FILIPE RIBEIRO FORMIGA TEIXEIRA

ANÁLISE DE FATORES APLICADA NA SELEÇÃO GENÔMICA 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
2015

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
T266a
2015

Teixeira, Filipe Ribeiro Formiga, 1989-
Análise de fatores aplicada na seleção genômica em suínos / Filipe Ribeiro Formiga Teixeira. – Viçosa, MG, 2015.
xi, 63f. : il. (algumas color.) ; 29 cm.

Inclui anexo.
Inclui apêndices.
Orientador: Moysés Nascimento.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Suíno - Melhoramento genético - Métodos estatísticos. 2. Teoria bayesiana de decisão estatística. 3. Análise de variância. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 631.4082

FILIPE RIBEIRO FORMIGA TEIXEIRA

Análise de fatores aplicada na seleção genômica 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 fevereiro de 2015

Ana Carolina Campana Nascimento (Coorientadora)	Débora Martins Paixão

Talles Eduardo Ferreira Maciel

Moysés Nascimento
(Orientador)

DEDICO

À minha família.

AGRADECIMENTOS

Agradeço primeiramente a Deus por ter me dado força e determinação para mais essa conquista, além de ter sempre posto coisas muito boas no meu caminho.

Aos meus pais, Manuel Benício e Luíza Helena por estarem sempre presentes e atuantes na minha vida, prestando sempre o suporte necessário em todos os aspectos, pelo incentivo que recebo todo dia, por me propiciarem uma boa educação e formação e por serem, acima de tudo, meus exemplos.

À minha irmã Mariana, pela companhia, conselhos, amizade, apoio e por se manter próxima a mim mesmo morando longe.

A toda a minha família, incluindo tios, primos e avós, que eu sei que estão sempre na torcida e que fazem muita falta no meu cotidiano.

Aos meus grandes amigos que considero como irmãos José Luíz, Lúcio, Ithalo e Eduardo pela amizade e parceria.

À minha querida Anamaria, por compartilhar comigo os momentos difíceis e momentos de alegria, por ser companheira, amiga e por tornar meus dias ainda melhores, mesmo estando longe.

Ao meu orientador Moysés Nascimento pelos ensinamentos, pela paciência, preocupação e por ter sempre me incentivado desde os primeiros dias de curso. Agradeço também por ser meu grande amigo e uma pessoa com quem posso contar, além de um exemplo como pessoa e como profissional.

Aos meus coorientadores Ana Carolina Campana Nascimento, Fabyano Fonseca e Silva e Cosme Damião Cruz, e também à professora Camila Azevedo, por contribuírem diretamente no meu aprendizado e pelas sugestões nos trabalhos até aqui realizados.

Aos professores e funcionários do departamento de estatística da Universidade Federal de Viçosa, que sempre se empenharam e se mostraram acessíveis, dispostos a compartilharem conhecimento e suporte para com os alunos.

Aos membros da banca, Ana Carolina Campana Nascimento, Débora Martins Paixão e Talles Eduardo Ferreira Maciel por aceitarem o convite e por estarem dispostos a dar suas contribuições para este trabalho.

Aos meus mestres e amigos da Universidade Federal do Piauí, pelo empenho que sempre tiveram para tornar cada vez melhor o curso de graduação, e por me ensinarem os primeiros conceitos da estatística.

Aos meus amigos de Viçosa Matheus, Lázaro, Rodolfo, Mayra, Laís e André pela amizade e pelo suporte, principalmente nos primeiros dias em que morei em Viçosa.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria pela oportunidade.

À CAPES, pela concessão da bolsa de estudos.

Por fim, a todos os que contribuíram direta ou indiretamente para a concretização deste trabalho.

BIOGRAFIA

FILIPE RIBEIRO FORMIGA TEIXEIRA, filho de Luíza Helena Ribeiro Formiga Teixeira e Manuel Benício Teixeira Neto, nasceu em Teresina, Piauí, em 20 de novembro de 1989.

Em março de 2009, ingressou no curso de Bacharelado em Estatística na Universidade Federal do Piauí, Teresina – PI, graduando-se em maio de 2013.

Em agosto do mesmo ano, iniciou o curso de mestrado do Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa da dissertação em 26 de fevereiro de 2015.

