J. Dairy Sci. 100:2057–2067
https://doi.org/10.3168/jds.2016-11951


# Comparison between direct and indirect methods for exploiting Fourier transform spectral information in estimation of breeding values for fine composition and technological properties of milk

**V. Bonfatti,*[1] D. Vicario,† L. Degano,† A. Lugo,‡ and P. Carnier***
*Department of Comparative Biomedicine and Food Science (BCA), University of Padova, 35020, Legnaro, Italy
†Italian Simmental Cattle Breeders Association (ANAPRI), 33100, Udine, Italy
‡Friuli Venezia Giulia Milk Recording Agency (AAFVG), 33033, Codroipo, Italy

## ABSTRACT

The aim of this study was to compare the common method of exploiting infrared spectral data in animal breeding; that is, estimating the breeding values for the traits predicted by infrared spectroscopy, and an alternative approach based on the direct use of spectral information (direct prediction, DP) to predict the estimated breeding values (EBV). Traits were pH, milk coagulation properties, contents of the main casein and whey protein fractions, cheese yield measured by micro-cheese making, lactoferrin, Ca, and fat composition. For the DP method, the number of spectral variables was reduced by principal components analysis to 8 latent traits that explained 99% of the original spectral variation. Restricted maximum likelihood was used to estimate variance components of the latent traits. (Co)variance components of the original spectral traits were obtained by back-transformation and EBV of all derived milk traits were then predicted as traits correlated with the genetic information of the spectra. The rank correlation between the EBV obtained for the infrared-predicted traits and those obtained from the DP method was variable across traits. Rank correlations ranged from 0.07 (for the content of saturated fatty acids expressed as g/100 g of fat) to 0.96 (for dry matter cheese yield, %) and, for most traits, was <0.5. This result can be explained by the nature of the principal components analysis: it does not take into account the covariance between the spectral variables and the reference traits but produces latent traits that maximize the spectral variance explained. Thus, the direct approach is more likely to be effective for traits more related to the main sources of spectral variation (i.e., protein and fat). More research is required to study spectral genetic variation and to determine the best way to choose spectral regions and the type and number of considered latent traits for potential applications.

**Key words:** infrared spectroscopy, fatty acid, protein fraction, breeding value

## INTRODUCTION

Fourier-transform infrared spectroscopy (**FTIR**) is a useful tool to predict individual phenotypes for traditional and innovative milk traits and a candidate method to replace gold standard methodologies, which are often not applicable for population-wide phenotyping due to high cost or other practical limitations. Infrared prediction of individual phenotypes relies on the availability of calibration equations developed using gold standard measures of traits of concern and FTIR spectra for a limited number of reference samples. Together with pedigree information and variance component estimates, predicted phenotypes can be used in BLUP to obtain EBV. This approach is referred to as the indirect prediction method (**IP**; Dagnachew et al., 2013b) because the spectral information is not directly used in EBV prediction procedures, although the spectra provide insights into the genetic variation in milk components (Soyeurt et al., 2010).

Starting from evidence that milk FTIR spectral variables exhibit tight correlations among each other (Soyeurt et al., 2010; Dagnachew et al., 2013a) and considering that direct genetic analysis on such correlated spectral variables may enhance the accuracy of genetic evaluation methods, a direct prediction (**DP**) approach has been proposed (Dagnachew et al., 2013b). In the DP approach, EBV prediction is performed using the milk FTIR spectral variables directly, and EBV for traits of concern are derived from the predicted EBV for the spectral variables (i.e., EBV for the traits of interest are predicted as EBV of traits correlated with the genetic component of the spectra).

Received September 1, 2016.
Accepted December 5, 2016.
[1]Corresponding author: valentina.bonfatti@unipd.it

The DP method has some benefits over the IP method: there is no need to predict phenotypes from the spectra to estimate the EBV for the traits, and EBV are predicted once (only for the spectra) and then used to derive the EBV of traits. This is particularly relevant when considering the high number of traits for which FTIR calibration equation are being developed (Bonfatti et al., 2016; Gengler et al., 2016). In addition, direct analysis of the spectral variables may increase the precision of the estimated genetic parameters and the accuracy of EBV predictions and genetic gains, particularly for low-heritability traits, as a consequence of exploiting the genetic relationships among many spectral variables (Dagnachew et al., 2013b).

Dagnachew et al. (2013b) compared DP and IP using goat milk spectra and reported very high rank correlations between the EBV provided by the 2 methods. In that study, infrared predictions of fat, protein, and lactose contents were used as phenotypes because no data from chemical analysis were available. The investigated traits were directly linked to spectral information, and the calibration equations developed by Dagnachew et al. (2013b) had very high predictive ability, with $R^2$ values ranging from 0.95 to 0.98. The DP and IP methods have not been compared for calibration equations developed using independent reference data obtained by chemical analysis and for traits predicted with intermediate to low accuracy.

The aim of this study was to compare DP and IP as methods for routine prediction of EBV in a dairy cattle population for a group of traits that describe the fine composition and technological properties of milk and that are predicted with variable accuracy using FTIR spectra.

## MATERIALS AND METHODS

### FTIR Spectra and Calibration Models

A total of 100,272 milk samples were collected (from February 2013 to June 2014) from 11,216 Italian Simmental cows (92 herds) during the routine milk recording operated in Italy in the Friuli Venezia Giulia region by the regional milk recording agency (AAFVG, Codroipo, Italy). On average, each cow provided 6.9 milk samples, with a minimum of 1 and a maximum of 12 samples. Samples were analyzed using a MilkoScan FT6000 (Foss Electric A/S, Hillerød, Denmark), and the generated FTIR absorbance spectral data (1,060 variables per spectrum) were recorded.

Calibration equations used in this study were the outcome of a research project (MilCo project, CPDA122982; University of Padova, Padova, Italy), which started in 2013 with the aim of developing procedures to estimate EBV for the Italian Simmental cattle population for FTIR predictions of detailed protein and FA composition, minerals, lactoferrin, coagulation properties, cheese yield, and curd composition. Reference data for the development of calibration equations were obtained for milk protein composition by reversed phase HPLC (Bonfatti et al., 2008), for FA composition by accelerated extraction (Dionex application note 345; Thermo Scientific Dionex, 2016) followed by 2-dimensional gas chromatography separation (Pellattiero et al., 2015), for minerals by inductively coupled plasma atomic emission spectroscopy (Soyeurt et al., 2009), for lactoferrin by ELISA (Soyeurt et al., 2007), for milk coagulation properties by lactodynamography (Bonfatti et al., 2016), and for cheese yield and curd composition by micro-cheese making (Bonfatti et al., 2016). In total, 92 traits were measured and calibration equations were developed using more than 1,000 samples for each of the investigated traits, with the exception of minerals (n = 689) and lactoferrin (n = 558). Details on procedures providing reference data for the traits investigated in this study can be found in Bonfatti et al. (2016).

