GMR

The Original by FUNPEC-RP

Comparison of variable selection methods in predictive models applied to near-infrared and genomic data

R.A. Ferreira[1,2] and L.A. Peternelli[1]

[1] Universidade Federal de Viçosa, Viçosa, MG, Brasil
[2] Instituto Federal de Minas Gerais, Governador Valadares, MG, Brasil

Corresponding author: R.A. Ferreira; L.A. Peternelli
E-mail: roberta.amorim@ifmg.edu.br; peternelli@ufv.br

Genet. Mol. Res. 20 (2): gmr18909
Received May 28, 2021
Accepted June 28, 2021
Published July 29, 2021
DOI http://dx.doi.org/10.4238/gmr18909

ABSTRACT. Many research areas have datasets that face the challenges of high dimensionality and multilinearity. Although existing methods are efficient for constructing a complete model, it is often necessary to select the most important explanatory variables to obtain more parsimonious models. We evaluated and built models using three methods of selection of variables applied to data of single nucleotide polymorphism (SNP) markers and near-infrared spectroscopy (NIR), in addition to assessing the improvement in prediction quality when compared to the use of complete data. These included ordered predictors selection associated with partial least squares regression (PLS-OPS), sparse partial least squares regression (SPLS), and Supervised BLasso, the latter being an adaptation of the Bayesian Lasso (BLasso) method for variables selection. We used simulated data sets evaluated in two scenarios, and three real data sets, composed of one set of SNPs and two sets of NIR data. The predictive quality of each model was evaluated based on the mean correlation coefficient between predicted and actual values, and the square root mean squared error. In the set of simulated data evaluated in the first scenario, in terms of predictive capacity, the models after variables selection were similar when compared to the use of the complete data model, whereas in the second scenario, on average, the models performed better after the selection of variables, with SPLS

being superior to the other methods. In the real SNPs dataset, the PLS-OPS had a good performance, attesting the usefulness of this method for this kind of data. In the NIR datasets, the predictive quality of models after variable selection were close to those obtained with the complete data. In general, when using the selection methods, the models maintained a good predictive capacity and became simpler due to the considerable reduction in the number of variables.

Key words: PLS regression; BLASSO; OPS; Chemometrics; Prediction power

INTRODUCTION

In general, any experiment to be evaluated can be modeled by a statistical function that contains the variables under study. When the models present many correlated (multicollinearity) explanatory variables associated with few observations (high dimensionality), traditional methods for constructing models for the prediction cannot be used. In these cases, specific techniques are employed for the model fitting.

Although existing multivariate methods are, in most cases, efficient for the model construction, it is often necessary to select the most critical variables that can guarantee more parsimonious models and, in some cases, with more predictive capacity (Cai et al., 2008). Some research areas originally have datasets with high dimensionality and multilinearity. We can cite the Genome-Wide Association (GWAS), which consists of evaluating the association between data of SNPs (*Single Nucleotide Polymorphisms*) markers to some phenotype of interest (Beck et al., 2014) and chemometrics, a discipline of chemistry that consists of analyzing sets of multivariate data of chemical origin using statistical methods (Ferreira et al., 1999). One way to get information about the chemical properties of samples is by using near-infrared spectroscopy (NIR) (Dardenne et al., 2000; Da Silva et al.,2017; De Lima et al., 2020).

NIR spectroscopy, coupled with multivariate statistical methods, such as partial least squares (PLS) regression, has been used to make predictions of some specific property of samples in place of high-cost laboratory methods (Cozzolino et al., 2004; Spahn et al., 2008). Its use is simple, fast, accurate, and does not generate waste in the environment (Valderrama et al., 2007; Morgano et al., 2008). For NIR data, the PLS regression method proved to be more efficient in dealing with experimental noises (Teófilo et al., 2009).

Among the methods of selecting variables used in chemometrics, the ordered predictors' selection (OPS) (Teófilo et al., 2009; Roque et al., 2019) proved to be efficient in the selection of NIR data variables (Costa and De Lima, 2013; Guimarães et al., 2016; Caliari et al., 2017; Castro et al., 2019; Yu et al., 2020; Oliveira et al., 2021). The OPS method combines PLS adjustment with some criteria for variable selection (PLS-OPS).

In the GWAS, after using criteria for variable selection, we can identify which SNPs are responsible for the variation of the phenotypic characteristic of interest. To perform the modeling of SNPs data, de Los Campos et al. (2009) suggest some statistical methods under Bayesian focus, such as the Bayesian Lasso (BLasso) (Park and Casella, 2008). The BLasso with a selection of variables (Supervised BLasso) functions as a method of fitting the data while selecting variables once a significance limit is established so that the regression coefficients below this limit are discarded.

In addition to the selection methods mentioned, we can highlight the sparse partial least squares (SPLS) regression, proposed by Chun and Keles (2010), used in different datasets (Feng et al., 2012; Colombani et al., 2012; Abdel-Rahman et al., 2014). SPLS can be applied to a large set of highly correlated data. The method was developed based on PLS and has the advantage of simultaneously reducing the dimensionality of the data and selecting the variables efficiently (Chun and Keles, 2010). Therefore, it also works as a selection method.

We evaluated and constructed models using three variable selection methods (PLS-OPS, supervised BLasso, and SPLS) on SNPs and NIR data. We ran the analysis on real and simulated data of SNP markers and on real NIR data. The effect of variable selection on the predictive quality of the reduced model compared to the full model was also evaluated.

MATERIAL AND METHODS

Obtaining synthetic data

We simulated datasets of SNP markers with N independent individuals and p SNPs. For each simulation, in addition to the X matrix of markers, the vector y of phenotypic observations was obtained. We built the scripts to generate the simulated data using the software R (R Core Team, 2017), following the algorithm presented in Feng et al. (2012).