SUMÁRIO

RESUMO..... ix
ABSTRACT..... xi
1. INTRODUÇÃO GERAL..... 1
2. REVISÃO DE LITERATURA..... 4
2.1.Análise de fatores..... 4
2.1.1. Definição..... 4
2.1.2. Modelo da análise de fatores..... 4
2.1.3. Adequabilidade da matriz de correlações..... 5
2.1.4. Estimação dos escores fatoriais..... 7
2.1.5. Rotação Varimax..... 7
2.2. Seleção genômica ampla..... 8
2.3. Inferência bayesiana..... 9
2.4. Inferência bayesiana na seleção genômica ampla..... 10
2.4.1. Bayes A..... 10
2.4.2. Bayes B..... 11
2.4.3. RR-BLUP Bayes..... 12
2.4.4. LASSO bayesiano..... 12
2.5. Estimação do mérito genético, capacidade preditiva, herdabilidade e acurácia.. 13
2.6. Validação cruzada..... 14
2.7. Coeficiente Cohen's Kappa..... 14
REFERÊNCIAS BIBLIOGRÁFICAS..... 16
CAPÍTULO 1: Determinação de fatores em características de suínos..... 19
RESUMO..... 19
1. Introdução..... 20
2. Material e métodos..... 21
3. Resultados e discussão..... 24
4. Conclusões..... 28
REFERÊNCIAS BIBLIOGRÁFICAS..... 29
CAPÍTULO 2: Análise de fatores aplicada na seleção genômica em suínos..... 31
RESUMO..... 31
1. Introdução..... 32

2. Material e métodos..... 33
3. Resultados e discussão..... 39
4. Conclusões..... 44
REFERÊNCIAS BIBLIOGRÁFICAS..... 45
CONSIDERAÇÕES FINAIS..... 48
APÊNDICES..... 49
Apêndice A – Algoritmos utilizados no estudo..... 49
Apêndice B – Tabela 2: Loadings obtidos na análise de fatores..... 59
Apêndice C – Resultados da acurácia para os demais fatores..... 61
ANEXOS..... 63
Anexo I: Teorema de Bayes..... 63

RESUMO

TEIXEIRA, Filipe Ribeiro Formiga, M.Sc., Universidade Federal de Viçosa, fevereiro de 2015. **Análise de fatores aplicada na seleção genômica em suínos.** Orientador: Moysés Nascimento. Coorientadores: Ana Carolina Campana Nascimento, Cosme Damião Cruz e Fabyano Fonseca e Silva.

A seleção genômica, ou seleção genômica ampla foi proposta com a finalidade de otimizar o processo de melhoramento genético utilizando simultaneamente dados fenotípicos e genotípicos por meio de marcadores SNP's, presentes em todo o genoma. Várias metodologias têm sido implementadas, principalmente utilizando regressão linear e considerando os efeitos dos marcadores fixos (BLUP, LS, etc.) ou considerando os efeitos aleatórios (Bayes A, Bayes B, LASSO Bayesiano, dentre outras). Em geral, tais metodologias analisam cada característica individualmente, ou seja, os resultados obtidos são válidos apenas para uma única variável. Entretanto, em programas de melhoramento o interesse recai em ganhos para mais de uma característica conjuntamente. Dessa forma, desenvolver uma abordagem que trabalhe com análises que considerem várias características simultaneamente pode ser interessante, visto que poderíamos estudar um conjunto de caracteres importantes conjuntamente. Uma metodologia possível de ser utilizada para este fim é a análise fatorial (ou análise de fatores - AF). Tal metodologia permite a obtenção de variáveis latentes (fatores comuns) que representam um conjunto das variáveis originais. A partir de então, análises posteriores podem ser realizadas utilizando as variáveis latentes criadas. Diante do exposto, o principal objetivo deste trabalho foi propor a utilização da análise de fatores para criação de variáveis latentes altamente associadas às variáveis fenotípicas originais, de modo que possamos estimar o mérito genético para os indivíduos considerando várias variáveis simultaneamente. Objetivou-se também comparar os resultados obtidos com aqueles advindos das análises individuais. Para tanto, foram utilizados dados fenotípicos e genotípicos provenientes de 345 suínos obtidos pelo cruzamento das raças Piau e Comercial, oriundos da Granja de Melhoramento de Suínos do Departamento de Zootecnia da Universidade Federal de Viçosa (UFV), no período de novembro de 1998 a julho de 2001, utilizando 237 marcadores SNP's e 41 variáveis fenotípicas. No primeiro capítulo foi aplicada a análise de fatores para verificar a estrutura de associação entre as variáveis e averiguar a que fator cada variável pertence, onde são identificados os fatores interpretáveis. No segundo capítulo, técnicas

bayesianas de seleção genômica ampla foram aplicadas nas variáveis latentes interpretáveis para predição do mérito genético e seleção dos indivíduos, comparando assim os resultados com o que foi encontrado para as variáveis fenotípicas individualmente. Os resultados obtidos indicam que a aplicação da análise de fatores para obtenção de variáveis latentes que representam um conjunto de caracteres para posterior uso em seleção genômica ampla se mostrou eficiente, visto que apresentou valores de acurácia semelhantes aos encontrados considerando as análises individuais das variáveis e alta concordância entre os indivíduos selecionados considerando a variável latente e as variáveis individuais.

