Comparison of dimensionality reduction methods to predict genomic breeding values for carcass traits in pigs

C.F. Azevedo[1], M. Nascimento[1], F.F. Silva[2], M.D.V. Resende[3], P.S. Lopes[2], S.E.F. Guimarães[2] and L.S. Glória[2]

[1]Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, MG, Brasil
[2]Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, MG, Brasil
[3]EMBRAPA Florestas/Departamento de Engenharia Florestal, Universidade Federal de Viçosa, Viçosa, MG, Brasil

Corresponding author: C.F. Azevedo
E-mail: camila.azevedo@ufv.br

Genet. Mol. Res. 14 (4): 12217-12227 (2015)
Received March 4, 2015
Accepted June 15, 2015
Published October 9, 2015
DOI http://dx.doi.org/10.4238/2015.October.9.10

ABSTRACT. A significant contribution of molecular genetics is the direct use of DNA information to identify genetically superior individuals. With this approach, genome-wide selection (GWS) can be used for this purpose. GWS consists of analyzing a large number of single nucleotide polymorphism markers widely distributed in the genome; however, because the number of markers is much larger than the number of genotyped individuals, and such markers are highly correlated, special statistical methods are widely required. Among these methods, independent component regression, principal component regression, partial least squares, and partial principal components stand out. Thus, the aim of this study was to propose an application of the methods of dimensionality reduction to GWS of carcass traits in an F_2 (Piau x commercial line) pig population. The results show similarities

between the principal and the independent component methods and provided the most accurate genomic breeding estimates for most carcass traits in pigs.

Key words: Partial least squares; Independent component regression; Principal component regression; Partial principal component

INTRODUCTION

Given the abundance of polymorphisms in certain molecular genetic markers, Meuwissen et al. (2001) idealized the use of genome-wide selection (GWS), which consists of incorporating genomic information directly into the predictions of individual genetic merit for traits of economic interest. However, the direct use of these markers in the selection of genetically superior individuals is still a challenge due to problems of dimensionality and multicollinearity arising from the large number of such markers in relation to the quantity of individuals genotyped. As a solution to such problems, Gianola et al. (2003) recommended the use of statistical methods that integrate both the selection of covariates and the regularization of the estimation process.

The most commonly used methods for GWS are the penalized regression under the frequentist (RR-BLUP and G-BLUP) and Bayesian (Bayes A and B and Bayesian Lasso) approaches. However, dimensionality reduction methods based on regression, such as uni- and multivariate partial least squares (UPLS and MPLS) and independent and principal component regression (ICR and PCR), are also of great applicability (Moser et al., 2009; Pintus et al., 2012; Azevedo et al., 2013a,b). Another methodology not yet used to compute genomic predictions of genetic merit and single nucleotide polymorphism (SNP) marker effects is partial principal components (PPC), and the theoretical detailed description of this method is described by Ferreira (2008).

The main difference between PLS and PCR is that the extracted components of PCR explain the variance of covariates (X) and the extracted components of PLS have higher covariance for the response variables (Y). The ICR method is similar to PCR but ICR uses single value decomposition and PCR uses spectral decomposition. While the proposed method, PPC, combines the purposes of PCR and PLS, it maximizes the variance of the covariates and the covariance with the response variables.

Several studies using these methodologies as a basis for breeding can be found in the literature. The study by Pintus et al. (2012) used principal component analysis to reduce the number of predictors for calculating genomic breeding values (GEBV) for dairy traits in Italian Brown and Simmental bulls. Colombani et al. (2012) used PLS and sparse PLS for predicting GEBV of French Holstein bulls. Recently, Azevedo et al. (2013b) proposed an ICR for the estimation of genomic values and of SNP marker effects for carcass traits in an F_2 pig population (Piau x commercial line). Although the literature present several comparisons between these methodologies (Moser et al., 2009; Colombani et al., 2012; Azevedo et al., 2013a,b), there are no comparisons that include independent, partial, and principal component methodologies in the same study.

