CAMILA FERREIRA AZEVEDO

MÉTODOS DE REDUÇÃO DE DIMENSIONALIDADE APLICADOS NA SELEÇÃO GENÔMICA PARA CARACTERÍSTICAS DE CARCAÇA EM SUÍNOS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2012

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T
A994m
2012

Azevedo, Camila Ferreira, 1988-
Métodos de redução de dimensionalidade aplicados na seleção genômica para características de carcaça em suínos / Camila Ferreira Azevedo. – Viçosa, MG, 2012.
50f. : il. ; 29cm.

Inclui apêndices.
Orientador: Fabyano Fonseca e Silva.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Genética - Métodos estatísticos . 2. Genômica.
3. Melhoramento genético. I. Universidade Federal de Viçosa.
II. Título.

CDD 22. ed. 567.50724

CAMILA FERREIRA AZEVEDO

MÉTODOS DE REDUÇÃO DE DIMENSIONALIDADE APLICADOS NA SELEÇÃO GENÔMICA PARA CARACTERÍSTICAS DE CARCAÇA EM SUÍNOS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 26 de julho de 2012.

Carlos Souza do Nascimento

Paulo Sávio Lopes

Marcos Deon Vilela de Resende
(Coorientador)

Fabyano Fonseca e Silva
(Orientador)

Ao meu pai, Sergio;
por ter dedicado a
vida a suas filhas.

AGRADECIMENTOS

A Deus por me amparar nos momentos difíceis, dar força interior para superar as dificuldades e indicar o caminho nas horas incertas.

A Universidade Federal de Viçosa e ao Programa de Pós–Graduação em Estatística Aplicada e Biometria, por proporcionar a realização de um curso de excelência.

Ao Departamento de Zootecnia da Universidade Federal de Viçosa, pela concessão dos dados utilizados na pesquisa.

Aos meus pais, Marilia e Sergio, pelo amor incondicional, pelos ensinamentos, pela dedicação e confiança.

À minha irmã, Lorena, pelas palavras de carinho e de motivação.

À minha família, por me apoiar em todos os momentos.

Ao Vitor, pela paciência, companheirismo, carinho e incentivo em todos os momentos.

Aos meus amigos, pela amizade, pelas palavras de conforto e por sempre estarem ao meu lado.

Ao professor e orientador Fabyano Fonseca e Silva, pelos ensinamentos, confiança, dedicação ao projeto, por contribuir para o meu crescimento profissional e por ser também um exemplo a ser seguido.

Ao Doutor e coorientador Marcos Deon Vilela de Resende, pelos saberes transmitidos, pela confiança e dedicação à pesquisa.

Ao coorientador e professor Luiz Alexandre Peternelli pela orientação da bolsa de estudos e pelos ensinamentos.

À coorientadora Simone Eliza Facioni Guimarães, pela disponibilidade e apoio.

Aos membros da banca examinadora, Doutor Carlos Souza do Nascimento, Doutor Marcos Deon Vilela de Resende e professor Paulo Sávio Lopes, pela disponibilidade e pelas sugestões para o enriquecimento deste trabalho.

Aos professores do Programa de Pós-Graduação em Estatística Aplicada e Biometria, por contribuírem para minha formação acadêmica.

Aos funcionários do Programa de Pós-Graduação em Estatística Aplicada e Biometria, pela prontidão.

Ao REUNI, pela concessão da bolsa de estudos.

Enfim, muito obrigada a todos aqueles que de certa forma contribuíram para o meu crescimento profissional e para a concretização deste trabalho.

BIOGRAFIA

CAMILA FERREIRA AZEVEDO, filha de Marilia Assis Ferreira Azevedo e de Sergio de Resende Azevedo, nasceu em Bom Jesus do Itabapoana, Rio de Janeiro, em 15 de abril de 1988.

Em maio de 2006, ingressou no curso de Licenciatura em Matemática na Universidade Federal de Viçosa, Viçosa-MG, graduando-se em julho de 2010.

Em agosto do mesmo ano, iniciou o curso de Mestrado no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa de dissertação em 26 de julho de 2012.

RESUMO

AZEVEDO, Camila Ferreira, M. Sc., Universidade Federal de Viçosa, julho de 2012. **Métodos de redução de dimensionalidade aplicados na seleção genômica para características de carcaça em suínos.** Orientador: Fabyano Fonseca e Silva. Coorientadores: Luiz Alexandre Peternelli, Marcos Deon Vilela de Resende e Simone Eliza Facioni Guimarães.

A principal contribuição da genética molecular no melhoramento animal é a utilização direta das informações de DNA no processo de identificação de animais geneticamente superiores. Sob esse enfoque, a seleção genômica ampla (*Genome Wide Selection* – GWS), a qual consiste na análise de um grande número de marcadores SNPs (*Single Nucleotide Polymorphisms*) amplamente distribuídos no genoma, foi idealizada. A utilização dessas informações é um desafio, uma vez que o número de marcadores é muito maior que o número de animais genotipados (alta dimensionalidade) e tais marcadores são altamente correlacionados (multicolinearidade). No entanto, o sucesso da seleção genômica ampla deve-se a escolha de metodologias que contemplem essas adversidades. Diante do exposto, o presente trabalho teve por objetivo propor a aplicação dos métodos de regressão via Componentes Independentes (*Independent Component Regression* – ICR), regressão via componentes principais (*Principal Component Regression* – PCR), regressão via Quadrados Mínimos Parciais (*Partial Least Squares* – PLSR) e RR-BLUP, considerando características de carcaça em uma população F_2 de suínos proveniente do cruzamento de dois varrões da raça naturalizada brasileira Piau com 18 fêmeas de linhagem comercial (Landrace × Large White × Pietrain), desenvolvida na Universidade Federal de Viçosa. Os objetivos específicos foram estimar Valores Genéticos Genômicos (*Genomic Breeding Values* – GBV) para cada indivíduo avaliado e estimar efeitos de marcadores SNPs, visando a comparação dos métodos.
Os resultados indicaram que o método ICR se mostrou mais eficiente, uma vez que este proporcionou maiores valores de acurácia na estimação do GBV para a maioria das características de carcaça.

ABSTRACT

AZEVEDO, Camila Ferreira, M. Sc., Universidade Federal de Viçosa, July, 2012. **Dimensionality reduction methods applied to genomic selection for carcass traits in pigs.** Adviser: Fabyano Fonseca e Silva. Co-advisers: Luiz Alexandre Peternelli, Marcos Deon Vilela de Resende and Simone Eliza Facioni Guimarães.

The main contribution of molecular genetics is the direct use of DNA information to identify genetically superior individuals. Under this approach, genome-wide selection (GWS) can be used with this purpose. GWS consists in analyzing of a large number of SNP markers widely distributed in the genome, and due to the fact that the number of markers is much larger than the number of genotyped individuals (high dimensionality) and also to the fact that such markers are highly correlated (multicollinearity). However, the use of methodologies that address the adversities is fundamental to the success of genome wide selection. In view of, the aim of this dissertation was to propose the application of Independent Component Regression (ICR), Principal Component Regression (PCR), Partial Least Squares (PLS) and Random Regression Best Linear Unbiased Predictor, whereas carcass traits in an F_2 population of pigs originated from the cross of two males from the naturalized Brazilian breed Piau with 18 females of a commercial line (Large White × Landrace × Pietrain), developed at the University Federal of Viçosa. The specific objectives were, to estimate Genomic Breeding Value (GBV) for each individual and estimate the effects of SNP markers in order to compare methods. The results showed that ICR method is more efficient, since provided most accurate genomic breeding values estimates for most carcass traits.

SUMÁRIO

INTRODUÇÃO GERAL

À partir do início do século XXI os avanços biotecnológicos na área de automação do processo de genotipagem permitiram o desenvolvimento de novas classes de marcadores moleculares (JENKINS & GIBSON, 2002), dentre os quais se destacam os SNPs (*Single Nucleotide Polymorphism*s). Diante da abundância destes marcadores, Meuwissen *et al.* (2001) idealizaram a seleção genômica ampla (*Genome Wide Selection* - GWS).

Devido a alta densidade dos marcadores SNPs no genoma, é possível assumir que alguns deles estejam em desequilíbrio de ligação com locos de características quantitativas (*Quantitative Trait Loci* – QTL), possibilitando sua utilização direta na estimação do valor genético de indivíduos sujeitos a seleção, inclusive de indivíduos que ainda não foram fenotipados. No âmbito da seleção genômica, uma vez que o número de marcadores é geralmente muito maior que o número de animais genotipados e tais marcadores são altamente correlacionados, métodos estatísticos baseados na redução de dimensionalidade são requeridos. Dentre estes destacam-se a Regressão via Componentes Principais, Quadrados Mínimos Parciais e Regressão via Componentes Independentes.

Diante do exposto, o presente trabalho tem por objetivo propor a aplicação e comparação entre os métodos de redução de dimensionalidade e o método RR-BLUP na seleção genômica para características de carcaça (espessuras de toucinho, bacon e rendimentos de carcaça) em uma população F_2 de suínos (Piau × Comercial) utilizando um painel de marcadores SNPs. Além de apresentar descrições metodológicas básicas a respeito dos métodos.

REVISÃO DE LITERATURA

1. Seleção Genômica Ampla

1.1. Definição e Importância

O principal atrativo da genética molecular, que beneficia o melhoramento animal, é a utilização direta das informações provenientes do DNA no processo de identificação de animais geneticamente superiores. O uso de marcadores moleculares

possibilita um aumento da eficiência seletiva e da rapidez na obtenção de ganhos genéticos desejáveis.

Os marcadores moleculares que mais se destacam são os SNPs (*Single Nucleotide Polymorphisms*) devido a sua baixa taxa de mutação, codominância e abundância. Meuwissen *et al.* (2001) idealizaram a seleção genômica ampla (*Genome Wide Selection* – GWS), a qual consiste na análise de um grande número de marcadores amplamente distribuídos no genoma, capturando os genes que afetam um caráter quantitativo.

Considerando que polimorfismos do DNA são a fonte de variação do mérito genético, marcadores SNPs em desequilíbrio de ligação com locos de características quantitativas (*Quantitative Trait Loci* – QTL) podem ser utilizados para predizer o valor genético genômico (*Genomic Breeding Value* – GBV) de cada indivíduo, apontar para diferenças genéticas não observáveis na matriz de parentesco médio (obtidas via *pedigree*) das equações de modelos mistos, além de ser um critério extra para identificação de indivíduos candidatos à seleção, o que aumentaria a acurácia na avaliação genética.

A aplicação prática destas informações é um desafio, pois geralmente não é possível a utilização adequada de métodos tradicionais baseados em quadrados mínimos (*Least Squares* – LS) para estimar o efeito de cada SNP no fenótipo, uma vez que geralmente o número de marcadores é muito maior que o número de animais genotipados.

Deste modo, um ponto chave para o sucesso da seleção genômica é a escolha adequada de metodologias a serem utilizadas e por esse motivo se faz necessário a comparação das mesmas.

1.2.Correção dos valores fenotípicos

Os valores fenotípicos utilizados nas análises de seleção genômica devem ser corrigidos para os efeitos dos genitores e desregressados, visando trabalhar basicamente com o mérito genético verdadeiro de indivíduos não aparentados (RESENDE *et al.*, 2010). O processo de correção dos fenótipos evitaria influências oriundas de genes de grande efeito presentes nos genitores. Além disso, esse dados não sofreriam duas regressões, uma baseada na matriz de parentesco e a outra na matriz de marcas, sendo a primeira menos precisa.

Em posse dos componentes de variância e dos valores genéticos de todos os indivíduos, devem ser empregados os procedimentos para a correção dos fenótipos, que são descritos por Garrick *et al.* (2009) e Resende *et al.* (2010) e detalhados a seguir.

A correção dos valores fenotípicos deve ser feita por meio de um sistema de equações,

$$\begin{bmatrix} Z'_{gm}Z_{gm}+4\lambda^* & -2\lambda^* \\ -2\lambda^* & Z'_iZ_i+2\lambda^* \end{bmatrix}\begin{bmatrix} \hat{g}_{gm} \\ \hat{g}_i \end{bmatrix}=\begin{bmatrix} y_{gm} \\ y_i \end{bmatrix},$$

em que:

$\hat{g}_i$ é o valor genético do indivíduo genotipado *i*, com $i = 1, 2, ..., I$;

$\lambda^*=(1-h^2)/h^2$, sendo h^2 a herdabilidade de cada característica;

$\hat{g}_{gm}=(\hat{g}_h+\hat{g}_k)/2$ é o valor genético médio dos genitores *h* e *k*;

$Z'_{gm}Z_{gm}$ é o conteúdo de informação associado à média dos genitores;

Z'_iZ_i é conteúdo de informação associado ao indivíduo incluindo seus descendentes;

y_{gm} e y_i dados fenotípicos corrigidos para os efeitos fixos associados à média dos genitores e ao indivíduo *i*, respectivamente.

A matriz de informação associada à média de genitores é uma quantidade desconhecida e pode ser obtida por meio da seguinte expressão:

$$Z'_{gm}Z_{gm}=\lambda^*(0,5\alpha-4)+0,5\lambda^*(\alpha^2+16/\delta)^{1/2},$$

em que:

$\alpha=1/(0,5-r^2_{gm})$;

$\delta=(0,5-r^2_{gm})/(1-r^2_i)$;

r^2_{gm} é a confiabilidade associada ao valor genético médio predito dos genitores *h* e *k* dada por $r^2_{gm}=(r^2_{gh}+r^2_{gk})/2$ e ainda r^2_i é a confiabilidade associada ao valor genético predito do indivíduo.