ABSTRACT

TEIXEIRA, Filipe Ribeiro Formiga, M.Sc., Universidade Federal de Viçosa, february, 2015. **Factor analysis applied to genomic selection in pigs.** Advisor: Moysés Nascimento. Co-advisors: Ana Carolina Campana Nascimento, Cosme Damião Cruz and Fabyano Fonseca e Silva.

Genomic selection, or genome-wide selection was proposed in order to optimize breeding process using both phenotypic and genotypic data by SNP's markers, present throughout the genome. Various methodologies have been implemented mostly using linear regression and considering the effects of fixed markers (BLUP, LS, etc.) or considering the random effects (The Bayes, Bayes B, Bayesian LASSO, among others). In general, such methods analyze each feature individually, or the results obtained are valid only for a single variable. However, in breeding programs interest falls in earnings to more than one feature together. Thus, developing an approach that works with analyzes that consider various features can be simultaneously interesting, because we could study a number of important characters together. One possible method to be used for this purpose is the factor analysis (or analysis of factors - AF). This methodology allows obtaining latent variables (common factors) that represent a set of original variables. Since then, further analysis can be performed using the latent variables created. Given the above, the main objective of this work was to propose the use of factor analysis for creating highly latent variables associated with the original phenotypic variables, so that we can estimate the genetic merit for individuals considering several variables simultaneously. The objective is to also compare the results with those arising from the individual reviews. Therefore, phenotypic and genotypic data from 345 pigs obtained by crossing the breeds were used Piau and Trade, coming from the Farm Improvement Pigs Department of Animal Science of the Federal University of Viçosa (UFV), from November 1998 to July 2001, using 237 markers SNP's and 41 phenotypic variables. In the first chapter was applied to factor analysis to assess the association structure between variables and determine the factor which each variable belongs, where interpretable factors are identified. In the second chapter, Bayesian techniques of genome-wide selection were applied in interpretable latent variables to predict genetic merit and selection of individuals, and compare it to what was found for the phenotypic variables individually. The results indicate that the application of the factor analysis to obtain latent variables representing a character set

for later use in genome wide selection proved to be efficient, since accuracy showed similar values to those found considering the analysis of individual variable and high agreement among the individuals selected considering the latent variable and individual variables.

1. INTRODUÇÃO GERAL

A carne suína é uma das mais consumidas do mundo, e o Brasil é um dos países que apresentam destaque em relação à produção e exportação deste tipo de carne segundo a ABPA (Associação Brasileira de Proteína Animal). Investimentos na suinocultura posicionaram o Brasil em quarto lugar no ranking de produção e exportação (10% do volume total) mundial de carne suína. De acordo com o MAPA (2014), nos últimos vinte anos, com a evolução genética da espécie, houve uma redução de 31% da gordura da carne, 10% do colesterol e 14% de calorias, e sua exportação acarreta em um lucro de cerca de US$ 1 bilhão por ano.

Devido à importância econômica da suinocultura no cenário mundial e nacional e com a evolução das metodologias utilizadas no melhoramento genético, pesquisas direcionadas a este ramo têm se tornado cada vez mais frequentes.

Dentre algumas publicações de trabalhos voltados para o melhoramento genético de suínos, podemos destacar análises de associação genômica (Genome-wide association) como, por exemplo, a realizada por Fan et al. (2011), em que foram encontrados QTL's (Quantitative trait loci) para características de toucinho e área do olho do lombo a partir de informações contidas em mais de 50.000 marcadores SNP's em uma população de 820 fêmeas comerciais. PAIXÃO et al. (2012) objetivaram o mapeamento de QTL's relacionados a características de carcaça e de qualidade baseado em uma população F2 de 684 suínos desenvolvida pelo cruzamento de dois reprodutores da raça brasileira Piau com 18 fêmeas comerciais (Landrace x Large White x Pietrain) utilizando 35 marcadores microssatélites. Já AZEVEDO et al. (2013) buscaram comparar, em termos de acurácia, a regressão PLS (Partial Least Squares, ou Mínimos Quadrados Parciais) univariada (UPLS) e multivariada (MPLS), técnicas de redução de dimensionalidade, na seleção genômica ampla (SGA) para características de carcaça em uma população F2 de suínos Piau x Comercial, obtendo melhores resultados para a abordagem multivariada.

Dentre as diversas abordagens apresentadas, a aplicada por Azevedo et al. (2013), denotada por Seleção Genômica Ampla (SGA), ou Genome Wide Selection (GWS) (Meuwissen et al., 2001) apresenta grande destaque, visto que tal metodologia possibilita incorporar informações do genoma diretamente na predição do mérito genético dos indivíduos, permitindo assim alta eficiência seletiva, grande rapidez na

obtenção dos ganhos genéticos com a seleção e baixo custo, em comparação com a tradicional seleção baseada em dados fenotípicos (RESENDE, 2012).

