J. Dairy Sci. 100:1–14
https://doi.org/10.3168/jds.2016-12203

Comparison of Bayesian regression models and partial least squares regression for the development of infrared prediction equations

V. Bonfatti,*[1] F. Tiezzi,† F. Miglior,‡§ and P. Carnier*
*Department of Comparative Biomedicine and Food Science, University of Padova, 35020, Legnaro, Italy
†Department of Animal Science, North Carolina State University, Raleigh 27695
‡Centre for Genetic Improvement of Livestock, University of Guelph, Guelph, N1G 2W1, Ontario, Canada
§Canadian Dairy Network, Guelph, N1K 1E5, Ontario, Canada

ABSTRACT

The objective of this study was to compare the prediction accuracy of 92 infrared prediction equations obtained by different statistical approaches. The predicted traits included fatty acid composition (n = 1,040); detailed protein composition (n = 1,137); lactoferrin (n = 558); pH and coagulation properties (n = 1,296); curd yield and composition obtained by a micro-cheese making procedure (n = 1,177); and Ca, P, Mg, and K contents (n = 689). The statistical methods used to develop the prediction equations were partial least squares regression (PLSR), Bayesian ridge regression, Bayes A, Bayes B, Bayes C, and Bayesian least absolute shrinkage and selection operator. Model performances were assessed, for each trait and model, in training and validation sets over 10 replicates. In validation sets, Bayesian regression models performed significantly better than PLSR for the prediction of 33 out of 92 traits, especially fatty acids, whereas they yielded a significantly lower prediction accuracy than PLSR in the prediction of 8 traits: the percentage of C18:1n-7 *trans*-9 in fat; the content of unglycosylated κ-casein and its percentage in protein; the content of α-lactalbumin; the percentage of α_{S2}-casein in protein; and the contents of Ca, P, and Mg. Even though Bayesian methods produced a significant enhancement of model accuracy in many traits compared with PLSR, most variations in the coefficient of determination in validation sets were smaller than 1 percentage point. Over traits, the highest predictive ability was obtained by Bayes C even though most of the significant differences in accuracy between Bayesian regression models were negligible.

Key words: infrared spectra, fatty acid, protein fraction, Bayesian regression

INTRODUCTION

Mid-infrared spectroscopy is a recognized tool for predicting novel milk traits for payment systems, management of dairy cows, and selective breeding purposes (Gengler et al., 2016). The utility of prediction (regression) equations for practical applications depends mostly on their accuracy. According to Soyeurt et al. (2011), equations with a coefficient of determination ($\mathbf{R^2}$) greater than 0.95 in cross-validation are useful in milk payment systems. For management purposes, the usefulness of equations depends on the correlation between the predicted traits and the management indicators (e.g., prevalence of metabolic disorders). For selective breeding, the usefulness of infrared predictions relies mostly on their heritability and on their genetic correlation with the breeding goal traits. Even though equations with medium to low accuracy might be successfully used for breeding purposes, more accurate prediction equations would lead to a faster genetic progress because a positive relationship exists between the R^2 in cross-validation and the estimated genetic correlation between measured and predicted traits (Rutten et al., 2010; Bonfatti et al., 2017). Thus, there is interest in finding chemometric methods that can increase the accuracy of prediction models.

To date, prediction equations have been developed mostly using partial least squares regression (**PLSR**). Recently, Bayesian models adopted for regression on high-dimensional genotypes have been reported to dramatically increase the prediction accuracy of infrared prediction equations for the prediction of 8 traits related to fatty acid (**FA**) composition and technological properties of milk (Ferragina et al., 2015). However, results might depend on the traits as well as on the conditions under which models are developed. For example, the presence of noise regions in the development of prediction models might affect the accuracy of PLSR (De Marchi et al., 2009; Bonfatti et al., 2011; Eskildsen et al., 2014) without influencing the accuracy of Bayesian regressions, as those methods allow shrinkage and

Received October 25, 2016.
Accepted May 4, 2017.
[1]Corresponding author: valentina.bonfatti@unipd.it

perform variable selection, down-weighting or excluding the uninformative variables (de los Campos et al., 2013).

