Proceedings, 10th World Congress on Genetics Applied to Livestock Production

Improving the Accuracy of Mid-infrared Prediction Models by Selecting the Most Informative Wavelengths through Uninformative Variable Elimination

M. De Marchi, P. Gottardo, M. Cassandro, and M. Penasa.
Department of Agronomy, Food, Natural resources, Animals and Environment, University of Padova, Italy

ABSTRACT: Mid-infrared spectroscopy (MIRS) is used to collect milk phenotypes at population level. The aim of this study was to test the ability of uninformative variable elimination (UVE) approach to select informative wavelengths before multivariate analysis. Reference values (n = 386) of milk titratable acidity were randomly selected from an existing database. The dataset was randomly divided into calibration (80%) and validation (20%) sets, and partial least squares (PLS) analysis was carried out before and after UVE procedure. After UVE procedure, 244 informative wavelengths were retained for subsequent PLS analysis. The elimination of uninformative variables before PLS regression led to an improvement of the accuracy of MIRS prediction models and it substantially reduced the computational time. Finally, dealing with much less variables would enhance the efficiency of MIRS models to predict phenotypes at population level.
Keywords: mid-infrared spectroscopy; milk acidity; partial least squares regression; phenotyping

Introduction

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, protein composition, coagulation ability, and body energy status of the cow (Soyeurt et al. (2006 and 2011); Bonfatti et al. (2011); De Marchi et al. (2013); McParland et al. (2012)).

The identification of the most informative spectra regions of a given trait is a crucial point to reduce computational time in statistical analysis, especially when dealing with large data (De Marchi et al. (2014)). Currently, MIRS prediction models for new milk traits (e.g., milk coagulation properties and acidity) are developed using the entire mid-infrared spectrum through multivariate models such as partial least squares (PLS) regression; only the typical spectrum regions characterized by high noise are removed before PLS analysis (Hewavitharana and Brakel (1997); De Marchi et al. (2009)). A method called uninformative variable elimination (UVE) has been proposed to reduce the number of variables in multivariate analysis (Centner et al. (1996)), so that only useful informative wavelengths are retained. The aim of this study is to test the ability of UVE approach to remove uninformative wavelengths and to evaluate the impact of this procedure on the accuracy of MIRS prediction models built through PLS analysis.

Materials and Methods

Data. Reference values and MIRS spectra (n = 386) of titratable acidity were randomly selected from an existing database of milk samples that were analyzed between 2011 and 2013 in the laboratory of the Breeders Association of Veneto region (ARAV, Padova, Italy). Titratable acidity was determined using Crison Compact D (Crison Instruments SA, Alella, Spain) and expressed in Soxhlet-Henkel degrees (°SH/50ml) as proposed by Anonymous (1963). Spectra information was collected over the spectral range of 4,000 to 900 cm^{-1} using a Milko-Scan FT6000 (Fourier transform infrared interferometer; Foss Electric A/S, Hillerød, Denmark). Reference values and spectra were edited by visual analysis, and outliers were identified following De Marchi et al. (2009) and Toffanin and De Marchi (2013).

Partial least squares (PLS) analysis. Ten calibration and validation sets were randomly selected from original data by including 80% of records (n = 309) in calibration and 20% of records (n = 77) in validation process. Therefore, PLS was carried out 10 times using R software (R Development Core Team (2006) to evaluate the robustness of the models, and an external validation was performed to validate PLS models. Before PLS analysis, the 1,060 wavelengths were reduced to 520 through the elimination of two spectra regions (1,601 to 1,717 and 3,052 to 5,011) which are known to be characterized by high noise level (Hewavitharana and Brakel (1997); De Marchi et al. (2013)). The statistical analysis consisted in the building of PLS models before and after UVE procedure. The optimal number of principal components used for the evaluation of PLS models was 11. This number was selected when the lowest root mean square error of prediction ($RMSE_P$) was reached. The coefficient of determination in external validation (R^2_V), the $RMSE_P$ and the ratio performance deviation (RPD) were used to evaluate the accuracy of PLS. According to Williams (1987), models with R^2_V between 0.83 and 0.90 are useful for practical application, and RPD values larger than 2 are considered adequate for analytical purposes (Karoui et al. (2006)).

Uninformative variable elimination. The UVE procedure was used following Centner et al. (1996) and Chen et al. (2007). Initially, a random matrix Z of 0s and 1s, and of exactly the same size of the X matrix was gener-

ated. The X matrix was multiplied by a small constant with the purpose of eliminating 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 that, in our case, was 10^{-4} according to Centner et al. (1996). This step was necessary to avoid large errors in the eigenvalues. This could also affect the regression coefficients (b) of the original matrix leading to the elimination of informative variables. The matrix X and the matrix Z were then merged into a XZ matrix of size n*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 in order to obtain a matrix containing the coefficients of all the equations used in the validation.

A stability criterion was adopted, according to Centner et al. (1996) and Chen et al. (2007):

$$E = bj / s(bj)$$

where E is the stability criterion, bj is the mean of the regression coefficients vector and s(bj) is the standard deviation of the vector. The absolute stability coefficient in the Z matrix was calculated and adopted as a filter into the original X matrix. The criterion for the construction of the filter was set to Ex < Ez, where Ex are the absolute values on matrix X and Ez are the absolute values on matrix Z. Finally, a new PLS model was performed using only the resulting informative wavelengths (n = 244).