Por conseguinte, a matriz de informação associada ao indivíduo Z'_iZ_i pode ser obtida através de $Z'_{gm}Z_{gm}$. Sendo assim Z'_iZ_i é dada por $\delta Z'_{gm}Z_{gm}+2\lambda^*(2\delta-1)$.

O vetor de fenótipos corrigidos para os efeitos fixos associados aos indivíduos é resultante dos sistemas de equações, ou seja, $\hat{y}_i = (-2\lambda^*)\hat{g}_{gm} + (Z'_i Z_i + 2\lambda^*)\hat{g}_i$.

Desta forma, o fenótipo será corrigido também para o valor genético médio de seus genitores. Após essa correção pode-se obter o valor genético desrregressado ($\hat{g}^*_i$), dado por $\hat{g}^*_i = \hat{y}_i / (Z'_i Z_i)$.

1.3. Método RR-BLUP

O método RR-BLUP (*Random Regression Best Linear Unbiased Predictor*) é tradicionalmente aplicado à seleção genômica. Esse método faz uso de preditores do tipo BLUP e assume que os efeitos dos marcadores SNPs são covariáveis de efeitos aleatórios. O nome mais apropriado deste método é Regressão Aleatória (*Random Regression*) do tipo BLUP aplicado à seleção genômica ampla (RR-BLUP/GWS), sendo um caso particular da regressão de cumeeira (RESENDE *et al*., 2010).

Os estimadores desse método são do tipo *shrinkage* que são funções penalizadas por uma função da quantidade λ. Assim, quando λ é desconhecido tem-se a regressão de cumeeira. Por outro lado, se o parâmetro de penalização está associado a $\lambda = \sigma^2_e / (\sigma^2_g / n_Q)$, tem-se a regressão aleatória BLUP, sendo σ^2_e a variância residual, σ^2_g a variância genética aditiva da característica, $n_Q = \sum_{j=1}^{J} 2p_j(1-p_j)$ (função do número J de marcas) e p_j a frequência alélica do marcador *j*.

A predição por meio do método RR-BLUP segue o modelo misto a seguir (RESENDE *et al.*, 2007, 2008):

$$y = Wb + Xm_{rr\text{-}blup} + e,$$

em que:

y é o vetor de dados fenotípicos com dimensão *I* × *1* sendo I o número de indivíduos;

b é o vetor de efeitos fixos, com matriz de incidência W;

$m_{rr\text{-}blup}$ é o vetor de efeitos aleatórios dos marcadores, com dimensão *J* × *1*, sendo J o número de marcas, associado a matriz de incidência X com valores 0, 1 e 2 para o número de alelos do marcador (ou suposto QTL).

As equações de modelos mistos genômicas visam predizer os efeitos dos marcadores e são dadas por:

$$\begin{bmatrix} W'W & W'X \\ X'W & X'X+I\dfrac{\sigma_e^2}{(\sigma_g^2/n_Q)} \end{bmatrix}\begin{bmatrix} \hat{b} \\ \hat{m}_{rr\text{-}blup} \end{bmatrix}=\begin{bmatrix} W'y \\ X'y \end{bmatrix},$$

em que se pressupõe a *priori* que todos os marcadores explicam quantidades iguais da variância genética aditiva (σ_g^2).

Desta forma, a predição do valor genético genômico (GBV) do indivíduo *j* via RR-BLUP é obtida pela expressão:

$$\hat{y}_i=\hat{\mu}+\sum_j X_j\hat{m}_{rr\text{-}blup_j},$$

sendo $\hat{\mu}$ a média da característica.

Apesar do método RR-BLUP ser amplamente usado na seleção genômica, outros métodos também têm sido usados com sucesso. Dentre estes se destacam a Regressão via Componentes Principais (*Principal Components Regression* – PCR) e o Quadrados Mínimos Parciais (*Partial Least Squares* – PLS) usados por Tyrisevä *et al.* (2011) e Moser *et al.* (2009) em gado leiteiro, os quais são denominados por Resende *et al.* (2010) como métodos de Redução de Dimensionalidade. Um método estatístico que pode ser enquadrado nesse contexto, porém até o momento ainda não foi empregado na seleção genômica é a Regressão via Componentes Independentes (*Independents Components Regression* – ICR). Desta forma, descrições metodológicas básicas a respeito dos métodos PCR, PLS, ICR são apresentadas a seguir.

2. Métodos de Redução Dimensional

2.1. Regressão via Componentes Principais

A regressão via componentes principais (*Principal Components Regression* – PCR) foi introduzida por Kendall (1957) e Hotelling (1957). Este método faz uso de alguns procedimentos empregados na análise de componentes principais (*Principal Component Analysis* – PCA) visando contornar os problemas apresentados pela regressão linear múltipla (*Multiple Linear Regression* – MLR). Uma vez que o PCA pode ser aplicado como um método de redução de dimensionalidade, soluciona os problemas existentes de multicolinearidade da matriz X e a necessidade de um

número excessivo de amostras para a construção de um modelo viável, sem acarretar na perda significativa de informações presentes nos dados (OTTO, 1999).

O processo se inicia com a obtenção das variáveis latentes Z_ν (ν=1,...,n_{pcr}), que são combinações lineares das variáveis explicativas ($X_1,...,X_J$) e determinadas por:

$$\hat{Z}=X\hat{P}, \tag{1}$$

sendo $\hat{P}$ a matriz de autovetores da matriz de covariância de X e $\hat{Z}$ a matriz cujas colunas são os componentes principais $\hat{z}_\nu$'s.

Os componentes principais também são denominados componentes ortogonais. Desta forma, as correlações entre as variáveis Z_ν e $Z_{\nu'}$ são suprimidas pelos componentes principais, ou seja, $cor(Z_\nu, Z_{\nu'}) = 0, \ \forall \ \nu \neq \nu'$.

Visando estabelecer a relação entre a variável dependente (Y) e os componentes principais Z_ν, utiliza-se a regressão linear múltipla obtendo, respectivamente, o modelo e a equação de predição a seguir:

$$y = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + ... + \alpha_{n_{pcr}} z_{n_{pcr}} + e_1;$$

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 \hat{z}_1 + \hat{\alpha}_2 \hat{z}_2 + ... + \hat{\alpha}_{n_{pcr}} \hat{z}_{n_{pcr}}, \tag{2}$$

em que e_1 é o vetor de resíduos, z_ν é o vetor coluna que compõe a matriz de componentes principais Z com estimativa igual a $\hat{z}_\nu$ e α_ν's são os coeficientes da regressão entre Y e Z com estimativas iguais a $\hat{\alpha}_\nu$'s, $\forall \nu$=1,...,n_{pcr}.

A metodologia do método PCR consiste, basicamente, em eliminar componentes que não contribuam na explicação da variância presente nos dados, o que reduz a dimensionalidade dos dados originais ($n_{pcr} \leq J$). Após a escolha do número ótimo de componentes a ser incluídos na equação de predição (2) os coeficientes podem ser determinados por meio do método dos quadrados mínimos ordinários (*Ordinary Least Squares*– OLS).

Como em qualquer regressão múltipla, a escolha do número de variáveis a serem incluídas no modelo é de extrema importância, tendo em vista a perda de informações relevantes. Segundo Roggo *et al.* (2007), cada componente descreve uma fração da variação total contida nos dados, tornando possível a determinação do número ótimo de componentes a serem incluídos na regressão.

Os coeficientes obtidos na equação de predição (2) não possuem significado biológico. Isso porque estão associados a variáveis transformadas. Assim, após obtenção desses coeficientes é possível, por meio das equações (1) e (2), encontrar uma estimativa para os coeficientes associados às variáveis originais (ω_{pcr}), dada por:

$$\hat{\omega}_{pcr}=\hat{P}\hat{\alpha},$$

sendo $\hat{\alpha}$ o vetor das estimativas dos coeficientes provenientes da regressão entre Y e Z.

Diante do exposto, para determinar a matriz P de projeção ortogonal das variáveis latentes no espaço coluna de X é necessário abordar os tópicos a seguir referentes à matriz de variância e covariância, autovetores e autovalores, respectivamente, descritos por Marcoulides e Hershberger (1997).

2.1.1.Matriz de Variância e Covariância

Sejam $X_1,...,X_J$ variáveis aleatórias com variâncias $\sigma_1^2,...,\sigma_J^2$, respectivamente, e com covariâncias $\sigma_{12},\sigma_{13},...,\sigma_{(j-1)j}$ (j=1,...,J). Desta forma, a matriz de (co)variância pode ser expressa por:

$$\text{Var(X)=V}=\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1J} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1J} & \sigma_{2J} & \cdots & \sigma_J^2 \end{bmatrix},$$

em que $\sigma_j^2=E[(X_j-E(X_j))^2], \forall\, j$ e $\sigma_{jj'}=E[(X_j-E(X_j))(X_{j'}-E(X_{j'}))], \forall\, j\neq j'$.

A matriz V é simétrica, assim, o elemento da posição *jj*, é a variância da variável aleatória X_j e o elemento da posição *jj'* (para $j\neq j'$) é a covariância entre as variáveis aleatórias X_j e $X_{j'}$.

Sob a notação matricial tem-se que a matriz V pode ser dada por:

$$V=Var(X)=E[(X-E(X))(X-E(X))'].$$

2.1.2.Autovetores e Autovalores

Seja T uma transformação linear dada por $T:V \to V$, em que V é um espaço vetorial qualquer. Assim, um vetor não nulo $v \in V$ é dito autovetor de T, se existe um número real λ tal que:

$$T(v)=\lambda v,$$

em que o escalar λ pode ser denominado autovalor de T associado a v.

Seja A uma matriz quadrada da transformação linear T, como visto anteriormente, tem-se que:

$$Av=\lambda v=\lambda Iv,$$

em que I é uma matriz identidade. Assim, tem-se: $\lambda Iv-Av=0$.

Desta forma,

$$(\lambda I-A)v=0.$$

Para uma solução não nula desse sistema de equações, o determinante deverá ser nulo, ou seja, $\det(\lambda I-A)=0$. Resolvendo esse determinante, obtêm-se os valores do escalar λ, assim substituindo em $Av=\lambda v$, torna-se possível a determinação dos autovetores.

2.2.Quadrados Mínimos Parciais

Quadrados Mínimos Parciais (*Partial Least Squares* – PLS) foi introduzido por Wold em 1975, sendo considerado útil a construção de equações de predições em situações nas quais se tem um grande número de variáveis explicativas e um número relativamente pequeno de dados amostrais (HOSKULDSSON, 1998). Resumidamente, a idéia geral do PLS é formar componentes que capturem a maior quantidade de informação possível disposta nas variáveis explicativas ($X_1,...,X_J$) para predizer a variável dependente (Y). Segundo Garthwaite (1994), o método PLS apresenta similaridades com o método de Regressão via Componentes Principais (PCR), sendo a maior diferença entre eles dada pelo fato do PCR levar em consideração apenas as variáveis explicativas na construção dos componentes, enquanto que o PLS também leva em consideração a variável dependente.

A metodologia PLS consiste na obtenção de um estimador para a variável resposta Y à partir de componentes T_ℓ ($\ell=1,...,n_{pls}$), que são combinações lineares

das variáveis explicativas $X_1,...,X_J$. Sob este enfoque, o modelo e a equação de predição são expressos por:

$$y = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + ... + \beta_{n_{pls}} t_{n_{pls}} + e_2$$

$$\hat{y} = \beta_0 + \beta_1 \hat{t}_1 + \beta_2 \hat{t}_2 + ... + \beta_{n_{pls}} \hat{t}_{n_{pls}} \quad , \qquad (3)$$

em que e_2 é o vetor de efeitos aleatórios, t_ℓ é o vetor coluna que compõe a matriz de componentes T com estimativa igual a $\hat{t}_\ell$ e β_ℓ's são os coeficientes da regressão entre Y e T com estimativas iguais a $\hat{\beta}_\ell$, $\forall \ell$=1,...,n_{pls}.

A correlação de qualquer par de componentes é igual a 0, isto é: $cor(T_\ell, T_{\ell'}) = 0, \ \forall \ \ell \neq \ell'$. Assim, o método PLS reduz o número de termos na equação de regressão, uma vez que o número de componentes na equação (3) geralmente é menor que o número de variáveis X.

Para simplificar os cálculos, utilizam-se variáveis centradas de Y e X_j denotadas respectivamente por U_1 e V_{1j}, em que:

$$U_1 = Y - \overline{Y} \text{ e } V_{1j} = X_j - \overline{X}_j \text{, para } j = 1,...,J, \qquad (4)$$

sendo seus vetores de valores dados por $u_1 = y - \overline{y}1$ e $v_{1j} = x_j - \overline{x}_j 1$, em que 1 um vetor coluna n-dimensional, $\{1,...,1\}'$.

Os componentes são determinados de forma sequencial, sendo que primeiramente é realizada uma regressão de U_1 em função de V_{11}, depois em função de V_{12}, e assim por adiante até V_{1J}. Tais regressões são independentes, portanto no desenvolvimento do método PLS as correlações entre os V_{1j}'s são ignoradas.

As equações de predição resultantes do método de quadrados mínimos ordinários para $j = 1,...,J$ são iguais a

$$\hat{U}_{1(j)} = \hat{b}_{1j} V_{1j}, \qquad (5)$$

em que $\hat{b}_{1j} = v'_{1j} u_1 / (v'_{1j} v_{1j})$, assim, cada uma das J equações fornece uma estimativa de $\hat{U}_{1(j)}$.