Carcass characteristics are very important in the pig industry, especially those related to a higher yield of meat and the smallest fat deposition, to meet the growing and demanding consumer market (Rosa et al., 2008; Zangeronimo et al., 2009). A larger quantity of carcass

meat of pigs has been the goal not only in the industry, because it improves profitability and decreases production costs. Thus, studies involving genomic selection for carcass traits in pigs are extremely important for improving the accuracy in estimating the genetic merit of individuals by considering genomic information.

In this study, we compared the accuracies of different dimensionality reduction methods for the computation of genomic prediction of genetic merit and of SNP marker effects for carcass traits in an F_2 pig population (Piau x commercial line).

MATERIAL AND METHODS

The F_2 pig population was generated by crossing two native Brazilian Piau boars with 18 commercial sows (Landrace x Large White x Pietran) selected for growth rate and backfat thickness. The F_1 generation consisted of 106 sows and 134 boars (Band et al., 2005). Twelve boars from different litters were randomly selected from the 134 F_1 boars and mated by natural breeding with 54 F_1 sows to produce the F_2 generation. The F_2 generation consisted of approximately 840 offspring divided into five batches according to the season in which they were born.

After slaughter, around 65 kg live weight (64.71 ± 0.24), the following carcass traits were evaluated in the animals of the F_2 generation: bacon depth (BCD), midline lower backfat thickness (L); midline backfat thickness after the last rib (LR); midline backfat thickness on the last lumbar vertebrae (LL), and backfat thickness after the last rib, 6.5 cm from the midline (ETO).

Details of the DNA extraction procedures used are described in Faria et al. (2006). Six primer pairs for microsatellite markers distributed on SSC7 (S0025, S0064, S0102, SW252, SW632, and S0212) were used. Amplifications were conducted in an MJ Research PTC 100-96® thermocycler, according to standard laboratory procedures (Faria et al., 2006). The amplified fragments were scored automatically by the GenScan software installed in an ABI PRISM 310 sequencer (Applied Biosystems). Annotation and genotype checking were conducted manually by two independent and previously trained technicians. The CRIMAP software (Lander and Green, 1987) was used to construct linkage maps of the related markers (S0025, S0064, S0102, SW252, SW632, and S0212), which were distributed at positions 0, 31, 65, 96, 108, and 136 cM, respectively.

All dimensionality reduction methods, besides enabling regularization in the estimation process, guarantee the removal of multicollinearity present in the data, once the correlation between any pair of components (linear combinations of SNPs) equals zero. In the dimensionality reduction methods, the X matrix is defined as the matrix of SNP markers x (values 0, 1, and 2 for the number of alleles of the SNP) and y is a vector of phenotypic variation corrected for fixed effects and deregressed.

PCR reduces dimensionality without resulting in significant loss of information present in the data (Otto, 1999). In this method, the components Z_v , $v = 1,...,n$, are linear combinations of the explanatory variables $X_1,X_2,...,X_j$. Thus, the following equation holds:

$$Z = X\hat{P} \qquad \text{(Equation 1)}$$

where $\hat{P}$ is the matrix of n_{pcr} first eigenvectors of the covariance matrices of the X and Z components. Aiming to establish the relationship between y and Z_v, multiple linear regression is used, obtaining the following prediction equation:

$$\hat{\mathbf{y}} = \hat{\alpha}_0 + \hat{\alpha}_1\hat{\mathbf{z}}_1 + \hat{\alpha}_2\hat{\mathbf{z}}_2 + ... + \hat{\alpha}_{n_{pcr}}\hat{\mathbf{z}}_{n_{pcr}} \quad \text{(Equation 2)}$$

where $\hat{\alpha}_v$ is the estimated regression coefficient, which has no biological interpretation. However, it is possible to estimate the coefficients associated to the original variables (SNPs) combining (1) and (2), using the following expression:

$$\hat{m}_{pcr} = \hat{P}\hat{\alpha}. \quad \text{(Equation 3)}$$

PLS is considered an appropriate method for data containing more covariates than observations (Hoskuldsson, 1988), as in GWS. This methodology also allows a multivariate approach considering multiple-dependent variables.

MPLS obtains estimators for the dependent variables (traits) using the component T_i ($i = 1,...,p$). Under this approach, the statistical model is expressed as follows:

$$Y_k = \beta_{k0} + \beta_{k1}T_1 + \beta_{k2}T_2 + ... + \beta_{kp}T_p + e \quad \text{(Equation 4)}$$

where Y_k is the dependent variable k ($k = 1,...,n$), β_{ki} is the regression coefficient, and e is the residual.