The simulated genotypes form the rows of the SNP markers matrix X. Genotypes are generated from two simulated independent haplotypes. Considering biallelic SNPs, these can be represented by 0 and 1. This situation can be exemplified in Table 1, in which an individual's genotype vector was constructed from haplotype vectors, considering p = 10 SNPs markers.

Table 1. Hypothetical example of an individual's genotype.

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
Vectors of	0	1	1	1	1	0	0	1	0	1
Haplotypes	1	0	0	1	1	1	0	1	0	0
Genotype	0/1	1/0	1/0	1/1	1/1	0/1	0/0	1/1	0/0	1/0

For defining each x_{ij} element of X, we considered the following encoding:

$$\mathbf{x_{ij}} = \begin{cases} -1 \text{ if the ith SNP has genotype } 0/0, \\ 0 \text{ if the ith SNP has genotype } 0/1 \text{ or } 1/0, \\ 1 \text{ if the ith SNP has genotype } 1/1 \end{cases}$$

Therefore, according to the example proposed in Table 1, the first row of matrix X will consist of:

$$[0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad -1 \quad 1 \quad -1 \quad 0]$$

Simulation description

The array of SNPs markers

The Bernoulli distribution was used, which provides us with the values 0 or 1 to define each haplotype vector's value. Therefore, considering $H^t = (H_1,\ldots,H_p)$ a vector of dimension p×1 of Bernoulli variables, that is, H_i~Bernoulli(μ_i) with i = 1,..., p, the means vector will be given by $\mu = (\mu_1,\ldots,\mu_p)$. We will summarize the simulation of the X matrix of markers in 4 steps:

STEP 1: Initially, 500 SNPs markers were defined for the construction of the dataset.

STEP 2: We obtained the vector of marginal means of each SNP. The vector of means $\mu_i = (\mu_1, \ldots, \mu_{500})$ with I = 1,…, 500 was obtained from the simulation based on the uniform distribution (0.1, 0.9).

STEP 3: We calculated the matrix of Covariances V of H. For this, considering H_i~Bernoulli(μ_i), a matrix was constructed in which the main diagonal contained the values V_{ii} = var (H_i) = μ_i (1- μ_i) and zero off-diagonal. Following that, we obtained the correlation matrix ρ_{ij}, i, j = 1, …, 500, in which ρ_{ij} = 1, when i = j. To ensure that closer SNPs are more correlated, we obtained the correlation matrix values from a uniform distribution accordingly to the distance between pairs of SNPs (Table 2):

Table 2. Correlation between simulated SNPs of uniform distribution with different intervals dependent on distance difference |i-j| between pairs of SNPs.

|i -j|	Range
1	(0.6-0.9)
2	(0.4-0.6)
3	(0.3-0.6)
4	(0.3-0.5)
5	(0.2-0.5)
6	(0.2-0.4)
7	(0.1-0.4)
8	(0.1-0.3)
9	(0.1-0.2)
10	(0-0.1)

Finally, we obtained the covariance matrix V of **H** by making its main diagonal composed of the values V_{ii} = var (H_i) = μ_i (1- μ_i) and the off-diagonal elements by $V_{ij} = V_{ji} = \mathrm{cov}(H_i, H_j) = \rho_{ij}\sqrt{V_{ii}V_{jj}}$, with $i \neq j$.

STEP 4: With the covariance matrix V and the SNPs mean vector, we simulated the haplotypes. The algorithm for finding the H_i haplotype vectors of individuals can be summarized as follows:

We generate H_1 from a Bernoulli distribution (μ_1);

For i = 2, ..., p, we consider V_{i-1} the covariance matrix of (H_1, H_2, ..., H_{i-1}) in which the index i-1 represents the first i-1 rows and columns of V. Let be a vector s_i of dimension i-1 given by $s_i = (\mathrm{cov}(H_1, H_i),\ldots, \mathrm{cov}(H_{i-1}, H_i))^t$, we can observe that s_i is precisely the first i-1 entry of the ith column of matrix V. So, H_i is generated from a Bernoulli (μ_i^*), where μ_i^* is a conditional average of Hi given the values of $(H_{1,\ldots,}H_{i-1})^t$. Given that $(H_{1,\ldots,}H_{i-1})^t = (h_1,\ldots,h_{i-1})^t$, the conditional mean of μ_i^* is given by:

$$\mu_i^* = P(\mathbf{H_i} = 1|h_1,\ldots,h_{i-1})$$
$$= \mu_i + \mathbf{V_{i-1}^{-1}s_i}[(h_1,\ldots,h_{i-1})^t - (\mu_1,\ldots,\mu_{i-1})^t]$$

Each individual's genotype was obtained from two independent haplotypes, as shown in Table 1. Thus, we get the X matrix of SNPs markers. For this study, we considered N = 100 individuals, corresponding to rows of matrix X.

Vector of phenotypic observations

The y vector of phenotypes is obtained from a multiple linear regression model. This model is described by:

$$y = X\beta + \varepsilon \quad \text{(Eq. 1)}$$

Being X the matrix of SNPs markers, β the regression coefficients and Ɛ the vector of independent random errors, with $\varepsilon_i \sim N(0, \sigma^2)$.

To find the vector y, we must use the X matrix of markers and get the vectors β_{500x1} and ε_{100x1}. Considering, $\sigma^2 = 1$, the vector Ɛ was generated from a normal distribution (0,1).

