Prediction and Heritability Estimation of Bovine Milk Composition Using Mid-Infrared Spectroscopy

Cameron Jakub

August 2019

Contents

1 Prediction of Fatty Acid Composition and Milk Fat Globule Size of Bovine Milk Samples Using Mid-Infrared Spectroscopy

1.1 Abstract

The purpose of this study was to use mid-infrared (MIR) spectroscopy on bovine milk samples to develop prediction equations to predict fatty acid composition as well as milk fat globule size. Traditional methods of fatty acid determination and MFG measurement are expensive and slow relative to MIR spectroscopy, which is already being used regularly in milk sample recording. This paper is based off of the methods implemented in Fleming et al. (2017a,b), but looks to improve upon the predictive ability of the models developed in that study. Fleming et al. (2017a,b) used partial least squares regression (PLSR) to develop prediction equations for 22 individual fatty acids, 7 fatty acid groups, and 2 milk fat globule measurements. In this study, PLSR was performed on the same fatty acids, fatty acid groups, and milk fat globule sizes, but two different regression methods were also tested: the least absolute shrinkage and selection operator (LASSO) and competitive adaptive reweighted sampling-partial least squares regression (CARS-PLSR). The regression methods were tested on fatty acids expressed as g/100g of fatty acid, g/100g of milk, ln (g/100g of milk + 1), and on randomized subsets of the natural logarithmic transformed dataset by removing excess samples with similar composition. Fatty acid determination was completed on 2,064 milk samples with recorded spectra by gas chromatography, and milk fat globule measurements were completed on 2,083 milk samples with recorded spectra by integrated light scattering techniques. The coefficient of determination of cross-validation (R^2_{cv}) was used to determine the predictive ability of each model. PLSR models developed in this study produced comparable R^2_{cv} values to those obtained by Fleming et al. (2017a,b). LASSO regression models typically offered very small improvements in predictive ability while only using a subset of the full set of 862 variables (wavelengths) considered. CARS-PLSR produced the highest R^2_{cv} values for all tests compared to PLSR and the LASSO. When compared to PLSR, CARS-PLSR offered an average of 14% increase in R^2_{cv} while only using an average of about 12% of the full set of 862 wavelengths used in PLSR.

1.2 Introduction

Determination of the fat composition of milk has become an important field of study because of the influence fat composition can have on the nutritious and technological properties of the milk, and because fat composition of milk samples can often serve as an indicator of bovine health. Mid-infrared (MIR) spectroscopy is already used regularly to determine total fat and protein contents of milk samples with enough accuracy to use for genetic selection and payment purposes (Ferrand-Calmels et al., 2013). However, it is also useful to know individual fatty composition and milk fat globule (MFG) size of milk samples. MFGs make up 95% of total milk fat by weight (W. Keenan and Mather, 2006), and over 400 fatty acids are present in bovine milk with 12 fatty acids being present in concentrations greater than 1% (Jensen, 2002).

The membrane of MFGs are rich in valuable lipids and glycoproteins, and milk samples with smaller MFGs contain more membrane material per unit of fat compared to large MFGs (Fleming et al., 2017b). Fatty acid composition plays a large role in determining the nutritious quality of milk, so it is important to be able to accurately and efficiently determine the fatty acid composition of milk to produce the most desirable product for consumers (Ferrand-Calmels et al., 2013). Both MFG size and fatty acid composition influence the technological properties of milk, which is important to determine which products can be made from a particular milk sample, such as whether it is to be for consumption in liquid form or made into a cheese, for example (Fleming et al., 2017a,b). Furthermore, changes in milk fatty acids can serve as an indicator of bovine health and energy balance (Fleming et al., 2017a). For these reasons, it is important that dairy producers have an effective and efficient method to determine fatty acid composition and MFG size of their milk samples.

Traditional methods of fatty acid determination and MFG size measurement are both slow and expensive relative to MIR spectroscopy (Fleming et al., 2017a,b). MIR spectroscopy technology used to measure the absorption spectra of milk samples is already implemented and in use, so it would be ideal if fatty acid composition and MFG size could be determined using MIR spectroscopy rather than slower and more expensive methods. This paper aims to predict fatty acid composition and MFG size of milk samples using

MIR spectroscopy.

The methods used in this paper are based on the studies performed by by Fleming et al. (2017a,b), but other regression methods are tested as well to try and improve prediction. Fleming et al. used partial least squares resgression (PLSR) to develop a prediction model for the fatty acid composition and MFG size of the milk samples. PLSR is a common regression method to use when working with spectroscopy data or data with a high number of highly correlated variables, but it has been predicted that wavelength selection prior to performing PLSR could improve regression models (Ferrand-Calmels et al., 2013). This paper compares two sparse regression methods which perform variable subset selection: the least absolute shrinkage and selection operator (LASSO), and competitive adaptive reweighted sampling-partial least squares regression (CARS-PLSR) to see if they yield any improvements in predictive ability relative to PLSR.

1.3 Statistical Theory

The following section provides a brief overview of the statistical theory behind the regression methods tested in this paper. All of the regression methods can be understood in terms of the classical linear model given by:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

Where:

$\boldsymbol{y}$ is a $n \times 1$ response vector given by

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

$\boldsymbol{X}$ is an $n \times (p+1)$ matrix of spectroscopy data, with the first column being all ones, and subsequent columns representing each of the p explanatory variables (wavelengths). Let $\mathbf{1}$ be a column wise vector consisting of n entries of 1. Then,

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_p \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}.$$

$\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of regression coefficients,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

$\boldsymbol{\epsilon}$ is an $n \times 1$ vector of residuals, $\boldsymbol{\epsilon} \sim MVN(0, \sigma^2 \boldsymbol{I}_n)$,

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Ordinary Least Squares (OLS)

Theory behind ordinary least squares (OLS) is adapted from Hastie et al. (2009).

OLS is a simple regression method that is not tested in this study, but is useful to know to understand the other regression methods tested. The idea behind OLS is to pick $\hat{\boldsymbol{\beta}}$ to minimize the residual sum of squares (RSS). i.e., minimize the Euclidean distance between the observed versus fitted vector:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2. \tag{2}$$

Equivalently,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\}. \tag{3}$$

The three other regression methods following typically offer better predictive power compared to OLS while also providing some form of model simplification, which OLS does not perform.

Partial Least Squares Regression (PLSR)

Theory behind PLSR is adapted from Hastie et al. (2009).

Spectroscopy data involves a high number of variables of which many are highly correlated. To help mitigate this issue, one strategy is to use PLSR. PLSR creates a small number of linear combinations of the original variables which can be viewed as orthogonal direction vectors in the p-dimensional input space. These direction vectors which will be referred to as "components" look for the directions that have high correlation with the response and for which the input matrix $\boldsymbol{X}$ displays high variance. The first component can be seen as the most informative component as it is the component which best satisfies the aforementioned conditions of displaying high variance and being highly correlated with the response. All subsequent components satisfy the conditions to a lesser degree than the previous and can be seen as less informative than the previous.

PLSR first computes

$$\hat{\varphi}_{1j} = \sum_j \langle \boldsymbol{x}_j, \boldsymbol{y} \rangle \qquad \text{for } j = 1, 2, \ldots, p \tag{4}$$

which it uses to calculate the first PLS component $\boldsymbol{z}_1$, given by

$$\boldsymbol{z}_1 = \sum_j \hat{\varphi}_{1j} \boldsymbol{x}_j \qquad \text{for } j = 1, 2, \ldots, p. \tag{5}$$

The response variable $\boldsymbol{y}$ is then regressed on $\boldsymbol{z}_1$ to obtain coefficient $\hat{\theta}_1$. $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ are then orthogonalized with respect to $\boldsymbol{z}_1$. This process is repeated $m \leq p$ times to obtain m orthogonal components $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m$. The response vector $\boldsymbol{y}$ is then regressed on $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m$ to obtain the partial least squares model given by

$$\hat{\boldsymbol{y}} = \bar{y}\mathbf{1} + \sum_{i=1}^{m} \hat{\theta}_i \boldsymbol{z}_m, \tag{6}$$

where $\bar{y}$ is the mean of $\boldsymbol{y}$.

PLSR offers many advantages when compared to the OLS method. PLS often offers greater prediction power, greater stability in the presence of highly correlated variables, and can produce useful equations even when the number of explanatory variables used is greater than the number of samples used for calibration

($n < p$) (Cramer, 1993). In this study, we are interested in achieving good predictions using spectroscopy data which often contains many highly correlated variables, so PLSR is preferred over OLS.

