Biometry, Modeling, and Statistics

DOI: http://doi.org/10.1590/1678-992X-2020-0397

Scientia Agricola

ISSN 1678-992X

Research Article

Determination of optimal number of independent components in yield traits in rice

Jaquicele Aparecida da Costa[1]*, Camila Ferreira Azevedo[1], Moysés Nascimento[1], Fabyano Fonseca e Silva[2], Marcos Deon Vilela de Resende[3], Ana Carolina Campana Nascimento[1]

[1]Universidade Federal de Viçosa – Depto. de Estatística, Av. Peter Henry Rolfs, s/n. – Campus Universitário – 36570-900 – Viçosa, MG – Brasil.
[2]Universidade Federal de Viçosa – Depto. de Zootecnia – Viçosa, MG – Brasil.
[3]Embrapa Café/Universidade Federal de Viçosa – Depto. de Engenharia Florestal – Viçosa, MG – Brasil.
*Corresponding author <jaquicele.costa@ufv.br>

Edited by: Thomas Kumke

Received December 21, 2020
Accepted August 27, 2021

ABSTRACT: The principal component regression (PCR) and the independent component regression (ICR) are dimensionality reduction methods and extremely important in genomic prediction. These methods require the choice of the number of components to be inserted into the model. For PCR, there are formal criteria; however, for ICR, the adopted criterion chooses the number of independent components (ICs) associated to greater accuracy and requires high computational time. In this study, seven criteria based on the number of principal components (PCs) and methods of variable selection to guide this choice in ICR are proposed and evaluated in simulated and real data. For both datasets, the most efficient criterion and that drastically reduced computational time determined that the number of ICs should be equal to the number of PCs to reach a higher accuracy value. In addition, the criteria did not recover the simulated heritability and generated biased genomic values.

Keywords: *Oryza sativa* L., genomic prediction, plant breeding, principal component regression, independent component regression

Introduction

The prediction process in Genome Wide Selection (GWS) (Meuwissen et al., 2001) presents statistical problems related to high dimensionality (number of markers greater than the number of individual phenotypic observations) and multicollinearity (highly correlated markers), which affect the accuracy of methods based on ordinary least squares (OLS) (Desta and Ortiz, 2014).

In this context, methodologies to solve such statistical challenges have gained prominence in GWS research. Resende et al. (2012) reported that the statistical methodologies applied to GWS could be divided into three groups: methods based on explicit regression, implicit regression, and the dimensionality reduction methods. Among these, the dimensionality reduction methods, the Principal Component Regression (PCR), and the Independent Component Regression (ICR) are highlighted when compared to the other methods applied to GWS as they present great applicability and relatively simple theory.

The PCR and ICR require the choice of the optimal number of components, which are linear combinations of the markers, to be inserted in the prediction equation. The statistical theory of PCR demonstrates that the first components represent most of the total data variability. Le Floch et al. (2012) presented the criterion for choosing the optima number based on this assertion.

In genomic selection, effective methodologies for the prediction process are desirable and accuracy is one of the main measurements of efficacy. Azevedo et al. (2014, 2015) chose the number of independent components (ICs) associated to greater accuracy; however, the execution of the analyses required a high computational effort, which often becomes impractical.

In this study we aimed to propose and evaluate, using simulated genomic data, seven decision criteria for the optimal number of components to be inserted into the template. We also evaluated seven criteria with real data in the genomic prediction of six rice yield traits to elucidate the importance of the procedures described in this study for breeding programs and the importance of genomic prediction for the Asian rice *Oryza sativa* L. (Grenier et al., 2015; Hassen et al., 2018; Spindel et al., 2015; Spindel et al., 2016).

Materials and Methods

The simulated dataset was generated as described by Azevedo et al. (2015, 2017). We simulated 2,000 equidistant Single Nucleotide Polymorphisms (SNPs) markers separated by 0.1 centiMorgan among ten chromosomes. The quantitative trait loci (QTLs) were randomly distributed in the regions covered by the SNPs. We genotyped and phenotyped 1,000 individuals from 20 families of full siblings. The simulations assumed absence of dominance and four scenarios were used in the analyses: two heritability levels in the restricted sense (about 0.20 and 0.30) × two genetic architectures (polygenic and mixed inheritance). The scenarios were analyzed considering the dimensionality reduction methods, ICR and PCR, and the criteria of choice of the components. Each type of population (or scenarios) was simulated ten times.

The real data set corresponded to the Asian rice and the database used in this study consisted of six yield traits referring to 370 accessions of rice, which were genotyped to 44,100 SNP markers. This dataset is free and is part of two projects, the OryzaSNP Project and the OMAP Project (Ammiraju et al., 2006; Zhao et al.,

2011) and it is available at https://ricediversity.org/data/. The six traits of rice yield used in this study were: (i) panicle number per plant, (ii) plant height, (iii) panicle length, (iv) primary panicle branch number, (v) seed number per panicle, and (vi) florets per panicle.

The linear model is given by:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{m}_a + \boldsymbol{e}, \tag{1}$$

where: $\boldsymbol{y}$ is the vector of phenotypic observations with dimension I × 1, where I is the number of individuals genotyped and phenotyped, μ is the overall mean of the trait, $\boldsymbol{m}_a$ is the vector of additive marker effects with incidence matrix $\boldsymbol{X}$ composed of values 0, 1, and 2 whose dimension is I × J. J is the number of markers and $\boldsymbol{e}$ is the vector of random errors with the structure of variance given by $\boldsymbol{e} \sim \boldsymbol{N}(\mathbf{0}, \mathrm{I}\sigma_e^2)$, where $\boldsymbol{I}$ is the identity matrix and σ_e^2 is the residual variance.

