J. Dairy Sci. 98:1–19
http://dx.doi.org/10.3168/jds.2014-9143

Bayesian regression models outperform partial least squares methods for predicting milk components and technological properties using infrared spectral data

A. Ferragina,* G. de los Campos,†‡ A. I. Vazquez,† A. Cecchinato,*[1] and G. Bittante*
*Department of Agronomy, Food, Natural Resources, Animals and Environment (DAFNAE), University of Padova, Viale dell'Università 16, 35020 Legnaro, Italy
†Epidemiology and Biostatistics Department, and
‡Department of Statistics and Probability, Michigan State University, East Lansing 48824

ABSTRACT

The aim of this study was to assess the performance of Bayesian models commonly used for genomic selection to predict "difficult-to-predict" dairy traits, such as milk fatty acid (FA) expressed as percentage of total fatty acids, and technological properties, such as fresh cheese yield and protein recovery, using Fourier-transform infrared (FTIR) spectral data. Our main hypothesis was that Bayesian models that can estimate shrinkage and perform variable selection may improve our ability to predict FA traits and technological traits above and beyond what can be achieved using the current calibration models (e.g., partial least squares, PLS). To this end, we assessed a series of Bayesian methods and compared their prediction performance with that of PLS. The comparison between models was done using the same sets of data (i.e., same samples, same variability, same spectral treatment) for each trait. Data consisted of 1,264 individual milk samples collected from Brown Swiss cows for which gas chromatographic FA composition, milk coagulation properties, and cheese-yield traits were available. For each sample, 2 spectra in the infrared region from 5,011 to 925 cm^{-1} were available and averaged before data analysis. Three Bayesian models: Bayesian ridge regression (Bayes RR), Bayes A, and Bayes B, and 2 reference models: PLS and modified PLS (MPLS) procedures, were used to calibrate equations for each of the traits. The Bayesian models used were implemented in the R package BGLR (http://cran.r-project.org/web/packages/BGLR/index.html), whereas the PLS and MPLS were those implemented in the WinISI II software (Infrasoft International LLC, State College, PA). Prediction accuracy was estimated for each trait and model using 25 replicates of a training-testing validation procedure. Compared with PLS, which is currently the most widely used calibration method, MPLS and the 3 Bayesian methods showed significantly greater prediction accuracy. Accuracy increased in moving from calibration to external validation methods, and in moving from PLS and MPLS to Bayesian methods, particularly Bayes A and Bayes B. The maximum R^2 value of validation was obtained with Bayes B and Bayes A. For the FA, C10:0 (% of each FA on total FA basis) had the highest R^2 (0.75, achieved with Bayes A and Bayes B), and among the technological traits, fresh cheese yield R^2 of 0.82 (achieved with Bayes B). These 2 methods have proven to be useful instruments in shrinking and selecting very informative wavelengths and inferring the structure and functions of the analyzed traits. We conclude that Bayesian models are powerful tools for deriving calibration equations, and, importantly, these equations can be easily developed using existing open-source software. As part of our study, we provide scripts based on the open source R software BGLR, which can be used to train customized prediction equations for other traits or populations.

Key words: infrared spectroscopy, Bayesian method, milk trait, fatty acid, cheese yield

INTRODUCTION

Infrared spectroscopy (**IRS**) is based on using different waves of the infrared region of the electromagnetic spectrum to excite molecules in relation to their rotational-vibrational structure (Karoui et al., 2010). The infrared spectrum of a sample is recorded after passing a beam of infrared light through it. When the frequency of the infrared wave is the same as the vibrational frequency of a chemical bond, absorption occurs; the spectrum therefore reflects the quantities and proportions of the various chemical bonds within the sample and hence its composition.

