
- Nørgaard, L., A. Saudland, J. Wagner, J. Nielsen, L. Munck, and S. Engelsen. 2000. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Appl. Spectrosc.* 54:413–419.
- Poulsen, N. A., F. Gustavsson, M. Glantz, M. Paulsson, L. B. Larsen, and M. K. Larsen. 2012. The influence of feed and herd on fatty acid composition in 3 dairy breeds (Danish Holstein, Danish Jersey, and Swedish Red). *J. Dairy Sci.* 95:6362–6371.

- Roger, J., F. Chauchard, and V. Bellon-Maurel. 2003. EPO-PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits. *Chemom. Intell. Lab. Syst.* 66:191–204.
- Rutten, M. J. M., H. Bovenhuis, K. A. Hettinga, H. J. F. van Valenberg, and J. A. M. van Arendonk. 2009. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. *J. Dairy Sci.* 92:6202–6209.
- Skov, T., D. Ballabio, and R. Bro. 2008. Multiblock variance partitioning: A new approach for comparing variation in multiple data blocks. *Anal. Chim. Acta* 615:18–29.
- Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. *J. Dairy Sci.* 89:3690–3695.
- Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. P. Berry, M. Coffey, and P. Dardenne. 2011. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. *J. Dairy Sci.* 94:1657–1667.
- Soyeurt, H., F. Dehareng, P. Mayeres, C. Bertozzi, and N. Gengler. 2008. Variation of  $\Delta^9$ -desaturase activity in dairy cattle. *J. Dairy Sci.* 91:3211–3224.
- Strang, G. 2006. Orthogonality. Pages 141–200 in *Linear Algebra and its Applications*. 4th ed. Thomson Learning Inc., Toronto, ON, Canada.

## APPENDIX: Cross-Validation of the Projection Procedure

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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