The Principal Component Regression (PCR) and the Independent Component Regression (ICR) can be used in any situation where there are problems of high dimensionality. The main difference between the methods is that in the PCR, the principal components $Z_m(m = 1, \ldots, n_{PCR})$ are orthogonal components and the first components explain much of the total variability. In the ICR, the components built are independent, that is, there is no functional relationship between the components that explain small parts and in different proportions the total data variability.

The Principal Component Regression determines that the PCs are defined as:

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{P}, \tag{2}$$

where: $\boldsymbol{X}$ is the incidence matrix of the markers and $\boldsymbol{P}$ is the matrix of the eigenvectors of the covariance matrix of $\boldsymbol{X}$. The first component is associated to the largest eigenvalue of the eigenvectors matrix and is the percentage of explanation of the j^th^ component given by:

$$\frac{\lambda_j}{\sum_{j=1}^{m} \sigma_j^2},$$

where λ_j is the corresponding eigenvalue. In order to perform the prediction of the genomic values, the vector of phenotypic observations ($\boldsymbol{y}$) is related to the components ($\boldsymbol{Z}$) and for this regression to be possible, the number of components to be inserted into the model (n_{PCR}) is less than or equal to (I, J) – 1. After this choice, the n_{PCR} first components, Z_1, Z_2, … , $Z_{n_{PCR}}$, are selected and the adjusted prediction equation is $\hat{y} = Z_1\hat{\alpha}_1 + Z_2\hat{\alpha}_2 + \cdots + Z_{n_{PCR}}\hat{\alpha}_{n_{PCR}}$, where $\hat{\alpha} = \left[\hat{\alpha}_1 \ \hat{\alpha}_2 \cdots \hat{\alpha}_{n_{PCR}}\right]$ is the vector of the estimated regression coefficients obtained by the OLS method.

The coefficients α are not related to the markers. The following expression is used to find the estimates of the effects of the markers:

$$\hat{m}_{PCR} = P_{n_{PCR}}\hat{\alpha}, \tag{3}$$

where: $\boldsymbol{P}_{n_{PCR}}$ is the matrix of associated eigenvectors to components Z_1, Z_2, …, $Z_{n_{PCR}}$.

The ICR decomposes the matrix $\mathbf{X}$ into $\boldsymbol{X} = \boldsymbol{SA'}$, where $\mathbf{S}$ ($I \times n_{ICR}$) is an ICs matrix. $\mathbf{A}$ ($J \times n_{ICR}$) is called the mix matrix, which is usually unknown, and n_{ICR} is the number of ICs chosen. To estimate matrix $\mathbf{A}$, the first step is to obtain matrix $\mathbf{K}$ (called the whitening matrix) by orthogonal decomposition of the covariance matrix of $\mathbf{X}$ to ensure that the covariance matrix of $\boldsymbol{XK}$ is equal to the identity matrix, the correlation between the columns of $\boldsymbol{XK}$ is equal to 0, and the variance is equal to 1. The orthogonal decomposition is applied to the covariance matrix of $\mathbf{X}$, denoted by $\boldsymbol{\Sigma}(J \times J)$ obtaining: $\sum = \boldsymbol{P}\Lambda^{-\frac{1}{2}}\boldsymbol{P'}$; where $\boldsymbol{P}$ is composed of the eigenvectors in its columns and Λ is a diagonal matrix of eigenvalues of the covariance matrix of $\mathbf{X}$. In regression, the matrix $\boldsymbol{K}$ ($J \times n_{ICR}$) is then defined as $\boldsymbol{P}_r\Lambda_r^{-\frac{1}{2}}$, where $\boldsymbol{P}_r$ is the matrix with n_{ICR} as the first columns of the matrix $\boldsymbol{P}$ (n_{ICR} first eigenvectors) and Λ_r the matrix with n_{ICR} as the first rows and columns of the matrix Λ (eigenvalues associated with these first eigenvectors). To achieve independence between the components, the algorithm proposed by Hyvärinen (1998), which is based on the principle of maximum entropy, is used to obtain a new matrix denoted by $\mathbf{R}$ ($n_{ICR} \times n_{ICR}$). After the algorithm, the ICs can be expressed by:

$$\boldsymbol{S} = \boldsymbol{XKR}. \tag{4}$$

Then, the prediction equation between the response variable Y and the ICs S_1, S_2, …, $S_{n_{ICR}}$ is given by $\hat{y} = \hat{\beta}_1 s_1 + \hat{\beta}_2 s_2 + \cdots + \hat{\beta}_{n_{ICR}} s_{n_{ICR}}$, where $\hat{\beta} = [\hat{\beta}_1 \ \hat{\beta}_2 \cdots \hat{\beta}_{n_{ICR}}]'$ is the vector of the regression coefficient estimates obtained by the OLS method. Analogously to the PCR, to find the estimates of the effects of the markers, it is enough to use the expression:

$$\hat{m}_{ICR} = \boldsymbol{KR}\hat{\beta}. \tag{5}$$

The simulated and real datasets were analyzed using two populations (estimation and validation population) according to both validation procedures. In the simulated data, the criteria were compared by means of an independent validation in which the first nine simulations were assumed as estimation populations and the 10th simulation was assumed as the validation population. The real data were evaluated under a ten-fold validation process. The use of different validation processes is justifiable, because the real dataset comprise a small number of individuals (370), which makes independent validation unviable and, in these cases, James et al. (2013) suggest a ten-fold validation.

The criteria analyzed aimed to determine the optimal number of ICs using the following procedures.