The estimated component T_i is dependent on two latent variables, V_{ij} and R_{ik}. Thus, it assumes that the components T_i ($i \geq 1$), the variables V_{ij} ($j = 1,...,m$), and R_{ik} ($k = 1,...,n$) have been determined. By definition, $V_{(i+1)j}$ is the residual from the regression between T_i and V_{ij}, and $R_{(i+1)k}$ is the residual from the regression between R_{ik} and T_i, respectively expressed by:

$$V_{(i+1)j} = V_{ij} - \{t'_i\, v_{ij}/(t'_i t_i)\}T_i \quad \text{(Equation 5)}$$

$$R_{(i+1)k} = R_{ik} - \{t'_i\, r_{ik}/(t'_i t_i)\}T_i \quad \text{(Equation 6)}$$

where $\mathbf{t}_i$ is the column vector of values of the i^{th} component, $\mathbf{v}_{ij}$ is the vector of the values of V_{ij}, $\mathbf{r}_{ik}$ is the vector of the values of R_{ik}, and $\mathbf{t}'_i\,\mathbf{v}_{ij}/(\mathbf{t}'_i\,\mathbf{t}_i)$ and $\mathbf{t}'_i\,\mathbf{r}_{ik}/(\mathbf{t}'_i\,\mathbf{t}_i)$ are regression coefficients. If $i = 1$, V_{ij} is the centered variable of X ($V_{1j} = X_j - \overline{X}_j$, to $j = 1,...,m$) and R_{1k} is the centered variable of Y_k ($V_{1k} = Y_k - \overline{Y}_k$, to $k = 1,...,n$). The process is successively repeated to obtain the matrices $R_{(i+1)} = (\mathbf{r}_{(i+1)1},...,\mathbf{r}_{(i+1)n})$ and $V_{(i+1)} = (\mathbf{v}_{(i+1)1},...,\mathbf{v}_{(i+1)m})$. Thus, using the matrices $R_{(i+1)}$ and $V_{(i+1)}$, a new vector, $\mathbf{u}_{i+1}$, is determined by multiplying $R_{i+1}\mathbf{c}_{i+1}$, where $\mathbf{c}_{i+1}$ is the corresponding eigenvector of the eigenvalue of $R'_{i+1}V_{i+1}V'_{i+1}R_{i+1}$ (Hoskuldsson, 1988). Garthwaite (1994) generally defines the component $T_{(i+1)}$ as a weighted average given by:

$$T_{(i+1)} = \sum_{j=1}^{m} w_{(i+1)j}\hat{b}_{(i+1)j}V_{(i+1)j}. \quad \text{(Equation 7)}$$

where $W_{(i+1)j}$ is a weight defined by $w_{(i+1)j} \propto \text{var}(V_{(i+1)j}) = (\mathbf{v}'_{(i+1)j}\mathbf{v}_{(i+1)j})/(n-1)$ and $\hat{b}_{(i+1)j}$ is

the estimated regression coefficient of $U_{(i+1)}$ in relation to $V_{(i+1)j}$, given by $\hat{b}_{(i+1)j} = \mathbf{v}'_{(i+1)j}\mathbf{u}_{(i+1)} / (\mathbf{v}'_{(i+1)j}\mathbf{v}_{(i+1)j})$.

The method is repeated to obtain $T_{(i+2)}, T_{(i+3)}, ..., T_p$ ($p \leq \min(m, q)$, where q is the number of observations). After obtaining the p components, the regression coefficients of model (4) are determined by the ordinary least squares method (OLS), obtaining the following prediction equation:

$$\hat{Y}_k = \hat{\beta}_{k0} + \hat{\beta}_{k1}\hat{T}_1 + \hat{\beta}_{k2}\hat{T}_2 + ... + \hat{\beta}_{kp}\hat{T}_p . \quad \text{(Equation 8)}$$

Analogous to the PCR method, it is possible to find the coefficients for each trait k, associated to the original variables, which in this context are the markers. Thus, it is necessary to combine Equations 7 and 8 to obtain the following equation:

$$\hat{m}_{pls} = \hat{B}W'\hat{\beta} \quad \text{(Equation 9)}$$

where W is the weight matrix, $\hat{B}$ is matrix whose elements are the coefficients from the regression between U_i and v_{ij}, and $\hat{\boldsymbol{\beta}}$ is the vector of coefficients $\hat{\boldsymbol{\beta}}_{ki}$ (i=1,...,p).