Segundo Garthwaite (1994), geralmente define-se T_1 como uma média ponderada, dada por:

$$\hat{T}_1 = \sum_{j=1}^{J} w_{1j} \hat{b}_{1j} V_{1j} \qquad (6)$$

em que w_{1j} é um peso, para o qual geralmente assume-se $w_{1j} \propto var(V_{1j}) = (v'_{1j}\ v_{1j})/(n\text{-}1)$. Assim, se $var(V_{1j})$ é pequena em relação à variância de X_j, então X_j é aproximadamente colinear com os componentes T_1 ,..., $T_{(\ell\text{-}1)}$. Além disso, o autor também comenta que a restrição $\sum_{j=1}^{J} w_{1j} = 1$ não é essencial ao desenvolvimento do método, mesmo que esse seja parte da equação de uma média ponderada.

De acordo com a metodologia PLS, sendo o componente T_1 uma média ponderada dos preditores de U_1, o mesmo é um preditor útil de U_1, e conseqüentemente de Y. No entanto, T_1 não contém todas as informações existentes em X_j. A informação ausente em T_1 pode ser estimada pelos resíduos obtidos na regressão entre X_j e T_1, os quais são idênticos aos resíduos provenientes da regressão entre V_{1j} e T_1. De modo análogo, a variabilidade em Y que não está sendo explicada por T_1 pode ser estimada pelos resíduos obtidos na regressão entre U_1 e T_1. Estes resíduos são denotados por V_{2j} para V_{1j} e por U_2 para U_1.

O segundo componente, T_2, é construído do mesmo modo que T_1, porém substituindo U_1 e V_{1j} por U_2 e V_{2j}, respectivamente. Desta forma, o procedimento se estende analogamente para os componentes T_2,..., $T_{n_{pls}}$.

Generalizando, suponha que $T_\ell\ (\ell \geq 1)$ seja construído à partir das variáveis U_ℓ e $V_{\ell j}$ com j=1,...,J. Para obter o componente $T_{\ell+1}$, as variáveis $V_{(\ell+1)j}$'s e $U_{(\ell+1)}$'s devem ser determinadas. Com este propósito, é feita uma regressão entre $V_{\ell j}$ e T_ℓ e, assim, $V_{(\ell+1)j}$ é definido por:

$$V_{(\ell+1)j} = V_{\ell j} - \{t'_\ell v_{\ell j}/(t'_\ell t_\ell)\} T_\ell \qquad (7)$$

sendo t_ℓ o vetor de valores de T_ℓ, $v_{(\ell+1)j}$ os resíduos da regressão e $t'_\ell v_{\ell j}/(t'_\ell t_\ell)$ o coeficiente da regressão entre $V_{\ell j}$ e T_ℓ.

De forma análoga, $U_{(\ell+1)}=U_\ell-\{t'_\ell u_\ell/(t'_\ell t_\ell)\}T_\ell$ e $u_{(\ell+1)}$ são os resíduos da regressão entre U_ℓ e T_ℓ. Assim, a j-ésima regressão entre $U_{(\ell+1)}$ e $V_{(\ell+1)j}$ resulta num coeficiente de regressão dado por:

$$\hat{b}_{(\ell+1)j}=v'_{(\ell+1)j}u_{(\ell+1)}/(v'_{(\ell+1)j}v_{(\ell+1)j}). \tag{8}$$

Analogamente a equação (6), define-se $T_{(\ell+1)}$ como sendo:

$$\hat{T}_{(\ell+1)}=\sum_{j=1}^{J} w_{(\ell+1)j}\hat{b}_{(\ell+1)j}V_{(\ell+1)j}, \tag{9}$$

sendo $w_{(\ell+1)j} \propto var(V_{(\ell+1)j})=(v'_{(\ell+1)j}\, v_{(\ell+1)j})/(n-1)$.

O método é repetido a fim de obter $T_{(\ell+2)}, T_{(\ell+3)}, ..., T_{n_{pls}}$, e após a obtenção dos n_{pls} componentes, os coeficientes da equação de predição (3) são determinados por meio do método dos quadrados mínimos ordinários.

Conforme já relatado, a principal característica do método PLS é que a correlação entre qualquer par de componentes é igual a 0, e isso se deve ao fato dos resíduos provenientes da regressão não serem correlacionados com o regressor, ou seja, $V_{(\ell+1)j}$ é não correlacionado com T_ℓ. Assim, como cada componente $T_{(\ell+1)},...,T_{n_{pls}}$ é uma combinação linear de $V_{(\ell+1)j}$, os componentes são não correlacionados com T_ℓ. Esta característica é de suma importância para a seleção genômica ampla, pois o método PLS torna-se uma alternativa eficiente para se obter preditores apesar da multicolinearidade presente nos dados de marcadores (covariáveis X_j). Além disso, garanta-se que os coeficientes da equação de predição (3) podem ser estimados por uma simples regressão feita entre Y e T_ℓ. Também, adicionando-se componentes na equação, os componentes anteriores não têm seus coeficientes alterados. Outra conseqüência é que $u_{(\ell+1)}$ e $v_{(\ell+1)j}$ podem ser vistos, respectivamente, como vetores de resíduos obtidos das regressões de Y e X_j com $T_1, ... ,T_{n_{pls}}$.

Como qualquer análise de regressão múltipla, um passo importante na execução do PLS é a determinação do número ótimo de componentes que devem ser incluídos na equação (3). Segundo Garthwaite (1994), geralmente utilizam-se métodos tradicionais de seleção de variáveis como Backward, Forward e Stepwise, porém no âmbito da seleção genômica ampla, devido a enorme quantidade de covariáves (marcadores), muitas vezes a aplicação de tais métodos torna-se inviável,

e como alternativa pode-se utilizar métodos de validação cruzada (RESENDE *et al.*, 2010). Deve-se ressaltar que após os coeficientes da equação (3) serem estimados, é possível usar as equações (4), (7) e (9) para expressar o modelo em relação a X_j ao invés dos componentes $T_{n_{pls}}$, ou seja, expressar o modelo em termos de suas variáveis originais e com interpretação biológica. Desta forma, tem-se:

$$\hat{\omega}_{pls} = \hat{B}W'\hat{\beta},$$

em que $\hat{B}$ é a matriz cujos elementos são estimativas dos coeficientes provenientes da regressão entre U_i e V_{ij}, W é a matriz de pesos e $\hat{\beta}$ é o vetor das estimativas dos coeficientes provenientes da regressão entre Y e T.

2.3.Análises de Componentes Independentes

Análises de Componentes Independentes (*Independent Component Analysis* – ICA) foi formalmente definida por Jutten e Hérault (1991) e Comon (1994) e assim, como os métodos PCR e PLS é um modelo de variáveis latentes. Resumidamente, o método consiste em decompor uma matriz de dados X (com *I* observações e *J* variáveis, cujos valores de cada variável são centrados em sua respectiva média) em duas matrizes, de forma que:

$$X' = A'S', \tag{10}$$

em que A é a matriz de misturas (relaciona os componentes independentes com as variáveis originais) e S é a matriz cujas colunas são componentes independentes.

O pressuposto do método ICA é que os dados apresentem distribuições não gaussianas. Possivelmente, por esse motivo, essa teoria ressurgiu tardiamente no campo científico. Tal suposição está fundamentada no fato de que a distribuição das variáveis gaussianas não é afetada por qualquer transformação ortogonal a qual o conjunto de dados sofre durante o desenvolvimento do método (SILVA *et al.*, 2011).

Supondo A uma matriz quadrada e conhecida, geralmente podemos calcular sua inversa (A^{-1}) e assim tem-se:

$$S = XA^{-1}. \tag{11}$$

Na maioria dos casos, A é uma matriz desconhecida que precisa ser estimada. Um procedimento útil, para esse problema é o processo de branqueamento (*whitening*), em que a matriz de dados X sofre uma transformação linear ortogonal

XK, sendo K denominada matriz de branqueamento obtida por meio da decomposição em autovetores (*Eigenvalue Decomposition* – EVD) da matriz de covariância de X. O branqueamento faz com que a nova matriz (XK) contenha variáveis não correlacionadas e que sua matriz de covariância seja unitária, ou seja, igual à identidade.

Além do processo de descorrelação, existem algoritmos especiais usados na ICA que tentam encontrar uma matriz R ortogonal que maximize a independência estatística das colunas de S. Uma medida quantitativa de independência entre as variáveis aleatórias pode ser fornecida por uma função de contraste. Existem diversos contrastes e algoritmos que maximizam essa independência, um deles foi desenvolvido por Hyvärinen (1998b). Os procedimentos desse algoritmo foram baseados no princípio da máxima entropia ($J(r)$)[1] e assume que a variável aleatória r é padronizada (possui média zero e variância igual a um). Desta forma, obtém-se a seguinte aproximação:

$$J(r) \approx \sum_{i=1}^{p} c_i [E\{G_i(r)\} - E\{G_i(v)\}]^2, \tag{12}$$

em que c_i é uma constante positiva, v é uma variável arbitrária gaussiana padronizada e G_i são funções não quadráticas.

Para casos em que é usada apenas uma função quadrática G, tem-se a seguinte proporcionalidade:

$$J(r) \propto [E\{G_i(r)\} - E\{G_i(v)\}]^2, \tag{13}$$

sendo G qualquer não função quadrática.

Segundo Hyvärinen e Oja (2000), a escolha de funções G's que não possuem crescimento rápido resulta em estimadores mais robustos para r. Desta forma, as funções G definidas abaixo se mostraram úteis para esse processo:

$$G_1(u) = \frac{1}{a_1} \log \cosh a_1 u, \quad G_2(u) = -\exp\left(-\frac{u^2}{2}\right), \tag{14}$$

em que u é uma variável e $1 \leq a_1 \leq 2$.

[1] A Entropia é uma medida de imprevisibilidade do sistema. Na estatística, a entropia pode ser vista como uma medida da aleatoriedade de distribuições de probabilidade. Quando a variável segue a distribuição gaussiana as medidas de entropia e variância são coincidentes.

O algoritmo para a obtenção de uma estimativa para r é iterativo. Assim, é possível obter aproximações com propriedades estatísticas desejáveis, principalmente robustez.

Após a obtenção da matriz R pode-se encontrar os componentes independentes, isso porque eles são dados por:

$$\hat{S}=XKR\,, \tag{15}$$

sendo a matriz KR uma forma aproximada de A'. A matriz A durante o processo de ICA torna-se ortogonal e a ortogonalização faz com que $A'A=I$ e $AA'=I^{*}$, sendo I e I^{*} matrizes identidades.

Uma vez obtida tais componentes, os mesmos podem ser usados sob o enfoque de regressão como sendo covariáveis independentes (análoga a regressão via componentes principais na equação (2)). Assim, o modelo e a equação de predição da Regressão de Componentes Independentes (*Independent Component Regression*– ICR) são expressos por:

$$\begin{aligned} y&= \gamma_0 + \gamma_1 s_1 + \gamma_2 s_2 + ... + \gamma_{n_{icr}} s_{n_{icr}} + e_3 \\ \hat{y}&= \hat{\gamma}_0 + \hat{\gamma}_1 \hat{s}_1 + \hat{\gamma}_2 \hat{s}_2 + ... + \hat{\gamma}_{n_{icr}} \hat{s}_{n_{icr}}\,, \end{aligned} \tag{16}$$

em que e_3 é o vetor de resíduos, s_κ é vetor coluna que compõe a matriz de componentes independentes S com estimativa igual a $\hat{s}_\kappa$ e γ_κ's são os coeficientes provenientes da regressão entre Y e S que podem ser determinados por meio do método dos mínimos quadrados ordinários, $\forall\kappa=1,...,n_{icr}$.

O número ótimo de componentes nesse método pode ser determinado por uma análise de PCA (CADAVID *et al.*, 2008) e desta forma utilizar o mesmo número ótimo de variáveis latentes que o PCR.

As estimativas dos coeficientes associados às variáveis originais podem ser determinados pela combinação das equações (15) e (16) e são dadas por:

$$\hat{\omega}_{icr}=KR\hat{\gamma}\,,$$

em que $\hat{\gamma}$ é o vetor das estimativas dos coeficientes provenientes da regressão entre Y e S.

2.3.1.Independência Estatística

A metodologia de ICA é baseada principalmente no conceito estatístico de independência entre variáveis aleatórias que serão descritos a seguir.

Duas variáveis se dizem dependentes, quando conhecer o valor de X_1 fornece algumas informações sobre X_2. Traduzindo-se em termos matemáticos, duas variáveis aleatórias contínuas X_1 e X_2 são dependentes estatisticamente se, e somente se

$$f_{x_2|x_1}(x_2|x_1)=\frac{f_{x_1x_2}(x_1,x_2)}{f_{x_2}(x_1)},$$

em que $f_{x_2|x_1}(x_2|x_1)$ é a função densidade de probabilidade condicional de X_2 dado X_1, $f_{x_1x_2}(x_1,x_2)$ é a função densidade de probabilidade conjunta de X_1 e X_2 $f_{x_1}(x_1)$ é a função densidade de probabilidade marginal de X_1.