Criterion 1 (Based on predictive ability or accuracy obtained through PCR fit): For each PC (m = 1, …, $min(I,J)$–1), the effects of the markers in estimating the population are estimated by the PCR and they are used in the validation population to estimate the genomic breeding values of the individuals of this population. Then, for the simulated data, we analyzed the accuracy $(r_{a\hat{a}})$, the correlation between the estimated genomic value and the real genomic value $(r_{a\hat{a}} = Cor(\hat{a}, a))$, and for the real data $(r_{y\hat{a}})$, the correlation between the estimated genomic value and the phenotypic value $(r_{y\hat{a}} = Cor(\hat{a}, y))$. This analysis ensures that the number of ICs is equal to the number of PCs whose genomic value leads to greater accuracy and predictive ability. Cadavid et al. (2008) and Azevedo et al. (2013), corroborated the use of PCR in the choice of the number of ICs.

Criterion 2 (Based on bias and predictive ability or accuracy obtained through the PCR fit): In Criterion 2, the same procedure in Criterion 1 was used, but the regression coefficient $(b_{y\hat{a}})$ is calculated between the phenotype and the estimated genomic value and, subsequently, the prediction bias given by $1 - b_{y\hat{a}}$. Thus, the number of ICs is determined as equal to the number of PCs whose genomic value leads to a smaller prediction bias.

Criterion 3 (Based on the percentage explanation of the total variation of the markers after obtaining the PCs): The percentage explanation of the total variation of X when using *m* PCs is given by:

$$p_m(\%) = \frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{J} \lambda_j} ,$$

where λ_j is the eigenvalue corresponding to the j^th^ eigenvector of the covariance matrix of X. Criterion 3 determines that the number of ICs is equal to the number of PCs that explain 80 % of the total variation of X, as recommended by Ferreira (2012). The researcher can also choose another threshold value and it must consider the explanation percentage of the data variation and the dimensionality reduction caused.

Criterion 4 (Based on the coefficient of determination obtained after the PCR fit): Using the coefficient of determination $(R^2 = Cor(y, \hat{a})^2 \times 100\%)$, the IC number is chosen as equal to the number of PCs explaining 80 % of the total variation of Y.

Criterion 5 (Based on the percentage of explanation of the total variation of markers after obtaining ICs): Assuming that the ICs have means equal to 0 and variances equal to 1, the variation explained by the k^th^ IC is given by:

$$\frac{I \sum_{j=1}^{J} a_{jk}^2}{\sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij}^2}$$

where a_{jk} is the element of the j^th^ row and the k^th^ column of the matrix of mixtures A, x_{ij} is the element of the i^th^ row and j^th^ column of the centered matrix of explanatory variables X (i = 1, 2, ..., I) (Bingham and Hyvärinen, 2000; Helwig and Hong, 2013). The number of ranked ICs that explain 80 % of the total variation of X is then chosen.

Criterion 6 (Based on the IC's Forward Selection algorithm): After the application of the ICA in matrix X, were determined which are predictors, ICs m = ((I, J) −1), to be included in the model. For this, based on the Forward Selection algorithm described by James et al. (2013), the M_0 model without ICs is considered. For the first iteration of the algorithm, the models with only one IC, denoted by M_{1i} (i = 1, ..., (I, J) −1) are constructed and R^2 is calculated for each model. Subsequently, the model with the highest R^2 is defined as model M_1. In the second iteration, the models with two ICs (all models must contain the component that is the model predictor M_1) denoted by M_{2i} are constructed and the model with the larger R^2 is denoted as M_2. This procedure is performed (I, J) −1 times for determinate the models M_1, M_2, ... $M_{(I,J)-1}$ with 1, ..., (I, J)−1 ICs, respectively, in each model. Among all these models, the model with the lowest BIC (Bayesian Information Criterion) was chosen. The present criterion determines which and how many ICs must be used in the chosen model.

Criterion 7 (Based on the IC Backward Elimination algorithm): Based on the Backward Elimination algorithm, as described by James et al. (2013), the complete model $M_{n_{(I,J)-1}}$ is considered, that is, the model with the maximum number of ICs built after the application of ICA. Subsequently, the models with (I, J)−2 components are defined as M (I, J)−2, which were constructed by removing one IC at a time and calculated the R^2 for each model. It is denoted as $M_{(I,J)-2}$, the model with the largest R^2. The process is repeated to determine the models $M_{((I,J)-3)}$, ..., M_1. The component that is not included into the model also does not participate in the following iteration. Then, from these (I, J)−1 models, only the model that features lower BIC. The present criterion determines which and how many ICs must be in the chosen model.

In the simulated data, efficacy measurements of genomic prediction were calculated for each replicate, such as: i) accuracy ($r_{\hat{a}a}$), $r_{\hat{a}a}$ is the correlation between the genomic estimated breeding values (GEBVs – denoted by $\hat{a}$) and the simulated genetic values (a); ii) prediction bias, which is defined as $1 - b_{y\hat{a}}$ being $b_{y\hat{a}}$ the regression coefficient between phenotype (y) and GEBVs; iii) additive genomic heritability (h_{aM}^2), given by:

$$h_{aM}^2 = \frac{\sigma_{aM}^2}{\sigma_{aM}^2 + \sigma_e^2} ,$$

where $\sigma_{aM}^2 = \sum_{j=1}^{J} 2 p_j q_j m_{aj}^2$ is the additive genomic variance, σ_e^2 is the residual variance, and p_i and q_i are the allelic frequencies of the j^th^ marker. After obtaining the efficacy measures for each replicate in each scenario, the results will be the mean and standard deviation of these values. In the real data, the efficacy measurements of genomic prediction were: i) predictive ability $\left(r_{y\hat{a}}\right)$, $r_{y\hat{a}}$ is the correlation between the GEBVs and phenotype; ii) prediction bias; iii) additive genomic heritability.