However, while PLSR typically works well for spectroscopy data, it is not necessarily the best method. PLSR can be sensitive to a low signal to noise ratio (Cramer, 1993). When there are a high number of explanatory variables of which many of them have low correlations to the response variable, PLSR is known to fail at identifying which variables have good correlations with the response (Cramer, 1993). As previously mentioned, PLSR does not perform any variable selection, meaning that in the case of spectroscopy data, wavelengths that may provide no useful information, noise, or non-linearity are all factored into the model (Frenich et al., 1995).

One way to solve the issue of uninformative/noisy variables being included in the model is to use a sparse regression method to only keep the most useful variables in the final model. This study aims to improve prediction of fatty acid composition in milk by testing regression methods which perform variable subset selection on the data. The LASSO and CARS-PLSR both perform variable subset selection and are tested in this study.

Least Absolute Shrinkage and Selection Operator (LASSO)

Theory behind the LASSO is adapted from Hastie et al. (2009).

When working with spectroscopy data with high numbers of correlated variables, a very large positive coefficient on a variable could be cancelled by a very large negative coefficient on one of the variables it is correlated to. This is often the case when using OLS. The LASSO can be viewed as a modified version of OLS which penalizes extremely positive or negative coefficients in the model. The idea behind the LASSO is to shrink the coefficients by introducing an l_1 penalty on the beta coefficients (excluding the intercept β_0). This will introduce bias in all the non-zero coefficients but will reduce their variance, which can lead to decreased error so long as the optimal value of the tuning parameter λ is chosen. Shrinkage methods such as the LASSO help to mitigate the issue of wildly large positive coefficients being cancelled by corresponding large negative coefficients. Since the LASSO constraint function is convex, the algorithm is able to zero out variables to simultaneously perform variable subset selection. The beta coefficients in LASSO regression are estimated by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{7}$$

where $\lambda \geq 0$ is a tuning parameter which controls the amount of shrinkage the LASSO performs. The coefficients obtained are then used in the classical linear model to obtain our final regression model.

Competitive Adaptive Reweighted Sampling - Partial Least Squares Regression (CARS-PLSR)

Theory behind CARS-PLSR is adapted from Li et al. (2009).

CARS-PLSR is a regression method which uses competitive adaptive reweighted sampling to perform variable subset selection, and then performs PLSR on the data using only the selected variables. CARS-PLSR works by creating s subsets of variables in s CARS sampling runs which are compared to see which subset performs the best.

First, a PLSR model is fit to the data and the values of the PLSR coefficients $\hat{\beta}_i,\ i = 1, 2, \ldots, p$ are calculated. Next, the weights for each coefficient are calculated as

$$w_i = \frac{|\hat{\beta}_i|}{\sum_{i=1}^{p} |\hat{\beta}_i|} \qquad \text{for } i = 1, 2, \ldots, p. \tag{8}$$

For each CARS sampling run, variable elimination is performed in 2 steps. In the first step, the ratio of variables to be retained is calculated using an exponentially decreasing function, given by

$$r_i = ae^{-ki} \tag{9}$$

where:

$$a = \left(\frac{p}{2}\right)^{\frac{1}{s-1}}, \tag{10}$$

$$k = \frac{\ln\left(\frac{p}{2}\right)}{s-1}. \tag{11}$$

Recall: s is the number of CARS sampling runs to be performed.

The ratio r_i calculated is used to determine the number of variables to retain in the first step of variable selection. The variables with the largest weights are retained while the rest are dropped.

In the second step of variable selection, adaptive reweighted sampling is used to further eliminate variables. Monte Carlo sampling runs are performed where the remaining variables are sampled with replacement, with each variables weight w_i determining its probability of being sampled. At the end of this step, variables that have not been sampled are eliminated and the final subset of variables is obtained for that CARS sampling run.

A PLSR model is then fit to the data using only the selected variables for the particular CARS sampling run and the root mean squared error of cross-validation (RMSECV) of that model is calculated and stored. This process is repeated in s CARS sampling runs to obtain s subsets of variables. Note that for each CARS sampling run, the value of i increases, causing the ratio of variables to retain in the first step of subset selection to become smaller. The subset of variables corresponding to the lowest RMSECV is chosen and used for the final PLS model obtained in CARS-PLSR.

Coefficient of Determination of Cross-Validation (R^2_{cv})

The coefficient of determination of cross-validation, R^2_{cv}, was used to evaluate the predictive ability of the models created in the study. High R^2_{cv} values (values close to 1) are considered an indicator of high predictive ability (Shen et al., 2016). In this study, we are interested in achieving R^2_{cv} above 2 certain thresholds: Soyeurt et al. (2011) proposed that regression models that achieve $R^2_{cv} > 0.95$ could be used for milk payment systems, while $R^2_{cv} > 0.75$ could be useful for genetic selection/breeding purposes. Furthermore, other studies have demonstrated that even with low R^2_{cv} values, high genetic correlation between observed and predicted values could be used to effectively improve milk coagulation properties (Cecchinato et al., 2009).

R^2_{cv} is calculated as

$$R^2_{cv} = 1 - \frac{SSECV}{SS_{tot}} \tag{12}$$

where:

$$SSECV = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

$\hat{y}_i$ are the predicted values obtained for each observation in cross-validation, and $\bar{y}$ is the mean of all the observed values of $\boldsymbol{y}$.

1.4 Data

For a full description on how the data used in this study was obtained for the fatty acid and MFG data, refer to Fleming et al. (2017a,b). The following sections provide a brief overview of the datasets and how they were obtained and edited.

Fatty Acid Data

CanWest DHI (Guelph, ON, Canada) and Valacta (Ste-Anne-de-Bellevue, QC, Canada) collected milk samples during routine recordings between March 2013 and October 2014 for CanWest DHI, and between December 2013 and May 2015 for Valacta. Milk sampling took place at a total 44 herds, sampling approximately 10 cows per herd, with roughly half in mid lactation and half at the beginning of lactation. Ayrshire, Brown Swiss, Holstein, and Jersey cows from Alberta, Ontario, and Quebec were all included in the sampling. Many cows were sampled multiple times during a lactation or during subsequent lactations. A portion of each milk sample was removed and ran through a spectrometer as per usual recordings, which recorded the absorption of each samples at 1060 data points in the MIR region from 5010-926 cm^{-1}. The rest of each milk sample was taken to the University of Guelph to measure its fatty acid concentration using gas chromatography. Fatty acid determination was completed on 2,064 milk samples from 374 cows.

Data preprocessing was performed to best recreate the steps done in Fleming et al. (2017a). Regions from 3105-3444 cm^{-1} as well as 1628-1658 cm^{-1} in the MIR spectra exhibited high noise due to the absorption of water and were therefore left out of the analysis. 862 wavelengths in the MIR region were left after removal. Individual fatty acid observations expressed in g/100 g of fat greater than or equal to 5 standard deviations away from the mean were considered outliers and removed. The entire record was removed if any of the fatty acid measurements in the record were considered outliers. Fatty acid measurements expressed in g/100 g of fat comprised the "Fat" dataset. The observations expressed in g/100 g of fat were multiplied by the fat percentage measurement of the milk sample to obtain fatty acid measurements expressed in g/100 g of milk. These observations comprised the "Milk" dataset. Fatty acids expressed in the Milk dataset showed more variation than those in the Fat dataset due to differences in the fat content of each of the milk samples. Fatty acid data was skewed for many of the fatty acids when expressed in g/100 g of milk. To try and make these distributions more Gaussian, the natural logarithm of the fatty acid measurements expressed in g/100 g of milk was taken after adding 1 to all of the observations to ensure that observations of 0 would not be undefined. These transformed observations comprised the "Ln" dataset. As per Fleming et al. (2017a), a random sampling procedure performed on the Ln dataset was implemented to obtain a subset of the Ln calibration set with a more uniform distribution of the response. These observations comprised the "Subset (Ln)" dataset. More information on the sampling procedure is given under the Methods section.

After data processing, there were 2015 observations in the Fat dataset, and 1903 observations in the Milk and Ln datasets. The spectra of all 2015 observations in the Fat dataset with noisy regions removed is shown in Figure 1.

Fatty acid analysis was performed on 22 individual fatty acids and 7 fatty acid groups. Fatty acid groups included saturated (SFA), unsaturated (UFA), monounsaturated (MUFA), polyunsaturated (PUFA), long-chain (LC), medium-chain (MC), and short chain (SC). Fatty acids with 4-10 carbon atoms were considered short-chain, 11-16 carbons were considered medium-chain, and 17-22 carbons were considered long-chain.