Por outro lado, duas variáveis são ditas independentes, quando a determinação do valor de uma não fornece nenhuma informação sobre o valor da outra. Traduzindo-se em termos matemáticos, duas variáveis aleatórias contínuas X_1 e X_2 são independentes estatisticamente se, e somente se

$$f_{x_1x_2}(x_1,x_2)=f_{x_1}(x_1)f_{x_2}(x_2),$$

em que $f_{x_2}(x_2)$ é a função de densidade probabilidade marginal de X_2.

As equações acima evidenciam que a propriedade de independência estatística é mais complexa que a de não–correlacionamento, uma vez que a última está associada à condição, $E(x_1,x_2)=E(x_1)E(x_2)$, relacionada à covariância entre as variáveis aleatórias. O único caso, em que a propriedade de não correlacionamento implica em independência estatística ocorre quando as variáveis aleatórias são gaussianas (PAPOULIS, 1991).

O conceito de independência se estende para casos em que há *J* variáveis aleatórias ($X_1,...,X_J$), tendo-se:

$$f_{x_1...x_J}(x_1,...,x_J)=\prod_{j=1}^{J} f_{x_j}(x_j).$$

3.Referências Bibliográficas

CADAVID, A. C.; LAWRENCE, J. K.; RUZMAIKIN, A.; KAYLENG–KNIGHT.Principal Components and Independent Component Analysis of Solar and Space Data. **Solar Phys**, v. 248, p. 247-261, 2008.

COMON, P. Independent component analysis – a new concept. **Sigmal Processing**, v. 45, p. 59-83, 1994.

GARRICK, D. J.; TAYLOR, J. F.; FERNANDO, R. L. Deregressing estimated breeding values and weighting information for genomic regression analyses. **Genetics Selection Evolution**, v. 41, p. 55, 2009.

GARTHWAITE, P.H. An Interpretation of Partial Least Squares. **Journal of the American Statistical Association**, v. 89, p. 122-127, 1994.

HOTELLING, H.The relations of the newer multivariate statistical me thods to factor analysis. **British Journal of Mathematical and Statistical Psychology**, v.10, p. 69-79, 1957.

HOSKULDSSON, P. PLS Regression Methods. **Journal of Chernometrics**, v. 2, p. 211-228, 1988.

HYVÄRINEN, A. New approximations of differential entropy for independent component analysis and projection pursuit. **In Advances in Neural Information Processing Systems,** v. 10, p. 273-279, 1998.

HYVÄRINEN, A.; OJA, E. Independent Component Analysis: Algorithms and Applications. **Neural Networks**, v. 13, p. 411-430, 2000.

JENKINS, S., GIBSON, N. High-throughput SNP genotyping. **Comparative and Functional Genomics**, v. 3, p. 57-66, 2002.

JUTTEN, C.; HERAULT, J. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. **Sigmal Processing**, v. 24, p. 1-10, 1991.

KENDALL,M. G. **A Course in Multivariate Analysis**. London: Griffin, 1957.

MARCOULIDES, G. A.; HERSHBERGER, S. L. **Multivariate Statistical Methods: A First Course**. Lawrence Erlbaum Associates, 322 p., 1997.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome wide dense marker maps. **Genetics**, v. 157, p. 1819-1829, 2001.

MOSER, G.; TIER, B.; CRUMP, R. E. ; KHATKAR, M. S.; RAADSMA, H. W. A comparison of five methods to predict genomic breeding values of dairy bulls from genome–wide SNP markers. **Genetics Selection Evolution**, v.41, p. 41-53, 2009.

OTTO, M. **Chemometrics**. Weinheim: Wiley, 328 p., 1999.

PAPOULIS, A. **Probability, Random Variables and Stochastic Processes**. McGraw Hill., 3rd ed, 1991.

RESENDE, M. D. V. Seleção genômica ampla (GWS) e modelos mistos. In: **Matemática e estatística na análise de experimentos e no melhoramento genético**. Colombo: Embrapa Florestas, p. 517-534, 2007.

RESENDE, M. D. V. **Genômica quantitativa e seleção no melhoramento de plantas perenes e animais**. Colombo: Embrapa Florestas, 330 p., 2008.

RESENDE, M. D. V.; RESENDE JUNIOR, M. F. R.; AGUIAR, A. M.; ABAD, J. I. M.; MISSIAGIA, A. A.; SANSALONI, C.; PETROLI, C.; GRATTAPALIA, D. **Computação da seleção genômica ampla (GWS)**. Colombo: Embrapa Florestas, 79p, 2010.

ROGGO, Y.; CHALUS, P.; MAURER, L.; LEMA–MARTINEZ, C.; EDMOND, A.; JENT, N. A review of near infrared spectroscopy and chemometrics in harmaceutical technologies. **Journal of Pharmaceutical and Biomedical Analysis**, v.44, p.683-700, 2007.

SILVA, L. R. da. **Aplicação da Decomposição em Valores Singulares e Análise de Componentes Independentes em dados de fMRI**. Dissertação, Universidade Federal do Pará, 52 p., 2011.

TYRISEVÄ, A. M.; MEYERS, K.; FIKSE, W. F.; DUCROCQ, V.; JAKOBSEN, J., LIDAUER, M. H., MÄNTYSAARI, E. A. Principal component and factor analytic models in international sire evaluation. **Genetics Selection Evolution**, v.43, p. 1–33, 2011.

Métodos de redução de dimensionalidade aplicados na seleção genômica para características de carcaça em suínos

RESUMO

A principal contribuição da genética molecular é a utilização direta das informações de DNA no processo de identificação de animais geneticamente superiores. Sob esse enfoque, idealizou-se a seleção genômica ampla (*Genome Wide Selection* – GWS), a qual consiste na análise de um grande número de marcadores SNPs (*Single Nucleotide Polymorphisms*) amplamente distribuídos no genoma. Devido ao grande número de marcadores, geralmente maior que o número de animais genotipados, e alta colinearidade entre os mesmos, métodos estatísticos de redução de dimensionalidade são requeridos. Dentre estes destacam-se os métodos de regressão via Componentes Independentes (*Independent Component Regression* – ICR), regressão via componentes principais (*Principal Component Regression* – PCR) e regressão via Quadrados Mínimos Parciais (*Partial Least Squares* – PLS). Diante do exposto, objetivou-se aplicar e comparar os métodos de redução de dimensionalidade e RR-BLUP (método tradicionalmente aplicado) para características de carcaça em uma população F_2 de suínos Piau×Comercial. Os resultados indicaram que o método ICR se mostrou mais eficiente, uma vez que este proporcionou maiores valores de acurácia na estimação do GBV para a maioria das características de carcaça.

Palavras-chave: Seleção genômica, redução dimensional e RR-BLUP.

ABSTRACT

The main contribution of molecular genetics is the direct use of DNA information to identify genetically superior individuals. Under this approach, genome-wide selection (GWS) can be used with this purpose. GWS consists in analyzing of a large number of SNP markers widely distributed in the genome, and due to the fact that the number of markers is much larger than the number of genotyped individuals and also to the fact that such markers are highly correlated special statistical methods are widely required. Among these stand out Independent Component Regression (ICR), Principal Component Regression (PCR), Partial Least Squares (PLS) and Random Regression Best Linear Unbiased Predictor. Thus, the

aim of this paper was to propose an application of the methods of dimensionality reduction and RR-BLUP (method traditionally applied) to GWS of carcass traits in an F_2 (Piau × commercial) pig population. . The results showed that ICR method is more efficient, since provided most accurate genomic breeding values estimates for most carcass traits.
Key words: Genomic selection, dimensionality reduction, RR-BLUP.

1.INTRODUÇÃO

As características de carcaça são muito importantes para o desenvolvimento da suinocultura, sobretudo aquelas relacionadas com maior rendimento de carne e menor deposição de gordura, para que se possa atender ao crescente e cada vez mais exigente mercado consumidor. Várias pesquisas foram desenvolvidas com a finalidade de utilizar marcadores moleculares no melhoramento de suínos para características de carcaça (PIRES *et al.*, 2006; RUSSO *et al., 2008;* SILVA *et al.*, 2011).

Considerando que polimorfismos do DNA são fonte da variação de mérito genético, marcadores SNPs (*Single Nucleotide Polymorphisms*) em desequilíbrio de ligação com locos de características quantitativas (*Quantitative Trait Loci* – QTL) podem ser utilizados como critério extra para identificação de indivíduos candidatos à seleção, o que aumentaria a acurácia na avaliação genética (MEUWISSEN *et al.*, 2001). Recentemente, tornam-se cada vez mais abundante informações genômicas individuais massivas para animais de produção, com dezenas ou centenas de milhares de marcadores SNPs (GODDARD & HAYES, 2007). Assim, a aplicação prática destas informações genômicas é um desafio, pois geralmente não é possível a utilização adequada de modelos funcionais que estimam o efeito de cada SNP no fenótipo, uma vez que geralmente o número de marcadores é muito maior que o número de animais genotipados e fenotipados. De acordo com Gianola *et al.* (2003), informações genômicas podem se constituir de dezenas ou centenas de milhares de covariáveis, possivelmente com alta colinearidade, o que demanda a utilização de metodologias estatísticas que considerem a seleção de covariáveis, a regularização do processo de estimação ou ainda abordagens não-paramétricas.

Assim, um ponto chave para o sucesso da seleção genômica ampla é a escolha adequada de metodologias a serem utilizadas, e, portanto, a comparação das

mesmas e o desenvolvimento de ferramentas computacionais que as contemplem representam uma importante linha de pesquisa no cenário moderno do Melhoramento Animal. Até o momento os métodos de Regressão Penalizada sob os enfoques frequentistas (RR-BLUP e G-BLUP) e bayesianos (Bayes A e B e LASSO Bayesiano – BL) têm sido os mais utilizados. Porém, existem outros métodos que têm sido aplicados a GWS com sucesso, dentre eles Quadrados Mínimos Parciais (PLS) e Regressão via Componentes Principais (PCR) denominados por Resende *et al.* (2010) como métodos de Redução de Dimensionalidade, que apresentam grande aplicabilidade e teoria relativamente simples.

Segundo Solberg *et al.* (2009), os métodos PCR e PLS fornecem uma análise rápida de grandes quantidades de dados que visam obter os valores genéticos genômicos (*Genomic Breeding Value* – GBV) dos indivíduos. Para isso, necessita-se apenas do pressuposto de aditividade dos efeitos de marcadores. Boulesteix e Strimmer (2006) comenta sobre alguns pontos positivos do método PLS como a flexibilidade, a versatilidade, a eficiência estatística e a rapidez computacional.

Ademais, o método estatístico Regressão via Componentes Independentes (*Independent Component Regression* – ICR) também pode ser enquadrado nesse contexto, porém até o momento não foi aplicado à seleção genômica. Essa metodologia pode ser vista também como um método de redução de dimensionalidade e como uma extensão da abordagem de PCR, incorporando uma estrutura estatística mais complexa do que o mesmo.

Diante do exposto, o presente trabalho teve por objetivo realizar um estudo inédito para comparar os métodos Quadrados Mínimos Parciais, Regressão via Componentes Principais, Regressão via Componentes independentes e o método RR-BLUP quanto à eficiência na estimação dos valores genéticos genômicos e dos efeitos de marcadores para características de carcaça de uma população F_2 de suínos (Piau × linhagem comercial) desenvolvida na Universidade Federal de Viçosa.

2. MATERIAIS E MÉTODOS

2.1.Descrições da população

Os dados utilizados no presente estudo são provenientes da Granja de Melhoramento de Suínos do Departamento de Zootecnia (DZO) da Universidade

Federal de Viçosa (UFV), em Viçosa, Minas Gerais, Brasil, no período de novembro de 1998 a julho de 2001. Neste experimento, a população F_2 foi composta de 345 suínos provenientes do cruzamento de dois varrões da raça local brasileira Piau, com 18 fêmeas de linhagem desenvolvida na UFV, pelo acasalamento de animais de linha comercial (Landrace × Large White × Pietrain).

Os detalhes dos procedimentos utilizados cuja extração do DNA foi realizada no Laboratório de Biotecnologia Animal do Departamento de Zootecnia da Universidade Federal de Viçosa, podem ser encontrados em Peixoto *et al.* (2006). A genotipagem foi realizada via tecnologia Golden Gate/Vera Code R, no Laboratório de Genética Animal (LGA), Embrapa Recursos Genéticos e Biotecnologia (CENARGEN), Brasília, DF, conforme descrito por Hidalgo *et al.* (2011). Os marcadores SNPs estão distribuídos da seguinte forma nos cromossomos de Sus scrofa: SSC1 (56), SSC4 (54), SSC7 (59), SSC8 (31), SSC17 (25), totalizando assim 237 marcadores.

As características de carcaça foram mensuradas após o abate (aproximadamente dos 105 dias), dentre elas as características escolhidas para as análises foram espessura de toucinho imediatamente após a última costela na linha dorso-lombar (ETUC); menor espessura de toucinho na região acima da última vértebra lombar, na linha dorso-lombar (ETL); espessura de toucinho medida entre a ultima e a penúltima vértebra lombar (ETUL); espessura de toucinho medida imediatamente após a última costela, a 6,5 cm da linha dorso-lombar (ETO);maior espessura de toucinho na região da copa, na linha dorso-lombar (ETSH); espessura de bacon (EBACON); comprimento de carcaça pelos métodos de classificação americano (MLC) e rendimento de carcaça (RCARC). Maiores detalhes podem ser encontrados em Hidalgo *et al.* (2011).