The objective of this study was to compare the accuracy of prediction equations obtained by PLSR and by Bayesian regression methods for predicting a large number of traits related to milk fine composition and technological properties. Comparison among models was performed considering different numbers of spectral variables and PLSR terms for prediction equation development.

MATERIALS AND METHODS

Reference Data

A total of 1,330 individual milk samples of Simmental cows were collected for reference analyses. Samples were collected during the morning milking in 21 herds located in northern Italy. Herd size ranged from 30 to 125 cows. Cows were between 5 and 484 DIM and ranged from 1 to 9 parities. All samples, with the exception of those that were lost during the analysis due to nonmatching sample numbers or poor preservation, were analyzed for pH, milk coagulation properties, micro-cheese yield, micro-cheese composition, and protein profile. Measures of pH and milk coagulation properties were available for 1,296 samples. Contents of α_{S1}-CN, α_{S2}-CN, β-CN, γ-CN, glycosylated and unglycosylated κ-CN, β-LG, and α-LA of individual milk samples were measured in 1,137 samples. Curd yield and composition were obtained for 1,177 samples using a micro-cheese making procedure and measuring DM content, protein content, and fat content in the micro-curds.

Samples from cows with known sire and dam (1,040 cows in 20 herds; daughters of 378 sires) were also analyzed for FA composition. Part of these samples, depending on budget constraints, was also analyzed for mineral contents and lactoferrin (**LF**). Contents of Ca, P, Mg, and K were analyzed in 689 samples from 15 herds, whereas LF was measured in 558 samples from 11 herds. All samples with measures of mineral contents, LF, and FA also had measures of pH, milk coagulation properties, protein profile, and cheese yield. In total, 92 traits were investigated. Details on the methods used to obtain the reference traits, as well as descriptive statistics of all the investigated traits, can be found in Bonfatti et al. (2016).

Milk Infrared Spectra

Infrared absorption spectra (1,060 variables) were collected on all samples by the Friuli Venezia Giulia Milk Recording Agency laboratory (Codroipo, Italy) using a MilkoScan FT6000 (Foss Electric A/S, Hillerød, Denmark). Spectra exhibiting a Mahalanobis distance from the population centroid greater than 3 were considered to be outliers and were discarded. Records with a trait value above 4 or below −4 SD from the mean were also excluded. For each trait, the number of outliers ranged from 5 to 11 samples. Spectra variables were standardized to a null mean and a unit variance before the analysis.

Due to the interference of water absorption, the O–H bending and stretching regions of the spectra (between 1,628 and 1,658 cm^{-1} and between 3,105 and 3,444 cm^{-1}, respectively) are assumed to contain no useful chemical information and to have very high coefficients of variation and a very low heritability (Soyeurt et al., 2010); for these reasons, they are often discarded before the chemometric analysis (De Marchi et al., 2009; Bonfatti et al., 2011; Eskildsen et al., 2014). Analyses were performed on the totality of the spectra variables and on the spectra variables cleared of the 2 water absorption regions of the spectra. In the latter case, prediction equations were developed using 872 spectra variables. The derivative of the spectral data is often used because it enhances resolution by sharpening the absorption bands and it removes baseline offset (Burns and Ciurczak, 2001). Models were fitted to both raw spectra and spectra transformed with a first derivative mathematical treatment. The results obtained from treated and untreated spectra were very similar, and only those obtained for the raw spectra are shown.

Prediction Equations

Prediction equations can be considered as a series of partial regression coefficients, in number equal to the number of wavelengths, providing a prediction of a measurable trait in a sample. In the present study, these regression coefficients were estimated using 6 different methods. The PLSR implemented in the R (R Development Core Team, 2013) package PLS (Mevik and Wehrens, 2007) was used as the reference method. The PLSR models were fitted by either (1) choosing the number of PLSR components that minimized the prediction error in the training set (**PLSR-M**) or (2) setting the maximum number of components to 16 (**PLSR-16**). The PLSR-16 was performed only on the totality of the spectra variables and was carried out to compare the results with previous literature estimates obtained using WinISI II software (InfraSoft International, State College, PA), for which the maximum number of terms is 16 by default. In addition, PLSR and modified PLSR (**MPLSR**; Shenk and Westerhaus,