GMR
The Original by FUNPEC-RP

A comparison of regression methods based on dimensional reduction for genomic prediction

J.A. da Costa[1], C.F. Azevedo[1], M. Nascimento[1], F.F. e Silva[2], M.D.V. de Resende[3] and A.C.C. Nascimento[1]

[1] Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, MG, Brasil
[2] Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, MG, Brasil
[3] Embrapa Café, Viçosa, MG, Brasil

Corresponding author: J.A. da Costa
E-mail: jaquicele.costa@ufv.br

Genet. Mol. Res. 20 (2): gmr18877
Received April 13, 2021
Accepted May 21, 2021
Published May 31, 2021
DOI http://dx.doi.org/10.4238/gmr18877

ABSTRACT. The quality of fit of a multiple linear regression model often encounters multicollinearity and high dimensionality problems, making it impossible to obtain stable estimates through the traditional method of estimation based on ordinary least squares. To overcome such challenges, dimensionality reduction methods have been proposed, because of their simple theory and easy application. We compared three dimensionality reduction methods: Principal Components Regression (PCR), Partial Least Squares (PLS), and Independent Components Regression (ICR). An important step for dimensionality reduction and prediction is selecting the number of components, as it affects the linear combinations of the explanatory variables. The linear combinations are inserted into the model to predict the response based on a reduced number of parameters. We examined the criteria for the selection of the number of components. The dimensionality reduction methods were applied to genomic and phenotype data. We evaluated 370 accessions of Asian rice, *Oryza sativa*, which were genotyped for 36,901 SNPs markers considered to predict the genomic values for the number of panicles per plant trait.

This data set presented multicollinearity and high dimensionality. The computational time for each method was also recorded. Among the methods, PCR and ICR gave the highest accuracy values, with ICR standing out for presenting estimates of the least biased genomic values. However, ICR required more computational time than the other methodologies.

Key words: Principal components; Partial least squares; Independent components; High dimensionality; Multicollinearity

INTRODUCTION

The linear relationship between a response variable Y and the numerous explanatory variables X ($X_1, X_2, , \ldots X_m$) is established using a multiple linear regression model. The main method for estimating the parameters of a statistical regression model is based on Ordinary Least Squares (OLS). The necessary assumptions for the estimation are linearity between Y and X variables and the values of these variables must be fixed and orthogonal with the errors, which must have zero mean, be homoscedastic, and independent. If these assumptions are verified, the OLS method will lead to BLUE (Best Linear Unbiased Estimator) type estimators: linear, non-biased, and minimal variance estimators (Puntanen and Styan, 1989)

However, frequently, the explanatory variables have some degree of linear association, called multicollinearity, which can be caused by several factors, including inadequate data collection, restrictions in the model or the sample population and super-parameterized models, as observed in real situations when there is high dimensionality (Montgomery et al., 2012). This effect increases considerably the variance associated with the parameter estimates, as the correlation coefficient of the explanatory variables approaches a high value, either positive or negative. Thus, the components of variance become high, the estimators become biased, and the statistical regression model is no longer appropriate for the prediction since it leads to unstable estimates (Gunst and Webster, 1975).

High dimensionality is detected when the number of observations (n) is less than the number of parameters (m) to be estimated in the statistical model. The existence of $m > n$ does not preclude the use of the method based on ordinary least squares, but the generalized inverse should be used instead of the classic one, which will lead to infinite possible estimates for the coefficients. Another alternative to use the ordinary least squares in these situations would be to estimate each effect in isolation and perform hypothesis tests to verify the statistical significance of these effects. However, such a practice is inefficient, as it causes an overestimation of the parameters' effects and consequently obtains a low predictive accuracy value (Resende et al., 2012).

The dimensionality problem can also be solved through the use of variable selection methods, including *Stepwise*, *Forward* and *Backward,* which use statistical tests to remove or maintain explanatory variables in the model. However, such methodologies are inadequate in some circumstances. For example, when the explanatory variables are genetic information that controls some trait, biologically, the response variable still depends on many variables and it does not make sense to remove them (James et al., 2013). Likewise,

in some situations, the objective is to estimate the parameters' effects, but also concomitantly, identify the variables that explain the response variable, such as the theoretical economic and genomic association models (Camarero et al., 2015; He and Lin, 2011). In these procedures, the order in which explanatory variables are included or removed from the model throughout the procedure's application is not associated with the degree of explanation of the variable concerning the response variable. The methods of selecting variables are also inappropriate since importance ranking of the variables is not related to the order in which they were included or removed during the method's application. Furthermore, in high dimensionality situations, for computational reasons, the selection of the best subset of variables may face statistical problems such as *overfitting* (over-training) and high variance of the coefficient estimates (Liu and Gillies, 2016). This explained by the fact that the larger the research space, the greater the probability of finding models that fit well to the training data, even if they do not have the power to predict future data.

Given the OLS method's statistical problems, it is necessary to search for methodologies that overcome these challenges and guarantee an efficient prediction in the face of more interpretable models. Several methods have been proposed for this purpose, mainly with emphasis on regularization methods such as Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996), Bayesian methods, such as the Bayesian version of LASSO denoted by BLASSO (Bayesian LASSO) (Kimeldorf and Wahba, 1970; Tibshirani ,1996) and dimensionality reduction methods, notedly on Principal Components Regression (PCR) (Kendall, 1957; Hotelling, 1957), Partial Least Squares (PLS) (Wold, 1975) and the Independent Components Regression (ICR) (Jutten and Hérault, 1991; Comon, 1994). Among these, the dimensionality reduction methods stand out for their great applicability and relatively simple theory when compared to the other methods applied in the estimation of parameters in the presence of multicollinearity and high dimensionality.