Os dados fenotípicos utilizados nas análises foram corrigidos para os efeitos fixos, dos genitores e desregressados (RESENDE *et al.*, 2010), visando trabalhar basicamente com o mérito genético verdadeiro de indivíduos não aparentados. Esses procedimentos foram realizados nos softwares ASREML® (GILMOUR *et al.*, 2000) e R (R Development Core Team, 2010).

2.2.Método RR-BLUP

O método RR-BLUP (*Random Regression Best Linear Unbiased Predictor*) faz uso de preditores do tipo BLUP e assume que os efeitos dos marcadores SNPs são covariáveis de efeitos aleatórios. A predição por meio do método RR-BLUP é baseada no modelo linear misto (RESENDE, 2007, 2008):

$$y = Wb + Xm_{rr\text{-}blup} + e,$$

em que:

y é o vetor de dados fenotípicos corrigidos e desrregressados, com dimensão $I \times 1$, sendo I o número de indivíduos genotipados e fenotipados;

b é o vetor de efeitos fixos com matriz de incidência W;

$m_{rr\text{-}blup}$ é o vetor de efeitos aleatórios dos marcadores com matriz de incidência X, com valores 0, 1 e 2 para o número de alelos do marcador, e dimensão $J \times 1$, sendo J o número de marcadores.

As equações de modelos mistos genômicas visam predizer os efeitos *m* dos marcadores e são dadas matricialmente, por:

$$\begin{bmatrix} W'W & W'X \\ X'W & X'X+I\dfrac{\sigma_e^2}{(\sigma_g^2/n_Q)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{m}_{rr\text{-}blup} \end{bmatrix} = \begin{bmatrix} W'y \\ X'y \end{bmatrix},$$

em que σ_e^2 é a variância do erro, σ_g^2 é a variância genética e $n_Q = \sum_{j=1}^{J} 2p_j(1-p_j)$ e p_j a frequência alélica do marcador *j*. Nesse método se pressupõe, a *priori*, que todos os marcadores explicam quantidades iguais da variância genética aditiva σ_g^2. Desta forma, a predição do valor genético genômico do indivíduo *j* via RR-BLUP é obtida pela expressão:

$$\hat{y}_i = \hat{\mu} + \sum_j X_j \hat{m}_{rr\text{-}blup_j}, \qquad (1)$$

sendo $\hat{\mu}$ a média da característica.

2.3. Quadrados Mínimos Parciais

Quadrados Mínimos Parciais (*Partial Least Squares* – PLS) é considerado um método apropriado para dados em que tem-se mais covariáveis do que observações

(HOSKULDSSON, 1998), conforme observado na seleção genômica. Além disso, o método garante que a correlação entre qualquer par de variáveis latentes, as quais são combinações lineares dos SNPs, seja igual a zero. Essas duas características são de suma importância para a seleção genômica devido a abundância de marcadores SNPs presente no genoma e a multicolinearidade existente nos dados de marcas.

A metodologia do PLS consiste em formar componentes que capturem a maior quantidade de informação possível disposta nas variáveis explicativas $(X_1,...,X_J)$ visando predizer a variável dependente (Y). Sob esse enfoque, as matrizes X (matriz de marcas) e Y (vetor de dados fenotípicos corrigidos e desrregressados) são decompostas simultaneamente, como nas equações a seguir:

$$X=TL'+E_1 \quad (2)$$

$$Y=Uq'+e_2 \quad (3)$$

em que T e U são as matrizes de componentes, L e q são matriz e vetor de carregamento, E_1 e e_2 são matriz e vetor de resíduos, respectivamente. A decomposição de ambas as matrizes não é independente, mas sim realizada simultaneamente , possibilitando estabelecer uma relação entre os componentes de X e Y de forma que para cada fator a seguinte relação é obtida:

$$\hat{u}_\ell=\hat{b}_\ell t_\ell \quad (4)$$

sendo b_ℓ $(\ell=1,...,n_{pls})$ o coeficiente de regressão entre os fatores que podem ser estimados via método dos mínimos quadrados ordinários (*Ordinary Least Squares* – OLS), $\hat{b}_\ell=(u'_\ell t_\ell)/(t'_\ell t_\ell)$. Os valores obtidos de b em todas as regressões são agrupados em uma matriz diagonal B. Desta forma, obtém-se o modelo e a equação de predição a seguir:

$$y=TBq'+e_3,$$

$$\hat{y}=\hat{T}\hat{B}\hat{q}'. \quad (5)$$

Os coeficientes Bq' associados aos componentes na equação de predição (5) não possuem interpretação biológica, porém é possível estimar os coeficientes associados às variáveis originais (efeitos dos marcadores) combinando a equação (5) e o modelo (3). Desta forma tem-se:

$$\hat{m}_{pls}=\hat{L}\hat{B}\hat{q}'. \quad (6)$$

2.4. Regressão via Componentes Principais

A metodologia de regressão via componentes principais (*Principal Components Regression* – PCR) é uma alternativa para contornar os problemas de multicolinearidade apresentados pela regressão linear múltipla (*Multiple Linear Regression* – MLR). A PCR pode ser aplicada como um método de redução de dimensionalidade sem acarretar na perda significativa de informações presentes nos dados (OTTO, 1999).

Os componentes Z_ν (ν=1,...,n_{pcr}) nesse método são combinações lineares das variáveis explicativas $X_1,...,X_J$, assim tem-se:

$$\hat{Z}=X\hat{P}, \tag{7}$$

sendo X a matriz de marcas, P a matriz de autovetores da matriz de covariância e Z a matriz cujas colunas são os componentes principais (componentes ortogonais).

Visando estabelecer a relação entre a variável dependente Y (vetor de dados fenotípicos corrigidos e desrregressados) e os componentes ortogonais Z_ν utiliza-se a regressão linear múltipla, obtendo o modelo e a seguinte equação de predição:

$$y = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + ... + \alpha_{n_{pcr}} z_{n_{pcr}} + e_4,$$

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 \hat{z}_1 + \hat{\alpha}_2 \hat{z}_2 + ... + \hat{\alpha}_{n_{pcr}} \hat{z}_{n_{pcr}}, \tag{8}$$

em que e_1 é o vetor de resíduos, z_ν é o vetor coluna que compõe a matriz de componentes principais Z com estimativa igual a $\hat{z}_\nu$ e α_ν's são os coeficientes da regressão entre Y e Z com estimativas iguais a $\hat{\alpha}_\nu$'s, $\forall \nu$=1,...,n_{pcr}.

O método PCR consiste basicamente em eliminar componentes que não contribuam na explicação da variância presente nos dados, ou seja $, n_{pcr} \leq J$. Os coeficientes associados às variáveis transformadas podem ser determinados por meio do método dos quadrados mínimos ordinários. De forma análoga ao método PLS pode-se encontrar os coeficientes associados às variáveis originais, que neste contexto são os marcadores. Desta forma deve-se combinar as equações (7) e (8) afim de obter a seguinte relação:

$$\hat{m}_{pcr}=\hat{P}\hat{\alpha}, \tag{9}$$

sendo $\hat{\alpha}$ o vetor das estimativas dos coeficientes provenientes da regressão entre Y e Z.

Segundo Garthwaite (1994), os métodos PLS e PCR são similares, sendo a maior diferença entre eles dada pelo fato do PCR levar em consideração apenas as variáveis explicativas na construção dos componentes, enquanto que o PLS também leva em consideração a variável dependente.

2.5. Regressão via Componentes Independentes

A Regressão via Componentes Independentes (*Independent Component Regression* – ICR) é baseada na metodologia de Análises de Componentes Independentes (*Independent Component Analysis* – ICA). Resumidamente, o método consiste em decompor a matriz de covariáveis X em combinações de componentes independentes, garantindo a retirada da multicolinearidade presente nos dados, além de reduzir a dimensionalidade. O pressuposto exigido nesta análise é de que os dados sejam provenientes de distribuições não gaussianas Tal suposição está fundamentada no fato de que a distribuição das variáveis gaussianas não é afetada por qualquer transformação ortogonal a qual o conjunto de dados sofre durante o desenvolvimento do método (SILVA *et al.*, 2011). Desta forma, ICA pode ser aplicado à seleção genômica ampla, isso porque a matriz de marcadores X está parametrizada com os valores 0, 1 e 2 (distribuição não gaussiana), que correspondem ao número de alelos do marcador. Sob esse enfoque, tem-se a decomposição dada por:

$$X' = A'S', \tag{10}$$

em que A é a matriz de misturas (relaciona os componentes independentes com os marcadores) e S é a matriz cujas colunas são os componentes independentes S_κ $\kappa = 1,...,n_{icr}$.

A matriz A é definida como uma $f(KR)$, sendo K uma matriz de ortogonalização de S e R uma matriz ortogonal. Sob o enfoque da seleção genômica, os algoritmos especiais usados na ICA tentam encontrar uma matriz R ortogonal que maximize a independência estatística das colunas de S. Para tanto, foram utilizadas funções de contraste, que são medidas quantitativas de independência. Existem diversos contrastes e algoritmos que maximizem essa independência. O algoritmo desenvolvido por Hyvärinen (1998b) é iterativo e foi baseado no princípio da máxima entropia ($J(r)$) assumindo que a variável aleatória r está padronizada. Desta forma, obtemos a seguinte aproximação:

$$J(r) \approx \sum_{p=1}^{P} c_i [E\{G_i(r)\} - E\{G_i(v)\}]^2, \quad (11)$$

em que c_i é uma constante positiva, v é uma variável padronizada e G_i são funções não quadráticas.

Segundo Hyvärinen e Oja (2000), a escolha de funções G's que não possuem crescimento rápido resulta em estimadores mais robustos para r. Desta forma, a função G utilizada no presente estudo é dada por:

$$G_1(u) = -\exp\left(-\frac{u^2}{2}\right), \quad (12)$$

sendo u uma variável.

Após o processo iterativo tem-se a matriz de componentes dada por:

$$\hat{S} = XKR, \quad (13)$$

sendo K uma matriz de branqueamento (matriz de ortogonalização) e KR uma aproximação de A'. Encontrada a matriz de componentes independentes S, obtêm-se o modelo ea equação de predição baseada na Regressão via Componentes Independentes (*Independent Component Regression* – ICR), que são expressos por:

$$y = \gamma_0 + \gamma_1 s_1 + \gamma_2 s_2 + \ldots + \gamma_{n_{icr}} s_{n_{icr}} + e_5;$$
$$\hat{y} = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{s}_1 + \hat{\gamma}_2 \hat{s}_2 + \ldots + \hat{\gamma}_{n_{icr}} \hat{s}_{n_{icr}}, \quad (14)$$

em que $\hat{s}_\kappa$ é o vetor coluna de estimativas que compõe a matriz de componentes independentes $\hat{S}$ e γ_κ é o coeficiente da equação de predição que é determinado por meio do método dos quadrados mínimos ordinários, $\kappa = 1,\ldots,n_{icr}$.

Similarmente aos outros métodos de redução de dimensionalidade pode-se obter os efeitos de marcas à partir dos coeficientes associados às variáveis transformadas combinando as equações (13) e (14), tendo-se as seguintes estimativas:

$$\hat{m}_{icr} = KR\hat{\gamma}, \quad (15)$$

em que $\hat{\gamma}$ é o vetor das estimativas dos coeficientes provenientes da regressão entre Y e S.

2.6. Número ótimo de variáveis latentes

Um passo importante dos métodos de Redução Dimensional é a escolha do número ótimo de variáveis latentes a serem inseridas no modelo. Não existe uma regra formal para determinar esse número, porém na literatura um critério de decisão para as metodologias PLS e PCR é adotar uma percentagem da variação total explicada (neste trabalho foi de 70%) que se deseja obter através dos métodos e usar o número de componentes que atinja esse valor pré-determinado. No entanto, no método ICR não é possível determinar a variância dos componentes independentes, então segundo Cadavid *et al.* (2008) deve-se aplicar primeiramente a análise de componentes principais (*Principal Component Analysis* – PCA) e desta forma utilizar o mesmo número ótimo de variáveis latentes do PCR.

2.7. Validação Cruzada

Os métodos de Redução Dimensional e o método RR-BLUP foram comparados por meio de um estudo de validação cruzada (RESENDE *et al.*, 2010), realizado separadamente para cada característica de carcaça. Nesta parte da análise, a população F_2 de suínos foi dividida em três populações distintas, cada qual com 115 indivíduos**.**

Assim, a cada repetição da análise, duas dessas populações foram assumidas como população de estimação (ou treinamento) e utilizadas para obter os efeitos dos marcadores SNPs. A outra população, denominada população de validação, foi utilizada para avaliar a concordância entre valores genéticos preditos via estimativas provenientes da população de treinamento e fenótipos corrigidos observados. O processo foi repetido de forma que a cada etapa uma das três populações constituísse a população de validação.

Dessa forma, a correlação entre o valor estimado nas três validações e o fenótipo corrigido e desrregressado constituiu a capacidade preditiva do método. A acurácia do método é dependente dessa correlação e equivale a razão da correlação pela raiz quadrada da herdabilidade, $r_{q,\hat{q}} = \frac{r_{y,\hat{y}}}{\sqrt{h^2_{sm}}}$, sendo h^2_{sm} a herdabilidade da

segregação mendeliana dada por $h^2_{sm} = \frac{0{,}5h^2}{0{,}5h^2 + (1-h^2)}$ e h^2 a herdabilidade do caráter estimada pelo método REML sobre fenótipos em um modelo unicaracterístico (RESENDE *et al.*, 2010).