The main difference between MPLS and UPLS is in the construction of the vector u_{i+1}. In the UPLS method, u_{i+1} is the vector of residuals from the regression between Y (single-dependent variable) and the components T_i ($i = 1, ..., p$). While the MPLS method applies a regression of each variable Y_k (several dependent variables) and the components T_i ($i = 1, ..., p$), and u_{i+1} is a linear combination between residual vectors.

PPC is considered an appropriate method for data containing many dependent variables of interest (Ferreira, 2008). The PPC method defines the principal components from the maximization of the residual variance matrix given by the difference between the total variation and the variation explained by the covariates.

The proposed method allows both key components for the dependent variable [$Y = (Y_1, Y_2, ..., Y_m)$] and the covariates [$X = (X_1, X_2, ..., X_n)$] to be obtained. So, it is necessary to obtain the joint covariance matrix of X and Y, defined by:

$$\Sigma = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix} \quad \text{(Equation 10)}$$

where Σ_{YY}, Σ_{XX}, and $\Sigma_{XY}=(\Sigma_{YX})'$ are covariance matrices of Y, X and between X and Y, respectively.

From the regression between X and Y, one can obtain the residual covariance matrix given by:

$$\Sigma_{Y.X} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$
$$\Sigma_{X.Y} = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} \quad \text{(Equation 11)}$$

where Σ_{YY} and Σ_{XX} represent the unconditional covariance matrices of X and Y, respectively, and consequently $\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$, $\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$, $\Sigma_{Y.X}$, $\Sigma_{Y.X}$, and $\Sigma_{X.Y}$ represent the dispersion of Y explained by X, the dispersion of X explained by Y, the covariance of Y|x, and the covariance of X| y, respectively.

Thus, we must maximize the variance $e_i \Sigma e_i$, subject to restriction $e_i e_i = 1$ to obtain principal components of the variables X and Y (Ferreira, 2008). Therefore, using the theorem for maximization of quadratic forms, we obtained:

$$(\Sigma_{Y.X} - \lambda_{1i} I) e_{1i} = 0$$
$$(\Sigma_{X.Y} - \lambda_{2i} I) e_{2i} = 0 . \quad \text{(Equation 12)}$$

Therefore, the partial principal components of X and Y are defined, respectively, by:

$$r = YP_Y$$
$$N = XP_X \quad \text{(Equation 13)}$$

where P_Y is the first eigenvector of $Ó_{Y.X}$ and P_X is the matrix of the n_{ppc} first eigenvectors of $Ó_{X.Y}$. Aiming to establish the relationship between the partial principal components of X(N) and Y(r), multiple-linear regression is used, obtaining the following equation:

$$r = N\hat{\varphi} \quad \text{(Equation 14)}$$

where $\hat{o}$ is the vector of the estimate coefficients from the regression between m and N. Similarly to other dimensionality reduction methods, marker effects can be obtained by combining Equations 13 and 14, resulting in the following estimates:

$$m_{ppc} = P_X \hat{\varphi} P_Y' (P_Y P_Y')^{-1} \quad \text{(Equation 15)}$$

ICR is the decomposition of the matrix X in linear combinations of completely independent components, in terms of both linear and non-linear relations. One advantage of ICR compared to PCR is the possibility of complete removal of any relationship of dependence between covariates. For this purpose, each independent component is built using the most representative SNPs chosen from a group of correlated SNPs.

Such analysis is also suitable for any distribution of the indicator variable in the matrix X, as long as it is a non-Gaussian distribution. Thus, ICR is well suited to GWS, since the matrix X of markers is parameterized with values 0, 1, and 2 (non-Gaussian distribution). In line with this, the decomposition is as follows:

$$X' = A'S' \quad \text{(Equation 16)}$$

where S is the matrix of independent components and A is the matrix of mixtures.