Apesar dos estudos de SGA apresentarem resultados satisfatórios, em geral, as análises levam em consideração apenas uma variável resposta de cada vez, ou seja, quando se deseja analisar um conjunto de variáveis, são realizadas várias análises univariadas. Deste modo, uma metodologia que pudesse trabalhar com uma quantidade menor de variáveis latentes que representam um conjunto de variáveis originais pode ser interessante, visto que possibilitaria a estimação do mérito genético e seleção para diversos caracteres simultaneamente e poderia reduzir o tempo computacional das análises.

Um método que visa criar variáveis latentes associadas às características originais é a análise de fatores (AF), ou análise fatorial. Este procedimento tem como principal objetivo descrever a variabilidade original do vetor aleatório **X**, em termos de um número menor de variáveis latentes, chamadas fatores comuns e que estão relacionadas com o vetor original **X** através de um modelo linear (MINGOTI, 2007). Visando a verificação da existência de QTL's para características de carcaça de suínos simultaneamente, Silva et al. (2011) utilizaram esta técnica, sob o enfoque bayesiano, e obtiveram resultados satisfatórios.

Uma alternativa semelhante seria a utilização de variáveis latentes em seleção genômica, visto que a sua utilização além de possibilitar estimação dos valores genéticos genômicos e a seleção de indivíduos para um conjunto de caracteres simultaneamente, simplificaria a análise e reduziria o tempo computacional para obtenção dos resultados, já que em geral as metodologias utilizadas na SGA demandam bastante tempo computacional.

Diante do exposto, este trabalho teve como objetivo propor a utilização de análise de fatores em estudos de seleção genômica visando estimar os efeitos de marcadores SNP's e assim acessar os valores genéticos genômicos (EGBV's) sobre uma variável latente que represente um conjunto de variáveis. Além disso, objetivou-se comparar os resultados com análises fenotípicas individuais.

Este trabalho está dividido basicamente em três partes: revisão de literatura, capítulo 1 e capítulo 2.

Na revisão de literatura, foram abordados conceitos que envolvem análise multivariada, seleção genômica ampla e inferência bayesiana, bem como definições e modelos estatísticos, que serão utilizados posteriormente nos capítulos 1 e 2.

No capítulo 1 foi realizada uma aplicação da análise de fatores para um conjunto de variáveis fenotípicas de interesse, com o intuito de reduzir a quantidade de variáveis a serem analisadas a uma quantidade menor de variáveis latentes que possuam interpretação prática e que possam ser nomeadas.

Ademais, no capítulo 2 foram apresentadas aplicações de metodologias bayesianas de seleção genômica ampla aos fatores interpretáveis, onde foram também feitas comparações entre as acurácias encontradas para o fator e para cada variável altamente associada ao mesmo.

Por fim, são apresentadas as conclusões e considerações finais.

2. REVISÃO DE LITERATURA

2.1. Análise de fatores

2.1.1. Definição

A análise fatorial exploratória teve seus primeiros conceitos sugeridos por Galton em 1888, quando apresentou métodos de regressão e coeficientes de correlação. Posteriormente, por volta de 1904, Spearman propôs a atual modelagem da estrutura fatorial em estudos de testes de escores na inteligência humana (FERREIRA, 2011).

Essa técnica tem como objetivo descrever a variabilidade original do vetor aleatório **X** (variáveis) em termos de um número menor, m, de variáveis aleatórias, chamadas de fatores comuns e que estão relacionadas com o vetor de variáveis **X** através de um modelo linear (MINGOTI, 2007).

Com a aplicação dessa metodologia, espera-se que as variáveis originais sejam agrupadas em subconjuntos, de novas variáveis não correlacionadas, denominados fatores ou variáveis latentes. Cada fator está altamente associado a um conjunto de variáveis semelhantes, podendo representar por si só esse grupo de caracteres.

2.1.2. Modelo da análise de fatores

O modelo fatorial adotado para uma variável X_i observável, com média μ_i pode ser representado da seguinte forma (FERREIRA, 2011):

$$X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1$$
$$\vdots$$
$$X_i - \mu_i = l_{i1}F_1 + l_{i2}F_2 + \cdots + l_{im}F_m + \varepsilon_i$$
$$\vdots$$
$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p$$

em que $i = 1,2,\dots,p$ e $m \leq p$, sendo p o número de variáveis originais observáveis; o coeficiente l_{ij} é chamado de carga fatorial da i-ésima variável sobre o j-ésimo fator comum, sendo $j = 1,2,\dots,m; F_1, F_2, \dots, F_m$ são denominados fatores comuns, variáveis aleatórias não observáveis e ε_i são os erros aleatórios que estão associados somente a i-ésima variável X_i, respectivamente.