Em posse do melhor método para cada característica, o efeito dos marcadores em valor absoluto foram estimados e padronizados considerando toda a população F_2 de suínos fazendo uso das equações (1), (6), (9) e (15), respectivamente, para os métodos RR-BLUP, PLS, PCR e ICR. À partir destas informações o *Manhattan plot* foi construído, sendo que cada ponto representa um marcador SNP, com o eixo X mostrando sua localização no cromossomo e o eixo Y a magnitude de seu efeito.

Todas as rotinas computacionais dos métodos foram implementadas no software R (R Development Core Team, 2010) utilizando os pacotes e as funções apresentados na Tabela 1.

Tabela 1 – Pacotes e funções do *software* R utilizados nas análises.

Metodologias	**Pacotes**	**Funções**
RR-BLUP	*rrBLUP*	*mixed.solve*
Quadrados Mínimos Parciais	*pls*	*plsr*
Componentes Principais	*pls*	*pcr*
Componentes Independentes	*caret*	*icr*

3.RESULTADOS E DISCUSSÕES

As estimativas das correlações genéticas entre os pares de características de carcaça são apresentados na Tabela 2. As estimativas das correlações genéticas entre espessura de toucinho e as outras características de carcaça (RCARC e MLC) obtiveram valores negativos, o qual garante, possivelmente, eficiência na seleção.

Dentre as características, as que apresentaram maiores estimativas de herdabilidade (Tabela 2) foram às características relacionadas à espessura de toucinho, sugerindo um possível progresso no melhoramento genético das mesmas.

O rendimento de carcaça apresentou estimativa de herdabilidade próxima ao valor encontrado por Catalan (1986) que foi de 0,25 em suínos Landrace, Large White e Duroc. As quatro medidas de espessuras de toucinho evidenciaram herdabilidades que variam de 0,10 a 0,42, valores similares aos reportados por

Mendonça *et al.* (2012), utilizando também suínos em uma população F_2 Piau × linhagem comercial. Enquanto, a estimativa de herdabilidade obtida para o comprimento de carcaça assemelhou-se a encontrada por Oliveira *et al*. (1997).

Tabela 2 – Estimativas de herdabilidade e de correlações genéticas para as características de carcaça.

Caract.	RCARC	MLC	ETSH	ETUC	ETUL	ETL	ETO	EBACON
RCARC	**0,20**	-0,35	-0,25	-0,35	-0,24	-0,32	-0,13	-0,12
MLC		**0,17**	-0,28	-0,43	-0,25	-0,26	-0,50	-0,61
ETSH			**0,10**	0,55	0,71	0,79	0,60	0,50
ETUC				**0,35**	0,64	0,60	0,63	0,58
ETUL					**0,36**	0,88	0,60	0,46
ETL						**0,33**	0,66	0,55
ETO							**0,42**	0,80
EBACON								**0,34**

As estimativas de herdabilidade são apresentadas na diagonal; as estimativas de correlações genéticas são apresentadas acima da diagonal.

RCARC = rendimento de carcaça (%); MLC = comprimento de carcaça pelo Método Americano (cm); ETSH = maior espessura de toucinho na região da copa, na linha dorso-lombar (mm); ETUC = espessura de toucinho imediatamente após a última costela, na linha dorso-lombar (mm); ETUL = espessura de toucinho medida entre a última e a penúltima vértebra lombar (mm); ETL= menor espessura de toucinho na região acima da última vértebra lombar, na linha dorso-lombar (mm); ETO = espessura de toucinho medida imediatamente após a última costela, a 6,5 cm da linha dorso-lombar (mm); EBACON= espessura de bacon (mm).

Conforme apresentado, optou-se pelo número de variáveis latentes que explicassem 70% da variável X. Desta forma, o número determinado para o método PLS, PCR e ICR foi, em média, de 47 componentes (redução de 80% da dimensionalidade dos dados originais), 34 (redução de 85% da dimensionalidade), 34 (redução de 85% da dimensionalidade) componentes, respectivamente.

As acurácias obtidas pelos métodos de redução dimensional, para cada característica de carcaça, bem como pelo método RR-BLUP são apresentadas na Tabela 3. Tendo em vista esses resultados, o método RR-BLUP mostrou-se eficiente para as características MLC e ETL, enquanto o método ICR mostrou-se adequado para ETSH, ETUC, ETUL e EBACON. O método PCR se destacou para o caráter

ETO, porém mostrou resultado similar ao RR-BLUP na característica restante, RCARC.

Desse modo, o método que mais se destacou nessas análises foi o ICR. A principal vantagem observada em relação aos demais métodos de redução dimensional é que o mesmo considera a independência entre os componentes, que além de garantir a ausência de relação linear entre as variáveis latentes também garante a retirada de qualquer dependência funcional não linear.

Tabela 3 – Acurácias obtidas pelos métodos considerando cada característica de carcaça.

Características	**Acurácia**			
	PLS	**PCR**	**ICR**	**RR-BLUP**
RCARC	0,20	0,29	0,24	**0,30**
MLC	0,05	0,21	0,15	**0,25**
ETSH	0,11	0,35	**0,45**	0,34
ETUC	0,01	0,47	**0,51**	0,29
ETUL	0,03	0,27	**0,39**	0,27
ETL	0,03	0,42	0,45	**0,47**
ETO	0,11	**0,37**	0,29	0,24
EBACON	0,08	0,40	**0,48**	0,24

RCARC = rendimento de carcaça (%); MLC = comprimento de carcaça pelo Método Americano (cm); ETSH = maior espessura de toucinho na região da copa, na linha dorso-lombar (mm); ETUC = espessura de toucinho imediatamente após a última costela, na linha dorso-lombar (mm); ETUL = espessura de toucinho medida entre a última e a penúltima vértebra lombar (mm); ETL= menor espessura de toucinho na região acima da última vértebra lombar, na linha dorso-lombar (mm); ETO = espessura de toucinho medida imediatamente após a última costela, a 6,5 cm da linha dorso-lombar (mm); EBACON= espessura de bacon (mm).

Por outro lado, o método PLS apresentou os mais baixos valores de acurácia para todas as características de carcaça, possivelmente, porque o método em seu desenvolvimento não contempla a dependência existente entre as variáveis explicativas, que neste caso são os marcadores SNPs.

Sob o enfoque da seleção genômica, embora não existam relatos na literatura a respeito da utilização dos métodos de redução dimensional em melhoramento de suínos , em outras espécies os métodos PLS e PCR já foram utilizados, e de alguma

forma os valores de acurácia reportados para os mesmos podem ser usados como referência. Moser *et al.* (2009) realizou um estudo de comparação entre cinco métodos para dados de gado leiteiro, dentre eles o PLS e RR-BLUP, que obtiveram acurácias próximas para os dois métodos resultados diferentes aos encontrados no presente estudo. Por outro lado, Solberg *et al.* (2009) realizou um estudo de comparação entre os métodos PLS e PCR e observou acurácias similares para as metodologias, 0,47 e 0,45, discordando dos valores obtidos nesse trabalho em que o PCR foi extremamente superior ao PLS. No entanto, os valores de acurácia encontrados por Macciotta *et al.*(2010b) para o PCR foram similares (0.28-0.46) aos resultados encontrados.

As acurácias esperadas (Tabela 4) foram obtidas considerando o tamanho efetivo da população, o tamanho dos cromossomos (em Morgans), o número de locos e a herdabilidade de cada característica, conforme apresentado por Resende (2008) e aplicado por Grattapaglia e Resende (2010).

Tabela 4 – Acurácias esperadas e os desvios estimados para cada característica de carcaça considerando cada método.

Características	**Acurácia Esperada**	**PLS**	**PCR**	**ICR**	**RR-BLUP**
RCARC	0,40	0,20	0,11	0,16	**0,10**
MLC	0,38	0,33	0,17	0,23	**0,13**
ETSH	0,30	0,19	-0,05	**-0,15**	-0,04
ETUC	0,51	0,50	0,04	**0,00**	0,22
ETUL	0,52	0,49	0,25	**0,13**	0,25
ETL	0,50	0,47	0,08	0,05	**0,03**
ETO	0,54	0,43	**0,17**	0,25	0,30
EBACON	0,50	0,42	0,10	**0,02**	0,26

RCARC = rendimento de carcaça (%); MLC = comprimento de carcaça pelo Método Americano (cm); ETSH = maior espessura de toucinho na região da copa, na linha dorso-lombar (mm); ETUC = espessura de toucinho imediatamente após a última costela, na linha dorso-lombar (mm); ETUL = espessura de toucinho medida entre a última e a penúltima vértebra lombar (mm); ETL= menor espessura de toucinho na região acima da última vértebra lombar, na linha dorso-lombar (mm); ETO = espessura de toucinho medida imediatamente após a última costela, a 6,5 cm da linha dorso-lombar (mm); EBACON= espessura de bacon (mm).

O método ICR obteve menor desvio médio (0,09) entre as acurácias observadas e as esperadas (Tabela 4) evidenciando que o mesmo foi o método, em média, que mais se aproximou do esperado e do ideal, considerando as informações da população e das características. Enquanto, o método PLS apresentou o maior desvio médio (0,38) mostrando-se novamente ser um método não eficiente para a predição de valores fenotípicos.

A eficiência matemática obtida pelos métodos de redução de dimensionalidade em relação ao RR-BLUP (método tradicionalmente aplicado a GWS) considerando cada característica de carcaça é apresentada na Tabela 5. Em média, os métodos PCR e ICR demonstraram 19% e 26% a mais de eficiência ao serem comprados com o método RR-BLUP, respectivamente. Por outro lado, o PLS novamente demonstrou resultados inferiores aos demais, mostrando-se 75% menos eficiente, em média, ao RR-BLUP.

Tabela 5 – Eficiência matemática obtida pelos métodos de redução de dimensionalidade em relação ao RR-BLUP, considerando cada característica de carcaça.

Características	**Eficiência Matemática**		
	PLS	**PCR**	**ICR**
RCARC (%)	0,67	**0,97**	0,80
MLC (cm)	0,20	**0,84**	0,60
ETSH (mm)	0,32	1,03	**1,32**
ETUC (mm)	0,03	1,62	**1,76**
ETUL (mm)	0,11	1,00	**1,44**
ETL (mm)	0,06	0,89	**0,96**
ETO (mm)	0,30	**1,54**	1,21
EBACON (mm)	0,33	1,67	**2,00**

RCARC = rendimento de carcaça (%); MLC = comprimento de carcaça pelo Método Americano (cm); ETSH = maior espessura de toucinho na região da copa, na linha dorso-lombar (mm); ETUC = espessura de toucinho imediatamente após a última costela, na linha dorso-lombar (mm); ETUL = espessura de toucinho medida entre a última e a penúltima vértebra lombar (mm); ETL= menor espessura de toucinho na região acima da última vértebra lombar, na linha dorso-lombar (mm); ETO = espessura de toucinho medida imediatamente após a última costela, a 6,5 cm da linha dorso-lombar (mm); EBACON= espessura de bacon (mm).

Tabela 6 – Estimativas das correlações dos GBVs (Valores Genéticos Genômicos) provenientes dos métodos para as características de carcaça.

Características	Métodos	PLS	PCR	ICR	RR-BLUP
RCARC(%)	PLS	1	0,35	0,35	0,50
	PCR		1	0,97	0,87
	ICR			1	0,87
MLC(cm)	PLS	1	0,34	0,33	0,57
	PCR		1	0,95	0,87
	ICR			1	0,87
ETSH(mm)	PLS	1	0,35	0,35	0,50
	PCR		1	0,97	0,87
	ICR			1	0,87
ETUC(mm)	PLS	1	0,47	0,46	0,63
	PCR		1	0,98	0,90
	ICR			1	0,90
ETUL(mm)	PLS	1	0,48	0,47	0,64
	PCR		1	0,98	0,89
	ICR			1	0,89
ETL(mm)	PLS	1	0,44	0,43	0,61
	PCR		1	0,98	0,91
	ICR			1	0,91
ETO(mm)	PLS	1	0,43	0,42	0,63
	PCR		1	0,96	0,88
	ICR			1	0,88
EBACON(mm)	PLS	1	0,46	0,46	0,61
	PCR		1	0,95	0,89
	ICR			1	0,89

RCARC = rendimento de carcaça (%); MLC = comprimento de carcaça pelo Método Americano (cm); ETSH = maior espessura de toucinho na região da copa, na linha dorso-lombar (mm); ETUC = espessura de toucinho imediatamente após a última costela, na linha dorso-lombar (mm); ETUL = espessura de toucinho medida entre a última e a penúltima vértebra lombar (mm); ETL= menor espessura de toucinho na região acima da última vértebra lombar, na linha dorso-lombar (mm); ETO = espessura de toucinho medida imediatamente após a última costela, a 6,5 cm da linha dorso-lombar (mm); EBACON= espessura de bacon (mm).

O PLS é o método que mais se destoa dos demais, uma vez que as correlações dos resultados de acurácia (Tabela 3) obtidos entre ele e os métodos PCR, ICR e RR-BLUP, respectivamente, foram -0,38, -0,57 e -0,11. Por outro lado, os mais

concordantes foram os métodos PCR e ICR, com correlação de 0.864, pois o ICR pode ser considerado uma extensão do método PCA, além disso, para determinar o número de componentes do primeiro método é necessário realizar uma PCA.

De acordo com as estimativas das correlações dos valores genéticos genômicos estimados obtidos pelos métodos (Tabela 6), continua a observar que a maior correlação é encontrada entre os métodos ICR e PCR. Além disso, mais uma vez o método PLS apresenta menores correlações com os demais métodos. Também vale ressaltar, que as correlações entre RR-BLUP e ICR e PCR foram altas, variando de 0,87 a 0,91.