Regarding the interpretation of efficacy measures, we have: i) high accuracy values indicate that the GEBV is close to the real genomic value; ii) high predictive ability values indicate that the GEBV is close to the phenotype; iii) regression coefficients below 1 $\left(b_{y\hat{a}} < 1\right)$, it is understood that the GEBVs were overestimated, for regression coefficients above 1 $\left(b_{y\hat{a}} > 1\right)$, it is concluded that the GEBVs were underestimated, and for coefficients equal to 1 $\left(b_{y\hat{a}} = 1\right)$, it concludes that GEBVs are unbiased; iv) In simulated data, estimated genomic heritability should be close to simulated heritability. In real data, the estimated genomic heritability was compared to the heritability presented in other studies. The configuration of the computer used in the statistical analyses was: Intel (R) Core (TM) i7-6500 (CPU 2.50 GHz) processor with 16 Gb of RAM. All the computational routines of the methods used were implemented in GenomicLand (Azevedo et al., 2019) available at https://licaeufv.wordpress.com/genomicland/.

Results and Discussion

The mean results and the deviations from the simulations regarding the number of components, additive molecular heritability, accuracy, and bias considering the ICR and each criterion for choosing the optimal number of ICs are presented in Tables 1 and 2. In addition, the

Table 1 – The parametric additive heritability (h^2_{Mapar}), the number of components (N_c), additive heritability (h^2_{aM}), accuracy ($r_{a\hat{a}}$), and regression coefficient considering ($\hat{b}_{y\hat{a}}$) each criterion of choice for the number of independent components and the scenarios of polygenic inheritance.

Scenario	h^2_{Mapar}	Criterion	N_c	h^2_{aM}	$r_{a\hat{a}}$	$\hat{b}_{y\hat{a}}$
Scenario 1	0.20	Exhausting	36 ± 0.00	0.20 ± 0.03	0.71 ± 0.02	0.90 ± 0.05
		Criterion 1	66 ± 78	0.22 ± 0.07	0.70 ± 0.01	0.89 ± 0.11
		Criterion 2	5 ± 4	0.04 ± 0.03	0.34 ± 0.15	1.10 ± 0.10
		Criterion 3	130 ± 0.00	0.27 ± 0.03	0.70 ± 0.02	0.75 ± 0.04
		Criterion 4	730 ± 0.00	1.00 ± 0.00	0.48 ± 0.04	0.25 ± 0.03
		Criterion 5	780 ± 0.00	1.00 ± 0.00	0.46 ± 0.04	0.23 ± 0.02
		Criterion 6	630 ± 470	1.00 ± 0.00	0.58 ± 0.17	0.07 ± 0.06
		Criterion 7	630 ± 470	1.00 ± 0.00	0.58 ± 0.17	0.07 ± 0.06
Scenario 2	0.30	Exhausting	44 ± 0.00	0.22 ± 0.04	0.75 ± 0.02	0.92 ± 0.07
		Criterion 1	40 ± 25	0.21 ± 0.04	0.74 ± 0.02	0.94 ± 0.06
		Criterion 2	7 ± 8	0.08 ± 0.04	0.48 ± 0.13	0.98 ± 0.06
		Criterion 3	130 ± 0.00	0.28 ± 0.04	0.72 ± 0.02	0.78 ± 0.05
		Criterion 4	730 ± 0.00	1.00 ± 0.00	0.50 ± 0.04	0.25 ± 0.03
		Criterion 5	780 ± 0.00	1.00 ± 0.00	0.48 ± 0.04	0.23 ± 0.03
		Criterion 6	950 ± 33	1.00 ± 0.00	0.71 ± 0.01	0.04 ± 0.01
		Criterion 7	950 ± 33	1.00 ± 0.00	0.71 ± 0.01	0.04 ± 0.01
Scenario 3	0.20	Exhausting	277 ± 0.00	0.53 ± 0.06	0.77 ± 0.02	0.77 ± 0.05
		Criterion 1	260 ± 96	0.51 ± 0.14	0.77 ± 0.02	0.80 ± 0.10
		Criterion 2	16 ± 17	0.15 ± 0.09	0.53 ± 0.19	1.00 ± 0.03
		Criterion 3	130 ± 0.00	0.36 ± 0.04	0.75 ± 0.01	0.89 ± 0.06
		Criterion 4	730 ± 0.00	1.00 ± 0.00	0.66 ± 0.02	0.45 ± 0.02
		Criterion 5	780 ± 0.00	1.00 ± 0.00	0.63 ± 0.03	0.42 ± 0.03
		Criterion 6	960 ± 16	1.00 ± 0.00	0.72 ± 0.02	0.04 ± 0.01
		Criterion 7	960 ± 16	1.00 ± 0.00	0.72 ± 0.02	0.04 ± 0.01
Scenario 4	0.30	Exhausting	189 ± 0.00	0.48 ± 0.03	0.80 ± 0.02	0.83 ± 0.03
		Criterion 1	200 ± 110	0.50 ± 0.14	0.80 ± 0.02	0.83 ± 0.09
		Criterion 2	4 ± 3	0.11 ± 0.05	0.45 ± 0.11	1.00 ± 0.04
		Criterion 3	130 ± 0.00	0.42 ± 0.03	0.79 ± 0.02	0.87 ± 0.04
		Criterion 4	730 ± 0.00	1.00 ± 0.00	0.66 ± 0.03	0.45 ± 0.02
		Criterion 5	780 ± 0.00	1.00 ± 0.00	0.63 ± 0.03	0.42 ± 0.03
		Criterion 6	960 ± 30	1.00 ± 0.00	0.73 ± 0.01	0.04 ± 0.01
		Criterion 7	960 ± 30	1.00 ± 0.00	0.73 ± 0.01	0.04 ± 0.01

Number of independent components leading to: Independent Component Regression at higher accuracy (exhaustive); Principal Component Regression (PCR) at higher accuracy (Criterion 1); PCR at a lower bias value (Criterion 2); 80 % of the total variation of X explained by the principal components (Criterion 3); 80 % of the total variation of Y explained by the principal components (Criterion 4); 80 % of the total variation of X explained by the independent components (Criterion 5); Forward Selection (Criterion 6); Backward Elimination (Criterion 7).