The PCR and ICR regression methods are based, respectively, on the *Principal Component Analysis* (PCA) and *Independent Component Analysis* (ICA). These methodologies consist of building linear combinations of the explanatory variables to reduce the dimensionality of the problem studied. These latent variables are called orthogonal (uncorrelated) components in the PCA and independent in the ICA. Thus, in PCR and ICR, after building the components, in a number less than the number of observations and with the absence of multicollinearity, it is possible to perform a multiple linear regression between the variable Y and these components. In the PCA, the components are built to maximize their variance, while in the ICA, they are built to maximize the independence of the components. Unlike these methods that consider only the variables X in the construction of the components, the PLS method was developed as a regression methodology. Thus, it considers the explanatory variables X and the response variable Y to maximize the covariance between Y and the components.

Furthermore, the applications in biometrics, environmental and agricultural data of regression methods based on dimensionality reduction are observed in the genomic selection (Azevedo et al., 2013a; Azevedo et al., 2013b; Azevedo et al., 2014; Azevedo et al., 2015; Silveira et al., 2017), gene expression (Nascimento et al., 2017), spectroscopy analysis NIR (*Near Infrared*) (Morgano et al., 2005; Teófilo et al., 2009), sensory data (Westad, 2005), climatic data (Lim et al., 2015) and process control data (Han et al., 2003), among others. However, dimensionality reduction methods can be applied whenever

multicollinearity and high dimensionality are present, and the the study is aimed at regression.

Theoretical aspects related to obtaining estimators via the method of ordinary least squares and dimensionality reduction methods will be presented below, with focus on the main problems faced in the estimation process via OLS and the alternatives to determine the number of components to be inserted into the model.

MATERIAL AND METHODS

Estimation using the ordinary least squares method

Consider the linear model:

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

where y is the vector of observations of the response variable with dimension $n \times 1$, where is n the number of observations; μ is the overall average; X is a matrix whose columns contain the explanatory variables with dimension $n \times m$, where m is the number of parameters; e is the vector of random errors: $e \sim N(0, I\sigma_e^2)$ where I is the identity matrix $(n \times n)$ and σ_e^2 is the residual variance.

The estimates of the coefficients via OLS present several advantages, including easy understanding in the geometric and mathematical approach since the method is based on Euclidean distance. Also, it is not necessary to assume distributions for the parameter estimators. Estimates are obtained by minimizing the sum of squares of errors, given by: $S(\beta) = (y - X\beta)'(y - X\beta)$, which results in $(X'X)\hat{\beta} = X'y$. Thus, it is essential to obtain the classic inverse of the matrix X′X, so that by multiplying the factor $(X'X)^{-1}$ on both sides of the equation, we obtain a single estimate for the parameters: $\hat{\beta} = (X'X)^{-1}X'y$. In cases where the matrix X presents perfect linear dependence between variables, the determinant of the matrix X′X becomes null, and the inverse of this matrix cannot be found. Consequently, it is impossible to obtain the estimates $\hat{\beta}$ through the OLS method using the classical inverse. When the linear dependency is not perfect, the estimators' variance components become high, and the parameter estimators are biased.

In the sets of observations formed by molecular markers, NIR spectroscopy data, among others, the matrix X presents not only multicollinearity, but high dimensionality. In other words, the number of observations (n) is lower that the number of explanatory variables (m). Thus, in the linear model, expression (1), where we have the number of explanatory variables equal to m and the average, there are, in total, $m + 1$ parameters (number of explanatory variables + the general average) to be estimated. Whether we have only n observations $(n < m + 1)$, we have a mathematical problem of solving a system, with n equations with $m + 1$ unknown, which hinders the obtainment of single estimates for the parameters. Thus, the high dimensionality prevents the obtaining of parameter estimates by the method of ordinary least squares. Other methodologies that address these challenges of statistical analysis should be used. Alternatively, the dimensionality reduction methods described below are proposed.

Dimensionality reduction methods

The regression methods by dimensionality reduction are carried out in three steps: the first consists of transforming the explanatory variables into latent variables, the so-called components; the second consists of performing a regression between the response variable Y and the constructed components, and no longer the explanatory variables; and the third is used to estimate the coefficients associated with the explanatory variables X. The theory of each method guarantees the orthogonality between the components. In addition, the number of components is always less than the number of observations, which makes it possible to adjust high-dimensional and multicollinearity-free regression models.

In the context, the most prominent methodologies are regression via principal components, partial least squares, and regression via independent components. These methods differ from each other in the way they build the components. In the PCR, the components are built to maximize the variance of the X variables, in PLS, it aims to maximize the covariance between Y, and the components and in the ICA, the components are built to maximize component independence. The three methodologies are detailed later.

Principal Components Regression

The PCR was introduced by Kendall (1957) and Hotelling (1957) and defined the j-th principal component z_j as: $z_j = p_{j1}x_1 + p_{j2}x_2 + \cdots + p_{jm}x_m = p_j^{'}X$, where $x_j{'}$s are the matrix columns X and $p_j^{'}$ is an unknown vector that establishes the j−th linear combination, so that $j = 1,2,\ldots,m$. The PCA's main objective is to find the components $Z_j{'}$s through the variables X $(X_1, X_2, \ldots, X_m)$ that maximize the principal components' variability. For such, the variance of Z_j and the covariance of Z_j and Z_k $(j \neq k)$ are given, respectively, by:

$$Var(Z_j) = Var\left(p_j^{'}X\right) = p_j^{'}Var(X)p_j = p_j^{'}\Sigma p_j \quad \text{and} \qquad \text{(Eq. 2)}$$

$$Cov(Z_j, Z_k) = Cov\left(p_j^{'}X, p_k^{'}X\right) = p_j^{'}\,Var(X)p_k = p_j^{'}\Sigma p_k \qquad \text{(Eq. 3)}$$

where $Var(X) = \Sigma$, that is, the explanatory variables' variance and covariance matrix. Besides, to maximize the variance of Z_j, it is observed that the expression (2) corresponds to a quadratic form and according to the theorem of maximization of quadratic forms, if Σ is a symmetric matrix $(m \times m)$, so the maximum $p_j^{'}\Sigma p_j$ under restriction $p_j^{'}p_j = 1$ is given by the largest of the eigenvalues λ_j of Σ, $j = 1, 2, \ldots, m$, and the corresponding eigenvector p_j which are solutions of the system of homogeneous equations, $\left(\Sigma - \lambda_j I\right)p_j = 0$.

