J. Dairy Sci. 99:1–9
http://dx.doi.org/10.3168/jds.2016-10849

Variable selection procedures before partial least squares regression enhance the accuracy of milk fatty acid composition predicted by mid-infrared spectroscopy

P. Gottardo,* M. Penasa,*[1] N. Lopez-Villalobos,† and M. De Marchi*
*Department of Agronomy, Food, Natural Resources, Animals and Environment, University of Padova, Viale dell'Università 16, 35020 Legnaro (PD), Italy
†Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand

ABSTRACT

Mid-infrared spectroscopy is a high-throughput technique that allows the prediction of milk quality traits on a large-scale. The accuracy of prediction achievable using partial least squares (PLS) regression is usually high for fatty acids (FA) that are more abundant in milk, whereas it decreases for FA that are present in low concentrations. Two variable selection methods, uninformative variable elimination or a genetic algorithm combined with PLS regression, were used in the present study to investigate their effect on the accuracy of prediction equations for milk FA profile expressed either as a concentration on total identified FA or a concentration in milk. For FA expressed on total identified FA, the coefficient of determination of cross-validation from PLS alone was low (0.25) for the prediction of polyunsaturated FA and medium (0.70) for saturated FA. The coefficient of determination increased to 0.54 and 0.95 for polyunsaturated and saturated FA, respectively, when FA were expressed on a milk basis and using PLS alone. Both algorithms before PLS regression improved the accuracy of prediction for FA, especially for FA that are usually difficult to predict; for example, the improvement with respect to the PLS regression ranged from 9 to 80%. In general, FA were better predicted when their concentrations were expressed on a milk basis. These results might favor the use of prediction equations in the dairy industry for genetic purposes and payment system.

Key words: genetic algorithm, milk fatty acid, mid-infrared spectroscopy, variable selection

INTRODUCTION

Infrared technologies are fast, cheap, and largely used to determine a large number of milk characteristics (De Marchi et al., 2014; Visentin et al., 2015; McDermott et al., 2016a,b). Methods such as partial least squares (**PLS**) and principal component regression are useful to extract relevant information from the spectra and potentially build accurate and robust models using the whole spectrum (Geladi and Kowalski, 1986; Thomas and Haaland, 1990). These methods are considered almost insensitive to noise and, therefore, it is a common belief that variable selection is not necessary for models construction (Haaland and Thomas, 1988); however, in the last decade some researchers have suggested that an efficient variable selection is useful and sometimes necessary to obtain suitable prediction models for both routine and research purposes (Martens and Naes, 1989; Helland, 2001; Chun and Keleş, 2010). In fact, PLS regression is a projection-based method because it ignores the direction in the variable space where noisy and irrelevant variables are present; however, when the number of predictor variables is much greater than the number of observations, this property of the PLS estimator ceases to exist. Faber et al. (1995) studied the error propagation in principal component analysis and reported that the bias of the model is greatly influenced by the number of predictor variables and by the measurement error; if several uninformative variables are present (i.e., a model with a lot of noise) the final PLS estimators will be biased.

Prediction of concentration of individual fatty acids (**FA**) in cow milk using mid-infrared spectroscopy (**MIRS**) applying PLS is robust and precise for the major FA groups and for some individual FA, but the quality of the prediction decreases for FA that are present in low concentrations (Soyeurt et al., 2006, 2011; De Marchi et al., 2011). The application of a variable selection procedure before PLS regression could lead to

Received January 3, 2016.
Accepted July 3, 2016.
[1] Corresponding author: mauro.penasa@unipd.it

a better and less complex prediction model (Mehmood et al., 2012). The present study compared 2 different methods of variables selection, namely the uninformative variable elimination (**UVE**) procedure proposed by Centner et al. (1996) and a genetic algorithm (**GA**). The UVE procedure has been already tested with satisfactory results on titratable acidity, calcium content, and detailed protein composition of bovine milk (Gottardo et al., 2015; Niero et al., 2016); GA has been mainly used on traits predicted by near infrared spectroscopy, but also reviewed and developed for MIRS by Leardi and González (1998) and tested on milk FA composition by Ferrand et al. (2011). The main goal of the present study was to investigate which of the 2 methods is the best to select variables and thus to be practically implemented under field conditions.

