Bayesian and partial least square global forage calibrations models developed by an iterative procedure using R

A. Ferragina,* F. Benozzo and P. Berzaghi

University of Padova, Italy. E-mail: aferragina84@gmail.com

The aim of our study was to test an iterative process of validation implemented in the R software, assessing the accuracy of the best selected equations, developed using two different regression algorithms Partial Least Square (PLS) and Bayesian. A data set (Set_a) with 3187 records of 6 different types of forages was used. The calibrations were tested for Protein, Neutral Detergent Fiber and Acid Detergent Fiber. For each sample a spectrum was collected using a FOSS NIRSystem (1100–2498 nm). A subset composed of 20 samples for each type of forage (Set_{ext};120 samples) was randomly selected for a final validation of the best selected equations. The remaining samples ($Set_b = Set_a - Set_{ext}$) were used for the iterative calibration process. For each iteration the Set_b was randomly divided in a testing set (Set_{tst}; 10% of Set_b) and a training set ($Set_{trn} = Set_b - Set_{tst}$); 300 iterations were done. All of the computations were done in the R environment. The packages used were "pls" for the PLS, "BGLR" for the Bayesian, "prospectr" for the spectral treatments. In each iteration we used three spectral treatments (raw, 1 derivative, standard normal variate and detrend), two approaches for selection of the optimal number of PLS components and the Bayesian model. Nine types of equations were developed and tested in each iteration [(2 PLS techniques + 1 Bayesian) × 3 spectral treatments]. Among the 300 iterations, for each one of the 9 equation types, the best one (lowest RMSE) and the average of the best 25% (RMSE < 1 quartile) were selected and validated by forage type. R has demonstrated its potential when used for the chemiometric process on big data set and with complex statistical procedures. R^2 higher than 0.9 was obtained for almost all the calibrations. In the external validation the Bayesian models in many cases outperform the commonly used PLS, demonstrating that an alternative for the improvement of the prediction accuracy exists. The present work has demonstrated that iterative validation subsampling on big data can lead to the selection of proper equations, and it can be done using R.

Introduction

PLS is the most common regression model used for calibrations and its optimization is based on cross validation. A high number of external validation samples is important for the assessment of the prediction equations, but it is not always possible because of the data set (e.g. not enough samples) or the time needed using commercial software. The importance of a robust validation for a global equation is high when a multiproduct data set is used. Many software packages are available for the calibration process, among these R[1] is an open source program that gives a high variety of chemometric tools, included modelling packages based on the Bayesian approach. In recent years a new application of the Bayesian models for calibration was proposed.[2] The BGLR package[3] is commonly used for genomic analysis, and has demonstrated[2] that it has high potentiality when used also for chemometric purposes. The aims of this work was to compare the PLS and Bayesian models when used to develop a global equation on a multiproduct data set, through an iterative validation process developed in R.

Materials and methods

Sample set

Samples of six types of forages were used. The total number of samples used was 3116 [486 samples for corn silage

Correspondence
A. Ferragina (aferragina84@gmail.com)
doi: 10.1255/nir2017.051
Citation: A. Ferragina, F. Benozzo and P. Berzaghi, "Bayesian and partial least square global forage calibrations models developed by an iterative procedure using R", in *Proc. 18th Int. Conf. Near Infrared Spectrosc.*, Ed by S.B. Engelsen, K.M. Sørensen and F. van den Berg. IM Publications Open, Chichester, pp. 51–55 (2019). https://doi.org/10.1255/nir2017.051

ISBN: 978-1-906715-27-4

Table 1. Descriptive statistics.

	Protein			NDF			ADF		
	N	Mean (%)	SD	N	Mean (%)	SD	N	Mean (%)	SD
CSL	477	9.25	2.40	287	49.45	8.26	477	27.41	5.04
HAY	1207	17.10	5.05	544	54.34	11.38	1209	35.15	6.42
HLG	690	17.99	4.18	490	54.23	9.16	690	38.87	5.72
SGS	280	13.60	5.17	202	58.25	9.50	282	35.47	7.54
TMR	276	12.65	4.68	65	45.05	12.65	227	23.77	10.81
PRO	119	10.59	5.08	117	66.23	7.47	119	36.15	4.55
ALL	3049	15.10	5.55	1705	54.41	10.78	3004	33.98	8.01

CSL = Corn Silage; HLG = Haylage; SGS = Small Grain Silage; TMR = Total Mixed Ration; PRO = Experimental trial of different species; ALL = CSL+HAY+HLG+SGS+TMR+PRO.

(CSL), 1231 samples for alfalfa (HAY), 701 samples for haylage (HLG), 287 samples for small grain silage (SGS), 283 samples for total mixed ration (TMR), 128 samples of different species from an experimental trial (PRO)]. For each forage type, portions of the samples were analyzed, by the corresponding reference methods, for Protein, Neutral Detergent Fiber (NDF) and Acid Detergent Fiber (ADF; Table 1).