Thus, we define the latent variables Z_j $(j = 1,2,\ldots,m)$ as linear combinations of the explanatory variables X $(X_1, X_2, \ldots, X_m)$:

$$Z = XP, \qquad \text{(Eq. 4)}$$

where $P = [p_1 \quad p_2 \quad \cdots \quad p_m]'$ $(m \times m)$ is the eigenvector matrix of the covariance matrix of X $(n \times m)$, and Z $(n \times m)$ is the matrix whose columns are the principal components $Z_j{'}s$. The construction of the matrix P requires the approach of concepts related to the variance and covariance matrix, eigenvectors, and eigenvalues portrayed by Marcoulides and Hershberger (1997).

In addition, the system of homogeneous equations, provides: $\left(\Sigma - \lambda_j I\right)p_j = 0$, which implies, $\Sigma p_j = \lambda_j p_j$. By replacing the previous expression in the expressions of

variance (2) and covariance (3) and using the information that the eigenvectors of a symmetric matrix are orthogonal (that is, $p_j^{'}p_k = 0$) and the constraint ($p_j^{'}p_j = 1$), it follows that: $Var(Z_j) = p_j^{'}\Sigma p_j = p_j^{'}p_j\lambda_j = \lambda_j$ and $Cov(Z_j, Z_k) = p_j^{'}\Sigma p_k = p_j^{'}p_k\lambda_j = 0$. Therefore, the correlation between the components Z_j and Z_k ($j \neq k$), is null, that is: $Cor(Z_j, Z_k) = \frac{Cov(Z_j, Z_k)}{\sqrt{Var(Z_j)Var(Z_k)}} = 0$.

The PCR methodology consists of eliminating components that do not contribute considerably to explaining total variance present in the data, which reduces the dimensionality of the studied problem. For this reason, $n_{PCR} \leq min(n, m) - 1$. In the context of high dimensionality, $n_{PCR} \leq n - 1$. The selection of the number of components to be used in the model should not result in the loss of relevant information related to the original data (X) (Otto, 1998).

The criterion commonly used to select the number of principal components is based on the variability present in the explanatory variables. When spectral decomposition (Rencher and Christensen, 2002) is apllied in the variance matrix Σ, it is obtained: $\Sigma = P\Lambda P^{'}$, where P is composed of the eigenvectors of Σ and Λ is the diagonal matrix of eigenvalues of Σ. Thus, it follows that the trace of the matrix Σ is given by: $tr(\Sigma) = tr(P\Lambda P^{'}) = tr(\Lambda) = \sum_{j=1}^{m}\lambda_j$. Besides $tr(\Sigma) = \sum_{j=1}^{m}\sigma_j^2$ which leads to the conclusion that $\sum_{j=1}^{m}\lambda_j = \sum_{j=1}^{m}\sigma_j^2$. Therefore, it is determined that the percentage of explanation of the j-th component is given by $\frac{\lambda_j}{\sum_{j=1}^{m}\sigma_j^2}$.

Thus, under the context of regression, the total variability present in the original variables is only achieved using the maximum number of components built ($n_{PCR} = min(n, m) - 1$). However, most of this variability can often be explained by a small number of major components ($n_{PCR} < min(n, m) - 1$), since the first principal components explain most of the total variability of the variables X. Acoording to Ferreira (2012), the selection of the number of principal components can be based on the determination of the desired fraction of the total variation, which generally ranges between 70% and 80%. Under some conditions, the goal is to predict the response variable Y. Thus, an alternative criterion to determine the number of latent variables in the PCR is based on the variability present in the response variable Y. The explanation percentage of Y by the principal components is obtained through the coefficient of determination, that is, $Cor(y, \hat{y})^2 \times 100\%$.

Partial Least Squares

PLS was designed by Wold (1975) and, similarly to Garthwaite (1994), it has similarities with the PCR method. However, the PCR considers only the explanatory variables X in the construction of the components. At the same time, PLS also considers the variable response Y. In order to estimate the variable Y, the components associated with X are denoted by t_r ($n \times 1$, where $r = 1, 2, \ldots, n_{PLS}$ being $n_{PLS} \leq min(n, m) - 1$), and the components associated with Y are denoted by z_r.

To determine the first components t_1 and z_1, the variables Y and $X_j's$ are centered on the mean and define the variables U_1 and V_{1j}, as: $u_1 = y - \bar{y}$ and $v_{1j}{}' = x_j - \bar{x_j}$, for $j = 1, \ldots, m$. Subsequently, the variable is defined as S_1, so that $s_1 = V_1{}' u_1$ ($m \times 1$)

where $V_1 = [v_{11} \quad v_{12} \quad \dots \quad v_{1m}]$ $(n \times m)$, and the decomposition into singular values applies (SVD) portrayed by Härdle and Hlávka (2007) in the vectorss_1; $s_1 = L_1 k_1 q_1^{'}$, where L_1 is a unitary matrix $(m \times m)$ with the first column vector equal to $\frac{s_1}{||s_1||}$ (vector s_1 normalized), k_1 is a vector $(m \times 1)$ with the first value equal to$||s_1||$ (vector norm s_1) and q_1 is a scalar equal to 1. The components T_1 and Z_1 are respectively defined by:

$$t_1 = V_1 L_1 \text{ and } z_1 = u_1 q_1 \frac{Var(t_1)}{Cov(t_1, u_1)}. \quad \text{(Eq. 5)}$$

However, not all the information in the variables X_j $(j = 1,2,\dots,m)$ and variable Y are contained in the component T_1, defined above. Therefore, the missing information in T_1 can be estimated through the residuals of the regression between the variables X_j and T_1 or, equivalently, the regression between latent variables V_{1j} and T_1, since the residues of both are identical (Garthwaite, 1994). Likewise, the Y's variability that is not explained by T_1 can be estimated through the regression residuals between U_1 and T_1.Therefore, the variables are defined as U_2 and $V_{2(j)}$, respectively, $\hat{v}_{2(j)} = v_{1(j)} - t_1 \hat{r}_1^{'}$ and $\hat{u}_2 = u_1 - t_1 \hat{p}_1^{'}$, that is, U_2 and V_{2j} are the residuals, r_1 and p_1 are the coefficients obtained from the regression between $U_1$and T_1 and V_{1j} and T_1, in this order. A new variable is defined as S_2, so that $s_2 = V_2^{'} u_2$ and applies again to SVD, similarly to s_1 to build the component T_2 .