This study consisted of two scenarios that differ from each other according to the vector β used. In Scenario 1, the vector β was generated from a normal distribution (0,1). In this case, the values greater than the module of 1.6 were defined as significant, responsible for the variation of the phenotypic characteristic of interest. In Scenario 2, β was also generated from a normal distribution (0,1), and we chose some specific values ranging from 0.4 to 2.1 to be associated with significant SNPs. Once the vectors β, Ɛ and the X matrix of SNPs markers were obtained, the vector y of phenotypic observations was calculated, according to Equation (1). In both scenarios, we set values of the vector β not considered significant as 0.

The process described above was repeated 1000 times, thus producing 1000 datasets. At each dataset simulation, the matrix X and the vector Ɛ varied, making different vector y. The β vector was generated only once, and then its values were fixed, aiming to verify the frequency with which the methods select the SNPs associated with these fixed values.

In each simulated set, the samples/individuals of X and y (i.e., their rows) were separated into two subsets (training and testing). Twenty percent of the samples formed the test set, and the remaining (80% of the samples) the training set, according to the Kennard-Stone algorithm (Kennard and Stone, 1969).

Real SNP dataset

For this study, we used a dataset of corn production in irrigated conditions presented by Crossa et al. (2010). Thus, we considered the grain yield of 264 individuals as a quantitative characteristic under 1135 SNP markers.

Real NIR dataset

We use the fiber content (FIBER) and Lignin of Sugarcane obtained from an experiment carried out in the Sugarcane Genetic Breeding Program (PMGCA) of the Federal University of Viçosa (UFV), Vicosa, MG, Brazil. The spectra referring to FIBER were obtained in the middle third of the stalk. The data were arranged in an X matrix, with 168 rows and 3113 columns. In this study, the best pretreatments evaluated in previous

analyses were: smoothing (5-point window, polynomial degree = 2), multiplicative scatter correction (MSC), and mean centering.

A total of 256 leaf samples were used to predict the Lignin content in sugarcane. NIR spectra were obtained directly from the green leaf without a sample preparation procedure. The data referring to the spectra were arranged in an X matrix, with 256 rows and 1038 columns. The best pretreatments obtained from previous analyses were: baseline correction, second derivative, MSC, and mean centering. Pasquini (2003) presents detailed information on pretreatments and their importance in reducing noise intrinsic to spectra. Spectra graphs were obtained through the Matlab 7.9 software (Math Works, Natick, USA) at the Instrumentation and Chemometrics Laboratory (LINQ) of UFV.

Computational Resources

Sparse partial least squares regression (SPLS)

The SPLS model was performed in the R software via the spls() function of the SPLS package (Chung et al., 2019).

Ordered predictors selection associated with PLS Regression (PLS-OPS)

The computational scripts referring to the PLS-OPS method (Teófilo et al., 2009) were all specifically developed and implemented in the R software. For model fitting, we used the pls package (Mevik et al., 2019) in R.

Supervised BLasso

For fitting the model with the supervised BLasso method, we used the bglr() function of the BGLR package (Perez and De Los Campos, 2014) of software R. After a stage of convergence tests, it was decided to user 25000 iterations, of which 10000 were discarded (burn-in) to ensure the heating of the chain, and with the selection of one every three iterations (thin).

Since supervised BLasso causes many regression coefficients to be close to zero (Park and Casella, 2008), a selection criterion was initially created in which 80% of the less significant variables would be discarded and the remaining variables selected. For a better comparison between the methods proposed in Scenario 2 of the simulated data, in addition to the selection criterion described above, a second criterion was adopted for the supervised BLasso: we matched the number of variables that the supervised BLasso would select from the number of variables selected by the SPLS and PLS-OPS methods, respectively, according to the comparison of interest.

For each fitted selection model, we obtained the vector that contains the predicted values of the phenotypes ($\hat{y}$), and then we got the statistics that infer about prediction errors, as described below.

Data simulation

The simulated data matrices were obtained under two distinct scenarios (Figure 1):

Scenario 1

We generated the β vector from a $N(0,1)$ distribution; values greater than the module of 1.6 (corresponding to 52 elements) were chosen to be significant.

In this scenario, we implemented SPLS, PLS-OPS, and supervised BLasso (Criterion 1: 20% of the most significant variables were selected).

Scenario 2

We chose ten random values from β ~N(0,1) ranging from 0.4 to 2.1.

In this scenario, we implemented SPLS, PLS-OPS, and supervised BLasso (Criterion 1 and Criterion 2: we equaled the number of variables that the supervised BLasso would select to the number of variables selected by the SPLS and by the PLS-OPS methods, respectively).

The following methods were evaluated in the real dataset: SPLS, PLS-OPS, and supervised BLasso (under Criterion 1).

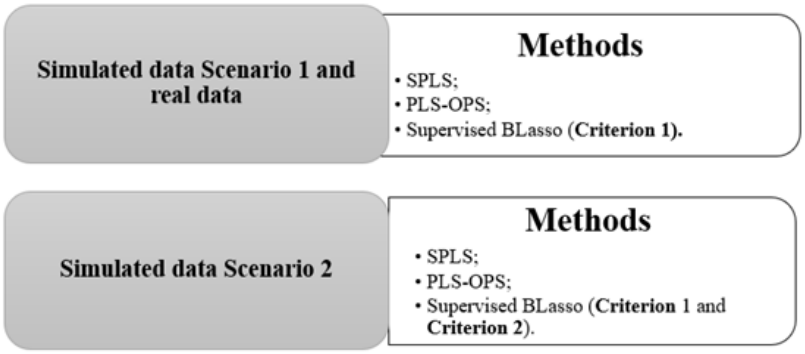

Figure 1. Schema of the methodologies implemented in the datasets.