A identificação de marcadores de grandes efeitos é de suma importância para a seleção genômica, pois possibilita determinar a posição desses SNPs a fim de verificar a existência de QTLs que afetam o caráter quantitativo nessas regiões. Assim, a identificação dos marcadores de maiores efeitos para cada característica foi realizada por meio das estimativas dos parâmetros do modelo de regressão (efeitos de SNPs) do método que obteve maior valor de acurácia. A fim de facilitar tal identificação, os gráficos Manhattan foram confeccionados e estão dispostos na Figura 1.

A característica que apresentou maior comportamento poligênico foi o caráter associados à espessura de toucinho ETL, uma vez que os efeitos se distribuíram de forma uniforme ao longo dos cromossomos. Por outro lado, a característica para a qual esse comportamento se tornou menos evidente foi o RCARC. Quanto aos métodos de redução de dimensionalidade ICR e PCR, não foi possível de forma expressiva discriminar os efeitos de marcadores.

O SNP de maior efeito (Figura 1) para a característica de espessura de toucinho ETSH encontra-se no terço inicial do cromossomo 4. Bem como, Gonçalves *et al.* (2005) que detectou dois QTLs significativos, nessa mesma região para espessura de toucinho, em suínos resultantes do cruzamento entre machos da raça Meishan e fêmeas Large White e Landrace.

Enquanto, os SNPs de maior efeito (Figura 1), nas características comprimento de carcaça MLC e espessura de toucinho ETUC, foram encontrados no terço final do cromossomo 7, concordando com os resultados reportados por Nezer *et al*. (2002), que encontrou QTL na região 99 cM, e Sousa *et al.* (2011), respectivamente. Também neste mesmo cromossomo, Silva *et al.*(2011) detectou QTL significativo em uma população F_2 (Piau ×linhagem comercial) para a

característica EBACON. Esse resultado foi concordante ao encontrado no presente estudo que apresentou SNPs mais relevantes nesta região.

No terço inicial do cromossomo 8 (Figura 1), encontra-se alguns dos SNPs mais significativos para a característica ETO. Assim como, Bidanel *et al.* (2001) que analisaram uma população F2 de Meishan e Large White e encontraram QTL significativo nessa mesma região pra essa característica. Também no cromossomo 8 encontrou-se SNPs relevantes para a característica ETUL tal como Hidalgo *et al.* (2011) que detectou um QTL nessa região cromossômica.

As estimativas dos efeitos dos marcadores das características em que o método ICR se mostrou mais eficiente foram comparadas aos efeitos estimados pelo método tradicional RR-BLUP (Figura 2), visando verificar a eficiência do mesmo. O gráfico das estimativas dos efeitos de SNPs para a característica ETUL no método ICR apresentou algumas divergências se comparado ao obtido por meio do método RR-BLUP. O gráfico do ICR apresentou efeitos de marcadores de maior magnitude além de discriminar os SNPs de maior relevância, enquanto o do RR-BLUP distribuiu os efeitos de SNPs de forma uniforme ao longo do cromossomo.

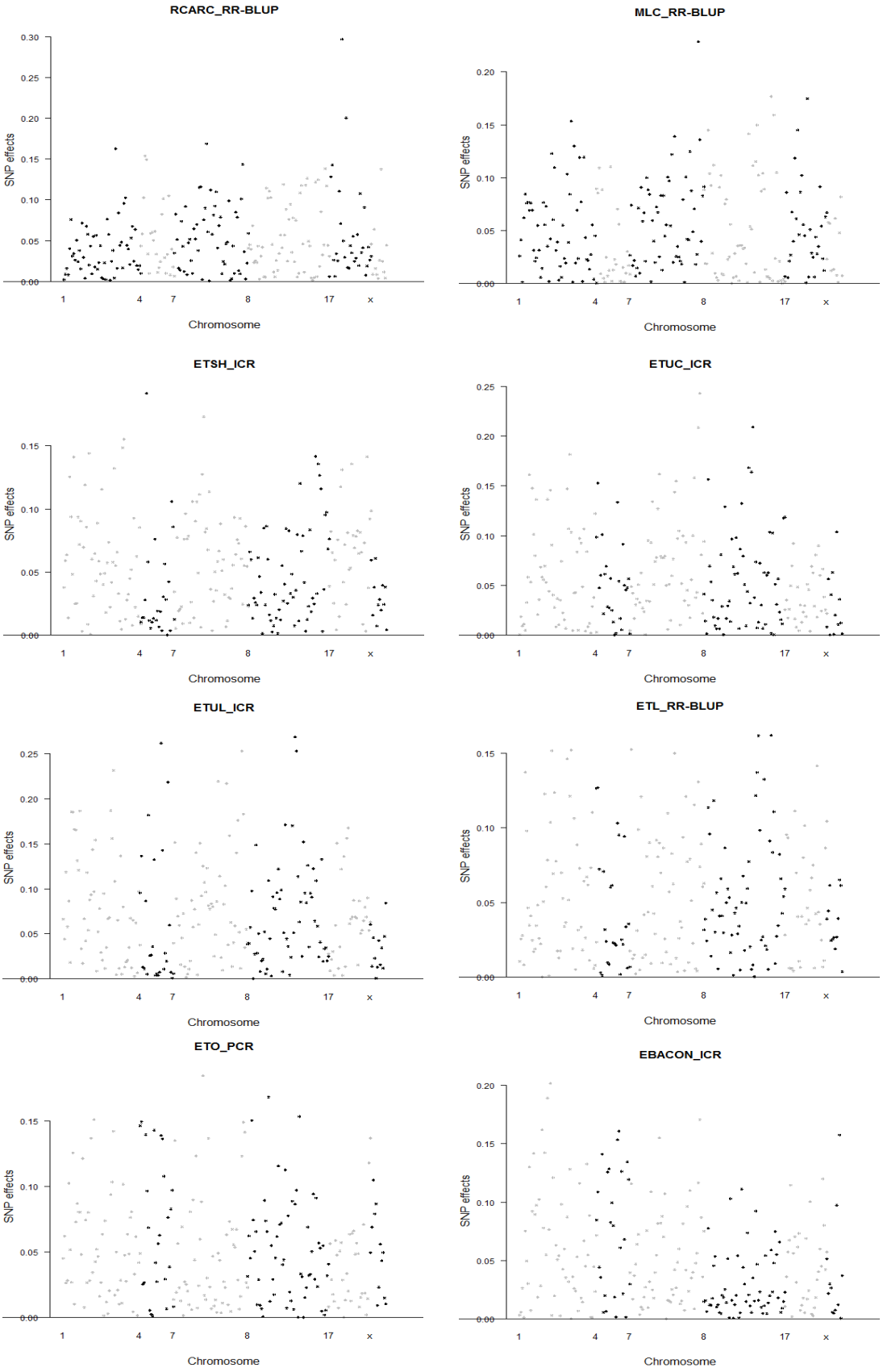

Figura 1- *Manhattan plot* dos efeitos de marcadores padronizados considerando o método mais acurado para cada característica.

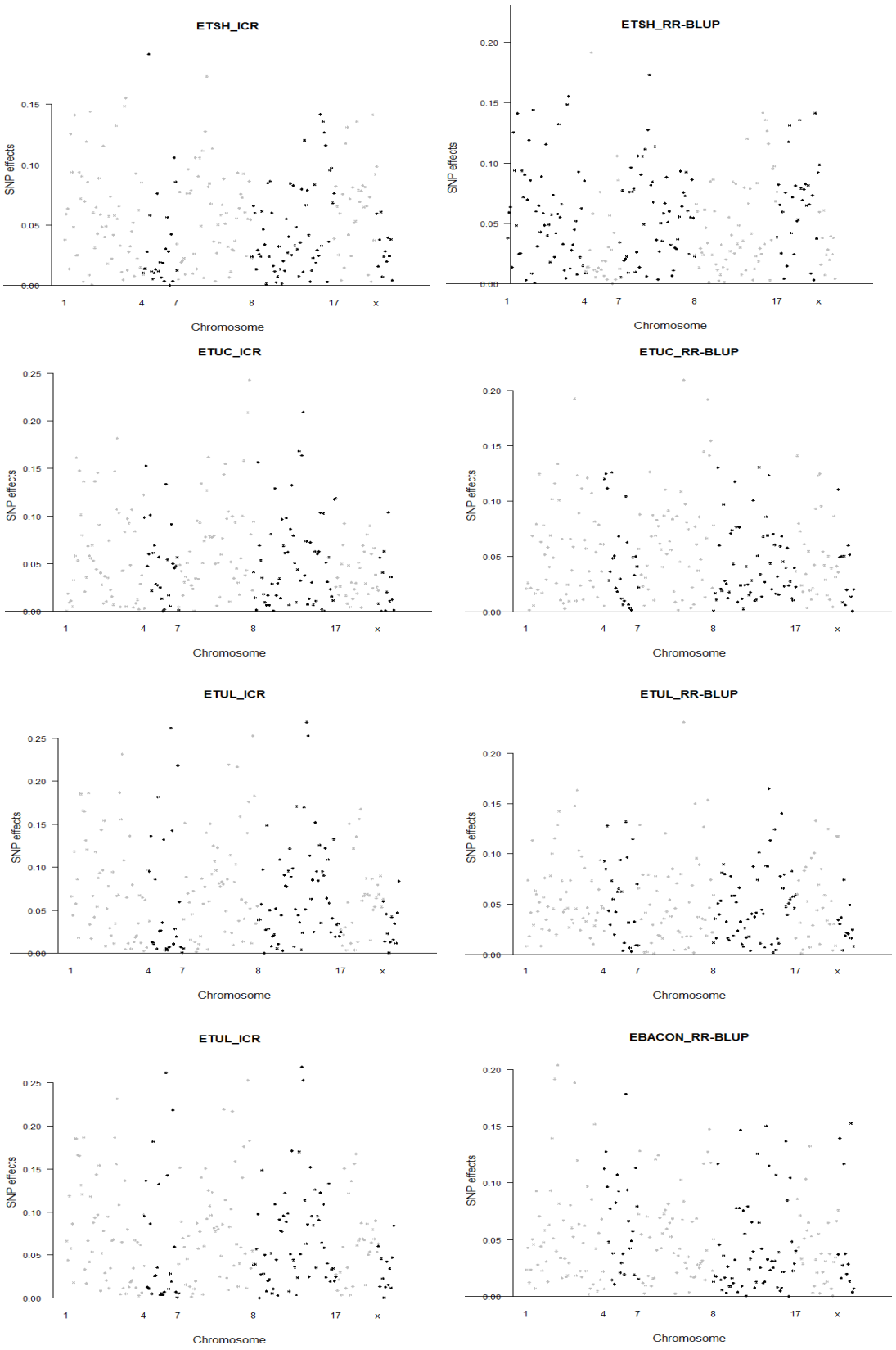

Figura 2- *Manhattan plot* dos efeitos de marcadores padronizados estimados por meio dos métodos ICR e RR-BLUP considerando apenas as características que o método ICR se mostrou eficiente.

CONCLUSÕES

De forma geral, nesse estudo, o método que apresentou maiores valores de acurácia e se mais mostrou eficiente matematicamente para a predição de valores fenotípicos foi o método proposto ICR. Por outro lado, o método PLS apresentou valores baixos de acurácia e mostrou-se ineficiente para a predição, além de não se destacar para nenhuma característica considerada.

Vale ressaltar também que, em sua maioria, os métodos possibilitaram a identificação de marcadores SNPs de maior relevância. Além disso, esses SNPs estão localizados em regiões cromossômicas previamente relatadas em outros estudos como sendo regiões relacionadas a presença de QTLs que afetam características de carcaça em suínos.

Os métodos que apresentaram maiores similaridades foram o ICR e o PCR. Por outro lado, o método que apresentou valores mais discrepantes em relação aos demais foi o PLS.

5. REFERÊNCIAS BIBLIOGRÁFICAS

BIDANEL, J.P.; MILAN, D.; IANNUCCELLI, N.; AMIGUES, Y.; BOSCHER, M. Y.; BOURGEOIS, F.; CARITEZ, J. C.; GRUAND, J.; LE ROY, P.; LAGANT, H.; QUINTANILLA, R.; RENARD, C.; GELLIN, J.; OLLIVIER, L.; CHEVALET, C. Detection of quantitative trait loci for growth and fatness in pigs. **Genetics Selection Evolution**, v.33, p.289-309, 2001.

BOULESTEIX, A. L.; STIMMER, K. Partial least squares: a versatile tool for the analysis of high–dimensional genomic data. **Briefings in Bioinformatics**, v. 8, p. 32-44, 2006.

CADAVID, A. C.; LAWRENCE, J. K.; RUZMAIKIN, A.; KAYLENG–KNIGHT.Principal Components and Independent Component Analysis of Solar and Space Data. **Solar Phys**, v. 248, p. 247-261, 2008.

CATALAN, G. **Estimativa de parâmetros genéticos e fenotípicos em suínos Landrace, Large White e Duroc, nas fases de crescimento e terminação**. Dissertação (Mestrado em Zootecnia) - Universidade Federal de Viçosa, 129p.,1986.

GARTHWAITE, P.H. An Interpretation of Partial Least Squares. **Journal of the American Statistical Association**, v. 89, p. 122-127, 1994.