The components $t_3,\dots,t_{n_{PLS}}$ $(1 \leq n_{PLS} \leq \min(n, m) - 1)$ are determined successively and similarly to the previous ones. Besides all components are surely orthogonal (Garthwaite, 1994). The correlation between $V_{2(j)}$ and T_1 is equal to 0 since they are, respectively, residuals and regressor. Thus, as each component $T_2, T_3, \dots, T_{n_{PLS}}$ is linear combination of $V_{2(j)}$ then, they are not correlated with T_1 either.

In order to determine the number of components (n_{PLS}) one can also use the criterion based on the data's variability. The variability present in the explanatory variables X explained by the components can be measured using $\frac{\sum_{i=1}^{n} r_{ij}^2 \sum_{i=1}^{n} t_{ij}^2}{\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}^2}$ and the variability present in the variable Y, that is explained by the components, is also given by the coefficient of determination.

Independent Components Regression

The ICR, proposed by Jutten and Hérault (1991) and Comon (1994), assumes that the data comes from a non-Gaussian distribution. Under the context of regression, the ICR decomposes the data matrix X as $X = SA'$, where the matrix X $(n \times m)$ is decomposed into a matrix of independent components S $(n \times min(n, m))$, and a matrix A $(m \times min(n, m))$ called a mixing matrix. If matrix A were known and square, we would easily obtain the independent components. However, since reduced dimensionality is generally desired, matrix A is not square and unknown. Without loss of generality, we can define the mixing matrix as the product of a whitening matrix K $(m \times min(n, m))$ and the matrix R $(min(n, m) \times min(n, m))$, which guarantees independence between the components.

The first process to estimate matrix A is the whitening process that makes the original variables $(X_1, X_2, \dots, X_m)$ uncorrelated and with unit variance, that is, the covariance matrix of the blanched data is the identity matrix. For whitening, orthogonal

decomposition is applied to the covariance matrix of X, denoted by Σ $(m \times m)$, which results in: $\Sigma = P\Lambda^{\left(-\frac{1}{2}\right)}P'$ where P is composed of eigenvectors in their columns and Λ is a diagonal eigenvalue matrix of the covariance matrix of X. Under the context of regression, the matrix K is then defined as $P_r\Lambda_r^{\left(-\frac{1}{2}\right)}$, where P_r is the matrix with the $min(m,n)$ first columns of the matrix P with dimension $m \times min(m,n)$ ($min(m,n)$ first eigenvectors) and Λ_r $(min(m,n) \times min(m,n))$ is a matrix with the $min(m,n)$ first rows and columns of the matrix Λ (eigenvalues associated with these first eigenvectors). Thus, the whitened data (A), will be obtained through XK $(n \times m)$.

Since the data come from a non-Gaussian distribution, the non-correlation between the variables does not imply statistical independence. Independence is only achieved through the process described below. Thus, the second process to guarantee independence between the S's columns is done based on the maximization of non-Gaussianity (normality). This process is based on the central limit theorem, which states that the sum of N independent and identically distributed random variables, for N large enough and satisfying certain general conditions, will have an approximate Gaussian distribution. Under the ICR context, since the variables X are a linear combination of the components, it can be concluded that the components present a more distant distribution than Gaussian. Therefore, we can obtain the independent components by maximizing the non-Gaussianity of the whitened data matrix.

The main non-Gaussian maximization algorithms are based on kurtosis and negentropy. Hyvärinen (1998) proposed the most used algorithm, which was denominated FastICA, and is based on negentropy. Negentropy is defined as:

$$J(R) = H\left(R_{gaussiana}\right) - H(R), \qquad \text{(Eq. 6)}$$

where $H(R) = -\int_R f_R(r)\, lnf_R(r)dr$ is the entropy of a random variable R with probability density function $f_R(.)$ and $H\left(R_{gaussiana}\right)$ is the entropy of a random variable R with a Gaussian distribution. Statistically, entropy is a measure of the average uncertainty associated with the observation of a random variable. Therefore, the greater the entropy, the more unpredictable the variable's observation. When the variable has a Gaussian distribution, the entropy and variance are coincident. Considering the variables R and $R_{gaussiana}$ with the same variance, $H\left(R_{gaussiana}\right)$ is the maximum entropy value found is that a Gaussian random variable has greater entropy than any other random variable of the same variance (Hyvarinen et al., 2001). Thus, negentropy can quantify the degree of non-Gaussianity of a random variable, and its measurement will always be a non-negative value.

According to Hyvärinen (1998), the maximization of negentropy leads to the estimation of independent components. However, the negentropy-based algorithm is hindered by calculation of entropy. Thus, approximations must be used in the expression (6), such as: $J(R) \propto \left\{E[G(R)] - E\left[G\left(R_{gaussiana}\right)\right]\right\}^2$, where G is a non-quadratic function, and the choice of function G influences the approach of negentropy (Hyvärinen, 1999), and the functions which are G most employed for this purpose are: $G_1(r) = \frac{1}{a}\log\cosh(ar)$ and $G_2(r) = -exp\left(-\frac{r^2}{2}\right)$, where a is a constant $(1 \leq a \leq 2)$ and the selected G function cannot grow very quickly.