Criteria for Comparison of Methodologies

The efficiency of the constructed models can be verified using: correlation coefficient (r) and the root mean squared error (RMSE), given by:

$$r = \frac{[\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})]}{\sqrt{\left[\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2\right][\sum_{i=1}^{n}(y_i - \bar{y})^2]}} \quad \text{(Eq. 2)}$$

and

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad \text{(Eq. 3)}$$

where:

y_i e $\hat{y}_i$ are the observed and the predicted values, respectively, $\bar{\hat{y}}$ is the mean predicted value and n is the number of samples belonging to each prediction subset.

Evaluation of the performance of the methods in simulated datasets.

To evaluate the performance of the built models, as well as their predictive ability, we used two criteria:

The frequency with which the proposed methods selected significant SNPs in the simulated datasets;

The mean correlation values and RMSE variation intervals in the prediction subset.

RESULTS AND DISCUSSION

Simulated datasets (Scenario 1)

In Scenario 1, the vector of regression coefficients representing the actual SNPs' effects consisted of 52 elements, each corresponding to a significant marker. The remaining 448 SNPs received their effects β_i equal to zero.

The mean values of the correlation coefficient (r) and the range of variation of the root mean squared error (RMSE) of prediction using BLasso and PLS methods, on the complete dataset (a) and by the selection methods (supervised BLasso, SPLS, and PLS-OPS) (b), appear in Table 3.

Table 3. Mean correlation coefficient (r) and root mean squared error (RMSE) variation interval between predicted value ($\hat{y}_p$) and actual (y) belonging to the prediction subset in the 1000 simulations, evaluated in scenario 1. (a) BLasso and PLS methods applied on the complete dataset. (b) supervised BLasso, SPLS, and PLS-OPS applied on the dataset after variable selection.

	(a) Complete dataset		(b) Dataset after selection		
	BLasso	**PLS**	**Supervised BLasso**	**SPLS**	**PLS-OPS**
r	0.688	0.702	0.703	0.690	0.696
RMSE	4.22 to 12.02	3.71 to 18.51	4.05 to 11.27	4.27 to 18.64	3.44 to 17.44

BLasso: Bayesian Lasso Regression; PLS: Partial least squares regression; SPLS: Sparse partial least squares regression; PLS-OPS: Ordered predictors selection associated with partial least squares regression.

The BLasso and PLS methods are similar in their predictive ability (0.688 and 0.702, respectively; and minimum RMSE values equal to 3.71 and 4.22) (Table 3-a). Similar r and RMSE values also occurred after selecting variables by the three methods evaluated (Table 3-b).

To better understand the statistics presented (RMSE and the r), the simulated set number 3 was randomly taken as an example. In this simulation, the RMSE and r values for supervised BLasso were 6.91 and 0.75, respectively. Referring to SPLS, we found RMSE = 6.20 and r = 0.75; and referring to PLS-OPS, RMSE = 6.56 and r = 0.75. According to Ferreira (2015), the model is considered adequate when the RMSE is much lower than the test set's standard deviation or when the ratio between the standard deviation of the test set and the RMSE value is a number around 10. The standard deviation of the original data in the test set was 9.03, and the ratio between the standard deviation and the RMSE was approximately 1.45 for the supervised BLasso, 1.42 for the SPLS, and 1.4 for the PLS-OPS. Thus, empirically, we infer that the prediction model is still not adequate for the chosen dataset, despite the relatively good correlation obtained.

In Figure 2, the x-axis shows the 500 SNPs evaluated. In the y-axis, there are the frequencies with which the models via supervised BLasso (Figure 2-a), SPLS (Figure 2-b), and PLS-OPS (Figure 2-c) selected each of the SNPs. In other words, Figure 2 shows how many times each SNP was chosen among the 1000 models built, according to the selection model.

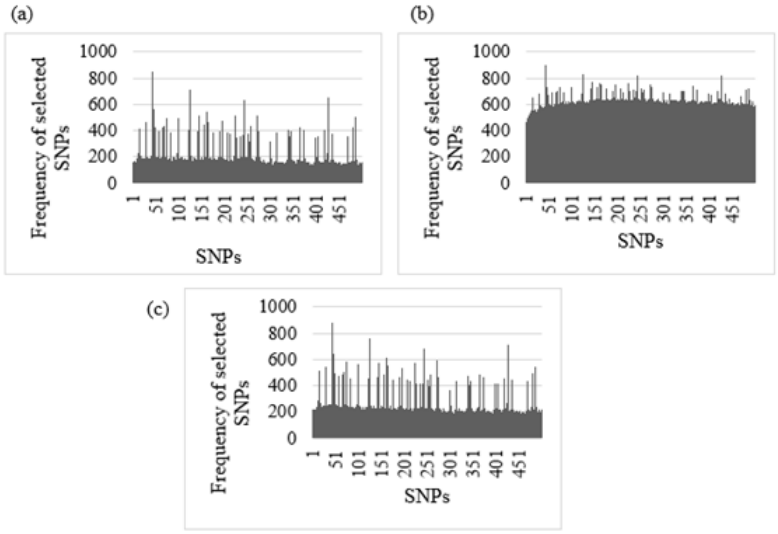

Figure 2. Selection frequency for the 500 SNPs in the 1000 datasets simulated. (a) supervised BLasso, (b) SPLS, and (c) PLS-OPS as selection methods evaluated in Scenario 1.

In this work, we simulated the correlation matrix from a uniform distribution that varied according to the distance between SNPs (Table 2). According to Feng et al. (2012), closer SNPs are more correlated to each other. Because of the simulation process, some of the more distant SNPs may have stronger associations than the closest ones, justifying the fact that in addition to the initially significant variables (SNPs), the methods also selected other variables, especially those more comparable to those that were considered significant in the simulation process. By following the frequencies with which the models are selecting the actual SNPs (Figure 2), we can observe that the SPLS selects the non-significant SNPs more often than the supervised Blasso or the PLS-OPS.