MATERIALS AND METHODS

Sample Collection and MIRS Acquisition

Individual milk samples of 63 Holstein-Friesian, 24 Brown Swiss, and 25 Jersey cows from parity 1 to 7 and from 7 to 408 DIM were collected in 4 herds between February and March 2015 during the morning milking. Preservative (Bronopol, 2-bromo-2-nitropropan-1,3-diol; Grunenthal Prodotti & Farmaceutici Formenti, Milan, Italy) was immediately added to milk, which was then transferred at 4°C to the laboratory of the South Tirol Dairy Association (Bolzano, Italy) and analyzed for milk chemical composition using a MilkoScan FT6000 (Foss Electric A/S, Hillerød, Denmark). For each sample, the absorbance spectrum contained 1,060 infrared data points over the spectral range from 900 to 5,000 cm^{-1}. An aliquot of each sample was transferred to the laboratory of the Department of Agronomy, Food, Natural Resources, Animals and Environment of the University of Padova (Legnaro, Italy) for FA analysis.

Milk FA Analysis

Milk lipids were determined with accelerated solvent extraction method using Dionex ASE 350 system (Thermo Scientific, Dreieich, Germany) with petroleum ether in isopropanol (2:1) as solvent. Methyl esterification of FA was carried out according to Palmquist and Jenkins (2003) with a basic or acid reaction. Fatty acid separation and quantification were performed by a Agilent 7820A GC System equipped with an automatic sampler G4567A (Agilent Technologies, Santa Clara, CA) and flame ionization detector. The column used was a Supelco Omegawax capillary column (30 m of length, 0.25 mm of inner diameter and a film thickness of 0.25 μm; Supelco, Bellefonte, PA). Temperatures of injector and flame ionization detector were set at 250°C. Oven temperature was initially 50°C for 2 min, increased at 4°C/min to 220°C, and held for 18 min. Hydrogen was the carrier gas and its flow was set at 1 mL/min with average speed of 21 cm/s. Fatty acid standard Supelco FAME mixC4–C24 #18919–1AMP (Sigma-Aldrich, Castle Hill, Australia) was analyzed before GC analysis for FA identification. Determination of FA values was obtained using GC ChemStation software (Agilent Technologies, Santa Clara, CA) and were expressed both as total identified FA and on a milk basis. Individual FA were C4:0, C6:0, C8:0, C12:0, C14:0, C16:0, C16:1n7, C18:0, and C18:1n9, and groups of FA were SFA, UFA, MUFA, and PUFA.

Statistical Analysis

Spectral Information. Spectral data were transformed to absorbance using the $\log_{10}$ of transmittance value and the 1,060 wavenumbers were reduced to 480 through the elimination of 2 spectra regions (1,601 to 1,717 and 3,052 to 5,011 cm^{-1}) known to be related to water absorption and, thus, characterized by high noise (Hewavitharana and van Brakel, 1997). Prediction equations were derived based on PLS regression using the ChemometricsWithR package (Wehrens, 2011) and the possible presence of outliers was checked using the robust Mahalanobis distance procedure. Following this procedure, no outliers were detected in the data set. Due to the quite low spectral number of observations, we decided to perform a leave-one-out cross-validation instead of an external validation. Wavenumbers to be included in as dependent variables of the PLS regression were selected based on 2 methods, UVE and GA. Both UVE and GA were run using R (64 bit) statistical software (R Core Team, 2015). The UVE procedure was performed using a homemade script whereas GA using the rgba.bin function implemented in the genalg package of R.

UVE Procedure. The UVE procedure was proposed by Centner et al. (1996) and involves the addition of artificial noise variables to the original wavelength matrix of predictors to create a filter. All the original variables having less stability than the noisy artificial variables are eliminated. The procedure is repeated until a stop criterion is reached. The optimal number of principal components must be set for each trait before starting the PLS procedure and we decided to use 10 of them for each trait. This choice was taken to avoid possible overfitting issues due to the low number of samples in