After its convergence, it is possible to find a matrix R that makes the matrix columns XK independent and, consequently, the S columns, since the independent components can be obtained by:

$$S = XKR. \quad \text{(Eq. 7)}$$

Basically, in the analysis of independent components, it is impossible to determine the variance of independent components. However, it is feasible possible when the independent components are assumed to have unit variance and mean equal to 0 (Hyvärinen, 1999). Based on this assumption, the variability of explanatory variables X explained by the components can be measured using $\frac{n\sum_{j=1}^{m} a_{jr}^2}{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}^2}$, where a_{jk} is the element of the i-th row and j-th column of the mixing matrix ($j = 1,2,\dots,m$ e $r = 1,\dots,min(n,m)$), x_{ij} is the element of the i-th row and j-th column of the matrix centered on explanatory variables X ($j = 1,2,\dots,m$) and n is the number of observations (Bingham and Hyvärinen, 2000; Helwig and Hong, 2013). Unlike the principal components, each independent component explains a small part of the data's total variance. Besides, it is impossible to determine the order in which independent components are extracted. The coefficient of determination also gives the variability present in the Y variable explained by the components.

Predictions by PCR, PLS and ICR

Multiple linear regression is performed between the Y variable and the components Z, T and S, obtained by PCR, PLS, and ICR, respectively. Then we have the following predictions:

$$\hat{y} = Z\hat{\alpha}, \quad \text{(Eq. 8)}$$

$$\hat{y} = T\hat{\beta} \quad \text{(Eq. 9)}$$

$$\hat{y} = S\hat{\gamma}, \quad \text{(Eq. 10)}$$

where $\hat{\alpha}_m$ ($m = 1,\dots,n_{PCR}$), $\hat{\beta}_m$ ($m = 1,2,\dots,n_{PLS}$) and $\hat{\gamma}_m$ ($m = 0,1,\dots,n_{ICR}$ so that $1 \leq n_{PCR}, n_{PLS}, n_{ICR} \leq min(2J, I) - 1$) are the estimates of the coefficients obtained by the ordinary least squares method.

Since the coefficients $\hat{\alpha}_m$, $\hat{\beta}_m$ and $\hat{\gamma}_m$, previously obtained, are not associated with the original variables, that is, they do not have a practical interpretation, the estimates of the effects associated with variables X are given by:

$$\widehat{m}_{PCR} = P\hat{\alpha} \quad \text{(Eq. 11)}$$

$$\widehat{m}_{PLS} = L(R'L)^{-1}\hat{\beta} \quad \text{(Eq. 12)}$$

$$\widehat{m}_{ICR} = KR\,\hat{\gamma} \quad \text{(Eq. 13)}$$

where L is the matrix whose columns are $L_1, L_2, \dots, L_{n_{PLS}}$, called loading matrix X, and R is the matrix whose columns contain the coefficients $r_1, r_2, \dots, r_{n_{PLS}}$. The expression (11) is obtained by combining the expressions (4) and (8) and the expression (13) by combining expressions (7) and (10). The estimates of the original PLS coefficients are not trivially obtained since the columns of the matrix V ($V_{1(j)}, V_{2(j)}, \dots, V_{2J(j)}$) are not directly compared to X as observed in PCR and ICR, since they are successively deflated. According to Wold (1975), they can be obtained by expression (12).

Genomic data for the application of the methods

The three methodologies (PCR, PLS, and ICR) were applied to the public data set of Asian rice, *Oryza sativa*. This data set is public and is part of two projects, the OryzaSNP Project and the OMAP Project (Ammiraju et al., 2006; Zhao et al., 2011), and is available on the website https://ricediversity.org/data/. The database used in this study is composed of the number of panicles per plant trait (y), regarding 370 accessions of rice, which were genotyped for 36,901 SNPs markers (Single Nucleotide Polymorphism). Thus, the matrix X, with dimension 370 × 36,901, denotes the matrix of incidence of the markers, where x_{ij} is encoded by 0, 1, or 2; corresponds to the number of alleles of the j-th marker for the i-th access.

Comparison between methodologies

The methods PCR and PLS were applied for each number of components, and the effects of markers on the estimation population are estimated by PCR (or PLS, or ICR) and these are used in the validation population to estimate the genomic values of individuals in this population. Then, the predictive ability is calculated by the correlation between the estimated genomic value and the phenotypic value ($r_{y\hat{y}}$) . Thus, we selected the number of components that leads the genomic value to an increased predictive ability.

The data were evaluated under a validation process *k-fold* with $k = 5$, and the efficiency of the methods PCR, PLS, and ICR were evaluated according to the predictive ability $r_{y\hat{y}}$ and the regression coefficient, which is defined as one minus the regression coefficient between the phenotype and the estimated genomic value ($1 - b_{y\hat{y}}$). For regression coefficients below 1 ($b_{y\hat{y}} < 1$), the genomic values were overestimated. For the regression coefficients above1 ($b_{y\hat{y}} > 1$), the genomic values were underestimated. For the regression coefficients equal to 1 ($b_{y\hat{y}} = 1$), the genomic values are not biased.

All implementations of the methods used were performed in the R software system 2018 (R Development Core Team) under the GenomicLand interface. The analysis was carried out using the packages (and functions) *pls* (*pcr*), *caret* (*icr*) e *leaps* (*regsubsets*) for methodologies PLS, PCR, and ICR, according (Mevik et al., 2011; Kuhn, 2017; Lumley, 2017), respectively.

In Figures 1 and 2, it is observed that in the PCR and PLS methods, the first components explain the explanatory variable in greater proportions X, while in ICR (Figure 3), the variability present in explains in X is explained in smaller proportions, and there is no ranking of the order of the components. The fact that the first components explain much of the total variability in the case of PCR and ICR contributes to better visualization of the data in two- or three-dimensional graphs, when these are summarized in two or three components, respectively; facilitating the identification of some patterns present in the data (Yao et al., 2012).