In general, the SNPs taken as responsible for the phenotypic variation were the most selected, especially the SNP43, SNP124, SNP242, and SNP425, which were the most frequent in all methods used (Table 4). We can highlight that these most selected SNPs had the most significant effects.

Neither the supervised BLasso nor the PLS-OPS methods selected the model containing all significant SNPs evaluated in this scenario. However, in 5 times, 31 of the 52 actual SNPs were selected by supervised BLasso, and in 7 times, 38 real SNPs were selected by PLS-OPS. The SPLS, on the other hand, chose the exact model 168 times.

On average, the models constructed by the SPLS method selected about 310 variables for each simulation, while by PLS-OPS, they chose on average 124 variables. The supervised BLasso according to the selection criteria adopted (20% of the most significant variables selected), for each simulation the models selected 100 variables.

Table 4. Frequency with which the models built using the three variable selection methods (Supervised BLasso, SPLS and PLS-OPS) selected some SNPs of higher effects evaluated in the first scenario in a total of 1000 simulations.

	Frequency		
	Supervised BLasso	**SPLS**	**PLS-OPS**
SNP 43	847	896	876
SNP 124	713	826	766
SNP 242	636	817	685
SNP 425	648	820	713

In general, after using the methods to select the "best" variables, we observed that the RMSE and r values remained very close compared to the complete model (Complete data: BLasso: r = 0.688 and RMSE: 4.22 to 12.02; and PLS: r = 0.702 and RMSE: 3.71 to 18.51. Data with selection: supervised BLasso: r = 0.703 and RMSE: 4.05 to 11.27; SPLS: r = 0.690 and RMSE: 4.27 to 18.64; and PLS-OPS: r = 0.696 and RMSE: 3.44 to 17.44). The great advantage of these variable selection models would be identifying more influential regions in the variable under study and working with a small number of variables, making the models more parsimonious.

Simulated datasets (Scenario 2)

In Scenario 2, ten SNPs (SNP2, SNP5, SNP10, SNP34, SNP49, SNP 73, SNP76, SNP139, SNP153, SNP199) were randomly chosen from the whole set of SNPs to contribute to the phenotypic variation, with regression coefficients given respectively by β = (0.4; 0.6; 0.7; 0.9; 1.2; 1.5; 1.6; 1.9; 2; 2.1). The remaining 490 SNPs received their effects β_i (i= 1 to 500) equal to zero (Ferreira, 2018).

Scenario 2 differs from Scenario 1 because in Scenario 1 the regression coefficients corresponded to values above the module of 1.6, while in Scenario 2 they corresponded to specific values ranging from 0.4 to 2.1. We can notice that some lower effects values of regression coefficients were chosen to verify the ability of the different models to detect those corresponding SNPs.

Table 5 shows the variation intervals for r and RMSE in the test sets using the PLS method and the BLasso, on all variables of the data matrix, in the 1000 simulations (Table 5a) and the results related to the use of the supervised BLasso selection methods (selection of 20% of the variables), SPLS and PLS-OPS (Table 5b), evaluated in Scenario 2.

Table 5. Mean correlation coefficient (r) and RMSE variation interval between predicted value ($\hat{y}_p$) and actual (y) belonging to the test subset in 1000 simulations, evaluated in Scenario 2, by BLasso and PLS methods, on the complete dataset (a) and using the selection methods: supervised BLasso (selection of 20% of variables), SPLS and PLS-OPS (b).

	(a) Complete data		**(b) Data with selection**		
	BLasso	**PLS**	**Supervised BLasso**	**SPLS**	**PLS-OPS**
r	0.665	0.632	0.753	0.846	0.705
RMSE	1.24 to 3.88	1.30 to 5.90	1.13 to 3.35	0.92 to 4.66	0.86 to 5.32

In terms of predictive ability, we can see that SPLS outperformed the other methods. The standard deviations in the test population of the data ranged from 1.78 to 5.71. The ratios between the standard deviations and the RMSE values for each method are not yet adequate for a satisfactory prediction model, according to the empirical criterion suggested by Ferreira (2015).

In this scenario (Scenario 2), two criteria were used to choose the number of variables selected by the supervised BLasso. In Criterion 1 we selected 20% of the whole set of variables, and in Criterion 2 the number of variables chosen was equal to the number of variables selected by the SPLS and PLS-OPS methods, respectively. Figure 3 shows the graphs of the frequency of SNPs selected from each method. The supervised BLasso was evaluated in the first criterion.

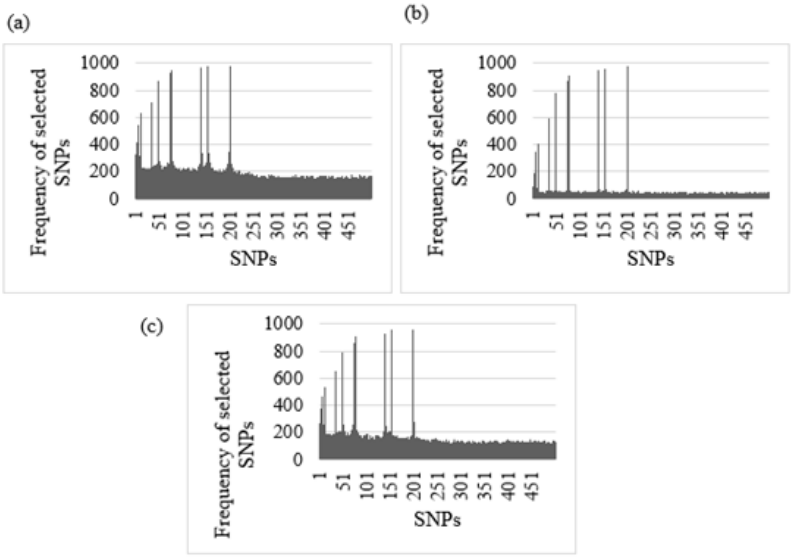

Figure 3. Selection frequency for the 500 SNPs in the 1000 simulated datasets obtained from the selection methods evaluated in the second scenario: (a) supervised BLasso (selection of 20% of variables), (b) SPLS, and (c) PLS-OPS.