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 $cm^{-1}$ and 3,444 $cm^{-1}$, respectively) were removed from each spectrum, as suggested by Soyeurt et al. (2010). Spectral outliers were identified based on the standardized Mahalanobis distance (Burns and Ciurczak, 2007).

Calibration equations to be used in this study were developed using the remaining 872 spectral variables and modified partial least square (**MPLS**; Shenk and Westerhaus, 1991) regression procedures, as implemented in the software WinISI II (Infrasoft International Inc., State College, PA), and were cross-validated using a 10-random-segments procedure.

### Estimates of EBV Under the DP and IP Methods

Sixteen traits were used to compare the EBV obtained from application of IP and DP. The traits were selected from the 92 traits investigated in the MilCo project to compare the 2 methods under scenarios in which FTIR predictions had variable accuracy. Performance of the calibration equations used in the prediction of the 16 selected traits is reported in Table 1. Values of the coefficient of determination of cross-validation ($R^2_{CV}$) and the ratio of performance to deviation (i.e., the ratio of the SD of a trait to the standard error in cross-validation) ranged from 0.35 to 0.86 and from 1.24 to 2.85, respectively.

A schematic representation of IP and DP methods is depicted in Figure 1. The estimation of genetic and

**Table 1**. Parameters[1] describing the predictive ability of calibration equations used in the study and descriptive statistics for the measured reference traits

| | Descriptive statistics | | | Calibration performance | | |
|---|---|---|---|---|---|---|
| Trait | Mean | Minimum | Maximum | $SE_{CV}$ | $R^2_{CV}$ | RPD |
| Technological traits | | | | | | |
| pH | 6.75 | 6.34 | 7.32 | 0.04 | 0.80 | 2.23 |
| Coagulation time (min) | 18.62 | 7.65 | 59.18 | 3.14 | 0.71 | 2.00 |
| Raw cheese yield (%) | 26.58 | 0.65 | 65.67 | 4.05 | 0.62 | 1.66 |
| DM cheese yield (%) | 7.56 | 0.17 | 14.04 | 0.45 | 0.86 | 2.85 |
| Curd moisture (%) | 70.93 | 43.15 | 83.93 | 2.66 | 0.59 | 1.63 |
| Major protein fractions (g/L) | | | | | | |
| Casein | 31.73 | 22.66 | 46.31 | 1.39 | 0.85 | 2.60 |
| $\alpha_{S1}$-CN | 13.52 | 8.27 | 21.56 | 0.76 | 0.76 | 2.06 |
| $\beta\gamma$-CN | 10.46 | 6.08 | 17.38 | 1.16 | 0.59 | 1.56 |
| $\kappa$-CN | 3.58 | 1.76 | 6.88 | 0.61 | 0.35 | 1.24 |
| Lactoferrin (μg/mL) | 120.04 | 9.42 | 441.88 | 71.05 | 0.41 | 1.30 |
| Ca (mg/kg) | 1,206 | 700 | 2,068 | 131 | 0.53 | 1.49 |
| Fatty acids (g/100 g of fat) | | | | | | |
| SFA | 74.15 | 53.73 | 82.75 | 1.76 | 0.76 | 2.06 |
| MUFA | 21.95 | 14.54 | 42.01 | 1.54 | 0.76 | 2.08 |
| PUFA | 3.90 | 2.20 | 7.34 | 0.53 | 0.59 | 1.57 |
| n-3 | 0.55 | 0.13 | 1.17 | 0.10 | 0.62 | 1.63 |
| CLA *cis*-9,*trans*-11 | 0.35 | 0.02 | 0.86 | 0.07 | 0.42 | 1.37 |

[1]$SE_{CV}$ = standard error of cross-validation; $R^2_{CV}$ = coefficient of determination in cross-validation; RPD = ratio of prediction to deviation, calculated as the ratio of the SD to the $SE_{CV}$ for a trait.

nongenetic (co)variance components for the 872 spectral variables, which would be needed to perform a BLUP analysis of the 872 spectral variables, was unfeasible due to computer memory constraints. As suggested by Soyeurt et al. (2010) and Dagnachew et al. (2013a), the dimensionality of the analysis was reduced by principal component analysis (**PCA**). The procedure described by Dagnachew et al. (2013b) was used to obtain the predictions of EBV under DP. Briefly, a principal component decomposition of milk FTIR spectra can be represented as follows:

$$\mathbf{w} = \left(\mathbf{I}_n \otimes \mathbf{P}\right) \otimes \mathbf{t} + \boldsymbol{\varepsilon}, \quad [1]$$

where $\mathbf{w}$ is a vector of observed spectral variables (with the spectrum of one sample above the other), $\mathbf{I}_n$ is an identity matrix of size $n$, where $n$ is the number of milk samples, $\otimes$ denotes the Kronecker product operator, $\mathbf{P}$ is a matrix of loadings of size $m \times a$, where $m$ is the number of spectra variables and $a$ is the number of principal components, $\mathbf{t}$ is a vector of scores to be estimated (with the scores of one sample on top of the other), and $\boldsymbol{\varepsilon}$ is a vector of residuals. The scores ($\mathbf{t}$) are referred to as latent traits (**LT**).

The first 8 LT, corresponding to the 8 highest eigenvalues of the correlation matrix of the original spectral data and accounting for 99% of the spectral variance, were considered for further analysis. The LT were adjusted for the effect of herd-test-date (4,343 levels), parity (1, 2, 3, and 4 and later parities), and stage of lactation (15-d classes, up to 360 or more DIM). Only herd-test-date levels with at least 5 observations were retained. The following multi-trait animal model was used to estimate the EBV for the latent traits $\mathbf{t}$:

$$\mathbf{t} = \boldsymbol{\mu} + \mathbf{Z}\mathbf{u}_\mathbf{t} + \mathbf{H}\mathbf{h}_\mathbf{t} + \mathbf{e}_\mathbf{t}, \quad [2]$$

where $\boldsymbol{\mu}$ is a vector of 1, $\mathbf{u}_\mathbf{t}$ is a vector of animal additive genetic effects, $\mathbf{h}_\mathbf{t}$ is a vector of permanent environmental effects, $\mathbf{e}_\mathbf{t}$ is a vector of random residuals, and $\mathbf{Z}$ and $\mathbf{H}$ are design matrices relating observation in $\mathbf{t}$ to effects in $\mathbf{u}_\mathbf{t}$ and $\mathbf{h}_\mathbf{t}$, respectively.