Results and Discussion

The mean and variation of calibration (n = 10) and validation (n = 10) sets were similar (results not shown). The R^2_V in calibration and external validation ranged from 0.71 to 0.83, and the accuracy for titratable acidity was similar to that reported by De Marchi et al. (2009) and Toffanin and De Marchi (2013).

Table 1 summarizes the validation statistics for the PLS models before and after UVE approach. On average, R^2_V, $RMSE_P$, and RPD were better for prediction models after the UVE procedure. In particular, the average improvement of the accuracy after UVE was 6% for R^2_V, and 12% for $RMSE_P$ and RPD. Moreover, the prediction models developed using UVE approach were more robust than those from PLS analysis without UVE; in fact, the variation of $RMSE_P$ and RPD was lower. The improvement obtained using UVE was previously reported by Centner et al. (1996), who found similar results.

Table 1. Validation statistics[§] of partial least squares (PLS) models before and after uninformative variable elimination (UVE).

Validation PLS			Validation UVE-PLS		
R^2_V	$RMSE_P$	RPD	R^2_V	$RMSE_P$	RPD
0.71	0.15	1.88	0.83	0.13	2.22
0.80	0.15	2.23	0.83	0.14	2.49
0.81	0.14	2.33	0.83	0.13	2.58
0.82	0.15	2.36	0.83	0.13	2.66
0.77	0.17	2.11	0.77	0.14	2.54
0.79	0.15	2.20	0.83	0.13	2.40
0.77	0.14	2.10	0.81	0.13	2.31
0.71	0.15	1.86	0.84	0.12	2.30
0.80	0.15	2.24	0.87	0.12	2.79
0.82	0.13	2.36	0.85	0.13	2.40

[§] R^2_V = coefficient of determination in external validation; $RMSE_P$ = root mean square error of prediction; RPD = ratio performance deviation.

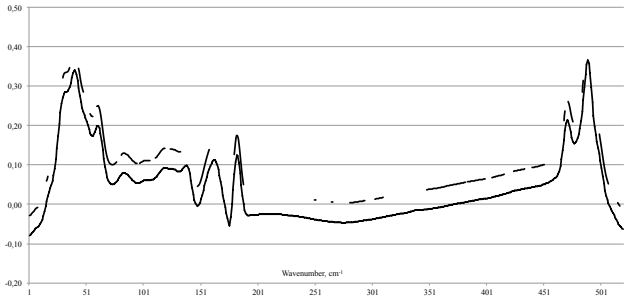

Figure 1. Example of spectra wavenumbers used in the partial least squares analysis before (---) and after (- - -) UVE.

Figure 1 depicts the most informative wavelengths extrapolated from the spectrum through the UVE approach. The number of wavelengths decreased from 520 to 244 after UVE, the accuracy of PLS improved, and computational time for statistical analysis decreased.

Conclusion

The application of UVE procedure allowed the elimination of uninformative wavelengths in the multivariate data before PLS. The elimination of specific uninformative wavelengths depends on the studied trait. The UVE seems to be a valid approach to improve the accuracy of prediction models compared with PLS. Moreover, the reduction of the number of wavelengths in the PLS analysis reduces the computational time and it would enhance the efficiency of MIRS models to predict phenotypes at population level.

Literature Cited

Anonymous. (1963). Milchwissenschaft 18:520.

Bonfatti, V., Di Martino, G., and Carnier, P. (2011). J. Dairy Sci. 94:5776–5785.

Centner, V., Massart, D. L., de Noord, O. E. et al. (1996). Anal. Chem. 68:3851-3858.

Chen, D., Cai, W., and Shao, X. (2007). Anal. Chim. Acta 598:19-26.

De Marchi, M., Fagan, C. C., O'Donnell, C. P. et al. (2009). J. Dairy Sci. 92:423-432.

De Marchi, M., Toffanin, V., Cassandro, M. et al. (2014). J. Dairy Sci. 97:1171-1186.

De Marchi, M., Toffanin, V., Cassandro, M. et al. (2013). J. Dairy Sci. 96:4707-4715.

Hewavitharana, A. K., and Brakel, B. V. (1997). Analyst 122:701-704.

Karoui, R., Mouazen, A. M., Dufour, E. et al. (2006). Eur. Food Res. Technol. 222:165–170.

McParland, S., Banos, G., McCarthy, B. et al. (2012). J. Dairy Sci. 95:7225–7235.

Soyeurt, H., Dardenne, P., Dehareng, F. et al. (2006). J. Dairy Sci. 89:3690–3695.

Soyeurt, H., Dehareng, F., Gengler, N. et al. (2011). J. Dairy Sci. 94:1657–1667.

Toffanin, V., and De Marchi, M. (2013). Ital. J. Anim. Sci. 12(Suppl. 1):37. (Abstr. C-086).

Williams, P.C. (1987). Implementation of near-infrared technology. Pages 143–167 in Near-Infrared Technology in the Agricultural and Food Industries. 2nd ed. P. C. Williams and K. Norris, ed. American Association of Cereal Chemists, St. Paul, MN.