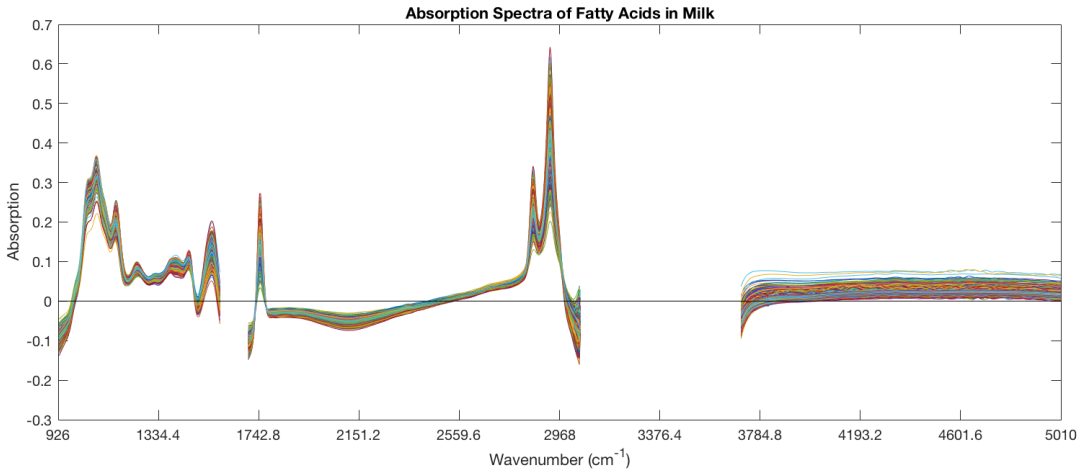

Figure 1: The absorption spectra of the 2015 samples used for fatty acid prediction model development in the study.

Milk Fat Globule Data

The same milk samples were used to create the MFG dataset as for the fatty acid dataset. The absorption spectra of each sample was recorded at 1060 wavelengths in the MIR region from 5010-926 cm^{-1}. A portion of each sample was taken to the University of Guelph to measure the mean MFG size by integrated light scattering for each record. Mean MFG diameter was measured as the volume moment mean (D[4,3]) and surface moment mean (D[3,2]), defined as:

$$D[k,z] = \frac{\sum N_i d_i^k}{\sum N_i d_i^z}, \tag{13}$$

where N_i is the number of globules in a size class of d_i.

MFG measurement was completed on 2,083 samples from 392 cows. Data preprocessing was performed to best recreate the steps done in Fleming et al. (2017b). Regions from 3105-3444 cm^{-1} as well as 1628-1658 cm^{-1} in the MIR spectra exhibited high noise due to the absorption of water and were therefore left out of the analysis. 862 wavelengths in the MIR region were left after removal. One record was removed because its D[3,2] measurement was greater than its D[4,3] measurement, and it was assumed to be a recording error. Records which had D[4,3] or D[3,2] measurements greater than 4 standard deviations away from their respective means were removed. After record removal, there were 2070 records with MFG measurements recorded.

1.5 Analysis Per Fleming et al. (2017a,b)

Fleming et al. (2017a,b) performed PLSR on 22 individual fatty acids, 7 fatty acid groups, and 2 MFG measurements. Each fatty acid/fatty acid group involved 4 separate regressions: one for each of the Fat, Milk, Ln, and Subset (Ln) datasets. For predicting MFG size using PLSR, Fleming et al. (2017b) achieved R^2_{cv} values of 0.51 and 0.54 for D[4,3] and D[3,2], respectively. Results for fatty acid predictions are given in Table 1.

	R^2_{cv}			
Fatty Acid	Fat	Milk	Ln	Subset (Ln)
C4:0	0.32	0.66	0.66	**0.73**
C6:0	0.18	0.38	0.37	**0.46**
C8:0	0.21	0.37	0.39	**0.40**
C10:0	0.52	0.66	0.67	**0.75**
C11:0	0.13	**0.21**	0.20	0.20
C12:0	0.61	0.71	0.72	**0.76**
C13:0	0.14	0.19	**0.36**	0.14
C14:0	0.60	0.80	0.80	**0.85**
C14:1	0.48	0.61	0.61	**0.68**
C15:0	0.42	0.61	0.61	**0.67**
C16:0	0.64	0.86	0.86	**0.91**
C16:1	0.39	0.62	0.63	**0.66**
C17:0	0.17	0.53	0.52	**0.58**
C17:1	0.14	0.31	0.30	**0.43**
C18:0	0.58	0.73	0.73	**0.80**
C18:1n-9 trans	0.55	0.60	0.61	**0.63**
C18:1n-9 cis	0.69	0.79	0.78	**0.83**
C18:2n-6 trans	0.14	**0.17**	0.14	0.13
C18:2n-6 cis	0.58	0.62	0.62	**0.68**
C18:3n-3	0.53	0.58	0.58	**0.61**
C18:2 cis-9, cis-12	0.62	0.65	0.65	**0.72**
C22:6n-3	0.16	**0.22**	0.21	0.16
Fatty Acid Group				
SFA	0.76	0.94	0.93	**0.96**
MUFA	0.75	0.84	0.83	**0.88**
PUFA	0.55	0.66	0.65	**0.72**
UFA	0.75	0.84	0.83	**0.87**
Short-chain	0.42	0.72	0.73	**0.78**
Medium-chain	0.72	0.90	0.89	**0.92**
Long-chain	0.72	0.83	0.81	**0.85**

Table 1: Results of PLSR performed by Fleming et al. (2017a). Bolded values represent the highest R^2_{cv} value for that particular fatty acid or fatty acid group.

1.6 Methods

For all 22 fatty acids, 7 fatty acid groups, and 2 MFG measurements, PLSR, the LASSO, and CARS-PLSR were performed. PLSR was included in this study to ensure that results were comparable to those achieved by Fleming et al. (2017a,b). LASSO and CARS-PLSR were tested to compare predictive ability and try to identify important spectral regions for prediction, since they also perform variable subset selection.

To create the Subset (Ln) datasets, steps were performed to best recreate what was implemented in Fleming et al. (2017a). According to Williams and Norris (2001), it is ideal to limit the calibration set to have roughly the same number of samples at uniform intervals of the response variable to create a more uniform distribution of the data. Calibrations made with data having a Gaussian distribution may cause future predictions to regress towards the mean of the calibration set. This is known as the "Dunne effect" (Dunne and Anderson, 1976). For calibration sets with Gaussian distribution, Williams and Norris (2001) suggest partitioning the data into 10 equally spaced bins which span the range of the data, and taking a uniform number of random samples from each of the bins to obtain a subset of the observations which are more uniformly distributed. For 10 bins, Williams and Norris (2001) suggests taking a maximum number of samples from each bin equal to 10% of the total sample size.

In this study, a similar strategy was implemented to try and obtain a more uniformly distributed calibration set. Fatty acid data in the Ln dataset was partitioned into 100 equally sized bins which spanned the range of the particular fatty acid or fatty acid group being analysed. A maximum of 18 observations (roughly 1% of the Ln dataset) were sampled uniformly at random from each bin without replacement to be included in the calibration set. If a particular bin had 18 or fewer samples, all of the samples from that bin were included. Outlier removal followed by PLSR, the LASSO, or CARS-PLSR were then performed on the resulting datasets. This process of creating a subset of the data, removing outliers, and then performing one of the regression methods was repeated 10 times for each dataset. The R^2_{cv} value and number of nonzero coefficients were taken as the mean of all the runs. The number of nonzero coefficients were rounded to the nearest whole number in this report. Identical seeds were set for every run so that the 10 Subset (Ln) calibration sets were the same for all 3 regression methods being compared. Across all of the fatty acids and fatty acid groups, there were on average 65.7 bins out of the 100 bins which included 18 or fewer samples. Therefore, all of the samples in these bins were always included in the 10 different randomized subsets used for calibration.

Outlier removal was implemented prior to performing the regression to best recreate what was done in Fleming et al. (2017a,b) for the fatty acid and MFG datasets. A PLSR model was fit to the data using 30 components and the root mean squared error (RMSE) of the fitted values were calculated for each observation. Observations with RMSE further than 3 standard deviations away from the mean were considered outliers and the corresponding record was removed from the calibration set. PLSR, LASSO, and CARS-PLSR models were then fit to the data and the results were recorded.