Special algorithms are used to try to find an orthogonal matrix R that maximizes the statistical independence of the columns of the S matrix using a quantitative measure of independence, which is a function of contrasts. The iterative algorithm developed by Hyvärinen (1998) is based on the maximum enthropy J(r) concept assuming that the variable r is standardized. According to this algorithm, the following approximation is obtained:

$$J(r) \propto [E\{G_i(r)\}-E\{G_i(v)\}]^2 \quad \text{(Equation 17)}$$

where, v is a standardized variable and $G_1(v) = -\exp(-v^2/2)$.

After the iterative process, the component matrix is obtained as follows:

$$S = XKR \quad \text{(Equation 18)}$$

where K is an orthogonalization matrix and KR is an approximation of A'. Thus, the equation of the predictions are obtained based on ICR and computed as:

$$\hat{\mathbf{y}}_k = \hat{\gamma}_{0k} + \hat{\gamma}_{1k}\hat{\mathbf{s}}_1 + \hat{\gamma}_{2k}\hat{\mathbf{s}}_2 + ... + \hat{\gamma}_{kn_{icr}}\hat{\mathbf{s}}_{n_{icr}} \quad \text{(Equation 19)}$$

where $\hat{\mathbf{y}}_k$ is a prediction vector of the k^{th} dependent variable and the $\hat{\gamma}_{\upsilon k}$ coefficients determined by the OLS method $\upsilon=1,...,n_{icr}$. Similarly to other dimensionality reduction methods, marker effects can be obtained by combining Equations 18 and 19, resulting in the following estimates:

$$m_{ppc} = KR\gamma \quad \text{(Equation 20)}$$

Dimensional reduction methods were compared using a cross-validation study (Resende et al., 2012), carried out separately for each trait. In this part of the analysis, the F_2 population of pigs was divided into three different populations, each one with 115 individuals.

Thus, for each analysis repetition, two of these populations were considered estimate (or trial) populations and used to obtain the effects of the SNP markers. The other population, denominated the validation population, was used to evaluate the agreement between predicted genetic values via estimates originating from the trial population and the corrected phenotypes observed. The process was repeated so that in each stage one of the three populations was the validating population.

Thus, the correlation between the estimated value in the three validations and the corrected and deregressed phenotypes constituted the predictive ability of the method. The method accuracy depends on this correlation and is equivalent to the ratio by the heritability square root:

$$r_{q,\hat{q}} = \frac{r_{y,\hat{y}}}{\sqrt{h^2_{sm}}} \quad \text{(Equation 21)}$$

where h^2_{sm} is the Mendelian segregation heritability computed as $h^2_{sm} = \frac{0.5h^2}{0.5h^2+(1-h^2)}$ and h^2 is the character heritability estimated by the REML (restricted maximum likelihood) method on phenotypes in a single-trait model (Resende et al., 2012).

Having the best method for each trait, the effects of the markers in absolute values were estimated and standardized considering the whole F_2 population of pigs using Equations 3, 9, 15, and 20, for methods PCR, PLS, PPC, and ICR, respectively. From this information, the Manhattan plot was constructed, where each point represents an SNP marker, the x-axis

shows location on the chromosome, and the y-axis shows the effect magnitude.

All computational routines were implemented in R (R Core Team, 2010) using the packages pls (PCR and PLS) and caret (ICR) and functions plsr (PLS), pcr (PCR), icr (ICR), which are freely accessible at http://www.det.ufv.br/~moyses/links.php.

RESULTS

The genetic correlations between pairs of carcass traits are presented in Table 1. Estimates of genetic correlations between carcass traits were high and positive, indicating an efficient selection through studies of multi-trais. The estimates of heritability for backfat thickness and bacon depth showed high values (Table 1), suggesting possible progress in the breeding of these traits.

Table 1. Estimates of heritability and genetic correlations between carcass traits in an F_2 (Piau x commercial line) pig population.