Utilizando notação matricial, o mesmo modelo pode ser representado da seguinte maneira (FERREIRA, 2011):

$$\boldsymbol{Y} - \boldsymbol{\mu} = \boldsymbol{\Gamma F} + \boldsymbol{\epsilon}$$

em que $\boldsymbol{\Gamma}$ é uma matriz pxm composta por coeficientes conhecidos como cargas fatoriais, que medem a associação entre cada variável e os fatores; $\boldsymbol{F}$ é o vetor $mx1$ que contém os escores fatoriais, os quais são os valores assumidos pelas novas variáveis latentes (não observáveis) estimadas por meio de uma metodologia específica e $\boldsymbol{\epsilon}$ é um vetor $px1$ que representa os erros aleatórios. A estimação da matriz $\boldsymbol{\Gamma}$ e do vetor de escores $\boldsymbol{F}$ segue nas seções posteriores.

No contexto da análise de fatores, são feitas algumas suposições adicionais acerca dos fatores comuns e das variáveis originais, como $E(\boldsymbol{Y}) = \mu$, $E(\boldsymbol{F}) = E(\boldsymbol{\epsilon}) = 0$, $Cov(\boldsymbol{F}) = \boldsymbol{I}_m$, $Cov(\boldsymbol{Y}) = \boldsymbol{\Sigma}$, $Cov(\boldsymbol{\epsilon}) = \boldsymbol{\psi}$ e $Cov(\boldsymbol{F}, \boldsymbol{\epsilon}) = 0$, sendo $\boldsymbol{\psi}$ uma matriz identidade, dada por:

$$\boldsymbol{\psi} = \begin{bmatrix} \psi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \psi_p \end{bmatrix}$$

2.1.3. Adequabilidade da matriz de correlação

Antes de aplicar a análise de fatores, é conveniente verificar a adequabilidade da matriz de correlação das variáveis em estudo. Um conjunto de variáveis que apresente uma matriz de correlações com altos valores certamente será mais adequado para aplicação, visto que as variáveis se agrupam de acordo com suas correlações. Com o objetivo de verificar essa adequabilidade, podemos destacar o critério de Kaiser-Meyer-Olkin (KMO) e o teste de esfericidade de Bartlett. O primeiro se refere a um índice proposto por Kaiser (1970), que pode apresentar valores entre 0 e 1. Esse coeficiente é dado por (MINGOTI, 2007):

$$KMO = \frac{\sum_{i \neq j} R_{ij}^2}{\sum_{i \neq j} R_{ij}^2 + \sum_{i \neq j} Q_{ij}^2}$$

em que R_{ij}e Q_{ij} são, respectivamente, as correlações amostral e parcial entre as variáveis i e j. A correlação parcial entre duas variáveis se trata da correlação que existe entre elas quando todas as outras $(p-2)$ variáveis são consideradas como constantes (KACHIGAN, 1991; JOHNSON& WICHERN, 2002). As correlações amostral (R_{ij}) e parcial (Q_{ij}) são dadas abaixo (RENCHER, 2008):

$$R_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

$$Q_{ij} = Q_{ij.rs\ldots q} = \frac{\sigma_{ij.rs\ldots q}}{\sqrt{\sigma_{ii.rs\ldots q}\sigma_{jj.rs\ldots q}}}$$

em que σ_{ij} é a covariância entre as variáveis i e j, e $\sigma_{ij.rs\ldots q}$ é a covariância entre as variáveis i e j sem o efeito das demais $(r, s, \ldots, q)$. Quanto maior o valor do índice KMO, melhor será a adequabilidade. A classificação segundo RENCHER (2002) segue abaixo (Tabela 1).

Tabela 1. Classificação do índice KMO.

KMO	Adequabilidade
0,9 – 1,0	Excelente
0,8 – 0,9	Ótima
0,7 – 0,8	Boa
0,6 – 0,7	Regular
0,5 – 0,6	Ruim
Abaixo de 0,5	Inadequado

O teste de esfericidade de Bartlett para matriz de correlação busca verificar se a matriz de correlações se aproxima da matriz identidade. Assim, possui as seguintes hipóteses:

$$\begin{cases} H_0: R_{pxp} = I_{pxp} \\ H_1: R_{pxp} \neq I_{pxp} \end{cases}$$

em que R_{pxp} é a matriz de correlação das p variáveis e I_{pxp} uma matriz identidade de ordem p. A estatística do teste é definida por (MINGOTI, 2007):

$$T = -\left[n - \frac{1}{6}(2p + 11)\right]\left[\sum_{j=1}^{p} \ln(\hat{\lambda}_i)\right]$$

em que $\ln(\hat{\lambda}_i)$ é o logaritmo neperiano do i-ésimo autovalor da matriz de correlação amostral R_{pxp}. A estatística T, sob a hipótese nula e para grandes amostras, possui uma distribuição aproximadamente qui-quadrado com $\frac{1}{2}p(1-p)$ graus de liberdade. Existe adequabilidade dos dados quando rejeitamos a hipótese nula.