PLSR models were fit with a maximum of 100 components, and the number of components corresponding to the minimum RMSECV were included in the final PLSR model. Leave-one-out cross-validation was used for the Fat, Milk, and Ln datasets while 10-fold cross validation was used for the Subset (Ln) datasets. All LASSO models were fit using 10-fold cross-validation. The optimal λ value for the LASSO models were chosen as the value which minimized the RMSECV. All CARS-PLSR models were fit using 10-fold cross-validation. 50 Monte Carlo sampling runs were performed, and a maximum of 50 components were used to fit a PLSR model. The number of components achieving the lowest RMSECV was used for the final model. CARS-PLSR was performed 25 times for the Fat, Milk, and Ln datasets, and 10 times for the Subset dataset - 1 run for each of the 10 subsets created. All results were averaged.

Packages and Software

PLSR models were fit using the plsregress function in MATLAB R2019a (The MathWorks, Inc., 2019). All LASSO models were fit using the cvglmnet function of the glmnet package (Qian et al., 2013) in MATLAB R2019a. CARS-PLSR models were fit using the carspls function of the libPLS 1.98 package (Li et al., 2018) in MATLAB R2019a.

1.7 Results

For all fatty acids, fatty acid groups, and MFG measurements, CARS-PLSR achieved the highest R^2_{cv} values. LASSO regression typically performed marginally better or the same as PLSR in terms of predictive ability. On average, the optimal LASSO models selected included less than half of the 862 wavelengths in the dataset, and CARS-PLSR typically included even fewer variables in the model in comparison to LASSO regression. The results of fatty acid prediction are given in Table 2 and Table 3. Table 2 displays the results obtained using the Fat and Milk datasets, and Table 3 displays the results obtained from the Ln and Subset (Ln) datasets. Results of MFG prediction are given in Table 4. In all tests, CARS-PLSR achieves the greatest R^2_{cv} values compared to PLSR and the LASSO.

	R^2_{cv} (Number of nonzero coefficients)					
	Fat			Milk		
Fatty Acid	PLSR	LASSO	CARS-PLSR	PLSR	LASSO	CARS-PLSR
C4:0	0.33 (862)	0.33 (248)	**0.39 (102)**	0.63 (862)	0.63 (330)	**0.66 (100)**
C6:0	0.19 (862)	0.21 (306)	**0.28 (116)**	0.39 (862)	0.41 (331)	**0.46 (117)**
C8:0	0.22 (862)	0.22 (178)	**0.29 (98)**	0.38 (862)	0.39 (221)	**0.43 (94)**
C10:0	0.56 (862)	0.58 (282)	**0.62 (111)**	0.70 (862)	0.71 (296)	**0.75 (108)**
C11:0	0.17 (862)	0.17 (126)	**0.21 (86)**	0.25 (862)	0.25 (164)	**0.28 (96)**
C12:0	0.66 (862)	0.68 (205)	**0.72 (126)**	0.76 (862)	0.78 (290)	**0.81 (107)**
C13:0	0.18 (862)	0.19 (178)	**0.25 (109)**	0.26 (862)	0.28 (221)	**0.33 (97)**
C14:0	0.64 (862)	0.65 (261)	**0.69 (89)**	0.82 (862)	0.83 (427)	**0.84 (105)**
C14:1	0.44 (862)	0.45 (362)	**0.50 (115)**	0.58 (862)	0.60 (375)	**0.64 (109)**
C15:0	0.44 (862)	0.45 (367)	**0.49 (103)**	0.63 (862)	0.64 (379)	**0.67 (105)**
C16:0	0.62 (862)	0.64 (437)	**0.69 (122)**	0.86 (862)	0.87 (444)	**0.88 (136)**
C16:1	0.31 (862)	0.34 (215)	**0.41 (107)**	0.59 (862)	0.60 (305)	**0.65 (97)**
C17:0	0.17 (862)	0.19 (227)	**0.25 (93)**	0.58 (862)	0.58 (296)	**0.61 (86)**
C17:1	0.21 (862)	0.22 (303)	**0.28 (84)**	0.38 (862)	0.40 (206)	**0.45 (75)**
C18:0	0.53 (862)	0.55 (456)	**0.61 (105)**	0.72 (862)	0.73 (416)	**0.77 (111)**
C18:1n-9 trans	0.57 (862)	0.58 (337)	**0.62 (132)**	0.60 (862)	0.62 (344)	**0.66 (127)**
C18:1n-9 cis	0.71 (862)	0.71 (387)	**0.75 (115)**	0.79 (862)	0.79 (383)	**0.82 (106)**
C18:2n-6 trans	0.13 (862)	0.14 (131)	**0.21 (103)**	0.20 (862)	0.21 (152)	**0.29 (94)**
C18:2n-6 cis	0.54 (862)	0.57 (354)	**0.62 (110)**	0.61 (862)	0.65 (356)	**0.68 (114)**
C18:3n-3	0.43 (862)	0.49 (350)	**0.56 (106)**	0.48 (862)	0.53 (237)	**0.60 (106)**
C18:2 cis-9, cis-12	0.59 (862)	0.60 (482)	**0.64 (114)**	0.63 (862)	0.63 (373)	**0.67 (117)**
C22:6n-3	0.21 (862)	0.21 (119)	**0.30 (81)**	0.23 (862)	0.25 (242)	**0.34 (92)**
Fatty Acid Group						
SFA	0.80 (862)	0.80 (382)	**0.82 (127)**	0.94 (862)	0.94 (423)	**0.95 (122)**
MUFA	0.76 (862)	0.76 (329)	**0.80 (118)**	0.84 (862)	0.84 (364)	**0.87 (114)**
PUFA	0.55 (862)	0.58 (375)	**0.62 (95)**	0.66 (862)	0.67 (301)	**0.71 (120)**
UFA	0.79 (862)	0.79 (439)	**0.81 (119)**	0.85 (862)	0.86 (507)	**0.88 (119)**
Short-Chain	0.38 (862)	0.40 (318)	**0.47 (122)**	0.69 (862)	0.70 (261)	**0.74 (110)**
Medium-Chain	0.75 (862)	0.75 (506)	**0.78 (129)**	0.91 (862)	0.90 (491)	**0.92 (129)**
Long-Chain	0.76 (862)	0.76 (539)	**0.79 (129)**	0.86 (862)	0.86 (542)	**0.88 (136)**

Table 2: Table of results from PLS, LASSO, and CARS-PLS regression for the Fat and Milk calibration sets. Bolded values represent the highest R^2_{cv} value between the 3 regression methods tested for each dataset of each fatty acid/ fatty acid group.