Table 1 shows the results of the efficiency of the three methods. Since the exhaustive criterion aims to choose the number of components associated with a higher predictive ability, the PLS needed fewer components than the other methods, as reported by Du et al. (2018). It is observed that the methods presented similars predictive abilities values, as reported by Azevedo et al. (2015). This leads us to believe that the explanatory

variables have only a linear relationship. All three methods overestimated the genomic genetic values, and these values were less biased in the PCR, that is, closer to 1.

Table 1. Number of components (Nc), computational time (CT), predictive ability ($r_{y\hat{y}}$) and prediction bias ($\hat{b}_{y\hat{y}}$) considering Principal Components Regression (PCR), Partial Least Squares (PLS) and the Independent Components Regression (ICR) applied to the matrix of molecular markers to predict the genomic values ($\hat{y} = Z\hat{\alpha}$,) of rice accessions for the trait number of panicles per plant (y).

Methods	Nc	CT (hours)	$r_{y\hat{y}}$	$\hat{b}_{y\hat{y}}$
PCR	104	8.70	0.82	0.97
PLS	4	10.48	0.81	0.95
ICR	81	12.98	0.82	0.99

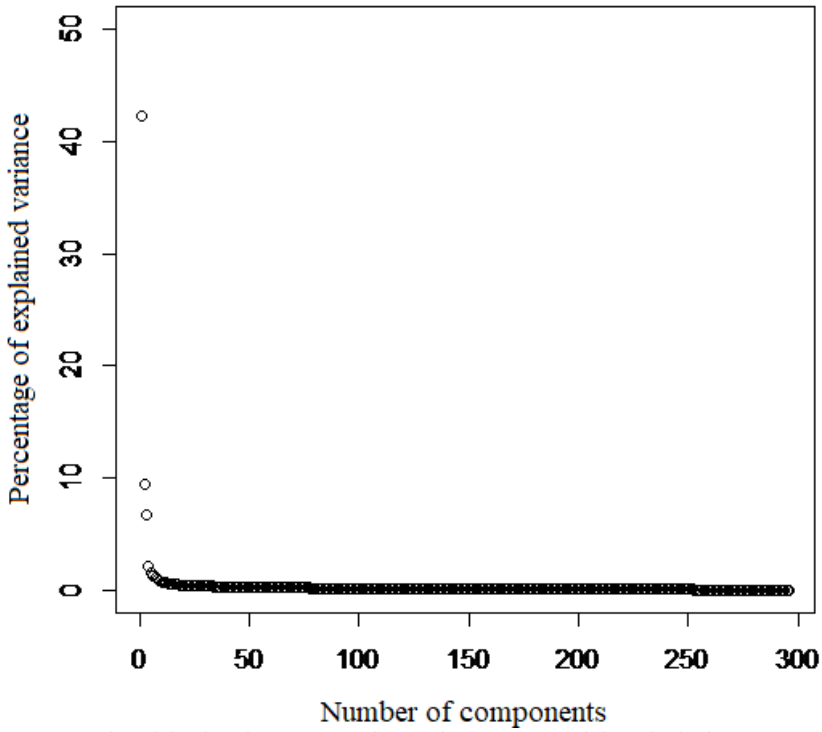

Figure 1: Percentage of explained variance concerning each component of the principal component regression (PCR) considering the matrix of incidence of the markers (the number of alleles of the marker), composed by 36,901 SNPs markers (single nucleotide polymorphisms), to predict the genetic value of 370 rice accessions, *Oryza sativa,* for the trait number of panicles per plant. This genomics data and is part of the OryzaSNP Project and the OMAP Project.

In the analyzes considered the 5-folds cross-validation procedure for each method, it is observed that the ICR time is longer concerning the PCR and PLS time (Table 1). One of the disadvantages of the ICR method is the requirement for a high demand for the computational time when selecting the number of components that leads to greater accuracy or predictive ability (Azevedo et al., 2014; Azevedo et al., 2015), which is aggravated when

the data set has high dimensionality. The studies developed by Costa et al. (2020), as suggested by Cadavid et al. (2007) and Azevedo et al. (2013b), show some criteria based on PCR that can reduce computational time to choose the optimum number of independent components.

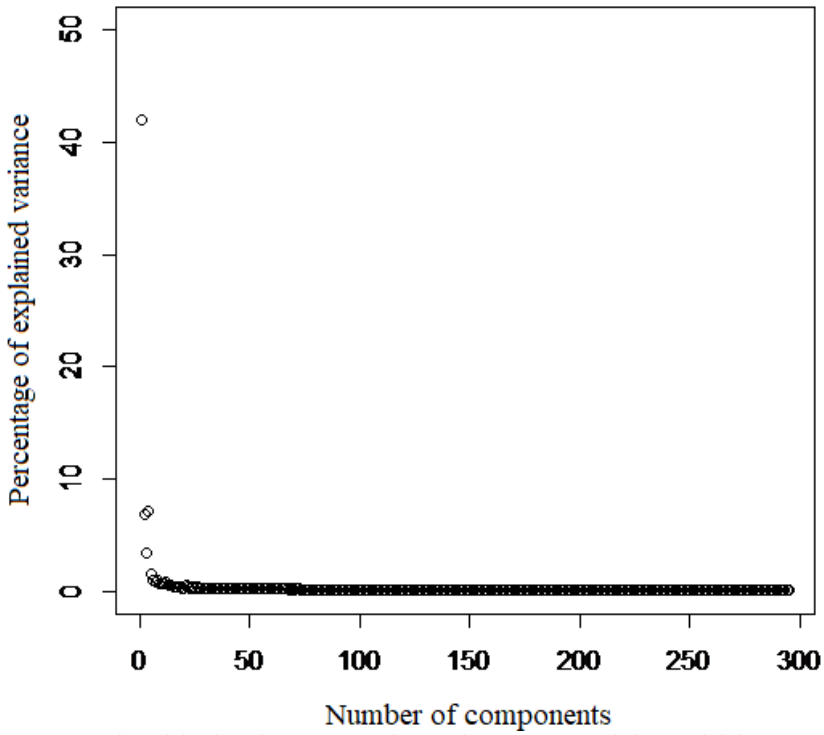

Figure 2. Percentage of explained variance concerning each component of the partial least squares (PLS) considering the matrix of incidence of the markers (the number of alleles of the marker), composed by 36,901 SNPs markers, to predict the genetic value of 370 rice accessions, *Oryza sativa,* for the trait number of panicles per plant. This genomics data and is part of the OryzaSNP Project and the OMAP Project.