2.1.4. Estimação dos escores fatoriais

Após a verificação da adequabilidade da matriz de correlação, deve-se obter a matriz dos loadings ($\widehat{\Gamma}$), conhecidos também como cargas fatoriais, e as unicidades ($\hat{\psi}_{pxp}$), matriz diagonal contendo as variâncias específicas. Suas estimativas podem ser obtidas por meio do método dos componentes principais (MINGOTI, 2007), aplicado com base na decomposição espectral da matriz de correlação, como segue:

$$\widehat{\Gamma} = \left[\sqrt{\hat{\lambda}_1}\hat{e}_1\sqrt{\hat{\lambda}_2}\hat{e}_2 \ldots \sqrt{\hat{\lambda}_m}\hat{e}_m\right] = \begin{bmatrix} l_{11} & \cdots & l_{1m} \\ \vdots & \ddots & \vdots \\ l_{p1} & \cdots & l_{pm} \end{bmatrix}$$

e

$$\hat{\psi}_{pxp} = diag(R_{pxp} - \widehat{\Gamma}\widehat{\Gamma}^T)$$

em que $\hat{\lambda}_i$ é o i-ésimo autovalor da matriz de correlação R_{pxp}.

Uma vez estimados $\widehat{\Gamma}$ e $\hat{\psi}$, podemos obter os escores fatoriais, que são os valores assumidos pelas novas variáveis (não observáveis). Usualmente utiliza-se o método da regressão (FERREIRA, 2011), onde a estimação dos escores para o j-ésimo indivíduo é dado por:

$$\hat{F}_j = \widehat{\Gamma}^T \left(\widehat{\Gamma}\widehat{\Gamma}^T + \hat{\psi}\right)^{-1} (Y_j - \bar{Y})$$

em que $\widehat{\Gamma}_{pxm}$ é a matriz dos loadings, $\hat{\psi}$ a matriz das unicidades, que representa o erro aleatório, Y_j é o vetor referente aos valores assumidos pelo conjunto de variáveis do j-ésimo indivíduo ($j = 1,2,\ldots,n$) e $\bar{Y}$ o vetor de médias referente as variáveis avaliadas.

A alocação das variáveis em cada fator é feita através dos loadings (l_{ij}). Assim, quanto maior o valor do loading (em módulo), mais relacionada a variável fenotípica será com o respectivo fator.

2.1.5. Rotação Varimax

Após determinarmos os escores fatoriais, juntamente com a matriz $\widehat{\Gamma}_{pxm}$, a ideia é encontrar uma rotação ortogonal dos fatores com a finalidade de facilitar a interpretação (FERREIRA, 2011), buscando uma estrutura mais simples para a matriz de cargas fatoriais. A aplicação da rotação se trata de uma transformação na matriz dos loadings de modo que simplifique a identificação das variáveis latentes.

O critério Varimax foi proposto por Kaiser (1958) e têm sido o critério de rotação fatorial mais usual, visto que esse método tem como objetivo buscar fatores com grandes variabilidades nos loadings, ou seja, para cada fator, identificar um grupo de variáveis altamente correlacionadas e outro grupo que tenha correlação desprezível ou moderada (MINGOTI, 2007). Utilizando esse procedimento, para cada fator fixo, a solução é obtida a partir da maximização da variação (v) dos quadrados dos loadings da matriz $\hat{\Gamma}$. O valor de v é definido por (FERREIRA, 2011):

$$v = \frac{1}{p^2}\sum_{j=1}^{m}\left[p\sum_{i=1}^{p}x_{ij}^4 - \left(\sum_{i=1}^{p}x_{ij}^2\right)^2\right]$$

Sendo $x_{ij} = \frac{l_{ij}}{\sqrt{\sum_{j=1}^{m} l_{ij}^2}}$ a ij-ésima carga dividida pela raíz quadrada da sua respectiva comunalidade, que é a proporção de variância de cada variável que é explicada pelas variáveis latentes.

2.2. Seleção Genômica Ampla

A Seleção Genômica Ampla (SGA), ou Genome Wide Selecion (GWS) foi proposta por Meuwissen et al. (2001) com o objetivo de utilizar informações diretas do DNA na seleção e predição do mérito genético, de forma a permitir alta eficiência seletiva, grande rapidez na obtenção de ganhos genéticos com a seleção de baixo custo em comparação com a seleção tradicional baseada em dados fenotípicos (RESENDE, 2012). A seleção genômica ampla é fundamentada em marcadores moleculares, que têm efeitos estimados a partir de uma metodologia específica e aplicados para predição do mérito genético e seleção de animais, vegetais, logo após o nascimento, pois identifica através de marcadores moleculares os alelos associados a uma determinada característica de interesse, permitindo a seleção precoce.

Atualmente, os marcadores moleculares mais usuais são os SNP's (polimorfismo de um único nucleotídeo, ou single nucleotide polymorphisms), pois é a forma mais abundante de variação do DNA nos genomas, sendo preferidos em relação a outros marcadores genéticos devido a sua baixa taxa de mutação e facilidade de genotipagem, aliados ao baixo custo (RESENDE, 2012). Esses marcadores se distribuem em todo o genoma em grande quantidade, o que facilita a detecção de desequilíbrios de ligação que estão associados a uma determinada característica de interesse.