For all the samples the spectra were collected using a FOSS NIRSystem 5000 in the range between 1100 nm and 2498 nm (9090–4003 cm^{-1}) every 2 nm.

Statistical analysis

All the statistical analysis was done using the open source software R. Prior to data analysis the outlier spectra were detected using the Mahalanobis distances. The outlier detection was done by product and the spectra showing a distance higher than the mean plus three standard deviations were considered outliers and deleted from the data set.

Two regression techniques were used, the Bayesian model (i.e. Bayes B) implemented in the R package "BGLR" and the partial least square regression (PLS) implemented in the R package "pls".[4]

The package "prospectr"[5] was used for the mathematical treatments of the spectra. The standard normal variate and detrend (snvd) and the first order derivative (1sgd) were used. The models were fitted also on raw spectra.

After the outlier detection and before the calibration process, 20 samples for each product (120 samples) were random selected as external validation set (SET_{EXT}) and used for the validation of the selected equations at the end of an iterative process.

Iterative process of testing. The calibration was based on an iterative procedure, where for each round the data set (SET_B) was divided in a training set (SET_{TRN}) used to generate the equation and a testing set (SET_{TST}) for a first validation and the optimization of the PLS components number. Three hundred iterations were

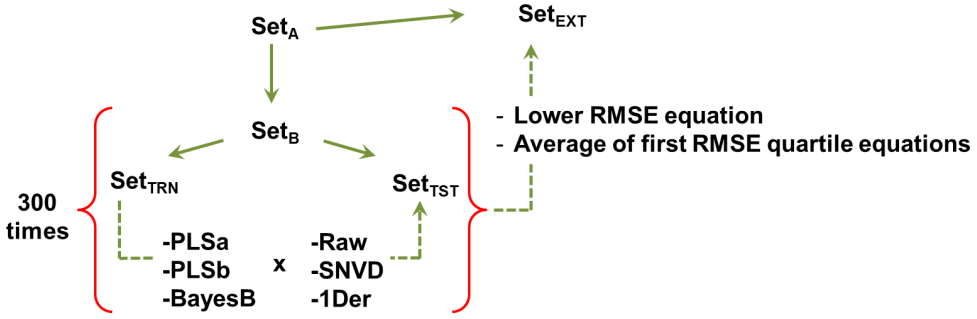

Figure 1. Diagram of the data analysis.

performed and for each iteration all the models were tested on the same TRN and TST sets. The number of samples included in the SET_{TST} was 10 % of all the samples (Figure 1).

Bayesian model. The Bayes B (BB) model implemented in the "BGLR" package of R was used. A detailed description of the model and algorithms can be found in Pérez and de los Campos[6] as well as the default parameters used. As example for the application of BGLR for infrared calibration can be found in Ferragina *et al.*[2] A low number of Bayesian iterations and burn-in were used, 15,000 and 5000, respectively.

PLS. The PLS was fitted using the statement "plsr" included in the package "pls". The maximum number of principal components tested in each PLS was 20, and 10 cross-validation segments were used.

The plsr gives as result a number of equations equal to the maximum number of principal components tested, thus, the choice of the optimum number of principal components (ONPC) was done applying each equation on the SET_{TST} and calculating for each one of the 20 equations the root mean squared error of validation ($RMSE_{TST}$) and using two intuitive algorithms (a and b). The algorithms were both based on a penalization system where the choice of increasing the number of the components was done according to the $RMSE_{TST}$.

The first choice of the ONPC (a) was done according to the first component. The $RMSE_{TST}$ of all the components were penalized and compared with that of the first. The first algorithm was specified as follow:

$$Dif = \mathrm{RMSE}_{\mathrm{TST}_1} - \mathrm{RMSE}_{\mathrm{TST}_{i+1}}$$

$$Pen = \left(\frac{\%pen \times i}{100}\right) \times \mathrm{RMSE}_{\mathrm{TST}_1}$$

Vet = TRUE for $Dif > Pen$ and Vet = FALSE for $Dif < Pen$

where $\mathrm{RMSE}_{\mathrm{TST}_1}$ is the $RMSE_{TST}$ for the equation of the first component; $i = 1, \ldots, 19$; *%pen* is the percentage of penalization (5 %); *Vet* is a logical vector of dimension $i + 1$ (20), where at position 1 *Vet* = True. The ONPC was the highest dimension of TRUE in the vector *Vet*.