The following (co)variance structure for the latent traits in $\mathbf{t}$ was assumed: $\text{var}(\mathbf{u}_\mathbf{t}) = \mathbf{G}_\mathbf{t} \otimes \mathbf{A}$, $\text{var}(\mathbf{h}_\mathbf{t}) = \mathbf{Q}_\mathbf{t} \otimes \mathbf{I}_\mathbf{h}$, and $\text{var}(\mathbf{e}_\mathbf{t}) = \mathbf{R}_\mathbf{t} \otimes \mathbf{I}_n$, where $\mathbf{G}_\mathbf{t}$ is the genetic (co)variance matrix for the latent traits, $\mathbf{A}$ is the additive relationship matrix among animals, $\mathbf{Q}_\mathbf{t}$ is the permanent environmental (co)variance matrix, $\mathbf{R}_\mathbf{t}$ is the residual (co)variance matrix, and $\mathbf{I}_\mathbf{h}$ and $\mathbf{I}_n$ are identity matrices of appropriate order. Restricted maximum likelihood estimates of the (co)variance matrices for the latent traits $\hat{\mathbf{G}}_\mathbf{t}$, $\hat{\mathbf{Q}}_\mathbf{t}$, and $\hat{\mathbf{R}}_\mathbf{t}$ were obtained using VCE software (version 6.0; Groeneveld et al., 2010).

Pedigree information was supplied by the Italian Simmental Cattle Breeders Association (ANAPRI, Udine, Italy) and included all animals with spectral data. The pedigree file included 46,870 animals.

The EBV for LT were back-transformed to the EBV of the original 872 spectral variables using the following linear transformation of Equation [1]:

$$\widetilde{\mathbf{u}_w} = \left(\mathbf{I}_n \otimes \mathbf{P}\right) \cdot \widetilde{\mathbf{u}_t}, \quad [3]$$

where $\widetilde{\mathbf{u}_w}$ are the predicted EBV for the spectral variables (genetic components of FTIR spectra). Breeding values for a milk trait, $\mathbf{y}_i$, were then predicted as correlated traits from the genetic component of FTIR spectra $\left(\tilde{\mathbf{u}}_w\right)$:

$$\tilde{\mathbf{u}}_i^* = \left(\mathbf{I}_n \otimes \hat{\mathbf{B}}_{PLS_i}{}'\right) \cdot \widetilde{\mathbf{u}_w} = \left(\mathbf{I}_n \otimes \hat{\mathbf{B}}_{PLS_i}{}'\right) \cdot \left[\left(\mathbf{I}_n \otimes \mathbf{P}\right) \cdot \tilde{\mathbf{u}}_t\right], [4]$$

where $\tilde{\mathbf{u}}_i^*$ is a vector of predicted EBV for a trait $\mathbf{y}_i$ using DP and $\hat{\mathbf{B}}_{PLS_i}$ is the vector of the MPLS regression coefficients of the calibration equation developed for trait $i$.

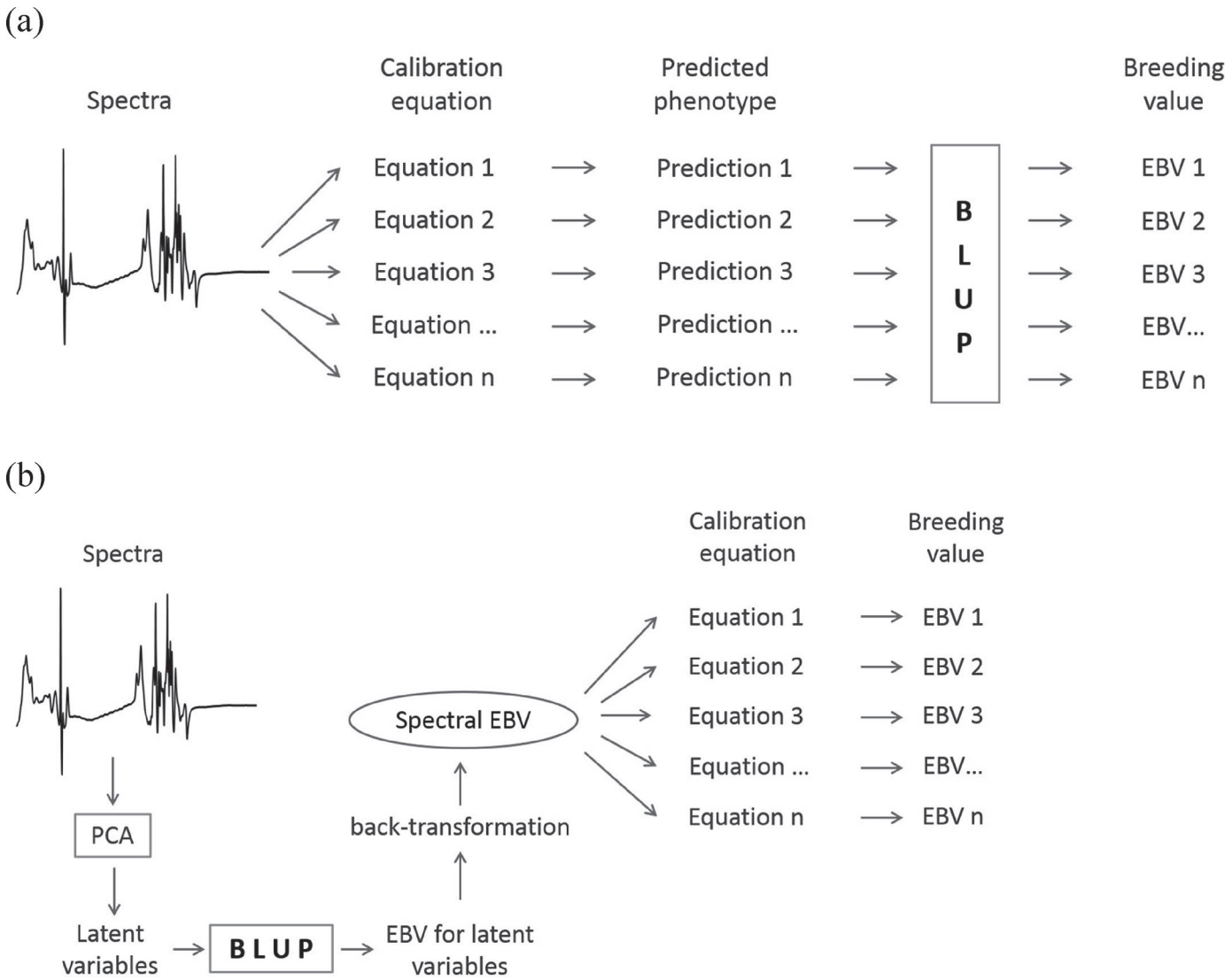


**Figure 1**. Schematic representation of (a) the indirect prediction method, and (b) the direct prediction method. In the indirect prediction method, calibration equations are applied to the spectra to predict phenotypes and BLUP estimates are obtained for each trait to derive the breeding values. In the direct prediction method, principal components analysis (PCA) is performed on the spectra and a multi-trait animal model is applied on the 8 latent traits having the greatest eigenvalues. Breeding values estimated for the latent traits are then back-transformed to the number of original spectra variables, and calibration equations are applied to the back-transformed EBV to obtain the EBV for single traits.