From Figure 3, we can notice that when the values of the coefficients are lower, the SPLS selects fewer irrelevant variables compared to the other methods. In a simulation study conducted by Feng et al. (2012), aiming to compare the SPLS and selection operator (LASSO) methods for the selection of SNPs of lower effects, the results obtained were similar to those found in this study, with the SPLS selecting fewer irrelevant variables (Figure 3b). Figure 4 refers to the results from the supervised BLasso evaluated in Criterion 2, in which we fixed the number of variables that the supervised BLasso would select to the number of variables selected by the SPLS and PLS-OPS methods.

When the supervised BLasso uses the number of variables selected by the SPLS and PLS-OPS methods, it selects fewer irrelevant variables when evaluated under Criterion 1 (20% of variables), especially when using the SPLS. The value of the correlation coefficient of the supervised BLasso turned to 0.80, with the RMSE ranging from 0.74 to 3.25, when we adopted the number of variables of the SPLS, selecting the correct model (models that

selected exactly the regression coefficients that we took as significant) 31 times. Analyzing the supervised BLasso with the PLS-OPS criterion, we obtained r = 0.77 and RMSE ranging from 0.79 to 3.76 and selected the correct model 68 times.

Comparing the complete models with the selection models, we can notice that in all methods the mean correlation coefficient was increased for both the complete data (BLasso: r = 0.665; and PLS: r = 0.632), as for those with selection (supervised BLasso: r = 0.753; SPLS: r = 0.846; and PLS-OPS: r = 0.705). In addition, the RMSE variation interval has decreased, indicating that making selection when the effects of SNPs are of smaller magnitudes is a good alternative (Ferreira, 2018).

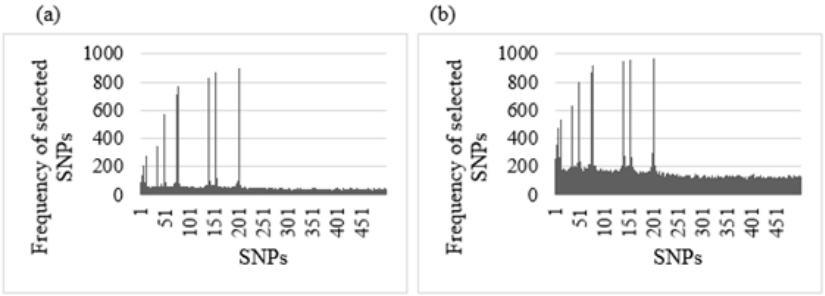

Figure 4. Selection frequency for the 500 SNPs in the 1000 simulated datasets obtained from supervised BLasso as a selection method, adopting the same number of variables selected in each simulation by methods (a) SPLS and (b) PLS-OPS.

Actual SNPs dataset

Table 6-a shows the r and RMSE values in the test sets using the BLasso and PLS methods on all variables (1135 columns) of the data matrix. In Table 6-b, we evaluated the same statistics after using the proposed selection methods.

Table 6. Correlation coefficient (r) and RMSE between the predicted ($\hat{y}_p$) and the actual (y) values belonging to the test subset in the actual SNPs dataset (data of corn production in irrigated conditions). (a) The BLasso and PLS methods were used on all variables (1135 variables). (b) The selection methods: supervised BLasso, SPLS and PLS-OPS selected 227, 1011 and 26 SNPs respectively.

	(a) Complete data		(b) Data with selection		
	BLasso	**PLS**	**Supervised BLasso**	**SPLS**	**PLS-OPS**
r	0.525	0.491	0.56	0.445	0.552
RMSE	0.73	2.63	0.56	2.76	2.96

Analyzing the correlation coefficient (r), the results obtained when using the supervised BLasso and the PLS-OPS were higher than those of the SPLS. Crossa et al. (2010) originally analyzed this same dataset in which several statistical methods were applied, among them the BLasso method, which presented better results (r = 0.525 and RMSE = 0.73) considering the complete model identical to that shown in our work.

We can note that when we applied the supervised BLasso evaluated under Criterion 1 (20% of the variables) and the PLS-OPS method, we obtained correlation values equal to

0.56 and 0.552, respectively (Table 6). The correlation value (r = 0.525) obtained by Crossa et al. (2010) was surpassed by the methods used in this study, affirming the use of supervised BLasso and PLS-OPS methods in SNPs data.

For this dataset, the standard deviation in the test population was 0.89, while the RMSE values for the supervised BLasso, SPLS, and PLS-OPS were, respectively, 0.56, 2.76, and 2.96. Therefore, the RMSE value = 0.56, with the ratio between the standard deviation and the RMSE approximately equal to 1.6 for the supervised BLasso, is the best prediction model, indicating that the best method for modeling this dataset, considering the selection of variables, is the supervised BLasso.

Real NIR dataset

As described in detail in the Material and Methods, we evaluated two sets of NIR data: DATA 1: Sugarcane fiber content, DATA 2: Lignin content of sugarcane. The NIR spectra of DATA 1 and 2, in the range of 4000 to 10000 cm^{-1}, are shown in Figure 5.