The second algorithm (b) was based on the comparison of the $RMSE_{TST}$ of each equation with the next. In the second case the algorithm was specified as follows:

$$Dif = \mathrm{RMSE}_{\mathrm{TST}_i} - \mathrm{RMSE}_{\mathrm{TST}_{i+1}}$$

$$Pen = \left(\frac{\%pen}{100}\right) \times \mathrm{RMSE}_{\mathrm{TST}_i}$$

Vet = TRUE for $Dif > Pen$ and Vet = FALSE for $Dif < Pen$

where $\mathrm{RMSE}_{\mathrm{TST}_i}$ is the $RMSE_{TST}$ for the equation of the i^{th} component; $i = 1, \ldots, 19$; *%pen* is the percentage of penalization (5 %); *Vet* is a logical vector of dimension $i + 1$ (20) (with TRUE for the first component). The ONPC was the highest dimension of TRUE in the vector *Vet*.

External validation. For each model 300 equations were generated (2700 equations for each trait), and among these the equation with the lowest $RMSE_{TST}$ and those included in the first quartile of the $RMSE_{TST}$ were selected and used to predict the samples in the SET_{EXT}. The predictions obtained using the first quartile equations for each sample of the SET_{EXT} were averaged. The R^2 and RMSE were calculated.

Results and discussion

In Table 1 the descriptive statistics by product are reported. The number of samples for each product was not balanced and a big variability was shown for all the traits among products.

The prediction results obtained in the iteration process are reported in Table 2. The R^2_{TST} and $RMSE_{TST}$ shown are the average of the 300 equations tested for each model. High R^2_{TST} values were found in validation for all the predicted traits. The R^2_{TST} was similar for almost all the models. According with the $RMSE_{TST}$ the PLS gave the best results for all the traits, ranging from 0.77 (PLS, a, 1sgd) to 4.92 (BB, raw) for protein, from 2.58 (PLS, a, 1sgd) to 14.65 (BB, raw) for NDF, from 1.66 (PLS, a snvd) to 12.76 (BB, raw) for ADF. On average the PLS b uses a lower ONPC, and the total explained variance near 100 % for both the PLS (data not shown). An important effect of the spectra treatment was shown on the BB, were the error was drastically reduced by implementing spectral treatments.

Among all the equations, the one with a lower $RMSE_{TST}$, and the equations with a $RMSE_{TST}$ in the first quartile were selected and validated on the external set by product. When the equations of the first quartile were used, the predictions were averaged. Only the results of the external validation for all the products together are reported in Table 3.

Table 2. Fitting statistics of validation in the iteration process.

Model	Treatment	Protein		NDF		ADF	
		R^2_{TST}	$RMSE_{TST}$	R^2_{TST}	$RMSE_{TST}$	R^2_{TST}	$RMSE_{TST}$
PLS	a, raw	0.98	0.85	0.93	2.86	0.93	2.13
PLS	a, 1sgd	0.98	0.77	0.94	2.58	0.96	1.69
PLS	a, snvd	0.98	0.79	0.94	2.60	0.96	1.66
PLS	b, raw	0.97	0.93	0.93	2.93	0.94	1.97
PLS	b, 1sgd	0.98	0.87	0.93	2.73	0.95	1.80
PLS	b, snvd	0.97	0.89	0.94	2.70	0.95	1.69
BB	raw	0.91	4.92	0.70	14.65	0.69	12.76
BB	1sgd	0.97	1.95	0.93	4.78	0.94	3.90
BB	snvd	0.95	3.12	0.89	7.37	0.87	6.75

R^2_{TST} and $RMSE_{TST}$ = average of the 300 iterations.

The selection of the best equations gives a lower RMSE respect to the $RMSE_{TST}$, and in general no big differences were found among the results of the lower RMSE equation and the first RMSE quartile equations. Among PLS and BB no big differences were found, furthermore in many cases BB outperform PLS comparing the results of the single products (data not shown).

In Figure 2 an example of the estimated coefficients of PLS and BB, and the correlation among the trait (NDF) and each wavelength are shown. In general the coefficients do not follow the correlation trend as could be expected. The estimated coefficients of PLS are highly represented along all the spectral range with homogenous peaks. Differently, the effect of the variable selection of BB, tends to select important groups of coefficients reducing the others toward zero.

Conclusions

R has demonstrated its potentiality when used for the chemometric process on big data set and with complex statistical procedures. In the external validation the Bayesian models outperform PLS, demonstrating that an alternative for the improvement of the prediction accuracy exists. The present work has demonstrated that iterative validation subsampling on big data can lead to the selection of proper equations, and it can easily be done using R.