Traits	LR	LL	L	ETO	BCD
LR	**0.35**	0.64	0.60	0.63	0.58
LL	-	**0.36**	0.88	0.60	0.46
L	-	-	**0.33**	0.66	0.55
ETO	-	-	-	**0.42**	0.80
BCD	-	-	-	-	**0.34**

Heritability estimates are presented on the diagonal; estimates of genetic correlations are presented above the diagonal. Midline backfat thickness after the last rib (LR); midline backfat thickness on the last lumbar vertebrae (LL); midline lower backfat thickness (L); backfat thickness after the last rib, 6.5 cm from the midline (ETO); bacon depth (BCD).

According to Figure 1, it is evident that the curve of the predictive ability of each method relative to the number of components has not reached a plateau. Thus, the number of components for each method, considering each trait, was reported that best predictor, i.e., the number of components corresponds to the peak of the curve shown in Figure 1.

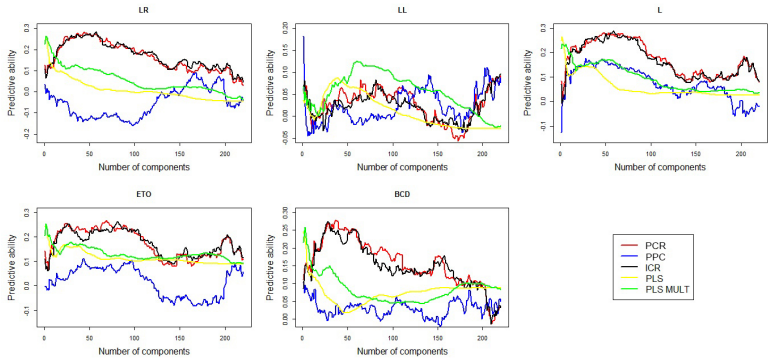

Figure 1. Curve of the predictive capacity of dimensionality reduction methods relative to the number of components for each trait from an F_2 (Piau x commercial line) pig population. Midline backfat thickness after the last rib (LR); midline backfat thickness on the last lumbar vertebrae (LL); midline lower backfat thickness (L); backfat thickness after the last rib, 6.5 cm from the midline (ETO); bacon depth (BCD); principal component regression (PCR); partial principal components (PPC); independent component regression (ICR); univariate partial least squares (PLS); multivariate partial least squares (PLS MULT).

The accuracy of the dimensional reduction methods for each carcass trait is presented in Table 2.

Table 2. Accuracy and bias of the dimensionality reduction methods for each trait in an F_2 (Piau x commercial line) pig population.

	Method	LR	LL	L	ETO	BCD
Predictive ability	PCR	0.63	0.63	0.23	0.62	0.75
	PPC	0.20	0.40	0.45	0.25	0.27
	ICR	0.62	0.65	0.25	0.60	0.75
	PLS	0.56	0.58	0.23	0.57	0.67
	MULT PLS	0.58	0.54	0.30	0.57	0.69
Bias	PCR	0.59	0.53	0.02	0.49	0.65
	PPC	0.63	1.17	>10	>10	>10
	ICR	0.59	0.54	0.03	0.45	0.82
	PLS	0.39	0.47	0.04	0.55	0.58
	MULT PLS	0.57	0.50	0.06	0.54	0.47

Midline backfat thickness after the last rib (LR); midline backfat thickness on the last lumbar vertebrae (LL); midline lower backfat thickness (L); backfat thickness after the last rib, 6.5 cm from the midline (ETO); bacon depth (BCD); principal component regression (PCR); partial principal components (PPC); independent component regression (ICR); univariate partial least squares (PLS); multivariate partial least squares (PLS MULT).

DISCUSSION

The heritability estimates found for different measurements of backfat thickness ranged from 0.33 to 0.42, values similar to those reported by Mendonça et al. (2012), also using pigs from a commercial strain x Piau F_2 population. Costa et al. (2001) observed estimates of 0.34, 0.43, and 0.50 for the breeds Duroc, Large White, and Landrace, respectively. Research by Torres Jr. et al. (1998) demonstrated estimates of 0.37 and 0.51 for Large White and Landrace, respectively, and Barbosa et al. (2008a,b) observed an estimate of 0.44 for Large White. In the current study, the heritability estimate for the thickness of the bacon was 0.24, lower than that found by Mendonça et al. (2012).