	R^2_{cv} (Number of nonzero coefficients)					
	Ln			Subset (Ln)		
Fatty Acid	PLSR	LASSO	CARS-PLSR	PLSR	LASSO	CARS-PLSR
C4:0	0.62 (862)	0.62 (302)	**0.65 (109)**	0.70 (862)	0.71 (251)	**0.75 (92)**
C6:0	0.38 (862)	0.39 (385)	**0.44 (113)**	0.46 (862)	0.44 (33)	**0.53 (74)**
C8:0	0.40 (862)	0.41 (251)	**0.46 (104)**	0.44 (862)	0.45 (143)	**0.52 (82)**
C10:0	0.72 (862)	0.72 (348)	**0.76 (113)**	0.77 (862)	0.72 (24)	**0.82 (77)**
C11:0	0.27 (862)	0.28 (129)	**0.31 (79)**	0.34 (862)	0.35 (118)	**0.41 (79)**
C12:0	0.78 (862)	0.79 (278)	**0.82 (110)**	0.84 (862)	0.78 (23)	**0.87 (74)**
C13:0	0.26 (862)	0.28 (197)	**0.33 (103)**	0.31 (862)	0.31 (40)	**0.40 (73)**
C14:0	0.81 (862)	0.82 (353)	**0.84 (92)**	0.84 (862)	0.85 (409)	**0.88 (85)**
C14:1	0.59 (862)	0.60 (372)	**0.64 (104)**	0.67 (862)	0.64 (29)	**0.72 (94)**
C15:0	0.65 (862)	0.66 (431)	**0.69 (93)**	0.71 (862)	0.73 (288)	**0.78 (86)**
C16:0	0.85 (862)	0.86 (447)	**0.87 (123)**	0.88 (862)	0.89 (370)	**0.91 (97)**
C16:1	0.59 (862)	0.60 (266)	**0.64 (116)**	0.62 (862)	0.65 (280)	**0.71 (104)**
C17:0	0.58 (862)	0.59 (320)	**0.61 (90)**	0.66 (862)	0.63 (21)	**0.71 (83)**
C17:1	0.29 (862)	0.31 (277)	**0.36 (75)**	0.56 (862)	0.54 (30)	**0.66 (79)**
C18:0	0.67 (862)	0.69 (403)	**0.74 (97)**	0.71 (862)	0.74 (297)	**0.79 (68)**
C18:1n-9 trans	0.60 (862)	0.62 (388)	**0.65 (131)**	0.64 (862)	0.67 (282)	**0.72 (106)**
C18:1n-9 cis	0.74 (862)	0.74 (356)	**0.78 (94)**	0.81 (862)	0.77 (22)	**0.86 (87)**
C18:2n-6 trans	0.17 (862)	0.19 (147)	**0.26 (86)**	0.34 (862)	0.36 (53)	**0.47 (84)**
C18:2n-6 cis	0.59 (862)	0.62 (362)	**0.66 (121)**	0.70 (862)	0.61 (31)	**0.77 (74)**
C18:3n-3	0.51 (862)	0.56 (227)	**0.63 (101)**	0.59 (862)	0.62 (222)	**0.69 (105)**
C18:2 cis-9, cis-12	0.63 (862)	0.63 (507)	**0.67 (123)**	0.69 (862)	0.70 (249)	**0.75 (93)**
C22:6n-3	0.26 (862)	0.27 (284)	**0.36 (104)**	0.20 (862)	0.20 (45)	**0.39 (76)**
Fatty Acid Group						
SFA	0.94 (862)	0.94 (436)	**0.95 (129)**	0.95 (862)	0.95 (259)	**0.96 (95)**
MUFA	0.81 (862)	0.81 (292)	**0.84 (102)**	0.86 (862)	0.80 (19)	**0.89 (86)**
PUFA	0.65 (862)	0.66 (349)	**0.70 (110)**	0.70 (862)	0.72 (282)	**0.76 (83)**
UFA	0.83 (862)	0.83 (331)	**0.85 (109)**	0.86 (862)	0.79 (19)	**0.90 (84)**
Short-Chain	0.71 (862)	0.72 (383)	**0.75 (109)**	0.77 (862)	0.74 (22)	**0.82 (93)**
Medium-Chain	0.90 (862)	0.90 (422)	**0.91 (117)**	0.92 (862)	0.92 (415)	**0.93 (98)**
Long-Chain	0.82 (862)	0.83 (568)	**0.85 (114)**	0.84 (862)	0.79 (20)	**0.88 (88)**

Table 3: Table of results from PLS, LASSO, and CARS-PLS regression methods for the Ln and Subset (Ln) datasets. Bolded values represent the highest R^2_{cv} value between the 3 regression methods tested for each dataset of each fatty acid/ fatty acid group.

	R^2_{cv} (Number of nonzero coefficients)		
MFG Size	PLSR	LASSO	CARS-PLSR
D[4,3]	0.51 (862)	0.52 (335)	**0.57 (45)**
D[3,2]	0.54 (862)	0.55 (193)	**0.59 (41)**

Table 4: Coefficient of determination (R^2_{cv}) values obtained from PLS, LASSO, and CARS-PLS regression methods on the MFG size data. Bolded values represent the highest R^2_{cv} value between each of the 3 regression methods tested.

1.8 Discussion

As previously observed, PLSR does not perform variable selection and therefore included all 862 wavelengths/variables in the model, while LASSO regression typically eliminated more than half of the variables, and CARS-PLSR typically eliminated even more variables than LASSO regression. We also observed that LASSO regression typically outperformed PLSR, and CARS-PLSR outperformed both PLSR and LASSO regression in all tests. CARS-PLSR achieved R^2_{cv} values that were on average 14% higher than PLSR while only using an average of about 12% of the full set of 862 variables/wavelengths. It is likely that of the 862 wave-

lengths analysed, many of them exhibited noise and/or were uninformative for predictive purposes. It then could be hypothesized that CARS-PLSR typically did a better job of eliminating those noisy/uninformative variables from the model to improve prediction, and LASSO regression may not have eliminated enough of the noisy/uninformative variables to achieve the same predictive power as CARS-PLSR.

Results obtained for PLSR are comparable but not exactly the same as that obtained by Fleming et al. (2017a,b). Slight differences in the findings were expected, since Fleming et al. included a maximum of 30 components in the PLSR models, whereas this study included a maximum of 100 components. Furthermore, this study selected the optimal number of components to include in the model as the number of components that minimized the SSECV, whereas the analysis done by Fleming et al. selected the optimal number of components by van der Voet's randomization-based model comparison criteria (van der Voet, 1994). Randomization procedures involved in 10-fold cross validation and sampling for creating the subset data could also contribute to any discrepancies between the two analyses. It is also possible that more test day information was found for the dataset used by Fleming et al. (2017a,b).

We must take into consideration that fatty acid determination was only measured above a certain limit of detection (LOD). Because of this, many of the fatty acid datasets included observations below the LOD and were therefore recorded as 0. Table 5 displays all of the fatty acids which had left-censored observations in their calibration sets. It is possible that better models could be developed for fatty acids which typically are present in concentrations close to or lower than the LOD by either implementing a method to properly deal with left-censored observations, or by using equipment with a lower LOD to measure the fatty acid concentrations.

Fatty Acid	**# of observations**	**Highest R^2_{cv}***	**Dataset**
C6:0	3	0.46	Milk
C11:0	55	0.31	Ln
C13:0	413	0.33	Milk, Ln
C14:1	2	0.64	Milk, Ln
C17:1	281	0.45	Milk
C18:2n-6 trans	241	0.29	Milk
C18:3n-3	12	0.63	Ln
C18:2 cis-9, cis-12	6	0.67	Milk, Ln
C22:6n-3	1595	0.36	Ln

Table 5: Table of the number of left censored observations in the calibration set for all fatty acids affected. The highest R^2_{cv} values for all fatty acids in the table were achieved using CARS-PLSR on their respective dataset(s) listed in the table. *Note that the Subset (Ln) dataset was not considered for the highest R^2_{cv} value in this table as the process of creating the subsets often removed many if not all of the left-censored observations.

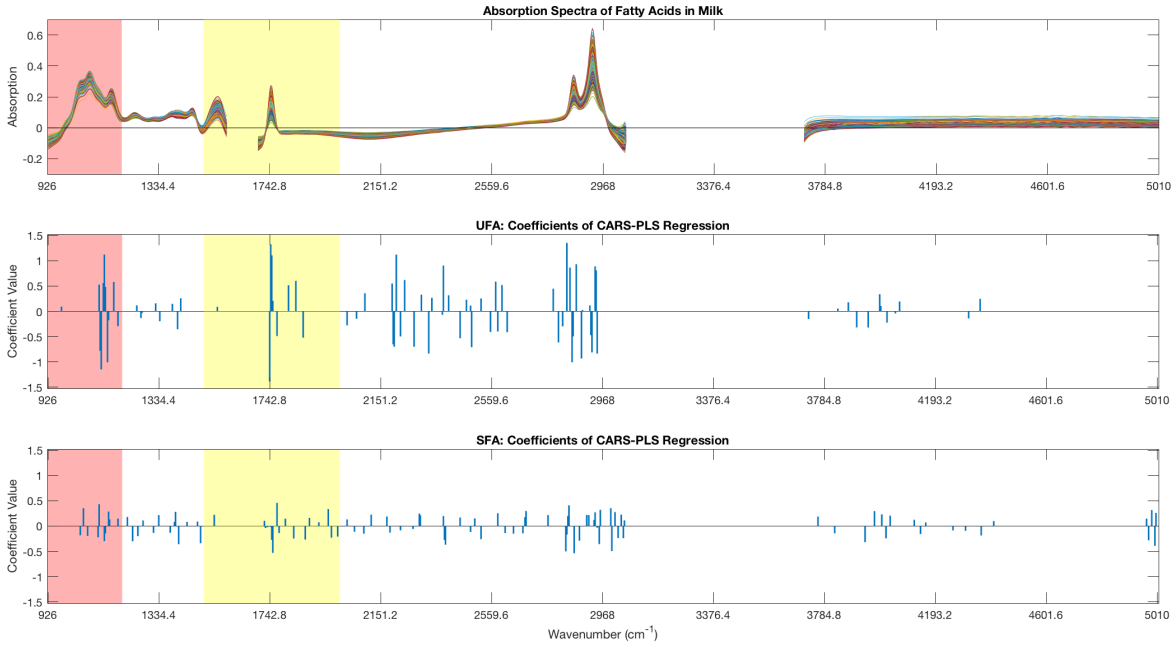

Figure 2: Plot of the regression coefficients for the CARS-PLS model on the Ln dataset for SFA and UFA groups. The region highlighted in yellow represents the C=C absorption region (Stuart, 2015), and the region highlighted in red represents the C−C absorption region (Naumann, 2001). Unsaturated fatty acids contain C=C bonds and saturated fatty acids do not, which could help explain why the regression coefficients for UFA in the C=C region are fewer, but larger in magnitude and less spread out in comparison to SFA.