The SNPs markers are widely used in Genome-Wide Selection (GWS), proposed by Meuwissen et al. (2001), to detect the genotypic information that is contributing to the phenotypic variability of individuals or accessions (Goddard e Hayes, 2007). The use of GWS allows an increase in the percentage of genetic gain concerning the cycle of selection, an increase in the accuracy of genomic values prediction, and a reduction in the interval between generations (Meuwissen et al., 2001). Therefore, there is a need to develop and discuss methods that effectively predict genetic values in the presence of multicollinearity and high dimensionality. De Los Campos et al. (2013) provides an overview of Bayesian methods applied to GWS. The Bayesian methodologies are based on Monte Carlo Markov Chain, and they need convergence analysis. They demand high computational time and effort. Bermingham et al. (2015) present the different contributions of GWS in animal and plant breeding and in supervised and unsupervised techniques to select markers and, consequently, decrease the explicative variables. The difference between the dimensionality reduction methods and these methods is that the PCR, ICR, and PLS are not select the

explanatory variables, and the final model presents all X variables. The dimensionality reduction methods are ideal for situations in which all variables explanatory models have an effect, even if small, on the response variable.

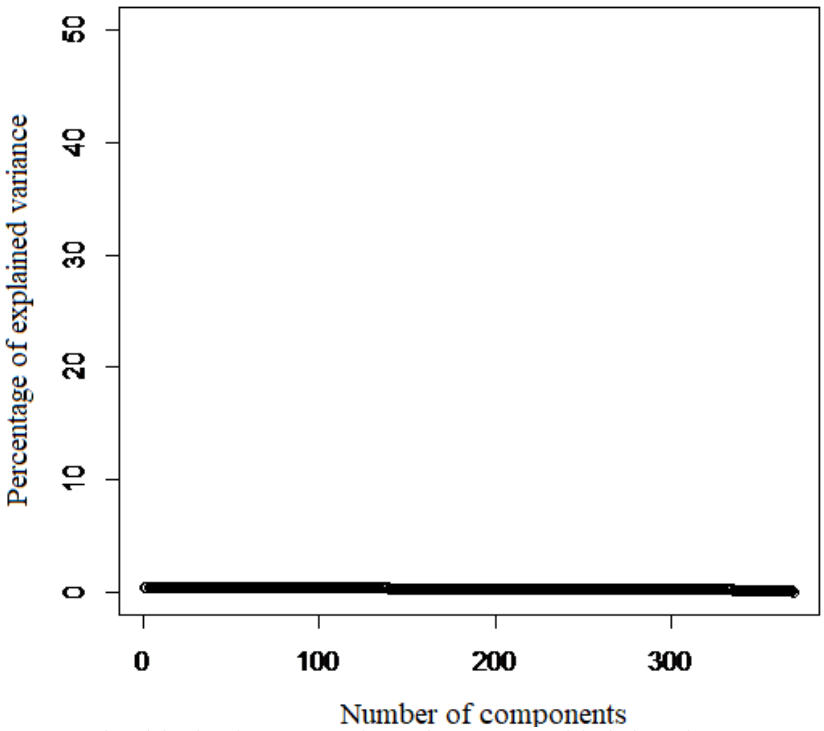

Figure 3. Percentage of explained variance concerning each component of the independent component regression (ICR) considering the matrix of incidence of the markers (the number of alleles of the marker), composed by 36,901 SNPs markers, to predict the genetic value of 370 rice accessions *Oryza sativa* for the trait number of panicles per plant. This genomics data and is part of the OryzaSNP Project and the OMAP Project.

CONCLUSION

Dimensionality reduction methods: Principal Components Regression (PCR), Partial Least Squares (PLS), and Independent Components Regression (ICR) have a relatively simple theory and present as interpretations for the parameter indications. They have the advantage of circumventing multicollinearity and high dimensionality and being efficient in the prediction process and identifying possible patterns present in the data. This class of methods has broad applicability for genomic data, as evaluated in this work to predict genomic values of rice accessions using 36,901 SNPs markers, in addition to environmental and agricultural data.

ACKNOWLEDGMENTS

We thank the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES) for financial support.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Ammiraju JSS, Luo M, Goicoechea JL, Wang W, et al. (2006). The Oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. *Genome Res.* 16: 140-147.

Azevedo CF, Nascimento M, Silva FF, de Resende MDV, et al. (2015). Comparison of dimensionality reduction methods to predict genomic breeding values for carcass traits in pigs. *Genet. Mol. Res*. 14: 12217-12227.

Azevedo CF, Resende MDVD, Silva FF, Lopes PS, et al. (2013b). Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. *Pesq. agropec. bras*. 48: 619-626.

Azevedo CF, Silva FF, de Resende MDV, Lopes MS, et al. (2014). Supervised independent component analysis as an alternative method for genomic selection in pigs. *J. Anim. Breed. Genet*. 131: 452-461.

Azevedo CF, Silva FF, Rezende MDVD, Peternelli LA, et al. (2013a). Quadrados mínimos parciais uni e multivariados aplicados na seleção genômica para características de carcaça em suínos. *Cienc. Rural* 43: 1642-1649.

Bermingham M, Pong-Wong R., Spiliopoulou A. et al. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Sci. Rep*. 5: 1-12.