The values of predictive ability obtained by ICR and PCR were similar considering the different number of components (Figure 1). These results suggest that such methods have similar, but not the same, statistical concepts. Specifically, the property of independence (ICR assumption) implies in the removal of non-linear and linear dependence between variables, while the PCR analysis guarantees only the removal linear dependence. Moreover, for all characteristics, except lumbar vertebrae, these methodologies presented higher peak values. The other methodologies did not present similar behavior when a different number of components were considered. The partial principal component methods showed lower results when compared to the other methods evaluated.

The peaks of the curves of predictive ability were 44 and 57 (PCR and ICR), 1 (PPC), 53 and 59 (PCR and ICR), 69 and 81 (PCR and ICR), 37 and 28 (PCR and ICR) for the components of the LR, LL, L, ETO, and BCD traits, respectively, providing a reduction of 81.4 and 75.9, 99.6, 77.6, and 75.1%, 70.9 and 65.8%, and 84.4 and 88.2% in the total number of original variables (237 SNPs), respectively.

Considering these results, which corroborate those obtained previously, the PCR and ICR methods were more efficient in predicting GEBV, with similar results for LR, LL, ETO, and BCD, and with an accuracy ranging from 0.60 to 0.75. The PLS method under the multivariate and univariate approach showed similar results, with accuracy values between 0.54

and 0.69. In contrast, the PPC method displayed the lowest predictive ability of values for these traits (0.20, 0.40, 0.25, and 0.27 for LR, LL, ETO, and BCD, respectively, and for the L, the PPC method had an accuracy (0.45) superior to other methods.

The use of dimensional reduction methods has only been reported by Azevedo et al. (2013a,b) and Azevedo et al. (2014) for genomic selection in carcass traits in the same population of pigs, but using a different criterion for choosing the number of components. Azevedo et al. (2013a) studied the performance of multivariate and univariate approaches of the PLS method, but the results disagreed with the values obtained in our study, where MPLS outperformed UPLS considerably. Azevedo et al. (2013b) compared the PLS, PCR, and ICR methods and reported results of lower accuracy, from 0.01 to 0.51, compared to those found in this study. Azevedo et al. (2014) compared the same methods of dimensional reduction beyond their supervised approaches (selection of covariates), showing different results to those reported in this study, where PLS had low performance compared to PCR and ICR.

Furthermore, dimensional reduction methods have been applied to other species and such results can be used as a reference. Moser et al. (2009) performed a study comparing five methods for dairy cattle data, including PLS, all of which displayed similar accuracies. This finding differs from the results obtained in the present study. In contrast, Solberg et al. (2009) performed a study comparing PLS and PCR and observed similar accuracy ability values (0.47 and 0.45, respectively), which agree with the values obtained in our study, where PLS has the same performance as PCR.

The regression coefficients between observed and predicted phenotypes or bias are presented in Table 2. The only method that obtained an estimate of the regression coefficient close to unity for a certain trait was PPC, indicating that the genetic evaluations are not biased and are effective in predicting the actual magnitudes of differences between individuals in the evaluation (Resende et al., 2010). However, the PPC method was the only method that showed bias values well above unity, indicating that the GEBVs were underpredicted, while the estimates for other biases were lower than unity, indicating that the GEBVs were overpredicted.

The PCR and ICR methods only differed significantly in estimates of bias for the ETO and BCD traits. However, the advantage of ICR for GWS compared to PCR is the fact that it considers complete independence of the components, guaranteeing the absence of both linear and nonlinear relationships between latent variables (Azevedo et al., 2013, 2014). Similar results observed between the PCR and the ICR methods are possibly due to the structure of association of the variables evaluated in this paper, which have linear behavior.

In conclusion, the similar methods of PCR and ICR presented the highest predictive ability values and were the most efficient for the prediction of phenotypic values. The proposed PPC method presented the highest predictive ability value for just a single carcass trait, but proved biased for most of the remaining traits. In contrast, MPLS and UPLS presented values with similar predictive ability, differing only in the results of bias.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by CAPES and CNPq.

REFERENCES

Azevedo CF, Silva FF, Resende MDV, Peternelli LA, et al. (2013a). Uni and multivariate partial least squares applied to genomic selection for carcass traits in pigs. *Ciênc. Rural* 43: 1642-1649.