Table 2 – The number of components (N_c), additive heritability (h^2_{aM}), predictive capacity ($r_{y\hat{a}}$), and prediction bias ($\hat{b}_{y\hat{a}}$) considering the exhaustive method and each criterion for choice of number of independent components.

Trait	Criterion	N_c	h^2_{aM}	$r_{y\hat{a}}$	$\hat{b}_{y\hat{a}}$
Panicle number per plant	Exhausting	117	0.69 ± 0.05	0.82 ± 0.06	0.98 ± 0.11
	Criterion 1	125	0.73 ± 0.03	0.82 ± 0.01	0.97 ± 0.03
	Criterion 2	105	0.74 ± 0.01	0.82 ± 0.01	0.97 ± 0.02
	Criterion 3	44	0.69 ± 0.03	0.82 ± 0.01	1.02 ± 0.03
	Criterion 4	55	0.80 ± 0.03	0.71 ± 0.02	0.79 ± 0.01
	Criterion 5	263	1.00 ± 0.00	0.70 ± 0.07	0.66 ± 0.10
	Criterion 6	295	1.00 ± 0.00	0.08 ± 0.13	0.00 ± 0.01
	Criterion 7	295	1.00 ± 0.00	0.08 ± 0.13	0.00 ± 0.01
Plant Height	Exhausting	175	0.65 ± 0.07	0.81 ± 0.05	1.00 ± 0.17
	Criterion 1	213	0.47 ± 0.01	0.78 ± 0.01	1.17 ± 0.01
	Criterion 2	5	0.24 ± 0.01	0.56 ± 0.01	1.14 ± 0.03
	Criterion 3	44	0.35 ± 0.01	0.72 ± 0.01	1.20 ± 0.02
	Criterion 4	55	0.38 ± 0.02	0.72 ± 0.01	1.18 ± 0.02
	Criterion 5	263	0.94 ± 0.06	0.71 ± 0.06	0.74 ± 0.07
	Criterion 6	295	1.00 ± 0.00	0.04 ± 0.09	0.00 ± 0.01
	Criterion 7	295	1.00 ± 0.00	0.04 ± 0.09	0.00 ± 0.01
Panicle length	Exhausting	157	0.52 ± 0.08	0.69 ± 0.06	0.92 ± 0.18
	Criterion 1	140	0.42 ± 0.02	0.68 ± 0.03	1.04 ± 0.05
	Criterion 2	152	0.44 ± 0.03	0.69 ± 0.02	1.05 ± 0.06
	Criterion 3	44	0.38 ± 0.02	0.66 ± 0.03	1.06 ± 0.06
	Criterion 4	55	0.39 ± 0.02	0.65 ± 0.03	1.04 ± 0.04
	Criterion 5	263	0.94 ± 0.06	0.51 ± 0.12	0.56 ± 0.19
	Criterion 6	295	1.00 ± 0.00	0.04 ± 0.12	0.00 ± 0.01
	Criterion 7	295	1.00 ± 0.00	0.04 ± 0.12	0.00 ± 0.01
Primary panicle branch number	Exhausting	123	0.46 ± 0.08	0.64 ± 0.08	0.90 ± 0.25
	Criterion 1	154	0.78 ± 0.06	0.51 ± 0.03	0.57 ± 0.04
	Criterion 2	3	0.19 ± 0.03	0.43 ± 0.06	0.97 ± 0.07
	Criterion 3	44	0.52 ± 0.05	0.54 ± 0.04	0.75 ± 0.08
	Criterion 4	55	0.60 ± 0.03	0.55 ± 0.02	0.70 ± 0.05
	Criterion 5	263	0.74 ± 0.26	0.51 ± 0.13	0.64 ± 0.26
	Criterion 6	295	1.00 ± 0.00	0.01 ± 0.07	0.00 ± 0.01
	Criterion 7	295	1.00 ± 0.00	0.01 ± 0.07	0.00 ± 0.01
Seed number per panicle	Exhausting	48	0.31 ± 0.08	0.56 ± 0.08	1.02 ± 0.13
	Criterion 1	27	0.31 ± 0.04	0.46 ± 0.03	0.82 ± 0.05
	Criterion 2	22	0.28 ± 0.04	0.47 ± 0.02	0.89 ± 0.07
	Criterion 3	44	0.37 ± 0.01	0.45 ± 0.04	0.74 ± 0.07
	Criterion 4	55	0.37 ± 0.02	0.46 ± 0.05	0.75 ± 0.08
	Criterion 5	263	0.89 ± 0.11	0.54 ± 0.09	0.60 ± 0.12
	Criterion 6	295	1.00 ± 0.00	0.01 ± 0.17	0.00 ± 0.01
	Criterion 7	295	1.00 ± 0.00	0.04 ± 0.12	0.00 ± 0.01
Florets per panicle	Exhausting	52	0.42 ± 0.09	0.66 ± 0.08	1.03 ± 0.13
	Criterion 1	140	0.69 ± 0.06	0.50 ± 0.05	0.72 ± 0.07
	Criterion 2	152	0.69 ± 0.02	0.60 ± 0.02	0.73 ± 0.02
	Criterion 3	44	0.31 ± 0.03	0.56 ± 0.02	1.00 ± 0.05
	Criterion 4	55	0.33 ± 0.02	0.56 ± 0.02	0.98 ± 0.06
	Criterion 5	263	0.73 ± 0.23	0.50 ± 0.08	0.60 ± 0.13
	Criterion 6	295	1.00 ± 0.00	0.01 ± 0.16	0.00 ± 0.01
	Criterion 7	295	1.00 ± 0.00	0.01 ± 0.16	0.00 ± 0.01

Number of independent components leading to: Independent Component Regression at higher accuracy (exhaustive); Principal Component Regression (PCR) at higher accuracy (Criterion 1); PCR at a lower bias value (Criterion 2); 80 % of the total variation of X explained by the principal components (Criterion 3); 80 % of the total variation of Y explained by the principal components (Criterion 4); 80 % of the total variation of X explained by the independent components (Criterion 5); Forward Selection (Criterion 6); Backward Elimination (Criterion 7).

results of the analyses of the calculating the number of components required to reach the maximum value of accuracy via ICR by the exhaustive method are also presented.