GRATTAPAGLIA, D.; RESENDE, M. D. V. Genomic selection in forest tree breeding. **Tree Genetics and Genomes**, Oct. 2010.

GIANOLA, D.; PEREZ–ENCISO, M.; TORO, M. A. On marker–assisted prediction of genetic value: beyond the ridge. **Genetics**,v.163, p. 347-365, 2003.

GILMOUR, A.R.; CULLIS, B.R.; WELHAM, S.J. THOMPSON, R. ASREML reference manual. **Orange: NSW Agriculture**, p. 218, 2000.

GODDARD, M. E.; HAYES, B. J. Genomic selection. **Journal of Animal Breeding and Genetics**, v. 124, p. 323-330, 2007.

GONÇALVES, T. de M.; OLIVEIRA, H. N. de; BOVENHUIS, H.; BINK, M.; ARENDONK, J. V. Modelos alternativos para detecção de locos de características quantitativas (QTL) de carcaça e crescimento nos cromossomos 4, 5 e 7 de suínos. **Revista Brasileira de Zootecnia**, v. 34, p. 1540-1552, 2005.

HIDALGO, A. M. et al. **Fine mapping and single nucleotide polymorphism effects estimation on pig chromossomes 1,4,7,8,17 and x**. Dissertation (Genetics and Breeding) - University Federal of Viçosa, 2011.

HOSKULDSSON, P. PLS Regression Methods. **Journal of Chemometrics**, v. 2, p. 211-228, 1988.

HYVÄRINEN, A. New approximations of differential entropy for independent component analysis and projection pursuit. **In Advances in Neural Information Processing Systems,** v. 10, p. 273-279, 1998.

HYVÄRINEN, A.; OJA, E. Independent Component Analysis: Algorithms and Applications. **Neural Networks**, v. 13, p. 411-430, 2000.

MACCIOTA N. P. P.; PINTUS, M. A.; STERI, R.; PIERAMATI, C.; NICOLAZZI, E. L., SANTUS, E.; VICARIO, D.; VAN KAAM, J. T.; NARDONE, A.; VALENTINI, A; AJMONE-MARSAN, P. Accuracies of direct genomic breeding values estimated in dairy cattle with a principal component approach. **Journal of Dairy Science**, v. 93, 532-533, 2010b.

MENDONÇA, P. T.; LOPES, P. S.; BRACCINI NETO, J.; CARNEIRO, P. L. S.; TORRES, R. de A.; GUIMARÃES, S. E. F.; VERONEZE, R. Estimação de parâmetros genéticos de uma população F_2 de suínos. **Revista Brasileira de Saúde e Produção Animal**, v. 13, p. 330-343, 2012.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome wide dense marker maps. **Genetics**, v. 157, p. 1819-1829, 2001.

MOSER, G.; TIER, B.; CRUMP, R. E. ; KHATKAR, M. S.; RAADSMA, H. W. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. **Genetics Selection Evolution**, v.41, p. 41-53, 2009.

NEZER, C.; MOREAU, L.; WAGENAAR, D.; GEORGES, M. Results of a whole genome scan targeting QTL for growth and carcass traits in a Pietrain × Large White intercross. **Genetics Selection Evolution**, v.34, p.371-387, 2002.

OLIVEIRA, A.I.G. **Aspectos genéticos das características físicas das carcaças de suínos em cruzamentos dialélicos**. Tese (Doutorado em Zootecnia) - Universidade Federal de Viçosa, 97 p., 1988.

OTTO, M. **Chemometrics**. Weinheim: Wiley, 328 p., 1999.

PEIXOTO, J. O.; GUIMARAES, S. E. F.; LOPES, P. S.; SOARES, M. A. M.; PIRES, A. V.; SILVA, M. V.; TORRES, R. A.; SILVA, M. A. E. Associations of leptin gene polymorphisms with production traits in pigs. **Journal of Animal Breeding and Genetics**, v. 123, p. 378-383, 2006.

PIRES, A. V. ; LOPES, P. S. ; GUIMARÃES, S. E. F. Mapeamento de Locos de Características Quantitativas (QTL) no Cromosomo 6 de Suínos, Associados às Características de Carcaça e de Órgão Internos. **Revista da Sociedade Brasileira de Zootecnia**, v. 35, p. 1660-1668, 2006.

R DEVELOPMENT CORE TEAM. *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2010.Disponivel em: <http://www.R–project.org>.

RESENDE, M. D. V. Seleção genômica ampla (GWS) e modelos mistos. In: **Matemática e estatística na análise de experimentos e no melhoramento genético**. Colombo: Embrapa Florestas, p. 517-534, 2007.

RESENDE, M. D. V. **Genômica quantitativa e seleção no melhoramento de plantas perenes e animais**. Colombo: Embrapa Florestas, 330 p., 2008.

RESENDE, M. D. V.; RESENDE JUNIOR, M. F. R.; AGUIAR, A. M.; ABAD, J. I. M.; MISSIAGIA, A. A.; SANSALONI, C.; PETROLI, C.; GRATTAPALIA, D. **Computação da seleção genômica ampla (GWS)**. Colombo: Embrapa Florestas, 79p, 2010.

RUSSO, V.; FONTANESI, L.; SCOTTI, E.; BERETTI, F.; DAVOLI, R.; COSTA, L. N.; VIRGILI, R; BUTTAZZONI, L. Single nucleotide polymorphisms in several porcine cathepsin genes are associated with growth, carcass, and production traits in Italian Large White pigs. **Journal of Animal Science**, v. 86, p. 3300-3314, 2008.

SILVA, F. F.; ROSA, G. J. M.; GUIMARÃES, S.E.F.; LOPES, P.S; DE LOS CAMPOS, G. Three-step Bayesian factor analysis applied to QTL detection in crosses between outbred pig populations. **Livestock Science**, v.42, p. 210-215, 2011.

SILVA, L. R. da. **Aplicação da Decomposição em Valores Singulares e Análise de Componentes Independentes em dados de fMRI**. Dissertação, Universidade Federal do Pará, 52 p., 2011.

SOLGERG, T. R.; SONESSON, A. K.; WOOLLIAMS, J. A.; MEUWISSEN, T. H. E. Reducing dimensionality for prediction of genome-wide breeding values. **Genetics Selection Evolution**, v. 41, p. 29, 2009.

SOUSA, K. R. S.; GUIMARÃES, S. E. F.; SILVA FILHO, M. I. da; LOPES, M. S.; PINTO, A. P. G.; VERARDO, L. L.; BRACCINI NETO, J.; LOPES, P. S. Mapeamento de locos de características quantitativas nos cromossomos 5, 7 e 8 de suínos. **Revista Brasileira de Zootecnia**, v. 40, p. 115-123, 2011.

APÊNDICE A – Estimativas dos coeficientes originais da regressão dos Métodos de Redução Dimensional

Um problema nos métodos de regressão de Redução Dimensional é que as equações de predições são expressas por componentes, assim, os coeficientes obtidos nesse modelo não possuem interpretação biológica. Deste modo, torna-se necessário a determinação dos coeficientes da regressão original. Para tanto, algumas considerações devem ser feitas.

Regressão via Componentes Principais

No método PCR, os componentes principais são definidos como combinações lineares das variáveis aleatórias X (I x J), ou seja,

$$\hat{Z}=X\hat{P} \quad . \qquad (1)$$

sendo P a matriz de autovetores da matriz de covariância de X e Z a matriz de componentes.

Por outro lado, a equação de predição expressa por componentes é dada por:

$$\hat{y}=\hat{Z}\hat{\alpha} . \qquad (2)$$

Desta forma, levando-se a definição (1) na equação (2), obtém-se:

$$\hat{y}=X\hat{P}\hat{\alpha} . \qquad (3)$$

Comparando a equação (3) com a equação de predição original $\hat{y}=X\hat{m}_{pcr}$, tem-se que $\hat{m}_{pcr}=\hat{P}\hat{\alpha}$ se, e somente se, X for uma matriz inversível.

Entretanto, na seleção genômica X não é inversível, pois X é retangular (número de marcadores maior que indivíduos genotipados). Deste modo, tem-se $\hat{P}\hat{\alpha}$ como sendo um candidato para a estimativa dos coeficientes originais m_{pcr} .

Quadrados Mínimos Parciais

O método PLS, consiste em decompor simultaneamente a matriz X e o vetor y, desconsiderando os erros inerentes ao processo (e) tem-se as equações a seguir:

$$X=TL' \qquad (4)$$

$$y=Uq' \qquad (5)$$

em que T e U são as matrizes de componentes, L é matriz de carregamento ortogonal e q é o vetor de carregamento ortogonal.

As matrizes de componentes T e U são regressados e dão origem a equação dada por:

$$\hat{U}=T\hat{B} \tag{6}$$

Combinando as equações (5) e (6), obtém-se a seguinte relação:

$$\hat{y}=T\hat{B}q'. \tag{7}$$

A matriz L é uma matriz ortogonal, assim por definição tem-se que:

$$L'L=I, \tag{8}$$

sendo I a matriz de identidade.

Considerando a definição (8) e rearranjando-a na equação (4), a matriz de componentes T pode ser expressa por:

$$T=XL. \tag{9}$$

Desta forma, levando-se a equação (9) na equação de predição (7), obtém-se:

$$\hat{y}=XL\hat{B}q'. \tag{10}$$

Comparando a equação (9) com a equação de predição original $\hat{y}=X\hat{m}_{pls}$, tem-se que $\hat{m}_{pls}=L\hat{B}q'$ se, e somente se, X for uma matriz inversível.

Entretanto, na seleção genômica X não é inversível, pois X é retangular (número de marcadores maior que indivíduos genotipados). Deste modo, tem-se $L\hat{B}q'$ como sendo um candidato para a estimativa dos coeficientes originais m_{pls}.

Regressão via Componentes Independentes

O método ICR decompõe a matriz X em duas matrizes, ou seja,

$$X'=A'S' \tag{11}$$

em que A é uma matriz de misturas e S é uma matriz de componentes independentes.

Os algoritmos de ICA por sua vez, determinam os componentes independentes, como sendo:

$$\hat{S}=XKR \tag{12}$$

em que K é a matriz de branqueamento e R a matriz que maximiza a independência estatística das colunas da matriz S.

A equação de predição da regressão ICR expressa por componentes é dada por:

$$\hat{y}=S\hat{\gamma}. \quad (13)$$

Desta forma, levando-se a equação (12) na equação de predição (13), obtém-se:

$$\hat{y}=XKR\hat{\gamma}. \quad (14)$$

Comparando a equação (13) com a equação de predição original $\hat{y}=X\hat{m}_{icr}$, tem-se que $\hat{m}_{icr}=KR\hat{\gamma}$ se, e somente se, X for uma matriz inversível.

Entretanto, na seleção genômica X não é inversível, pois X é retangular (número de marcadores maior que indivíduos genotipados). Deste modo, $KR\hat{\gamma}$ como sendo um candidato para a estimativa dos coeficientes originais m_{icr}.

APÊNDICE B– Rotinas computacionais implementadas

As rotinas computacionais dos métodos foram implementadas no software R (R Development Core Team, 2010) e estão descritas a seguir.

Regressão via Componentes Principais

```
######################## Pacotes ######################
library(pls)
################## Leitura dos dados ##################
dados=read.table("all_gws.txt",h=T)

######################## RCARC ########################
Yrcarc=as.matrix(dados$rcarc_d+ mean(dados$RCARC))
nc_rcarc=34
snp_rcarc=t(matrix(pcr(as.matrix(Yrcarc)~as.matrix(dados[,–
(1:75)]))$coefficients[,,nc_rcarc]))
gbv_rcarc=as.matrix(dados[,–(1:75)])%*%t(snp_rcarc)
```

Quadrados Mínimos Parciais

```
######################## Pacotes ######################
library(pls)

################## Leitura dos dados ##################
dados=read.table("all_gws.txt",h=T)

######################## RCARC ########################
Yrcarc=as.matrix(dados$rcarc_d+ mean(dados$RCARC))

nc_rcarc=47
snp_rcarc=t(matrix(plsr(as.matrix(Yrcarc)~as.matrix(dados[,–
(1:75)]))$coefficients[,,nc_rcarc]))
gbv_rcarc=as.matrix(dados[,–(1:75)])%*%t(snp_rcarc)
```

Regressão via Componentes Independentes

```
###################### Pacotes ####################
library(caret)
library(fastICA)

################ Leitura dos dados #################
dados=read.table("all_gws.txt",h=T)

###################### RCARC ######################

nc_rcarc=47
model_rcarc=icr(as.matrix(dados[,–(1:75)]),as.matrix(dados$rcarc_d+
mean(dados$RCARC)),n.comp=nc_rcarc)
Ainv_rcarc= model_rcarc$ica$ica$K%*% model_rcarc$ica$ica$W
betaS_rcarc=model_rcarc$model$coefficients[–1]
alfa_rcarc=Ainv_rcarc%*%betaS_rcarc
gbv_rcarc=as.matrix(dados[,–(1:75)])%*%alfa_rcarc
```

Método RR-BLUP

```
###################### Pacotes ####################
library(rrBLUP)

################ Leitura dos dados #################
dados=read.table("all_gws.txt",h=T)

############## RCARC – RRBLUP ######################

Yrcarc=as.matrix(dados$rcarc_d+ mean(dados$RCARC))
snp_rcarc=t(mixed.solve(Yrcarc,Z=as.matrix(dados[,–(1:75)]))$u)
gbv_rcarc=as.matrix(dados[,–(1:75)])%*%t(snp_rcarc)
```