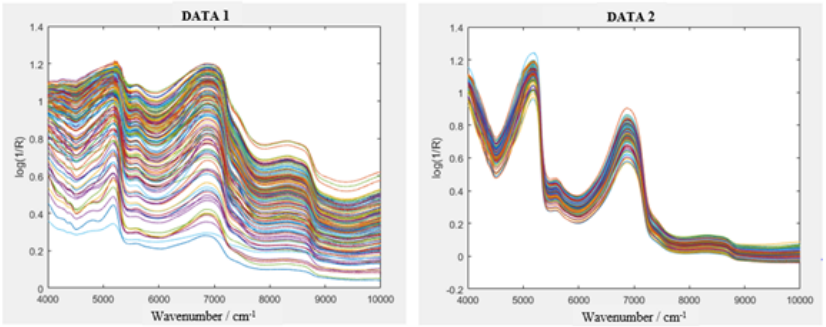

Figure 5. NIR SPECTRA: Sugarcane fiber content (DATA 1); Lignin content of sugarcane (DATA 2).

Table 7-a shows the values of r and RMSE in the test sets using the BLasso and PLS methods on all variables of the data matrix 1 and 2. Table 7-b presents the same statistics using supervised BLasso (20% of the selected variables), SPLS, and PLS-OPS, respectively, on the datasets studied.

Table 7. Correlation coefficient (r) and RMSE between the predicted ($\hat{y}_p$) and actual (y) values belonging to the test subset of the two sets of real NIR spectroscopy data (Sugarcane fiber content (DATA 1); Lignin content of sugarcane (DATA 2). The BLasso and PLS methods were used on all variables (DATA 1: 3113 variables and DATA 2: 1038 variables) (a). The selection methods: supervised BLasso, SPLS, and PLS-OPS selected 623, 2275, and 42 variables respectively in DATA 1 and 208, 1035, and 117 in DATA 2 (b).

		(a) Complete data		(b) Data with selection		
Data		**BLasso**	**PLS**	**Supervised BLasso**	**SPLS**	**PLS-OPS**
	r	0.69	0.68	0.69	0.679	0.676
1	RMSE	1.74	1.75	2.17	2.38	2.83
	r	0.96	0.93	0.956	0.83	0.946
2	RMSE	0.67	0.89	0.66	2.35	0.77

In the DATA 1 (Table 7-b) set, the selection methods work similarly, with r and RMSE values very close; in terms of predictive capacity, the methods are similar. The standard deviation value of the data in the test set was 2.35. The supervised BLasso presented RMSE = 2.17, lower than that obtained in the other proposed methods. However, the ratio of 2.35/2.17 does not meet the criteria defined by Ferreira (2015) to classify the model as adequate. The model with the selection of variables did not have an improved predictive ability than the complete model (Table 7).

In dataset 2, the best selection methods were the supervised BLasso and PLS-OPS. In terms of predictive ability, these models were similar to those obtained with the complete data. However, they have the advantage of having fewer variables.

CONCLUSIONS

The methods supervised BLasso and PLS-OPS provided similar prediction ability. The supervised BLasso ensured more parsimonious models, selecting fewer variables than PLS-OPS. When the effects of the variables were of lower magnitudes, the SPLS outperformed the other methods, assigning few irrelevant variables. The final models became simpler by using the selection methods than the respective models with the complete data since the number of variables decreased significantly in all datasets studied, without significant loss of prediction power.

ACKNOWLEDGMENTS

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) - Finance Code 001. We also thank the Foundation for Research of the State of Minas Gerais (FAPEMIG) for the financial support of research projects and the National Council for Scientific and Technological Development (CNPq) for the research scholarships. Finally, we thank RIDESA, the Inter-University Network for the Development of the Sugarcane Industry in Brazil, for providing the dataset.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Abdel-Rahman EM, Mutanga O, Odindi J, Adam E, et al. (2014). A comparison of partial least squares (PLS) and sparse PLS regressions for predicting yield of Swiss chard grown under different irrigation water sources using hyperspectral data. *Comput. Electron. Agric.*106: 11-19. DOI: 10.1016/j.compag.2014.05.001.

Caliari IP, Barbosa MHP, Ferreira SO and Teófilo RF (2017). Estimation of cellulose crystallinity of sugarcane biomass using near infrared spectroscopy and multivariate analysis methods. *Carbohydr. Polym.* 158: 20-28. DOI: 10.1016/j.carbpol.2016.12.005.

Cai W, Li Y and Shao X (2008). A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemometr. Intell Lab. Syst.* 90(2): 188-194. DOI: 10.1016/j.chemolab.2007.10.001.

Castro CADO, Nunes ACP, Roque JV, Teófilo RF, et al. (2019). Optimization of Eucalyptus benthamii progeny test based on Near-Infrared Spectroscopy approach and volumetric production. *Ind. Crops Prod.* 141: 111786. DOI: 10.1016/j.indcrop.2019.111786.

Chun H and Keles S (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat Soc. Series B. Stat. Methodol.* 72(1): 3-25. DOI: 10.1111/j.1467-9868.2009.00723.x.

Chung D, Chun H and Keles S (2019). Spls: Sparse Partial Least Squares (SPLS) Regression and Classification. R package version 2.2-3. Available at: https://CRAN.R-project.org/package=spls. Accessed November 3, 2018. DOI: 10.1111/j.1467-9868.2009.00723.x.

Crossa J, De Los Campos G, Pérez P, Gianola D, et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*. 186(2): 713-724. DOI: 10.1534/genetics.110.118521.

Colombani C, Croiseau P, Fritz S, Guillaume F, et al. (2012). A comparison of partial least squares (PLS) and sparse PLS regressions in genomic selection in French dairy cattle. *J. Dairy Sci.* 95(4): 2120-2131. DOI: 10.3168/jds.2011-4647.

Costa RC and De Lima KMG (2013). Prediction of parameters (soluble solid and pH) in intact plum using NIR spectroscopy and wavelength selection. *J. Braz. Chem. Soc.* 24(8): 1351-1356. DOI: 10.5935/0103-5053.20130172.