Among the seven criteria analyzed, criteria 1, 3, 6, and 7 presented the values of accuracy closer to the maximum accuracy value obtained by the exhaustive method, considering the four scenarios. Although criteria 6 and 7 presented high accuracy values, both criteria overestimated the genomic values, which can be observed in the regression coefficient, revealing that the estimates found have variability beyond the simulated ones. Criteria 1 and 3 present values closer to unity than those obtained by the exhaustive method, highlighting Criterion 1 even more as it presents high accuracy and low bias. The bias property is relevant because that selection involves individuals of many generations using effects of estimated markers in a single generation, which is desirable not only to select individuals, but also to determine the genomic merits of individuals (Resende et al., 2014).

No criterion were adequate to estimate heritability in scenarios 2, 3, and 4, since the values do not recover the simulated heritability. However, these values are close to the heritability attained by the exhaustive criterion, considering the maximum value of accuracy via the ICR. In criteria 4, 5, 6, and 7, heritability estimates equal 1 and these criteria are associated to the largest number of components in the model. Likewise, we evaluated the number of components influencing the heritability estimation and the extent to which components are included in the model where heritability tends to 1. This can be explained by the ICR method assuming the SNPs as fixed effects; since according to Resende et al. (2014), when the markers are assumed to be fixed effects, heritability is implicitly assumed to equal 1.

Regarding the Forward Selection and Backward Elimination criteria (criteria 6 and 7, respectively), the variable selection methods aimed to remove variables that are not relevant or those not closely related to the dependent variable (James et al., 2013). In the case of the ICR, the components were independent (the components were uncorrelated and without any functional relation to each other) and thus more variables were needed, that is, more components to explain the response variable in criteria 6 and 7. The prediction of genomic values using these selection criteria was not adequate since the criteria associated to the largest biases (coefficient values close to 0) were not adequate.

Other criteria have been proposed, such as Akaike Information Criterion (AIC), BIC, coefficient of determination mean square of the residues, and adjusted coefficient of determination. However, the application of these suggested criteria was not feasible, since the computational time would have been the same as in the exhaustive method. Similarly, using the Stepwise Selection method, the number of variables selected resulted in the complete model (considering 999 components) that was associated to low accuracy values.

The number of components, additive molecular heritability, predictive capacity, and prediction bias for the six rice traits are shown in Table 2, considering each criterion (Criterion 1 – Based on predictive ability or accuracy obtained through PCR fit, Criterion 2 – Based on bias and predictive ability or accuracy obtained through the PCR fit, Criterion 3 – Based on the percentage explanation of the total variation of the markers after obtaining the PCs, Criterion 4 – Based on the coefficient of determination obtained after the PCR fit), Criterion 5 – Based on the percentage of explanation of the total variation of markers after obtaining ICs, Criterion 6 – Based on the IC's Forward Selection algorithm and Criterion 7 – Based on the IC Backward Elimination algorithm) for choosing the optimal number of ICs. Likewise, the number of ICs required by the exhaustive model is also shown in Table 2. The results for the six traits corroborate the findings obtained in the simulated data.

For the real data, Criterion 1 presented values of predictive capacity closer to the maximum for the traits panicle number per plant, plant height, and panicle length. In this context, Criteria 2 and 3 were also significant for the traits panicle number per plant and panicle length, while Criterion 4 did not show prominence for any trait. The analyses of the regression coefficient showed that all the criteria were biased and Criteria 6 and 7 considerably overestimated the genomic values for all traits, as observed in the analyses of the simulated data.

In relation to the traits of plant height, panicle length, and seed number per panicle, Bisne et al. (2009) reported that heritability values oscillate between high and medium, indicating success in selection. Thus, considering the real dataset, the heritability values found in our study and other studies are presented in Table 3. Heritability presented by Akinwale et al. (2011) and Seyoum et al. (2012) was estimated via pedigree and, in the context of our study, genomic heritability was considered. In addition, Ogunbayo et al. (2014) reported a high heritability value for number of panicles in the primary panicle, which is justifiable, since these authors considered heritability in the broad sense.

The computational times associated to the simulated and real data in s and h are presented in Table 4. The computational time for the exhaustive method of the simulated dataset, considering a replicate of each scenario, required high computational time. This can also be observed in the real dataset using a high number of molecular markers. However, using Criterion 1, the reduction in time was drastic. This time would be substantially greater when we consider that 500,000 and 600,000 SNPs are identified in bovine and ovine genotyping (Brito et al., 2017; Wilkinson et

Table 3 – Heritability values and heritability observed in the literature for each trait.