Azevedo CF, Resende MDV, Fonseca F, Lopes PS, et al. (2013b). Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. *Pesqui. Agropecu. Bras*. 48: 619-626.

Azevedo CF, Silva FF, 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.

Band GDO, Guimarães SEF, Lopes PS, Peixoto JDO, et al. (2005). Relationship between the Porcine Stress Syndrome gene and carcass and performance traits in F2 pigs resulting from divergent crosses. *Genet. Mol. Biol*. 28: 92-96.

Barbosa L, Lopes PS, Carneiro PCS, Regazzi AJ, et al. (2008a). Comparação entre modelos para estimação de parâmetros genéticos em características de desempenho em suínos da raça Large White. *Rev. Ceres* 55: 60-65.

Barbosa L, Lopes PS, Regazzi AJ, Torres RA, et al. (2008b). Estimação de parâmetros genéticos em suínos usando Amostrador de Gibbs. *R. Bras. Zootec*. 37: 1200-1206.

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: 2120-2131.

Costa ARC, Lopes PS, Torres RA, Regazzi AJ, et al. (2001). Estimação de parâmetros genéticos em características de desempenho de suínos das raças Large White, Landrace e Duroc. *R. Bras. Zootec.* 30: 49-55.

Faria DA, Guimarães SEF, Lopes PS, Pires AV, et al. (2006). Association between G316A growth hormone polymorphism and economic traits in pigs. *Genet. Mol. Biol*. 29: 634-640.

Ferreira DF (2008). Estatística multivariada. 1st edn. UFLA, Lavras.

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

Gianola D, Perez-Enciso M and Toro MA (2003). On marker-assisted prediction of genetic value: beyond the ridge. *Genetics* 163: 347-365.

Hoskuldsson P (1988). PLS Regression Methods. *J. Chemometr.* 2: 211-228.

Hyvärinen A (1998). New approximations of differential entropy for independent component analysis and projection pursuit. Proceedings of the 1997 conference on Advances in Neural Information Processing Systems 10, 273-279.

Lander ES and Green P (1987). Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. U.S.A*. 84: 2363-2367.

Mendonça PT, Lopes PS, Braccini Neto J, Carneiro PLS, et al. (2012). Estimação de parâmetros genéticos de uma população F2 de suínos. *Rev. Bras. Saúde Prod. Anim*. 13: 330-343.

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

Moser G, Tier B, Crump RE, Khatkar MS, et al. (2009). A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genet. Sel. Evol*. 41: 56.

Otto M (1999). Chemometrics: statistics and computer application in analytical chemistry. Wiley-VCH, Weinheim.

Pintus MA, Gaspa G, Nicolazzi EL, Vicario D, et al. (2012). Prediction of genomic breeding values for dairy traits in Italian Brown and Simmental bulls using a principal component approach. *J. Dairy Sci*. 95: 3390-3400.

R Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org/.

Resende MDV, Resende Junior MFR, Aguiar AM, Abad JIM, et al. (2010). Computação da seleção genômica ampla (GWS). Embrapa Florestas, Colombo.

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 e estatística espacial. 1st edn. Universidade Federal de Viçosa, Viçosa.

Rosa AF, Gomes JDF, dos Reis Martelli M, do Amaral Sobral PJ, et al. (2008). Características de carcaça de suínos de três linhagens genéticas em diferentes idades ao abate. *Ciênc. Rural* 38: 1718-1724.

Solberg TR, Sonesson AK, Woolliams JA and Meuwissen TH (2009). Reducing dimensionality for prediction of genome-wide breeding values. *Genet. Sel. Evol*. 41: 29.

Torres Jr RAA, Silva MA, Lopes PS, Regazzi AJ, et al. (1998). Estimativas de componentes de (co)variância para características produtivas de suínos Landrace e Large White pelo método da máxima verossimilhança restrita. *R. Bras. Zootec*. 27: 283-291.

Zangeronimo MG, Fialho ET, Lima JADF, Girao LVC, et al. (2009). Desempenho e características de carcaça de suínos dos 20 aos 50 kg recebendo rações com reduzido teor de proteína bruta e diferentes níveis de lisina digestível verdadeira. *Ciênc. Rural* 39: 1507-1513.