1.9 Future Models to Consider

Three different regression methods were tested in this study: PLSR, the LASSO, and CARS-PLSR. It was observed that CARS-PLSR offered the greatest prediction power out of the 3 methods tested, but there are other regression methods not tested in this study that have potential to offer even greater predictive power.

Stability Competitive Adaptive Reweighted Sampling-Partial Least Squares Regression (SCARS-PLSR)

Zheng et al. (2012) proposed a modification to the CARS-PLSR algorithm used in this study named stability competitive adaptive reweighted sampling-partial least squares regression (SCARS-PLSR), which has potential to improve upon the predictive ability of CARS-PLSR. The main difference between CARS-PLSR and SCARS-PLSR is that variable selection in the SCARS algorithm is based on the index of stability of each variable, rather than the absolute value of each variable. The index of stability in SCARS-PLSR is defined as the absolute value of the regression coefficient divided by its standard deviation. SCARS-PLSR showed to have better predictive ability in cross validation than CARS-PLSR for multiple tests using spectroscopic data on organic samples to predict features of its composition (Zheng et al., 2012).

Sampling Error Profile Analysis-LASSO (SEPA-LASSO)

Another regression method to consider is sampling error profile analysis combined with the LASSO (SEPA-LASSO), proposed by Zhang et al. (2018). The SEPA-LASSO uses many loops involving Monte Carlo sampling and least angle regression to develop many LASSO sub-models of the same dimension. A vote rule is then implemented to determine which of the variables are the most important, and that is used to create different subsets of important variables. The algorithm then uses the error profile of each of the subsets of variables obtained to determine which is the optimal subset (Zhang et al., 2018). In the analysis performed by Zhang et al., SCARS was able to generate better predictions than PLSR, LASSO, and SCARS-PLSR

(mentioned above) when working with spectroscopic data for the purpose of predicting composition of organic materials.

Doubly Sparse Regularized Regression Incorporating Graphical Structure Among Predictors (DSRIG)

Doubly sparse regularized regression incorporating graphical structure among predictors (DSRIG), proposed by Stephenson (2018), is a regularized model that takes into account the undirected graph structure of the predictor variables included in the calibration when creating the model. DSRIG recognizes the groups created by the graph structure of the predictor variables and encourages sparsity both within and among the predictor groups to create a sparse regression model (Stephenson, 2018). This regression method is highly flexible and creates sparse models which have been shown to have better predictive performance compared to OLS and the LASSO (Stephenson, 2018).

Convolutional Neural Networks

The deep learning algorithms known as convolutional neural networks have also been applied to spectroscopy data and have shown to be effective for predictive purposes. One study compared convolutional neural networks to PLSR in the prediction of soil properties from raw soil spectra and found that convolutional neural networks decreased error by 87% when compared to PLSR (Padarian et al., 2019). However, convolutional neural networks are data-hungry algorithms and would likely require many more samples than what was used in this study in order to offer significant improvements in predictive performance (Padarian et al. (2019) used roughly 20,000 samples in their analysis). In the future, if absorption spectra and fatty acid composition/MFG size are obtained for many more samples, convolutional neural networks could potentially offer significant improvements in prediction power compared to those tested in this paper.

2 Heritability Estimation of Mid-Infrared Predicted Bovine Milk Sample Composition

2.1 Abstract

The purpose of this study was to estimate the heritability of 5 fatty acid group concentrations and 2 MFG size measurements in bovine milk samples. This study is based on the methods performed by Fleming et al. (2018). The concentration of 5 fatty acid groups (saturated (SFA), unsaturated (UFA), monounsaturated (MUFA), polyunsaturated (PUFA), long-chain (LC), medium-chain (MC), and short chain (SC)) on a per milk basis as well as the volume moment mean (D[4,3]) and surface moment mean (D[3,2]) of 49,127 milk samples were predicted from their MIR spectra recordings using CARS-PLSR prediction equations. Predicted fatty acid concentrations and MFG size measurements were used in a univariate mixed effects animal model to estimate the variance of each random component. Heritability of each fatty acid group and MFG size were estimated based on the variances of the random components in the model. Heritability estimates suggested that saturated fatty acids are more heritable than unsaturated fatty acids, short- and medium-chain fatty acids are more heritable than long-chain fatty acids, and D[3,2] is more heritable than D[4,3]. These results are consistent with those found by Fleming et al. (2018), but the estimated heritabilites in this study were lower.

2.2 Introduction

Fatty acid composition and MFG size of milk samples is important for determining the nutritious properties of milk, monitoring bovine health, and determining the technological properties of the milk (Fleming et al., 2017a,b). Fat composition of milk plays a large role in both determining nutritious qualities of the milk and also determining which consumer goods can be made with the milk. It follows that breeders would like to select for cows which produce milk with desirable qualities. For example, breeders interested in selling their milk to be used for cheese production would likely want to select for a certain MFG size, because the surface material of MFGs play an important role in the finished product of cheeses (Marie-Caroline Michalski et al., 2003). Fatty acid composition of milk can be influenced by both the cow's diet and the cows genetics (Fleming et al., 2018). MFG size is also influenced by the cow's diet, and it was discovered that variation in MFG diameter can exist within individuals in a herd, suggesting that MFG size is a trait that could possibly be selected for (Fleming et al., 2017c; Logan et al., 2014).

In order to implement a genetic selection program for fatty acid composition and/or MFG size, there must exist a fast way to cheaply and accurately determine the composition of many bovine milk samples. The previous section investigated using MIR spectroscopy to predict individual and group fatty acid concentrations as well as MFG size of milk samples. The prediction equations developed in the previous section were implemented in this study to predict the composition of a large number of milk samples based on their recorded MIR spectra, which were then used to estimate the heritability of certain fatty acid groups and MFG sizes. This study uses univariate animal models including fixed and random effects to estimate the heritability of each fatty acid group and MFG size studied.

2.3 Animal Model Overview

Theory behind animal models is adapted from de Villemereuil (2012).

Phenotypic traits of living organisms (such as fatty acid composition of mammal's milk as studied in this paper) exhibit variance in the population. The phenotypic variance of any given trait can be decomposed as the sum of the genetic and environmental variance within the population. The genetic variance can be further decomposed into different components. For our purposes, we will assume 2 components contribute to the total genetic variance: an additive component and a non-additive component. The additive component accounts for the additive effects of transmitted alleles, while the non-additive component accounts for other effects such as dominance effects and epistasis. Knowing this, we can write

$$V_P = V_A + V_{NA} + V_E, \tag{14}$$

where V_P represents the total phenotypic variance in the population, V_A represents the additive genetic variance, V_{NA} represents the non-additive genetic variance, and V_E represents the environmental variance. The additive genetic variance, V_A, is the component of the total phenotypic variability which can be transmitted to descendants in a population. Natural selection relies on a portion of the total phenotypic variation being transmittable.

From here we can discuss heritability. Heritability (h^2) is the proportion of phenotypic variability in a given population that can be explained by the genetic variation between members of the population. Based on this definition, heritability can be seen as a measure of the likelihood that the phenotypic variability within a population will be transferred to its offspring. Mathematically we can write this as:

$$h^2 = \frac{V_A}{V_P}, \qquad 0 \leq h^2 \leq 1. \tag{15}$$

Animal models are mixed models which can include many fixed and/or random factors to try and reduce bias introduced by non-additive genetic variance. In order to accurately estimate heritability in our model, we aim to include fixed and/or random factors to try to account for as much of the non-additive genetic variance (V_{NA}) as possible, so that it does not get mistaken for additive genetic variance (V_A). Animal models use pedigree files of the the whole population to try and account for phenotypic resemblance between samples taken from siblings. Another strategy to try and account for non-additive genetic variance is to include animal identification numbers in the model as a random effect since samples taken from the same animal are likely to be phenotypically similar.