Dentre as diversas metodologias aplicadas em seleção genômica, aquelas baseadas em inferência bayesiana apresentam destaque e são utilizadas em diversos estudos (MEUWISSEN et al., 2001; FAN et al., 2011). No trabalho realizado por Meuwissen et al. (2001) foram propostos os métodos bayesianos Bayes A e Bayes B e comparados com o BLUP (Best Linear Unbiased Predictor) (HENDERSON, 1974) e LS (Least Squares) (LANDE & THOMPSON, 1990), obtendo-se melhores resultados em termos de acurácia. Fan et al. (2011) também utilizaram metodologias bayesianas para análise de associação genômica, obtendo resultados satisfatórios.

Na próxima seção será introduzida a definição de inferência bayesiana e posteriormente, os conceitos referentes a algumas das suas aplicações no contexto da seleção genômica ampla.

2.3. Inferência Bayesiana

Diferentemente da inferência frequentista, na estimação de parâmetros através da inferência bayesiana o parâmetro de interesse θ é considerado uma variável aleatória, e não uma constante desconhecida. Sob este enfoque, admite-se um modelo de probabilidade para os dados amostrais (função de verossimilhança) e também é suposto que o(s) parâmetro(s) de interesse se distribui(em) de acordo com algum modelo probabilístico conhecido (distribuições à priori). O principal objetivo da inferência bayesiana é encontrar a distribuição à posteriori, a qual é utilizada para obter uma estimativa pontual do parâmetro de interesse. Seguem abaixo os procedimentos utilizados na determinação da distribuição à posteriori.

Primeiramente, assume-se uma distribuição de probabilidade para os dados $(Y_1, Y_2, \dots, Y_n)$, obtendo-se a partir desta a função de verossimilhança $f(\boldsymbol{y}|\theta)$, sendo θ o parâmetro de interesse. Em seguida, admite-se que o(s) parâmetro(s) se distribua(m) de acordo com algum modelo à priori $\pi(\theta)$. A distribuição à posteriori dos dados, denotada por $\pi(\theta|\boldsymbol{y})$ depende dos dados amostrais e pode ser calculada da seguinte maneira, fundamentada no teorema de Bayes (Anexo I):

$$\pi(\theta|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\theta)\pi(\theta)}{\int_{\Theta} f(\boldsymbol{y}|\theta)\pi(\theta)d\theta}$$

$$\pi(\theta|\boldsymbol{y}) \propto f(\boldsymbol{y}|\theta)\pi(\theta)$$

A inferência bayesiana está ligada a ferramentas computacionais, visto que em algumas situações podemos lidar com problemas matemáticos complicados dependendo da distribuição à priori e da função de verossimilhança. Dentre os métodos computacionais mais conhecidos, podemos destacar o amostrador de Gibbs e algoritmo de Metrópolis-Hastings, via cadeias de Markov (MCMC), que tem por finalidade de gerar amostras aleatórias da distribuição à posteriori. Maiores detalhes podem ser vistos em GAMERMAN (2006).

2.4. Métodos Bayesianos na Seleção Genômica

Os procedimentos bayesianos foram introduzidos na seleção genômica ampla por Meuwissen et al. (2001) a partir dos métodos Bayes A e Bayes B. Tais ferramentas foram comparadas aos procedimentos usuais BLUP (Best Linear Unbiased Predictor) (HENDERSON, 1974) e LS (Least Squares) (LANDE & THOMPSON, 1990) e os resultados foram satisfatórios, visto que apresentaram maiores acurácias.

O modelo geral para seleção genômica, proposto por Meuwissen et al. (2001) é dado por:

$$\boldsymbol{y} = 1\mu + \boldsymbol{X\beta} + \boldsymbol{e},$$

em que $\boldsymbol{y}$ é o vetor coluna que contém os valores da variável Y para cada indivíduo; 1 é o vetor coluna composto por 1's de dimensão $nx1$; $\boldsymbol{\mu}$ é a média da variável y; $\boldsymbol{\beta}$ é o vetor que contém os efeitos dos marcadores SNP's; $\boldsymbol{X}$ é a matriz de incidência que relaciona efeitos de SNP's aos valores da característica $\boldsymbol{y}$; e $\boldsymbol{e}$ é o vetor dos erros.

Na prática, a principal diferença entre os métodos bayesianos na SGA está nas suposições a respeito das distribuições à priori dos efeitos de marcadores (β_i) e das suas respectivas variâncias $(\sigma^2_{\beta i})$. A seguir são descritas, de maneira sucinta, algumas metodologias e suas conjunturas.