Cozzolino D, Kwiatkowski MJ, Parker M, Cynkar WU, et al. (2004). Prediction of phenolic compounds in red wine fermentations by visible and near infrared spectroscopy. *Anal. Chim. Acta.* 513(1): 73-80. DOI: 10.1016/j.aca.2003.08.066.

Beck T, Hastings RK, Gollapudi S, Free RC, et al. (2014). GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur. J. Hum. Genet.* 22(7): 949-952. DOI: 10.1038/ejhg.2013.274.

Da Silva EE, Da Silva LM, Wadt PG and Marchão RL (2017). Espectroscopia de infravermelho próximo na predição de propriedades químicas e físicas de solos de Roraima. *Biota Amazônia.* 7(2): 31-35. DOI: 10.18561/2179-5746/biotaamazonia.

Dardenne P, Sinnaeve G and Baeten V (2000). Multivariate calibration and chemometrics for near infrared spectroscopy: which method?. *J. Near Infrared Spectrosc*. 8 (4): 229-237. DOI: 10.1255/jnirs.283.

De Lima ABS, Batista AS, De Jesus JC, De Jesus Silva J, et al. (2020). Fast quantitative detection of black pepper and cumin adulterations by near-infrared spectroscopy and multivariate modeling. *Food Control*. 107: 106802. DOI: 10.1016/j.foodcont.2019.106802.

De Los Campos G, Naya H, Gianola D, Crossa J, et al. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*. 182(1): 375-385. DOI: 10.1534/genetics.109.101501.

Ferreira MMC (2015). Quimiometria: Conceitos, Métodos e Aplicações. Editora Unicamp, Campinas, Brazil. 1th edition. P.493.ISBN: 9788526810631

Ferreira MM, Antunes AM, Melgo MS and Volpe PL (1999). Chemometry I: multivariate calibration, a tutorial. *Química Nova*. 22(5): 724-731. DOI: 10.1590/S0100-40421999000500016.

Ferreira RA (2018). Comparação de métodos de seleção de variáveis em regressão aplicados a dados genômicos e de espectroscopia NIR. Tese de mestrado. A Universidade Federal de Viçosa, Viçosa. Available at: [https://locus.ufv.br//handle/123456789/20073].

Feng ZZ, Yang X, Subedi S and McNicholas PD (2012). The LASSO and sparse least squares regression methods for SNP selection in predicting quantitative traits. *IEEE/ACM Trans Comput Biol Bioinform*. 9(2): 629-636. DOI: 10.1109/TCBB.2011.139.

Guimarães CC, Assis C, Simeone MLF and Sena MM (2016). Use of near-infrared spectroscopy, partial least-squares, and ordered predictors selection to predict four quality parameters of sweet sorghum juice used to produce bioethanol. *Energ. Fuel.* 30(5): 4137-4144. DOI: 10.1021/acs.energyfuels.6b00408.

Kennard RW and Stone LA (1969). Computer Aided Design of Experiments. *Technometrics*. 11:137-148.

Mevik BH, Wehrens R and Liland KH (2019). Pls: Partial Least Squares and Principal Component Regression. R package version 2.7-1. Available at: https://CRAN.R-project.org/package=pls. Accessed November 3, 2020.

Morgano MA, Faria CG, Ferrão MF, Bragagnolo N, et al. (2008). Determinação de umidade em café cru usando espectroscopia NIR e regressão multivariada. *Ciênc. Tecnol. Aliment*. 28: 12-17. DOI: 10.1590/S0101-20612008000100003.

Oliveira UF, Costa AM, Roque JV, Cardoso W, et al. (2021). Predicting oil content in ripe Macaw fruits (*Acrocomia aculeata*) from unripe ones by near infrared spectroscopy and PLS regression. *Food Chem*. 351: 129314. DOI: 10.1016/j.foodchem.2021.129314.

Park T and Casella G (2008). The Bayesian Lasso. *J. Am. Stat. Assoc.* 103(482): 681-686. DOI: 10.1198/016214508000000337.

Pasquini C (2003). Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. *J. Braz. Chem. Soc.* 14: 198-219. DOI: 10.1590/S0103-50532003000200006.

Perez P and De Los Campos G (2014). Genome-Wide Regression and Prediction with the BGLR Statistical Package. *Genetics*. 198 (2): 483-495. DOI: 10.1534/genetics.114.164442.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/. Accessed November 7, 2020.

Roque JV, Cardoso W, Peternelli LA and Teófilo RF (2019). Comprehensive new approaches for variable selection using ordered predictors selection. *Anal. Chim. Acta.* 1075: 57-70. DOI: 10.1016/j.aca.2019.05.039.

Spahn G, Plettenberg H, Nagel H, Kahl E, et al. (2008). Evaluation of cartilage defects with near-infrared spectroscopy (NIR): an ex vivo study. *Med. Eng. Phys*. 3: 285-292. DOI: 10.1016/j.medengphy.2007.04.009.

Teófilo RF, Martins JPA and Ferreira MMC (2009). Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. *J. Chemom.* 23: 32-48. DOI: 10.1002/cem.1192.

Valderrama P, Braga JWB and Poppi RJP (2007). Validation of Multivariate Calibration Models in the Determination of Sugar Cane Quality Parameters by Near Infrared Spectroscopy. *J. Braz. Chem. Soc*. 18: 259-266. DOI: 10.1590/S0103-50532007000200003.

Yu HD, Yun YH, Zhang W, Chen H, et al. (2020). Three-step hybrid strategy towards efficiently selecting variables in multivariate calibration of near-infrared spectra. *Spectrochim. Acta A. Mol. Biomol. Spectrosc*. 224: 117376. DOI: 10.1016/j.saa.2019.117376.