Animal models are typically of the form:

$$\boldsymbol{y} = \sum_{i=1}^{m} \boldsymbol{X}_i \boldsymbol{c}_i + \sum_{j=1}^{l} \boldsymbol{Z}_j \boldsymbol{d}_j + \boldsymbol{\epsilon}, \tag{16}$$

where $\boldsymbol{y}$ is a vector of the response, $\boldsymbol{c}_i,\ i = 1, ..., m$ are vectors of fixed class effects with corresponding incidence matrices $\boldsymbol{X}_i,\ i = 1, ..., m$, $\boldsymbol{d}_j,\ j = 1, ..., l$ are vectors of random class effects with corresponding incidence matrices $\boldsymbol{Z}_j,\ j = 1, ..., l$, and $\boldsymbol{\epsilon}$ is a vector of residuals.

2.4 Data

The dataset used in this section was the same as in Fleming et al. (2018). This section provides a summary of the data collection method implemented in their study.

Data samples used in this study were collected during routine recordings of milk MIR spectra conducted at CanWest DHI in Guelph, Ontario and Valacta in Saint-Anne-de-Bellevue, Quebec. Over the period from January 2013 to January 2015, 2,053,396 records were collected from Holstein cow milk samples. Records included but were not limited to the animal's identification number, herd, calving age, days in milk, test date, and MIR absorption of its corresponding milk sample at 1,060 wavelengths. Fleming et al. then used milk fatty acid prediction equations developed using PLSR on MIR spectra to predict the fatty acid composition on a per milk basis for all the samples. Records with spectral data which were considered outliers or dissimilar to the spectra used to create the calibration equations were removed. RMSE of the standardized predictors calculated using the prediction equations developed by Fleming et al. (2018) were used to determine spectral outliers. Further record removal and data edits were performed as per Narayana et al. (2016). The final dataset used in this study contained 49,127 records obtained from 10,029 first-parity Holstein cows from 810 herds. Canadian Dairy Network also provided a pedigree file containing 76,074 cows which was used to develop the models in this study.

2.5 Methods

The prediction equations obtained from performing CARS-PLSR on the Ln dataset were used to predict the fatty acid content of the 49,127 records used in our model. The predicted composition of five fatty acid groups

(SFA, UFA, SC, MC, and LC) expressed as $\ln(g/100g\ milk+1)$ were predicted for each observation and used in the models. The prediction equations obtained from using CARS-PLSR on the MFG size data were used for prediction of D[4,3] and D[3,2] in the milk samples. Soyeurt et al. (2011) suggested that prediction models achieving $R^2_{cv} > 0.75$ could be used for genetic selection. The prediction models used achieved R^2_{cv} values of 0.95, 0.85, 0.75, 0.91, and 0.85 for the SFA, UFA, SC, MC, and LC groups, respectively. For the MFG data, the prediction models achieved R^2_{cv} values of 0.57 and 0.59 for D[4,3] and D[3,2], respectively. However, it has been suggested that if there exists high genetic correlation between observed and predicted values, models with $R^2_{cv} < 0.75$ could still be used to effectively improve milk coagulation properties (Cecchinato et al., 2009).

Heritability estimates of the predicted fatty acid contents were obtained through fitting the data to univariate linear animal models. One model was fit for each of the 5 fatty acid groups considered in this study. The models were as follows:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}_m\boldsymbol{m} + \boldsymbol{X}_h\boldsymbol{h} + \boldsymbol{Z}_c\boldsymbol{c} + \boldsymbol{Z}_p\boldsymbol{p} + \boldsymbol{Z}_a\boldsymbol{a} + \boldsymbol{\epsilon} \tag{17}$$

Where $\hat{\boldsymbol{y}}$ is a vector of the predicted group fatty acid concentration or MFG size depending on which trait was being analysed, $\boldsymbol{m}$ is a vector of fixed class effects for days in milk, $\boldsymbol{h}$ is a vector of fixed class effects for herd-test day, $\boldsymbol{c}$ is a vector of random class effects for herd-calving age, $\boldsymbol{p}$ is a vector of random class effects for permanent environment, $\boldsymbol{a}$ is a vector of random class effects for additive genetic effect of the animal, and $\boldsymbol{\epsilon}$ is a vector of random errors. $\boldsymbol{X}_m, \boldsymbol{X}_h, \boldsymbol{Z}_c, \boldsymbol{Z}_p$, and $\boldsymbol{Z}_a$ are all matrices which assign observations to effects.

The variance of each of the random effect terms in the model were estimated with the animal model and used to estimate the heritability of each trait and the repeatability of the results. Heritability was calculated as

$$h^2 = \frac{\sigma^2_a}{\sigma^2_c + \sigma^2_p + \sigma^2_a + \sigma^2_\epsilon}, \tag{18}$$

where σ^2_c is the herd-calving age variance, σ^2_p is the permanent environmental variance, σ^2_a is the additive genetic variance, and σ^2_ϵ is the residual variance. Repeatability was calculated as

$$r = \frac{\sigma^2_c + \sigma^2_p + \sigma^2_a}{\sigma^2_c + \sigma^2_p + \sigma^2_a + \sigma^2_\epsilon}. \tag{19}$$

Packages and Software

All animal models were fit using the Average Information REstricted Maximum Likelihood (AI-REML) algorithm implemented in the DMU software package (Madsen et al., 2006).

2.6 Results

The estimated variances for each random component in the model are given in Table 6 and the estimated heritabilities of each fatty acid group and MFG measurement are given in Table 7. Estimated heritabilities are slightly lower that found by Fleming et al. (2018, 2017c).

Fatty Acid Group	σ_c^2	σ_p^2	σ_a^2	σ_ϵ^2
SFA	1.14×10^{-4}	3.76×10^{-3}	7.40×10^{-3}	7.71×10^{-3}
MUFA	7.16×10^{-4}	1.26×10^{-3}	9.05×10^{-3}	3.86×10^{-2}
PUFA	5.91×10^{-5}	1.66×10^{-3}	4.35×10^{-3}	1.06×10^{-2}
UFA	5.21×10^{-4}	1.16×10^{-3}	5.16×10^{-3}	2.96×10^{-2}
Short-Chain	1.51×10^{-4}	3.79×10^{-3}	8.91×10^{-3}	1.40×10^{-2}
Medium-Chain	6.87×10^{-5}	2.48×10^{-3}	4.60×10^{-3}	6.27×10^{-3}
Long-Chain	4.72×10^{-4}	1.94×10^{-3}	1.04×10^{-2}	3.77×10^{-2}
MFG Size				
D[4,3]	1.14×10^{-3}	4.35×10^{-2}	1.04×10^{-1}	1.36×10^{-1}
D[3,2]	2.72×10^{-4}	1.80×10^{-2}	3.92×10^{-2}	4.23×10^{-2}

Table 6: Herd-calving age, permanent environment, additive genetic, and residual variance estimates of each of the fatty acid groups and MFG measurements analysed.

Fatty Acid Group	h^2	r
SFA	0.390	0.594
MUFA	0.182	0.222
PUFA	0.262	0.365
UFA	0.142	0.188
Short-Chain	0.332	0.479
Medium-Chain	0.343	0.533
Long-Chain	0.206	0.254
MFG Size		
D[4,3]	0.366	0.523
D[3,2]	0.393	0.576

Table 7: Table of estimated heritability of fatty acid groups and MFG measurements along with repeatability estimates of the results.

2.7 Discussion

Biologically, it would be expected that short- and medium-chain fatty acids would be more heritable than long-chain fatty acids in milk, because short- and medium-chain fatty acids are synthesized de novo in cows whereas long-chain fatty acids are dietary and adipose derived (Fleming et al., 2018). This is consistent with the heritabilities found in this study for SC, MC, and LC fatty acids, as well as the analysis performed by Fleming et al. (2018). Furthermore, both this study and that performed by Fleming et al. (2018, 2017c) found SFA to be more heritable than UFA, and D[3,2] to be more heritable than D[4,3]. However, the heritabilities estimated in this study are all lower than that estimated by Fleming et al. MUFA and PUFA were not included in the analysis performed by Fleming et al (2018), but these findings suggest that PUFA is more heritable than MUFA.

It is important to highlight that prediction equations for D[4,3] and D[3,2] achieved R^2_{cv} values of 0.57 and 0.59, respectively, which are lower than the genetic selection threshold of 0.75 proposed by Soyeurt et al. (2011). This suggests that the prediction equations developed for MFG size are likely not adequate for genetic selection purposes, unless there is a high genetic correlation between observed and predicted values (Cecchinato et al., 2009). This study does not look into said genetic correlations.