For the IP method, the predicted phenotypes $\hat{\mathbf{y}}_i$ for a trait $i$ ($i = 1, \ldots, 16$) were predicted from the spectra as follows:

$$\hat{\mathbf{y}}_i = \left(\mathbf{I}_n \otimes \hat{\mathbf{B}}_{PLS_i}{}'\right) \cdot \mathbf{w} \qquad (i = 1, \ldots, 16). \qquad [5]$$

The FTIR-predicted traits were adjusted for the same effects considered for DP and EBV were then estimated through univariate analyses using the same linear model used for LT in DP (see Equation [2]).

### Predictions Obtained from Back-Transformed Spectra

Under DP, only 99% of the spectral variation (i.e., the amount of variation explained by the first 8 LT) was used in EBV prediction. To investigate the effect of using a reduced amount of the original spectral information in prediction of phenotypes (which are never obtained using DP) and EBV for milk traits, an intermediate method ($\mathbf{BT_8}$) between IP and DP was also used. This method, schematically depicted in Figure 2, used observations on the first 8 LT to back-transform the spectra to the original 872 variables. The IP method was then applied to the back-transformed spectra to obtain predictions of the phenotypes for the investigated traits, and these predicted phenotypes were used to estimate the EBV. The same approach was applied using the first 21 LT ($\mathbf{BT_{21}}$), which explained 99.9% of the original spectral variation.

Following the approach proposed by Dagnachew et al. (2013b) to compare different models with the same (co)variance structure, the variance components of individual traits in the IP approach were calculated from the variance components of the FTIR spectra using the following equation:

$$\hat{\sigma}^2_{a_i} = \hat{\mathrm{B}}_{\mathrm{PLS}_i}{}' \cdot \mathbf{P} \cdot \widehat{\mathbf{G_t}} \cdot \mathbf{P}' \cdot \hat{\mathrm{B}}_{\mathrm{PLS}_i}. \qquad [6]$$

Likewise, the permanent environmental $\sigma^2_{pe_i}$ and residual $\sigma^2_{e_i}$ variances were estimated using similar equations by replacing the relevant terms $\left(\text{i.e., } \sigma^2_{pe_i} = \hat{\mathrm{B}}_{\mathrm{PLS}_i}{}' \cdot \mathbf{P} \cdot \widehat{\mathbf{Q_t}} \cdot \mathbf{P}' \cdot \hat{\mathrm{B}}_{\mathrm{PLS}_i}\right.$ and $\left.\sigma^2_{e_i} = \hat{\mathrm{B}}_{\mathrm{PLS}_i}{}' \cdot \mathbf{P} \cdot \widehat{\mathbf{R_t}} \cdot \mathbf{P}' \cdot \hat{\mathrm{B}}_{\mathrm{PLS}_i}\right)$.

For each milk trait, the Spearman rank correlations ($\boldsymbol{\rho}$) between EBV estimated under IP and DP using $\mathrm{BT}_8$ or $\mathrm{BT}_{21}$ were calculated using the CORR procedure of SAS (version 9.3; SAS Institute Inc., Cary, NC). The Pearson product-moment correlations between the infrared predicted phenotypes obtained from the IP method and those obtained from $\mathrm{BT}_8$ and $\mathrm{BT}_{21}$ methods were calculated using the same SAS procedure.

## RESULTS AND DISCUSSION

### Genetic Variability of Mid-Infrared Spectral Data

The spectral regions ranging from 1,628 and 1,658 $\mathrm{cm}^{-1}$ and from 3,105 $\mathrm{cm}^{-1}$ and 3,444 $\mathrm{cm}^{-1}$ (i.e., the water absorbance regions) exhibited large phenotypic variation compared with the rest of the spectrum, and PCA performed on the unprocessed spectral data, cleared of those regions, resulted in 8 LT explaining approxi-

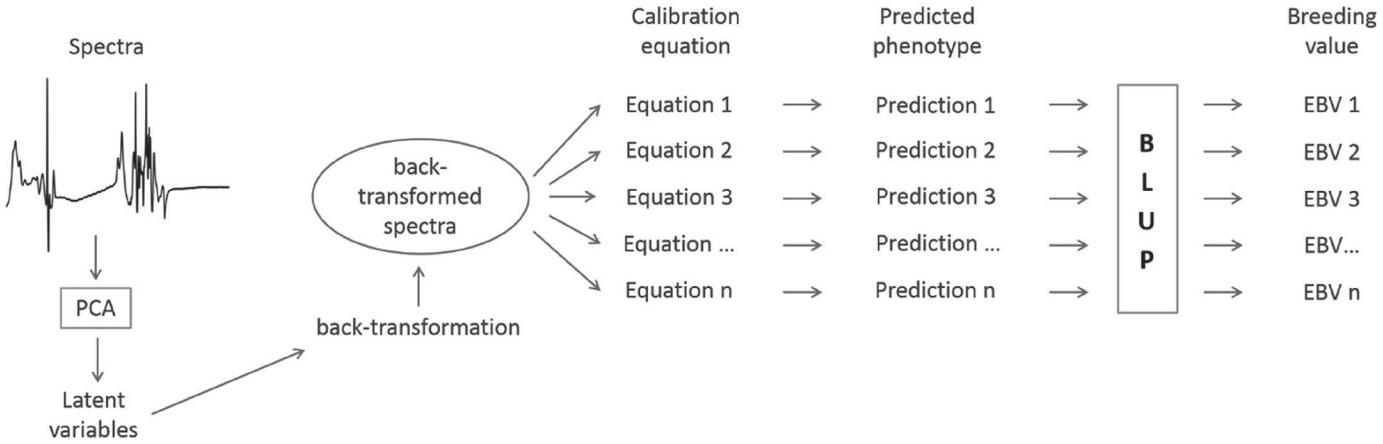


**Figure 2**. Schematic representation of the use of back-transformed spectra to obtain phenotypes and EBV using principal components analysis (PCA) applied to the spectra. The 8 latent traits having the greatest eigenvalues are used to back-transform the spectra to the number of original spectra variables. These correspond to spectra reconstructed using 99% of the original spectral variation. Calibration equations are used on the back-transformed spectra to obtain the prediction of phenotypes to be used in univariate animal models for the estimates of EBV.