Infrared spectroscopy is often used to predict the chemical composition of food and feed (Karoui et al.,

Received November 23, 2014.
Accepted July 6, 2015.
[1]Corresponding author: alessio.cecchinato@unipd.it

2010), but it is a secondary method needing prior calibration based on a training data set and validation based on a different data set, both obtained using samples analyzed according to reference methods. The Fourier-transform infrared (**FTIR**) spectroscopy, which measures transmission of a spectrum consisting of more than 1,000 different waves in the short-wave infrared region (**SWIR**, or near-infrared), the mid-wave IR (**MWIR**, or mid-infrared), and the long-wave IR (**LWIR**), is often used to predict the chemical composition of milk (Barbano and Lynch, 2006; Karoui and Baerdemaeker, 2007). Fourier-transform infrared spectroscopy has important advantages compared with traditional laboratory-based analysis techniques. Some advantages include the low time requirements, inexpensive except for the cost of the apparatus, ability to predict a large number of phenotypes for one sample carrying out only one analysis, ability to predict new phenotypes from stored spectra when a new prediction equation becomes available, and the feasibility of obtaining individual phenotypes for selection. In addition, FTIR spectroscopy is an accurate tool for predicting major milk component contents and is used internationally for the analysis of the fat, protein, casein, and lactose contents of cow milk from routine recording samples (ICAR, 2012). The prediction of milk components and technological traits (especially those that are difficult to analyze) is of particular interest in many areas, including milk payment systems, assessing technological properties of milk by the dairy industry, and direct (in relation to human health) or indirect (animal welfare, reproduction, methane production) prediction of some traits through milk FA content. All of these traits could be used in selection programs.

In recent years, several studies have used FTIR spectroscopy to predict the FA content of milk (Soyeurt et al., 2006; De Marchi et al., 2011). The FA show different prediction accuracy based on, for example, the amount of the individual FA in milk and the way of expressing that amount (on a milk or milkfat basis); for these reasons, these traits are considered to be difficult to predict and the level of accuracy is lower than when predicting major milk components (e.g., protein or fat). In part, this is because FA make up a smaller fraction of milk and many compounds with similar chemical composition are present (Stefanov et al., 2013; De Marchi et al., 2014). Calibration FTIR is even more difficult if an FA profile (i.e., each FA as a proportion of the sum of all FA) is to be predicted. Few studies have attempted to predict the FA profile of milk fat using FTIR spectroscopy, and the results are less accurate than those for the total FA content of milk (Soyeurt et al., 2006; Rutten et al., 2009).

Infrared spectroscopy technology is not very precise when used to predict the technological properties of food that only indirectly depend on the sample's chemical composition. In the case of milk, FTIR spectroscopy has been used to predict new phenotypes of significant economic interest to the dairy industry, such as milk coagulation properties (**MCP**; Cecchinato et al., 2009), cheese yield (**CY**) and curd recovery (**REC**) or whey loss of milk nutrients (Ferragina et al., 2013).

The IRS prediction of new phenotypes is of particular interest for its potential use in the selection of farm animal populations using existing samples and spectrometers, such as milk recordings for the genetic improvement of milk fat and protein. Several studies have estimated the genetic parameters of infrared-predicted phenotypes, such as FA content (Rutten et al., 2010; Bastin et al., 2011; Cecchinato et al., 2012a), MCP (Bittante et al., 2012), and CY and REC of different nutrients (Cecchinato et al., 2015). Heritability estimates of measured phenotypes are similar to, or lower than, the heritabilities of the predicted traits such as milk technological properties (e.g., RCT; Cecchinato et al., 2009, 2011b; Bittante et al., 2014). In the case of FA profiles, there is a higher variation of heritability estimates (Rutten et al., 2010). Importantly, the estimated genetic correlations between measured and FTIR-predicted values for all traits studied were greater than the phenotypic correlations between the same values. The biological basis of the potential of FTIR spectra for genetic improvement of farm animals lies in the fact that the absorbance of many individual waves (Bittante and Cecchinato, 2013) or their principal components (Soyeurt et al., 2010; Dagnachew et al., 2013) have been proven to be heritable.

The accuracy of predictions obtained with IRS is influenced by many factors, including the trait to be predicted, the quality of the reference data set and the spectra, the number of samples used to develop the prediction equations, and the amount of the analyzed substance in the samples (Rutten et al., 2009; Karoui et al., 2010). A distinction is needed between direct and indirect predictions, and this distinction plays an important role in the prediction accuracy. One trait can be considered directly predicted when it has a significant signal in the spectral data (e.g., protein content); otherwise, in the indirect prediction of one trait, the signal in the spectral data is related to traits having a relationship with the studied trait (e.g., cheese yield), and a greater number of samples in calibration set is needed for high prediction accuracy. A special role, however, is played by mathematical techniques known collectively as chemometrics, including the selection of wavelengths, the pretreatment of spectra data, and