Note that discrepancies between this paper and Fleming et al. (2018, 2017c) can be attributed to using different animal models run on different software, as well as differences in predicted milk composition properties due to different prediction equations used. Further studies could investigate how the CARS-PLSR prediction equations developed in this study affect the heritabilites of the various milk properties when the

data is used in an animal model of the same form as that used in Fleming et al (2018, 2017c). Improvements to this animal model could be made by first investigating other regression methods mentioned previously that have potential to offer more accurate predictions, and if significant improvements are made, using the new prediction equations to generate a new predicted dataset to use in an animal model. Alternatively, if MIR prediction does not offer sufficient accuracy, traditional methods of fatty acid and MFG size determination could be used to obtain much more accurate measurements for each milk sample which will help reduce error in the dataset used for running the animal models. However, this is unlikely due to the impracticality of traditional methods.

References

Alessio Cecchinato, Massimo De Marchi, Lizeth Gallo, Giovanni Bittante, and Paolo Carnier. Mid-infrared spectroscopy predictions as indicator traits in breeding programs for enhanced coagulation properties of milk. *J. Dairy Sci.*, 92:5304–13, 2009.

Richard D. Cramer. Partial least squares (pls): Its strengths and limitations. *Perspectives in Drug Discovery and Design*, 1(2):269–278, 1993.

Pierre de Villemereuil. Estimation of a biological trait heritability using the animal model: How to use the mcmcglmm r package. 10 2012.

William Dunne and J. Ansel Anderson. A system for segregating canadian wheat into subgrades of guaranteed protein content. *Canadian Journal of Plant Science*, 56(3):433–450, 1976.

M Ferrand-Calmels, I Palhière, Mickaël Brochard, O Leray, J.M. Astruc, M.R. Aurel, S Barbey, F Bouvier, P Brunschwig, H Caillat, M Douguet, F Faucon-Lahalle, Marine Gelé, G Thomas, J.M. Trommenschlager, and H Larroque. Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. *J. Dairy Sci.*, 97:17–35, 2013.

A. Fleming, F. S. Schenkel, J. Chen, V. Malchiodi, V. Bonfatti, R. A. Ali, B. Mallard, M. Corredig, and F. Miglior. Prediction of milk fatty acid content with mid-infrared spectroscopy in canadian dairy cattle using differently distributed model development sets. *J. Dairy Sci.*, 100:5073–5081, 2017a.

A. Fleming, F.S. Schenkel, J. Chen, F. Malchiodi, R.A. Ali, B. Mallard, M. Sargolzaei, M. Corredig, and F. Miglior. Variation in fat globule size in bovine milk and its prediction using mid-infrared spectroscopy. *J. Dairy Sci.*, 100(3):1640 – 1649, 2017b.

A. Fleming, F.S. Schenkel, A. Koeck, F. Malchiodi, R.A. Ali, M. Corredig, B. Mallard, M. Sargolzaei, and F. Miglior. Heritabilities of measured and mid-infrared predicted milk fat globule size, milk fat and protein percentages, and their genetic correlations. *J. Dairy Sci.*, 100(5):3735 – 3741, 2017c.

A. Fleming, F.S. Schenkel, F. Malchiodi, R.A. Ali, B. Mallard, M. Sargolzaei, J. Jamrozik, J. Johnston, and F. Miglior. Genetic correlations of mid-infrared-predicted milk fatty acid groups with milk production traits. *J. Dairy Sci.*, 101(5):4295 – 4306, 2018.

A. Garrido Frenich, D. Jouan-Rimbaud, D. L. Massart, S. Kuttatharmmakul, M. Martínez Galera, and J. L. Martínez Vidal. Wavelength selection method for multicomponent spectrophotometric determinations using partial least squares. *Analyst*, 120:2787–2792, 1995.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York, NY, 2 edition, 2009.

Robert G. Jensen. The composition of bovine milk lipids: January 1995 to december 2000. *Journal of Dairy Science*, 85(2):295 – 350, 2002.

H.-D. Li, Q.-S. Xu, and Y.-Z. Liang. libpls: an integrated library for partial least squares regression and discriminant analysis. *Chemom. Intell. Lab. Syst.*, 176:34–43, 2018.

Hong-Dong Li, Yi-Zeng Liang, Qingsong Xu, and Dong-Sheng Cao. Key wavelength screening using competitive adaptive reweighted sampling method for multivariate calibration. *Analytica Chimica Acta*, 648: 77–84, 2009.

A. Logan, M. Auldist, J. Greenwood, and L. Day. Natural variation of bovine milk fat globule size within a herd. *J. Dairy Sci.*, 97(7):4072 – 4082, 2014.

Per Madsen, Guosheng Su, Rodrigo Labouriau, and Ole Christensen. Dmu – a package for analyzing multivariate mixed models. *the proceedings of the 8th World Congress on Genetics Applied to Livestock Production; Brasil*, 01 2006.

Marie-Caroline Michalski, Jean-Yves Gassi, Marie-Hélène Famelart, Nadine Leconte, Bénédicte Camier, Françoise Michel, and Valérie Briard. The size of native milk fat globules affects physico-chemical and sensory properties of camembert cheese. *Lait*, 83(2):131–143, 2003.

S.G. Narayana, F.S. Schenkel, A. Fleming, A. Koeck, F. Malchiodi, J. Jamrozik, J. Johnston, M. Sargolzaei, and F. Miglior. Genetic analysis of groups of mid-infrared predicted fatty acids in milk. *J. Dairy Sci.*, 100 (6):4731 – 4744, 2017.

Dieter Naumann. Ft-infrared and ft-raman spectroscopy in biomedical research. *Applied Spectroscopy Reviews*, 36(2-3):239–298, 2001.

J. Padarian, B. Minasny, and A.B. McBratney. Using deep learning to predict soil properties from regional spectral data. *Geoderma Regional*, 16:e00198, 2019.

J. Qian, T. Hastie, J. Friedman, R. Tibshirani, and N. Simon. Glmnet for matlab, 2013. URL `https://web.stanford.edu/~hastie/glmnet_matlab/`.

Liang Shen, Dongsheng Cao, Qingsong Xu, Xin Huang, Nan Xiao, and Yizeng Liang. A novel local manifold-ranking based k-nn for modeling the regression between bioactivity and molecular descriptors. *Chemometr. Intell. Lab.*, 151:71–77, 2016.

H. Soyeurt, F. Dehareng, N. Gengler, S. McParland, E. Wall, D.P. Berry, M. Coffey, and P. Dardenne. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. *J. Dairy Sci.*, 94(4):1657 – 1667, 2011.

Matthew Stephenson. *Doubly Sparse Regularized Regression Incorporating Graphical Structure Among Predictors*. PhD thesis, University of Guelph, 2018.

Barbara Stuart. *Infrared Spectroscopy*, pages 1–18. American Cancer Society, 2015. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/0471238961.0914061810151405.a01.pub3`.

The MathWorks, Inc. Partial least squares regression and principal components regression – matlab & simulink example (r2019a), 2019. URL `https://www.mathworks.com/help/stats/examples/partial-least-squares-regression-and-principal-components-regression.html`.

H. van der Voet. Comparing the predictive accuracy of models using a simple randomization test. *Chemom. Intell. Lab. Syst.*, 25:313–323, 1994.

T W. Keenan and Ian Mather. *Intracellular Origin of Milk Fat Globules and the Nature of the Milk Fat Globule Membrane*, volume 2, pages 137–171. 2006.

P. Williams and K. Norris. *Near-infrared technology in the agricultural and food industries*. Amer. Assn. of Cereal Chemists, St. Paul, MN, 2 edition, 2001.

Ruoqiu Zhang, Feiyu Zhang, Wanchao Chen, Heming Yao, Jiong Ge, Shengchao Wu, Ting Wu, and Yiping Du. A new strategy of least absolute shrinkage and selection operator coupled with sampling error profile analysis for wavelength selection. *Chemometr. Intell. Lab.*, 175:47–54, 2018.

Kaiyi Zheng, Qingqing Li, Jiajun Wang, Jinpei Geng, Peng Cao, Tao Sui, Xuan Wang, and Yiping Du. Stability competitive adaptive reweighted sampling (scars) and its applications to multivariate calibration of nir spectra. *Chemometr. Intell. Lab.*, 112:48 – 54, 2012.