J. Dairy Sci. 98:4168–4173
http://dx.doi.org/10.3168/jds.2014-8752

Technical note: Improving the accuracy of mid-infrared prediction models by selecting the most informative wavelengths

P. Gottardo, M. De Marchi,[1] M. Cassandro, and M. Penasa
Department of Agronomy, Food, Natural Resources, Animals and Environment, University of Padova, Viale dell'Università 16, 35020 Legnaro (PD), Italy

ABSTRACT

Mid-infrared spectroscopy (MIRS) is widely used to collect milk phenotypes at the population level. The aim of this study was to test the ability of the uninformative variable elimination (UVE) method to select and remove uninformative wavelength variables before partial least squares (PLS) analysis. Milk titratable acidity (TA) and Ca content were used as examples to illustrate the procedure. Reference values and MIRS spectra (n = 208) of TA and Ca were retrieved from an existing database. The data set was randomly divided into calibration (70% of data) and validation (30% of data) sets, and PLS analysis was carried out before and after the UVE procedure. The UVE procedure selected 244 and 113 informative wavelengths for TA and Ca, respectively, from a total of 1,060. The elimination of uninformative variables before PLS regression increased the accuracy of MIRS prediction models, and it substantially reduced the computation time. Dealing with fewer variables is expected to enhance the efficiency of MIRS models to predict phenotypes at population level.

Key words: mid-infrared spectroscopy (MIRS), partial least squares regression, uninformative variable selection, titratable acidity, calcium content

Technical Note

Recently, De Marchi et al. (2014) reviewed the application of mid-infrared spectroscopy (**MIRS**) as a rapid and cost-effective tool for recording phenotypes on a large scale. Mid-infrared spectroscopy has been used to predict milk traits such as fatty acid profile (Soyeurt et al., 2006; De Marchi et al., 2011), body energy status of the cow (McParland et al., 2012), and milk coagulation ability (De Marchi et al., 2013).

Received August 17, 2014.
Accepted February 11, 2015.
[1]Corresponding author: massimo.demarchi@unipd.it

Multivariate calibration methods such as partial least squares (**PLS**) and principal component regressions are widely used in MIRS analysis. The accuracy of a calibration model is related not only to the quality of gold-standard measurements but also to the quality of wavelength variables and the number of factors used during generation of the PLS model (Li and Jing, 2014). The MIRS spectra contain a huge number of wavelength variables, some of which contain irrelevant information for multivariate analyses. The identification and removal of uninformative spectra regions of a given trait is crucial in reducing computation time in statistical analysis, especially when dealing with large data sets (De Marchi et al., 2014). Currently, MIRS prediction models are developed using PLS regression, and only the spectral regions typically characterized by high noise are removed before PLS analysis (Hewavitharana and van Brakel, 1997; De Marchi et al., 2013).

Several methods for variable selection have been proposed to remove unnecessary or redundant wavelengths from regression models. For example, Leardi (2000) and Ferrand et al. (2011) used a genetic algorithm combined with PLS regression to reduce the number of uninformative wavelengths and improve the accuracy of predictions. Among methods of variable selection, uninformative variable elimination (**UVE**) based on the analysis of PLS regression coefficients is of particular interest (Centner et al., 1996). The aim of this study was to test the ability of the UVE approach combined with PLS regression to remove uninformative wavelengths and to evaluate the effect of this procedure on the accuracy of MIRS calibration models.

Reference values and MIRS spectra (n = 208) of titratable acidity (**TA**) and Ca content were retrieved from an existing database of milk samples that were analyzed in 2011. Titratable acidity was determined using Crison Compact D (Crison Instruments SA, Alella, Spain) and expressed in Soxhlet-Henkel degrees (°**SH**/50 mL) as proposed by Anonymous (1963) in the laboratory of the Breeders Association of Veneto region (ARAV, Padova, Italy). Calcium content was determined after mineralization with nitric acid in

closed vessels by a microwave system using inductively coupled plasma optical emission spectrometry (full details on the procedure to determine Ca can be retrieved from Toffanin et al., 2015). Spectral information was collected over the spectral range of 5,011 to 900 cm^{-1} using a Milko-Scan FT6000 (Foss, Hillerød, Denmark).