Traits	Estimated heritability	References
Panicle number per plant	0.73^{C1}, 0.74^{C2}, 0.63^{C3}, (0.69)	(0.59) Akinwale et al. (2011) and (0.50) Seyoum et al. (2012)
Plant Height	0.47* (0.65)	Grenier et al. (2015), Spindel et al. (2015) and Spindel et al. (2016)
Panicle length	0.42^{C1}, 0.44^{C2}, 0.38^{C3}, (0.52)	(0.53) Akinwale et al. (2011)
Primary panicle branch number	0.78^{C1}, 0.52^{C2}, (0.46)	(0.76) Ogunbayo et al. (2014)
Seed number per panicle	0.31^{C1}, 0.28^{C2}, (0.31)	(0.70) Akinwale et al. (2011)
Florets per panicle	$0.69^{C1, C2}$, (0.42)	(0.60) Seyoum et al. (2012) and (0.61) Akinwale et al. (2011)

C1 = Estimated heritability by Criterion 1; C2 = Estimated heritability by Criterion 2; C3 = Estimated heritability by Criterion 3; ( ) = Estimated heritability by the exhaustive method.

Table 4 – Computational time in s (h) considering the simulated data and real data and each criterion for choosing the number of independent components.

Data	Criterion	Computational Time
Simulated	Exhausting	587,776.39 (163.27)
	Criterion 1	224.80 (0.06)
	Criterion 2	197.67 (0.05)
	Criterion 3	197.22 (0.05)
	Criterion 4	888.99 (0.25)
	Criterion 5	2,235.96 (0.62)
	Criterion 6	1,351.34 (0.38)
	Criterion 7	1,350.24 (0.38)
Real	Exhausting	3,189,757.20 (886.04)
	Criterion 1	5,945.90 (1.65)
	Criterion 2	5,029.09 (1.40)
	Criterion 3	2,179.76 (0.61)
	Criterion 4	2,524.28 (0.70)
	Criterion 5	19,827.66 (5.51)
	Criterion 6	3,260.44 (0.91)
	Criterion 7	3,262.01 (0.91)

Number of independent components leading to: Independent Component Regression at higher accuracy (exhaustive); Principal Component Regression (PCR) at higher accuracy (Criterion 1); PCR at a lower bias value (Criterion 2); 80 % of the total variation of X explained by the principal components (Criterion 3); 80 % of the total variation of Y explained by the principal components (Criterion 4); 80 % of the total variation of X explained by the independent components (Criterion 5); Forward Selection (Criterion 6); Backward Elimination (Criterion 7).

al., 2017), that is, hundreds of thousands of marker effects to be estimated considering only the additive model. It was also vrified that the computational time is drastically reduced considering Criterion 1.

Conclusion

In general, Criterion 1, the number of ICs equals to the number of PCs that leads to a higher value of accuracy, presented an effective and computationally feasible alternative compared to the exhaustive method, both for simulated data and for the traits of real data. Criterion 3 had high accuracy values for simulated data and for some traits of real data, but essentially lower values compared to Criterion 1. Criteria 6 and 7 had high accuracy values for real and simulated data, but they overestimate the genomic breeding values. Criteria 2 and 4 had low accuracy values. None of the criteria were capable of capturing the heritability values that were simulated.

Acknowledgments

To Coordination for the Improvement of Higher Level (CAPES) and Brazilian National Council for Scientific and Technological Development (CNPq), for financial support (Finance code 001).

Authors' Contributions

Conceptualization: Costa, J.A.; Azevedo, C.F.; Nascimento, M. **Data acquisition:** Resende, M.D.V. **Data analysis**: Costa, J.A.; Azevedo, C F.; Nascimento, M. **Design of methodology:** Costa, J.A.; Azevedo, C.F.; Nascimento, M.; Silva, F.F. **Software development:** Costa, J.A.; Azevedo, C.F. **Writing and editing: Costa**, J.A.; Azevedo, C. F.; Nascimento, M.; Resende, M.D.V.; Silva, F.F.; Nascimento, A.C.C.

References

Akinwale, M.G.; Gregorio, G., Nwilene, F.; Akinyele, B.O.; Ogunbayo, S.A.; Odiyi, A.C. 2011. Heritability and correlation coefficient analysis for yield and its components in rice (Oryza sativa L). African Journal of Plant Science 5**:** 207-212.

Ammiraju, J.S.S.; Luo, M.; Goicoechea, J.L.; Wang, W.; Kudrna, D.; Mueller, C.; Talag, J.; Kim, H.; Sisneros, N.B.; Blackmon, B.; Fang, E.; Tomkins, J.B.; Brar, D.; Mackill, D.; Maccouch, S.; Kurata, N.; Lambert, G.; Galbraith, D.W.; Arumuganathan, K.; Rao, K.; Walling, J.G.; Gill, N.Y.U.Y.; Sanmiguel, P.; Soderlund, C.; Jackson, S.; Wing, R.A. 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 Research 16: 140-147.

Azevedo, C.F.; Nascimento, M.; Fontes, V.C.; Silva, F.F.; Resende, M.D.V.; Cruz, C.D. 2019. GenomicLand: software for genome-wide association studies and genomic prediction. Acta Scientiarum. Agronomy 41: e45361.

Azevedo, C.F.; Resende, M.D.V.; Nascimento, M.; Viana, J.M.S.; Valente, M.S.F. Population structure correction for genomic selection through eigenvector covariates. 2017. Crop Breeding and Applied Biotechnology 17: 350-358.

Azevedo, C.F.; Resende, M.D.V.; Silva, F.F.; Lopes, O.S.; Guimarães, S.E.F. 2013. Independent component regression applied to genomic selection for carcass traits in pigs. Pesquisa Agropecuária Brasileira 48: 619-626.

Azevedo, C.F.; Resende, M.D.V.; Silva, F.F.; Viana, J.M.S.; Valente, M.S.F.; Resende Junior, M.F.R.; Muñoz, P. 2015. Ridge, LASSO and bayesian additive-dominance genomic models. BMC Genetics 16: 1-13.

Azevedo, C.F.; Silva, F.F.; Resende, M.D.V.; Lopes, M.S.; Duijvesteijn, N.; Guimarães, S.E.F.; Lopes, P.S.; Kelly, M.J.; Viana, J.M.S.; Knol, E.F. 2014. Supervised independent component analysis as an alternative method for genomic selection in pigs. Journal of Animal Breeding and Genetics 131: 452-461.