**Table 2**. Estimates and standard errors of variance ratios of additive genetic, permanent environmental, and residual effects to total phenotypic variance[1] for the first 8 latent variables (LT) obtained from the principal components analysis of the spectral data

| Latent variable | Cumulative variance explained by LT (%) | Heritability | | Permanent environment | | Residual | |
|---|---|---|---|---|---|---|---|
| | | Estimate | SE | Estimate | SE | Estimate | SE |
| LT1 | 43.73 | 0.161 | 0.005 | 0.106 | 0.004 | 0.732 | 0.003 |
| LT2 | 76.56 | 0.218 | 0.005 | 0.240 | 0.005 | 0.542 | 0.004 |
| LT3 | 92.20 | 0.088 | 0.003 | 0.101 | 0.003 | 0.811 | 0.003 |
| LT4 | 96.06 | 0.140 | 0.008 | 0.203 | 0.007 | 0.657 | 0.004 |
| LT5 | 97.26 | 0.101 | 0.005 | 0.117 | 0.004 | 0.782 | 0.003 |
| LT6 | 98.26 | 0.316 | 0.014 | 0.188 | 0.011 | 0.496 | 0.005 |
| LT7 | 98.63 | 0.409 | 0.016 | 0.198 | 0.013 | 0.394 | 0.005 |
| LT8 | 98.99 | 0.349 | 0.015 | 0.187 | 0.013 | 0.464 | 0.005 |

[1]Total phenotypic variance was calculated as the sum of additive genetic, permanent environmental, and residual variance.

mately 99% of the total spectral variation. Cumulative variance explained by the LT, as well as estimates and standard errors of variance ratios for additive genetic, permanent environmental, and residual effects on LT, are reported in Table 2. The first LT explained 44% of the total spectral variation and the first 3 LT explained more than 90% of the total spectral variation, indicating marked correlations among the original spectral variables. This is in agreement with results reported by Soyeurt et al. (2010) and Dagnachew et al. (2013a).

Heritability values for the 8 LT ranged from 0.09 and 0.4, and the last 3 LT exhibited the highest heritability estimates, indicating that a large proportion of the variation in the spectral information is not of genetic origin. This is in agreement with findings of other studies (Soyeurt et al., 2010; Dagnachew et al., 2013b).

Variance ratios for the permanent environment were between 0.10 and 0.24, and ratios for the residual variance ranged from 0.39 to 0.81. These results are consistent with those reported by Soyeurt et al. (2010) and Dagnachew et al. (2013b) for cow and goat milk spectra, respectively. The genetic correlations among LT are reported in Table 3. One property of the LT is that they are phenotypically uncorrelated but genetic relationships between LT may exist. Indeed, the magnitude of the additive genetic correlations between LT ranged from −0.72 to 0.78. Non-null genetic correlations between LT, ranging from 0.013 to 0.512, were reported also by Dagnachew et al. (2013b). Because of the large number of milk samples processed by FTIR, standard errors of heritabilities and correlation coefficients were very small.

### *Correlation Between EBV Obtained by IP and by DP*

Values of ρ between the EBV predicted using IP and DP are reported in Table 4. The ρ between EBV obtained from IP and those from DP was variable across traits, but all estimated ρ values were significantly different from zero ($P < 0.001$). Values of ρ ranged from 0.07 (for the content of SFA measured as g per 100 g of fat) to 0.96 (for dry matter cheese yield, %). For most traits, ρ was <0.5. For dry matter cheese yield, contents of casein, $\alpha_{S1}$-CN, $\beta\gamma$-CN, and lactoferrin, the EBV estimated by DP were strongly correlated ($\rho >$ 0.93) with those yielded by IP.

In contrast with our results, in Dagnachew et al. (2013b), the ρ between the EBV obtained by IP and DP was greater than 0.93 for all investigated traits (fat, lactose, and protein contents of milk). Some re-ranking of individuals was observed, but, according to those authors, it was attributable to the decrease in the EBV prediction error variance provided by DP compared with that of IP.

**Table 3**. Genetic correlations between latent variables (LT)[1]

| | LT2 | LT3 | LT4 | LT5 | LT6 | LT7 | LT8 |
|---|---|---|---|---|---|---|---|
| LT1 | −0.48 | −0.33 | −0.12 | −0.12 | −0.18 | 0.23 | 0.23 |
| LT2 | | −0.43 | −0.13 | 0.26 | 0.02 | −0.05 | 0.00 |
| LT3 | | | −0.17 | −0.34 | 0.41 | −0.35 | −0.22 |
| LT4 | | | | −0.55 | 0.13 | 0.25 | 0.44 |
| LT5 | | | | | −0.72 | 0.17 | −0.38 |
| LT6 | | | | | | −0.59 | 0.01 |
| LT7 | | | | | | | 0.78 |

[1]Standard errors ranged between 0.015 and 0.038.

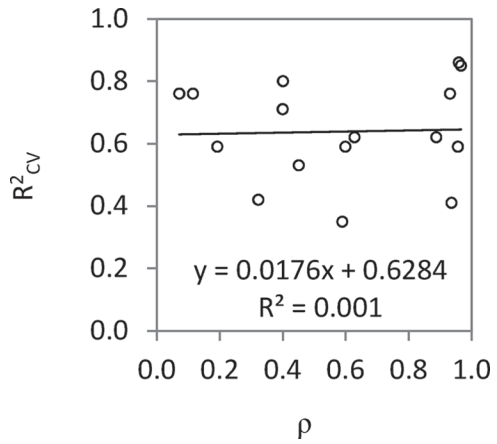


**Figure 3**. Relationship of the Spearman rank correlation between the EBV obtained by the indirect prediction and the direct prediction method (ρ) with the $R^2$ of cross-validation ($R^2_{CV}$) of the calibration equations.

As depicted in Figure 3, the magnitude of ρ was not related to the accuracy of the calibration equation that generated the predicted phenotypes (for IP) or the EBV (for DP). As a consequence, the high ρ values found by Dagnachew et al. (2013b) between EBV yielded by the 2 approaches were unlikely the result of the high accuracy of the calibration models used in that study.

Milk protein, fat, and lactose are the main sources of variation of spectral variables and, therefore, are also the traits mostly related to the LT. Consequently, it is reasonable to hypothesize that phenotypes strongly correlated with the major sources of variation of the spectra are also the traits for which variability is better explained by the LT with the highest eigenvalues. All traits having the highest ρ between IP and DP were among the traits having the strongest correlation with milk protein ($r > 0.52$; data not reported in tables), although the content of κ-CN (i.e., the trait predicted with the lowest accuracy; Table 1) exhibited a low ρ between EBV obtained by IP and DP, despite being highly correlated with milk protein content ($r = 0.83$; data not reported in tables). In general, however, a positive relationship was observed between ρ and the correlation of the traits with total milk protein or fat (Figure 4), confirming that the traits highly correlated with the major sources of variation of the spectra are also the traits for which variability is better explained by the first 8 LT and for which the DP approach would provide EBV highly correlated with those obtained by the IP method. This can explain the promising results obtained with DP on protein, fat, and lactose reported by Dagnachew et al. (2013b).