The same calibration and validation procedures were used for both traits; 10 calibration and validation sets were randomly selected from original data by including 70% records in calibration and 30% records during the validation process. Therefore, PLS was carried out 10 times using the ChemometricsWithR package (Wehrens, 2011) of R software (R Development Core Team, 2006) using an external validation. The statistical analysis consisted of generating PLS models before and after the UVE procedure. In the PLS analysis before UVE procedure, the 1,060 wavelengths were reduced to 520 through the elimination of 2 spectral regions (1,601 to 1,717 and 3,052 to 5,011 cm^{-1}) that are known to be related to water and thus characterized by a high noise level (Hewavitharana and van Brakel, 1997; De Marchi et al., 2013). The optimal number of principal components used for the evaluation of PLS models was 11 and 12 for TA and Ca, respectively; this number was selected when the lowest root mean square error of prediction ($\mathbf{RMSE_P}$) was reached. The coefficient of determination in external validation ($\mathbf{R^2_P}$), the $RMSE_P$, and the ratio performance deviation (**RPD**, calculated by dividing the SD by the $RMSE_P$ of the trait) were used to evaluate the accuracy of PLS models: in particular, an R^2_P value between 0.83 and 0.90 is considered useful for practical application (Williams, 1987), and RPD values >2 are considered adequate for analytical purposes (Sinnaeve et al., 1994; Karoui et al., 2006).

A homemade script implemented in R software (R Development Core Team, 2006) was created following Centner et al. (1996) and Chen et al. (2007). Initially, a random matrix **Z** with values between 0 and 1, and of exactly the same size as the **X** matrix, was generated. The **X** was a rectangular matrix of order 208 (number of samples) × 1,060 (number of wavelengths). The random matrix **Z** was multiplied by a small constant in order to eliminate any possible interaction with the original variables of the **X** matrix. The constant value should be lower than the level of inaccuracy of the instrument, which in our case was 10^{-10}, according to Centner et al. (1996). This step was necessary to avoid large errors during the calculation of eigenvalues. This could also affect the regression coefficients (b) of the original matrix, leading to the erroneous elimination of informative variables. The matrices **X** and **Z** were then merged into an **XZ** matrix of size $n \times 2p$, where n are the observations and p the wavelengths. A new PLS analysis was carried out on the **XZ** matrix through leave-one-out jackknife validation to obtain a matrix containing the coefficients of all the equations used in the validation. A stability criterion (E), as a standardized coefficient, was adopted, according to Centner et al. (1996) and Chen et al. (2007): $E = bj/s(bj)$, for $j = 1$ to p, where bj is the mean of the regression coefficients vector for the jth variable, and $s(bj)$ is the standard deviation of this vector. The absolute stability coefficient in the **Z** matrix was calculated and used as a filter into the original **X** matrix. The criterion for the construction of the filter was set to $Ex < \max(Ez)$, where Ex are the absolute values on matrix **X** and Ez are the absolute values on matrix **Z**. Following this criterion, if j variables from the **X** matrix exhibited a stability criterion smaller than the maximum obtained for the artificial variables, then the variables were uninformative. Finally, a new PLS model was performed using only the resulting informative wavelengths.

Means (SD) of calibration and validation sets were similar within each data set and they ranged from 3.40 to 3.44 °SH/50 mL (0.33 to 0.37) and 3.35 to 3.45 °SH/50 mL (0.28 to 0.37) for TA, respectively (Table 1), and from 1,148 to 1,164 mg/kg (129 to 135) and 1,125 to 1,188 mg/kg (123 to 133) for Ca, respectively (Table 2).

Figure 1 depicts the most informative wavelengths for TA and Ca extrapolated from the spectrum after the UVE procedure. The number of wavelengths decreased from 520 to 244 and from 520 to 113 for TA and Ca, respectively. The selected peaks were often related to specific chemical bonds: peaks occurring at 1,115, 1,146, and 1,180 cm^{-1} are related to C–O and C–C stretching; the peak at 1,331 cm^{-1} is related to C–C–H stretching; the peak at 1,240 and 1,500 cm^{-1} can be attributed to the amide group; and peaks at 2,935 and 2,839 cm^{-1} could be associated with lipids (Hewavitharana and van Brakel, 1997).

Concerning PLS models, the coefficient of determination in calibration ($\mathbf{R^2_C}$) and root mean square error of calibration ($\mathbf{RMSE_C}$) ranged from 0.70 to 0.83 and 0.13 to 0.19 °SH/50 mL for TA, respectively (Table 1), and from 0.53 to 0.65 and 76.4 to 89.8 mg/kg for Ca content (Table 2). The R^2_C and R^2_P were ≥0.70 and ≥0.50 for TA and Ca, respectively, whereas $RMSE_C$ and $RMSE_P$ were ≤0.19 °SH/50 mL for TA (Table 1) and ≤96.9 mg/kg for Ca (Table 2), confirming the ability of MIRS to predict these traits. Overall, RPD values for TA were >2, suggesting good predictive ability of the models (Sinnaeve et al., 1994; Karoui et al., 2006), and they were around 1.5 for Ca, indicating moder-