Bingham E and Hyvärinen A (2000). A fast fixed-point algorithm for independent component analysis of complex valued signals. *Int. J. Neural Syst*. 10: 1-8.

Cadavid AC, Lawrence JK and Ruzmaikin A (2007). Principal Components and Independent Component Analysis of Solar and Space Data. In Solar Image Analysis and Visualization: Springer. Springer, New York.

Camarero M, Forte A, Garcia-Donato G, Mendoza Y, et al. (2015). Variable selection in the analysis of energy consumption–growth nexus. *Energy Econ*. 52: 207-216.

Comon P (1994). Independent component analysis, a new concept? *Signal Process*. 36: 287-314.

Costa JAD, Azevedo CF, Nascimento M, Resende MDVD, et al. (2020). Genomic prediction with the additive-dominant model by dimensionality reduction methods. *Pesq. agropec. Bras*. 55: e01713.

de Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, et al. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*. 193: 327-345.

Du C, Wei J, Wang S and Jia Z (2018). Genomic selection using principal component regression. *J. Hered*. 121: 12-23.

Ferreira DF (2012). Estatística Multivariada. Universidade Federal de Lavras, Lavras.

Garthwaite PH (1994). An interpretation of partial least squares. *J. Am. Stat. Assoc*. 89: 122-127.

Goddard ME and Hayes BJ (2007). Genomic selection. *J. Anim. Breed. Genet*. 124: 323-330.

Gunst RF and Webster JT (1975). Regression analysis and problems of multicollinearity. *Commun. Stat. - Theory Methods*. 4: 277-292.

Han IS, Kim M, Lee CH, Cha W, et al. (2003). Application of partial least squares methods to a terephthalic acid manufacturing process for product quality control. *Korean J. Chem. Eng.* 20: 977-984.

Härdle W and Hlávka Z (2007). Multivariate Statistics: Exercises and Soluctions. Springer, New York.

Helwig NE and Hong S (2013). A critique of tensor probabilistic independent component analysis: implications and recommendations for multi-subject fMRI data analysis. *J. Neurosci. Methods*. 213: 263-273.

He Q and Lin DY (2011). A variable selection method for genome-wide association studies. *Bioinformatics*. 27: 1-8.

Hotelling H (1957). The relations of the newer multivariate statistical methods to factor analysis. *Br. J. Stat. Psychol*. 10: 69-79.

Hyvarinen A, Karhunen J and Oja E (2001). Independent Component Analysis. John Wiley & Sons, Hoboken.

Hyvärinen A (1998). New approximations of differential entropy for independent component analysis and projection pursuit. *Adv. Neural Inf. Process. Syst*. 10: 273-279.

Hyvärinen A (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw*. 10: 626-634.

James G, Witten D, Hastie T and Tibshirani R (2013). An introduction to statistical learning. Springer, New York.

Jutten C and Herault J (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Process*. 24: 1-10.

Kendall MGA (1957). Course in Multivariate Analysis. New York: Hafner Pub. Co., London.

Kimeldorf GS and Wahba G (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Stat*. 41: 495-502.

Kuhn M (2015). Caret: classification and regression training. *ASCL*. ascl-1505.

Lim Y, Lee J, Oh HS and Kang HS (2015). Independent component regression for seasonal climate prediction: an efficient way to improve multimodel ensembles. *Theor. Appl. Climatol*. 119: 433-441.

Liu R and Gillies DF (2016). Overfitting in linear feature extraction for classification of high-dimensional image data. *Pattern Recognit*. 53: 73-86.

Lumley T (2017). leaps: Regression subset selection. *R package version 3.0* (based on Fortran code by Alan Miller).

Marcoulides GA and Hershberger SL (1997). Multivariate Statistical Methods: A First Course. Lawrence Erlbaum Associates, Mahwah.

Meuwissen TH, Hayes BJ and Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 157: 1819-1829.

Mevik BH, Wehrens R and Liland KH (2011). pls: Partial least squares and principal component regression. *R package version. 2*.

Montgomery DC, Peck EA and Vining GG (2012). Introduction to linear regression analysis. John Wiley & Sons, Hoboken.

Morgano MA, Faria CG, Ferrão MF, Bragagnolo N, et al. (2005). Determinação de proteína em café cru por espectroscopia NIR e regressão PLS. *Food Sci. Technol.* 25: 25-31.

Nascimento M, Silva FFE, Sáfadi T, Nascimento ACC, et al. (2017). Independent Component Analysis (ICA) based-clustering of temporal RNA-seq data. *PloS One.* 12: e0181195.

Otto M (1998). Chemometrics: Statistics and Computer Application in Analytical Chemistry. Wiley-VCH, Weinheim.

Puntanen S and Styan GP (1989). The equality of the ordinary least squares estimator and the best linear unbiased estimator. *Am Stat*. 43: 153-161.

Rencher AC and Christensen WF (2002). Methods of multivariate analysis. John Wiley & Sons, Hoboken.

Resende MDV, Silva FF, Lopes PS and Azevedo CF (2012). Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial. Universidade Federal de Viçosa, Viçosa.

Silveira FGD, Duarte DAS, Chaves LM, Duarte MDS, et al. (2017). The optimal number of partial least squares components in genomic selection for pork Ph. *Cienc. Rural*. 47: e20151563.

Teofilo RF, Martins JPA and Ferreira MM (2009). Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. *J. Chemom.* 23: 32-48.

Tibshirani R (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat Methodol*. *58*: 267-288.

Westad F (2005). Independent component analysis and regression applied on sensory data. *J. Chemom*. 19: 171-179.

Wold H (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. *J. Appl. Probab*. 12: 117-142.

Yao F, Coquery J and Lê Cao KA (2012). Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. *BMC Bioinformatics*. *13*: 1-15.

Zhao K, Tung CW, Eizenga GC, Wright MH, et al. (2011). Genome-wide association mapping reveals a rich geneticarchitecture of complex traits in *Oryza sativa*. *Nat. Commun.* 2: 1-10.