### *Correlation Between EBV Obtained from Raw and Back-Transformed Spectra*

The correlations between the EBV yielded by the IP, $BT_8$, and $BT_{21}$ methods is reported in Table 4. Values of ρ between EBV obtained by IP and $BT_8$ were very

**Table 4**. Spearman rank correlations between EBV and Pearson product-moment correlations between infrared-predicted phenotypes obtained using different methods[1]

| | EBV | | | | | Phenotype | |
|---|---|---|---|---|---|---|---|
| Trait | IP–DP | IP–$BT_8$ | IP–$BT_{21}$ | DP–$BT_8$ | DP–$BT_{21}$ | IP–$BT_8$ | IP–$BT_{21}$ |
| Technological traits | | | | | | | |
| pH | 0.40 | 0.41 | 0.55 | 0.97 | 0.79 | 0.53 | 0.72 |
| Coagulation time (min) | 0.40 | 0.41 | 0.57 | 0.98 | 0.77 | 0.41 | 0.62 |
| Raw cheese yield (%) | 0.89 | 0.95 | 0.99 | 0.97 | 0.88 | 0.82 | 1.00 |
| DM cheese yield (%) | 0.96 | 0.99 | 1.00 | 0.97 | 0.96 | 0.99 | 1.00 |
| Curd moisture (%) | 0.19 | 0.22 | 0.66 | 0.96 | 0.32 | 0.41 | 0.93 |
| Major protein fractions (g/L) | | | | | | | |
| Casein | 0.97 | 0.97 | 0.99 | 0.98 | 0.97 | 0.95 | 0.99 |
| $\alpha_{S1}$-CN | 0.93 | 0.93 | 0.97 | 0.98 | 0.95 | 0.88 | 0.96 |
| βγ-CN | 0.96 | 0.95 | 0.99 | 0.98 | 0.97 | 0.82 | 0.98 |
| κ-CN | 0.59 | 0.62 | 0.69 | 0.96 | 0.92 | 0.83 | 0.89 |
| Lactoferrin (μg/mL) | 0.94 | 0.95 | 0.98 | 0.98 | 0.95 | 0.90 | 0.99 |
| Ca (mg/kg) | 0.45 | 0.47 | 0.54 | 0.97 | 0.68 | 0.44 | 0.79 |
| Fatty acids (g/100 g of fat) | | | | | | | |
| SFA | 0.07 | 0.05 | 0.71 | 0.93 | 0.14 | 0.19 | 0.82 |
| MUFA | 0.11 | 0.10 | 0.78 | 0.93 | 0.17 | 0.23 | 0.84 |
| PUFA | 0.60 | 0.63 | 0.73 | 0.92 | 0.72 | 0.29 | 0.69 |
| n-3 | 0.63 | 0.59 | 0.74 | 0.91 | 0.67 | 0.28 | 0.73 |
| CLA *cis*-9,*trans*-11 | 0.32 | 0.34 | 0.66 | 0.98 | 0.41 | 0.34 | 0.72 |

[1]IP = indirect prediction method; DP = direct prediction method; $BT_8$ = IP method applied over spectra back-transformed using the 8 latent traits having the greatest eigenvalues; $BT_{21}$ = IP method applied over spectra back-transformed using the 21 latent traits having the greatest eigenvalues. All estimated correlations were significantly different from zero ($P < 0.001$).

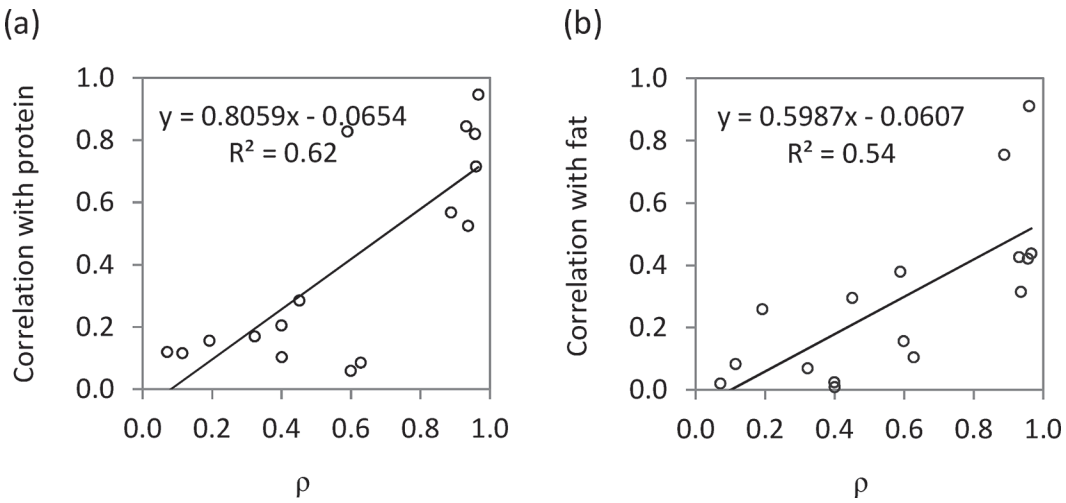


**Figure 4**. Relationship of the Spearman rank correlation between the EBV obtained by the indirect prediction and the direct prediction method (ρ) with the correlation of the traits with (a) milk protein, and (b) fat content.

similar to those observed between the EBV provided by IP and DP. The $BT_8$ method is analogous to the IP method for EBV estimation, in which only part of the original spectral information (the same as that used by the DP method) is used. Indeed, the ρ between EBV yielded by DP and $BT_8$ was generally very high ($\rho > 0.9$) for all the investigated traits (Table 4). This suggests that the use of the same amount of spectral information in a DP or IP approach yields consistent results. When the $BT_{21}$ approach was used, the rank correlation of the EBV with those obtained by IP improved for all investigated traits. However, ρ was greater than 0.95 only for 6 out of 16 traits. As the spectral information exploited in the DP approach was reduced compared with that exploited by the $BT_{21}$ method (8 vs. 21 LT), the ρ between EBV obtained by DP and $BT_{21}$ was lower than that between the EBV estimated from DP and $BT_8$ (Table 4).

### *Correlations Between Phenotypes Predicted from Raw and Back-Transformed Spectra*

The magnitude of the correlation between phenotypes predicted from the raw and the back-transformed spectra retaining 99% of the original spectral variability ($BT_8$) was variable across traits (Table 4). This could also explain why the ρ between EBV yielded by IP and DP was variable. For most traits, retaining 99% of the original spectral variability was not sufficient to guarantee good prediction accuracy compared with the accuracy obtained using the raw spectra. A considerable amount of the information needed to predict phenotypes is actually lost, which impairs the prediction of EBV from spectral information. Dagnachew et al. (2013b) suggested that the remaining 1% of the total spectral variation could also have relevant information for breeding and that the DP method could be improved to capture part of the remaining 1% of spectral variation.