Bingham, E.; Hyvärinen, A. 2000. A fast fixed-point algorithm for independent component analysis of complex valued signals. International Journal of Neural Systems 10: 1-8.

Bisne, R.; Sarawgi, A.K.; Verulkar, S.B. 2009. Study of heritability, genetic advance and variability for yield contributing characters in rice. Bangladesh Journal of Agricultural Research 34: 175-179.

Brito, L.F.; McEwan, J.C.; Miller, S.P.; Pickering, N.K.; Bain, W.E.; Dodds, K.G.; Schenkel, F.S.; Clarke, S.M. 2017. Genetic diversity of a New Zealand multi-breed sheep population and composite breeds' history revealed by a high-density SNP chip. BMC Genetics 18: 1-11.

Cadavid, A.C.; Lawrence, J.K.; Ruzmaikin, A. 2008. Principal components and independent component analysis of solar and space data. Solar Physics 248: 247-261.

Desta, Z.A.; Ortiz, R. 2014. Genomic selection: genome-wide prediction in plant improvement. Trends in Plant Science 19: 592-601.

Ferreira, D.F. 2012. Multivariate Statistics = Estatística Multivariada. Editora UFLA, Lavras, MG, Brazil (in Portuguese).

Grenier, C.; Cao, T.V.; Ospina, Y.; Quintero, C.; Châtel, M.H.; Tohme, J.; Courtois, B.; Ahmadi, N. 2015. Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding. PloS One 10: e0136594.

Hassen, M.B.; Cao, T.V.; Bartholomé, J.; Orasen, G.; Colombi, C.; Rakotomalala, J.; Bertone, C.; Biselli, C.; Volante, A.; Desiderio, F.; Jacquin, L.; Valè, G.; Ahmadi, N. 2018. Rice diversity panel provides accurate genomic predictions for complex traits in the progenies of biparental crosses involving members of the panel. Theoretical and Applied Genetics 131: 417-435.

Helwig, N.E.; Hong, S.A. 2013. Critique of tensor probabilistic independent component analysis: implications and recommendations for multi-subject fMRI data analysis. Journal of Neuroscience Methods 213: 263-273.

Hyvärinen, A. 1998. New approximations of differential entropy for independent component analysis and projection pursuit. Advances in Neural Information Processing Systems 10: 273-279.

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. 2013. An Introduction to Statistical Learning. Springer, New York, NY, USA.

Le Floch, É.; Guillemot, V, Frouin.; V, Pinel, P.; Lalanne, C.; Trinchera, L.; Tenenhaus, A.; Moreno, A.; Zilbovicius, M.; Bourgeron, T.; Dehaene, S.; Thirion, B.; Poline, J.B.; Duchesnay, É. 2012. Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. Neuroimage 63: 11-24.

Meuwissen, T.H.E; Hayes, B.J.; Goddard, M.E. 2001. Prediction of total genetic value using genome wide dense marker maps. Genetics 157: 1819-1829.

Ogunbayo, S.A.; Ojo, D.K.; Sanni, K.A.; Akinwale, M.G.; Toulou, B.; Shittu A.; Idehen, E.O.; Popoola, A.R.; Daniel, I.O.; Gregorio, G.B. 2014. Genetic variation and heritability of yield and related traits in promising rice genotypes (Oryza sativa L.). Journal of Plant Breeding and Crop Science 6: 153-159.

Resende, M.D.V.; Silva, F.F.; Azevedo, C.F. 2014. Mathematical, Biometric and Computational Statistics: Mixed, Multivariate, Categorical and Generalized Models (REML / BLUP), Bayesian Inference, Random Regression, Genomic Selection, QTL-GWAS, Spatial and Temporal Statistics, Competition, Survival = Estatística Matemática, Biométrica e Computacional: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência. Editora Suprema, Visconde do Rio Branco, MG, Brazil (in Portuguese).

Resende, M.D.V.; Silva, F.F.; Lopes, P.S.; Azevedo, C.F. 2012. Genomic Wide Selection (GWS) by Mixed Models (REML/BLUP), Bayesian Inference (MCMC), Multivariate Random Regression (RRM) and Spatial Statistics = Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada e Estatística Espacial. Editora UFV, Viçosa, MG, Brazil (in Portuguese).

Seyoum, M.; Alamerew, S.; Bantte, K. 2012. Genetic variability, heritability, correlation coefficient and path analysis for yield and yield related traits in upland rice (Oryza sativa L.). Journal of Plant Sciences 7: 13-22.

Spindel, J.E.; Begum, H.; Akdemir, D.; Collard, B.; Redoña, E.; Jannink, J.L.; McCouch, S. 2016. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity 116: 395-408.

Spindel, J.E.; Begum, H.; Akdemir, D.; Virk, P.; Collard, B.; Redoña, E.; Atlin, G.; Jannink, J.L.; Mccouch, S.R. 2015. Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLOS Genetics 11: e1004982.

Wilkinson, S.; Bishop, S.C.; Allen, A.R.; Mcbride, S.H.; Skuce, R.A.; Bermingham, M.; Woolliams, J.A.; Glass, E.J. 2017. Fine-mapping host genetic variation underlying outcomes to Mycobacterium bovis infection in dairy cows. BMC Genomics 18: 1-13.

Zhao, K.; Tung, C.W.; Eizenga, G.C.; Wright, M.H.; Ali, M.L.; Price, A.H.; Norton, J.G.; Islam, A.R.; Reynolds, A.; Mezey, J.; Mcclung, A.M.; Bustamante, C.D.; McClung, A.M. 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nature Communications 2: 1-10.