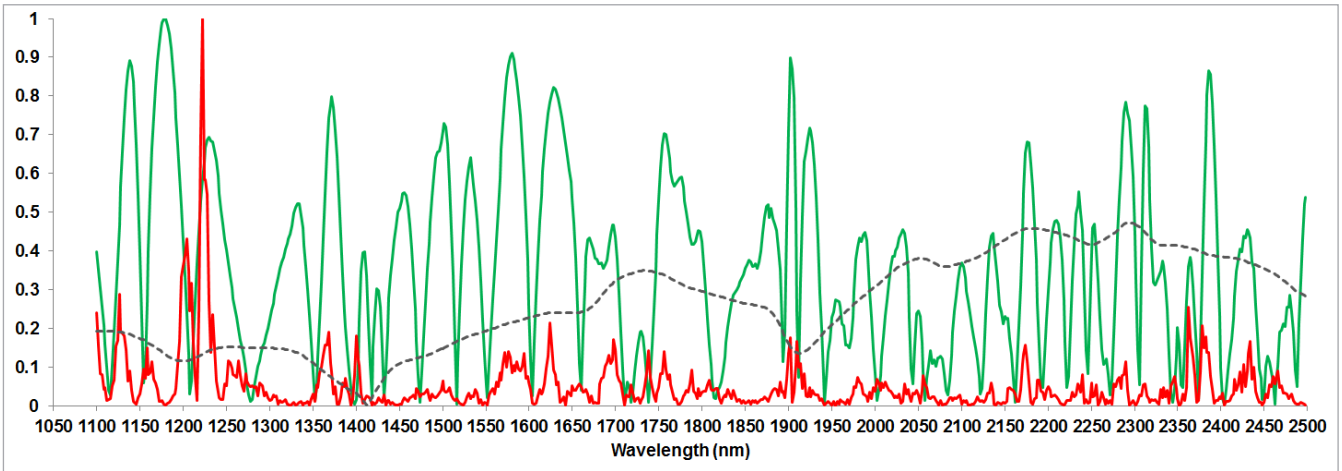

Figure 2. Absolute values of estimated coefficients for PLS (green line) and Bayes B (red line), and correlation of NDF and absorbances by wavelength (dashed curve).

Table 3. Statistics of the external validation for all the products.

		Protein				NDF				ADF			
		Low_{RMSE}		$1Quart_{RMSE}$		Low_{RMSE}		$1Quart_{RMSE}$		Low_{RMSE}		$1Quart_{RMSE}$	
Model	Treatment	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
PLS	a, raw	0.97	0.90	0.97	0.90	0.93	3.58	0.93	3.75	0.95	1.97	0.96	1.96
PLS	a, 1sgd	0.98	0.76	0.98	0.76	0.94	3.29	0.94	3.30	0.97	1.61	0.97	1.62
PLS	a, snvd	0.98	0.67	0.98	0.75	0.94	3.19	0.94	3.20	0.97	1.63	0.97	1.68
PLS	b, raw	0.97	0.87	0.97	0.87	0.94	3.43	0.93	3.67	0.97	1.73	0.96	1.76
PLS	b, 1sgd	0.98	0.73	0.98	0.83	0.93	3.46	0.93	3.43	0.97	1.58	0.97	1.65
PLS	b, snvd	0.98	0.70	0.98	0.78	0.94	3.19	0.94	3.10	0.97	1.58	0.97	1.64
BB	raw	0.92	1.53	0.93	1.50	0.81	5.16	0.82	5.05	0.81	3.99	0.78	4.25
BB	1sgd	0.98	0.79	0.98	0.75	0.94	3.13	0.94	3.17	0.97	1.50	0.97	1.50
BB	snvd	0.98	0.84	0.98	0.75	0.94	3.22	0.94	3.12	0.97	1.77	0.97	1.61

Low_{RMSE} = selected equation with the lowest $RMSE_{TST}$. $1Quart_{RMSE}$ = selected equations with the $RMSE_{TST}$ in the first quartile.

References

1. R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. (2015) http://www.R-project.org/
2. A. Ferragina, G. de los Campos, A.I. Vazquez, A. Cecchinato and G. Bittante, "Bayesian regression models outperform partial least squares methods for predicting milk components and technological properties using infrared spectral data", *J. Dairy Sci.* **98(11),** 8133–8151 (2015). https://doi.org/10.3168/jds.2014-9143
3. G. de los Campos and P. Perez Rodriguez, *BGLR: Bayesian Generalized Linear Regression*. R package version 1.0.4 (2015). http://CRAN.R-project.org/package=BGLR
4. B.-H. Mevik, R. Wehrens and K. Hovde Liland, *pls: Partial Least Squares and Principal Component Regression*. R package version 2.5-0 (2015). http://CRAN.R-project.org/package=pls
5. A. Stevens and L. Ramirez-Lopez, *An Introduction to the prospectr Package*. R package Vignette R package version 0.1.3 (2013).
6. P. Pérez and G. de los Campos, "Genome-wide regression and prediction with the BGLR statistical package", *Genetics* **198(2),** 483–495 (2014). https://doi.org/10.1534/genetics.114.164442