When spectra were back-transformed using an increased number of LT to guarantee that 99.9% of the original spectral variability was retained ($BT_{21}$), the correlation between the phenotypes predicted from the raw spectra and those predicted from the back-transformed spectra increased for all investigated traits. Nevertheless, the correlation in most cases was lower than 0.90. This indicates that even the information included in the 0.01% of the original spectral variability, which was neglected by using the 21 LT with the highest eigenvalues, was fundamental for the prediction of some of the investigated traits.

This result can be explained by the nature of PCA, which does not take into account the covariance between the spectral variables and the reference traits, but produces LT that maximize the spectral variance explained. If most of the spectral variance is related to sources of variation that are weakly associated with a trait, the use of the DP method for that trait would not be feasible. The DP approach is more likely to be effective for the traits more related to the main sources of spectral variation; that is, protein and fat. The proportion of the original standard deviation of each spectral variable that is preserved by using 8 or 21 LT is reported in Figure 5. The standard deviation of some spectral variables was greatly reduced when 8 LT were used for spectra back-transformation. If the

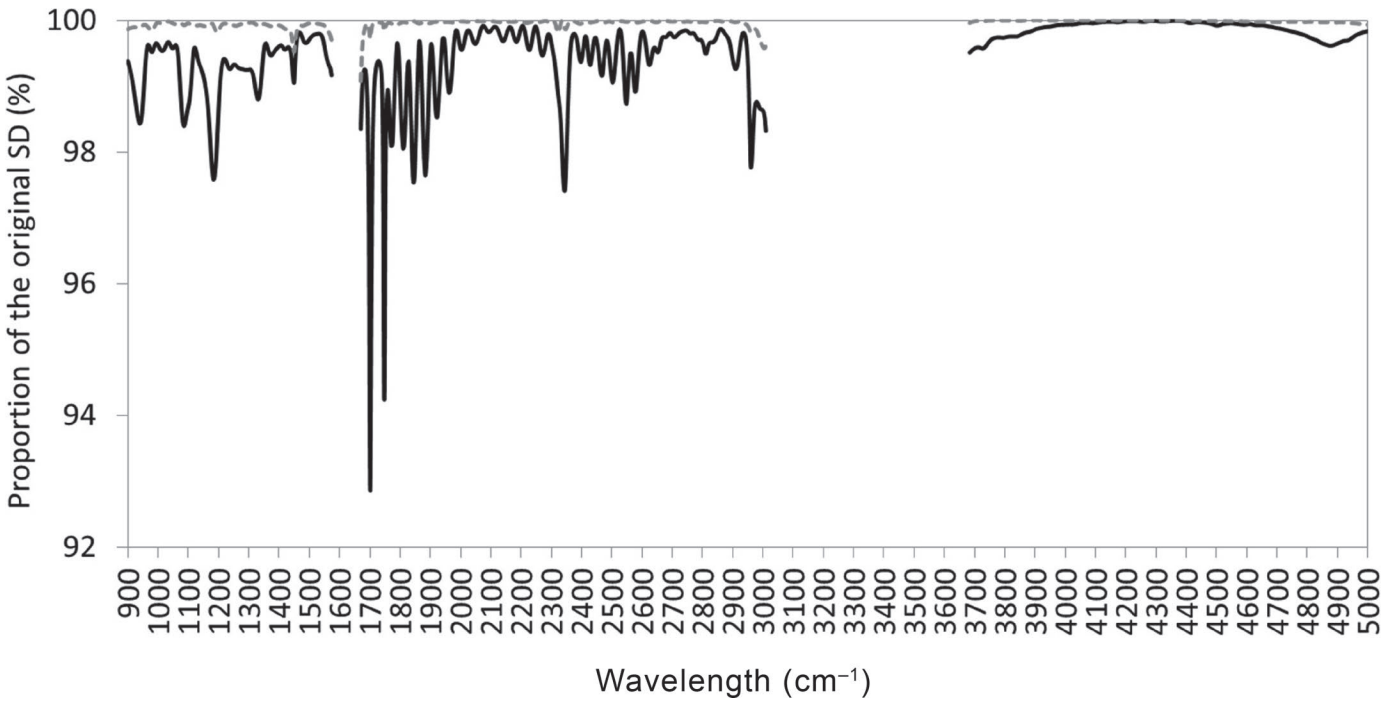


**Figure 5**. Proportion of the original spectral standard deviation obtained with the spectra back-transformation using 99% (solid line) or 99.9% (dashed line) of the original spectral variation.

spectral regions related to a particular trait were also the regions more affected by the loss of variability, the ρ for that particular trait would be low.

As it is not possible to estimate (co)variance components for all spectral variables (i.e., 872 spectral data points) simultaneously, there is a need to reduce the spectral dimension. Principal component analysis is the method that has been used to date (Soyeurt et al., 2010; Dagnachew et al., 2013a), but it extracts information focusing only on the magnitude of total variation explained by a component. This limits the ability of PCA to retain all the relevant variation for a specific trait.

### PCA Versus Partial Least Square Regression

To better investigate the ability of the PCA to retain information about the traits, calibration equations were developed by principal component regression implemented in the WinISI II software, using a maximum of 8 LT. For most of the investigated traits, the calibration based on principal component regression (i.e., a technique that is based on PCA) has a much lower predictive ability than that of the equation based on MPLS, whereas for traits that are easily predicted by FTIR spectroscopy (e.g., casein content or dry matter cheese yield), the 2 approaches are similar (data not reported in tables). The magnitude of the ρ was strongly correlated (r = 0.83) to the difference in the $R^2_{CV}$ obtained by MPLS and principal component regression (Figure 6).

### Use of the LT Most Correlated with the Traits

In the DP approach, the first 8 LT were used because they had the greatest eigenvalues, but, as stated above, the LT having the greatest eigenvalues are not neces-

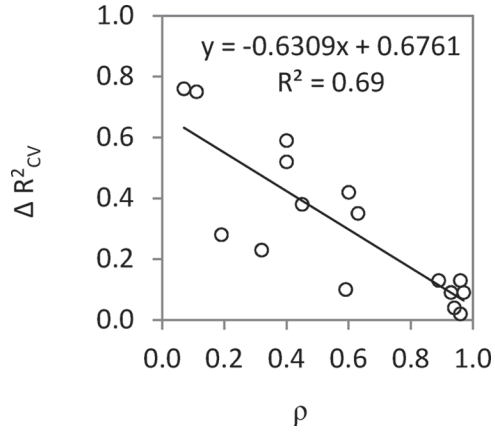


**Figure 6**. Relationship between the rank correlation between the EBV estimated by the direct and by the indirect prediction method (ρ) and the difference between the $R^2$ in cross-validation obtained by modified partial least square regression and principal component regression ($\Delta R^2_{CV}$).

**Table 5**. Latent traits (LT) most correlated with the content of SFA

| LT | Spectral variance explained (%) | Correlation with SFA |
|---|---|---|
| LT11 | 0.11 | 0.54 |
| LT9 | 0.23 | 0.49 |
| LT25 | 0.01 | 0.23 |
| LT5 | 1.21 | 0.23 |
| LT3 | 15.65 | 0.21 |
| LT22 | 0.01 | 0.18 |
| LT10 | 0.22 | 0.16 |
| LT1 | 43.73 | 0.16 |

sarily those retaining the greatest part of the genetic variation. Other LT might play a fundamental role in the prediction of some traits even if the amount of variance explained is limited. This problem might be addressed by using a supervised PCA, which is similar to conventional PCA except that it uses a subset of the LT based on their association with the trait (Bair et al., 2006). Also, partial least squares regression might yield a more accurate estimation of genetic parameters for the traits included in the model because it captures relevant variations of the spectra associated with those traits. However, it will not guarantee that information for other milk traits, which are not included in the partial least squares model, are retained in the LT.

To support our hypothesis, as an example, we estimated the EBV for SFA content using the DP method on the 8 LT most correlated with the trait instead of the 8 LT with the largest eigenvalues. The 8 LT most correlated with SFA content, as well as the proportion of spectral variance explained, are reported in Table 5. The ρ between the EBV yielded by the IP method and by the DP method using this approach was 0.68 ($P < 0.001$), a value much higher than that (0.07, $P < 0.001$) between the EBV yielded by IP and DP using the 8 LT having the greatest eigenvalues, and only slightly lower than the ρ observed between the EBV yielded by IP and $BT_{21}$ (0.71, $P < 0.001$).

The application of other methods for the reduction of spectral variables or for the selection of spectral regions should be investigated because the direct use of spectral information has a lot of potential. Soyeurt et al. (2010) suggested that direct use of the spectral information might be used to develop herd management tools, because metabolic disorders such as acidosis, ketosis, or mastitis influence many aspects of milk composition that could affect milk spectra. Based on the difference between the expected and observed values for the spectral traits, some disorders, currently undiscovered because of the limited number of studied milk components, could be detected (Soyeurt et al., 2010). In addition, reduction of the spectral dimension might reduce the requirements for storage of spectral variables, because only the storage of phenotypes of the few selected latent variables would be needed.

## CONCLUSIONS

We compared the common method of exploiting infrared predictions in animal breeding and an alternative approach based on the direct use of the spectral information to predict the EBV for several traits related to fine composition and technological properties of milk. Latent variables produced by PCA accounted for as much of the variability in the spectra as possible, but the covariance between the trait and the spectra was neglected. Principal components analysis was not suitable for reducing the number of spectral variables for the direct use of spectral information in genetic evaluation, except for traits that are highly correlated with fat and protein. All of the spectral variability associated with the traits should be retained in the latent variables, which means that an increased number of latent variables or methods to reduce the dimensionality of the spectra other than PCA should be used. More research is required to study spectral genetic variation and determine the best way to choose spectral regions and the type and number of latent traits for potential application.

## ACKNOWLEDGMENTS

The Friuli Venezia Giulia Milk Recording Agency (AAFVG, Codroipo, Italy) is gratefully acknowledged for milk samples collection. Financial support for this study was provided by the University of Padova (Progetto di Ateneo 2012, CPDA122982).

## REFERENCES

Bair, E., T. Hastie, D. Paul, and R. Tibshirani. 2006. Prediction by supervised principal components. J. Am. Stat. Assoc. 101:119–137.

Bonfatti, V., L. Degano, A. Menegoz, and P. Carnier. 2016. Short communication: Mid-infrared spectroscopy prediction of fine milk composition and technological properties in Italian Simmental. J. Dairy Sci. 99:8216–8221.

Bonfatti, V., L. Grigoletto, A. Cecchinato, L. Gallo, and P. Carnier. 2008. Validation of a new reversed-phase high-performance liquid chromatography method for separation and identification of bovine milk protein genetic variants. J. Chromatogr. A 1195:101–106.

Burns, D. A., and E. W. Ciurczak. 2007. Handbook of Near-Infrared Analysis. 3rd ed. CRC Press, Boca Raton, FL.

Dagnachew, B. S., A. Kohler, and T. Ådnøy. 2013a. Genetic and environmental information in goat milk Fourier transform infrared spectra. J. Dairy Sci. 96:3973–3985.

Dagnachew, B. S., T. H. E. Meuwissen, and T. Ådnøy. 2013b. Genetic components of Fourier-transform infrared spectra used to predict breeding values for milk composition and quality traits in dairy goats. J. Dairy Sci. 96:5933–5942.

Gengler, N., H. Soyeurt, F. Dehareng, C. Bastin, F. Colinet, H. Hammami, M.-L. Vanrobays, A. Lainé, S. Vanderick, C. Grelet, A. Vanlierde, E. Froidmont, and P. Dardenne. 2016. Capitalizing on

fine milk composition for breeding and management of dairy cows. J. Dairy Sci. 99:4071–4079.
Groeneveld, E., M. Kovac, and N. Mielenz. 2010. VCE User's Guide and Reference Manual. Version 6.0. Department of Animal Science, University of Illinois, Urbana.
Pellattiero, E., A. Cecchinato, F. Tagliapietra, S. Schiavon, and G. Bittante. 2015. The use of 2-dimensional gas chromatography to investigate the effect of rumen-protected conjugated linoleic acid, breed, and lactation stage on the fatty acid profile of sheep milk. J. Dairy Sci. 98:2088–2102.
Shenk, J. S., and M. O. Westerhaus. 1991. Population definition, sample selection, and calibration procedures for near infrared reflectance spectroscopy. Crop Sci. 31:469–474.
Soyeurt, H., D. Bruwier, J.-M. Romnee, N. Gengler, C. Bertozzi, D. Veselko, and P. Dardenne. 2009. Potential estimation of major mineral contents in cow milk using mid-infrared spectrometry. J. Dairy Sci. 92:2444–2454.
Soyeurt, H., F. G. Colinet, V. M.-R. Arnould, P. Dardenne, C. Bertozzi, R. Renaville, D. Portetelle, and N. Gengler. 2007. Genetic variability of lactoferrin content estimated by mid-infrared spectrometry in bovine milk. J. Dairy Sci. 90:4443–4450.
Soyeurt, H., I. Misztal, and N. Gengler. 2010. Genetic variability of milk components based on milk infrared spectral data. J. Dairy Sci. 93:1722–1728.
Thermo Scientific Dionex. 2016. All Application Notes, Updates, and Briefs. Accessed Jul. 22, 2016. http://www.dionex.com/en-us/documents/application-notes-updates/lp-84398.html.