2.4.1. Bayes A

No método Bayes A (MEUWISSEN et al., 2001) considera-se as variâncias dos marcadores heterogêneas, e os efeitos de marcadores são estimados considerando a informação combinada da função de verossimilhança (obtida através dos dados) e das distribuições à priori para as variâncias. A informação combinada pode ser obtida pelo

método do amostrador de Gibbs, via cadeias de Markov (MCMC). Os efeitos dos marcadores (β_i) são assumidos como amostras de uma distribuição normal com média zero e variância de cada marcador ($\sigma^2_{\beta i}$) dada por uma distribuição qui-quadrada inversa e escalonada (RESENDE, 2012), constituindo uma estrutura hierárquica em dois níveis, como segue abaixo, juntamente com a variância aditiva (σ^2_u):

$$\beta_i|\sigma^2_{\beta i} \sim N(0, \sigma^2_{\beta i})$$

$$\sigma^2_{\beta i} \sim \chi^{-2}(v_\beta, S^2_\beta)$$

$$\sigma^2_u = 2\sum_{i=1}^{m} p_i(1-p_i)\sigma^2_i,$$

em que v_β representa os graus de liberdade, S^2_β é o parâmetro da escala de distribuição e p_i e $(1-p_i)$ representam as frequências alélicas. Segundo Meuwissen et al. (2001) devemos considerar $v_\beta = 4{,}012$ ou $4{,}2$ e $S^2_\beta = 0{,}002$ ou $0{,}0429$. Sob o enfoque bayesiano, essas pressuposições indicam que um grande número de marcadores apresenta efeitos pequenos e poucos marcadores apresentam efeitos grandes (RESENDE, 2011).

2.4.2. Bayes B

O método Bayes B assume que um número de marcadores (com proporção π) tem efeito zero, e o restante dos marcadores, com proporção $1-\pi$, é amostrado com uma variância individual para cada marcador, considerando as mesmas prioris usadas no método Bayes A (CRUZ, 2013). Uma adversidade desse método é escolher o melhor valor para π, o que requer um conhecimento prévio sobre o genótipo a ser analisado ou uma metodologia específica para estimá-lo. Este procedimento se equivale ao Bayes A quando $\pi = 0$ (RESENDE, 2011). A estimação dos efeitos e variâncias dos marcadores e variância aditiva é dada abaixo:

$$\beta_i|\pi, \sigma^2_{i1} \sim N(0, \sigma^2_{\beta i})$$

$$\sigma^2_{\beta i} = 0 \text{ com probabilidade } \pi$$

$$\sigma^2_{\beta i} \sim \chi^{-2}(v_\beta, S^2_\beta) \text{ com probabilidade } 1-\pi$$

$$\sigma^2_u = 2\sum_{i=1}^{m} p_i(1-p_i)\sigma^2_i$$

em que v_β e S_β^2 assumem os mesmos valores do método Bayes A, citados anteriormente, baseados no estudo de Meuwissen et al. (2001) e p_i e $(1-p_i)$ são as frequências alélicas.

2.4.3. Bayesian Ridge Regression

A diferença essencial entre o método Bayesian Ridge Regression e o Bayes A é a suposição de homogeneidade das variâncias dos marcadores SNP's. Aqui, temos apenas um valor estimado para σ^2, e o modelo também é ajustado com estrutura hierárquica em dois níveis. Os parâmetros de efeito de SNP's, variância dos marcadores e variância aditiva seguem abaixo:

$$\beta_i|\sigma^2 \sim N(0, \sigma^2)$$
$$\sigma^2 \sim \chi^{-2}(v, S^2)$$
$$\sigma_u^2 = 2\sigma^2 \sum_{i=1}^{m} p_i(1-p_i).$$

em que os parâmetros v e S^2 são equivalentes a v_β e S_β^2 dos métodos Bayes A e Bayes B e p_i e $(1-p_i)$ denotam as frequências alélicas.

2.4.4. LASSO Bayesiano

No método LASSO (Least Absolute Shrinkage and Selection Operator) (TIBSHIRANI, 1996), o objetivo é encontrar os efeitos dos marcadores através da resolução da seguinte equação (CAMPOS et al., 2009):

$$min_\beta \left\{ \sum (y_i - \boldsymbol{x_i'\beta})^2 \ sujeito\ a \ \sum_j |\beta_j| \leq t \right\}$$

em que $\boldsymbol{x_i'}$ é o vetor das covariáveis; $\boldsymbol{\beta}$ o vetor que contém os coeficientes de regressão e t é uma constante positiva. Assumindo que os dados são centrados, o vetor de observações y_i terá média zero, o que faz a função acima ser equivalente a:

$$min_\beta \left\{ \sum (y_i - \boldsymbol{x_i'\beta})^2 + \lambda(t) \sum_j |\beta_j| \right\}$$

em que possui os mesmos parâmetros da equação anterior, com o acréscimo do parâmetro de encurtamento, ou suavização, λ.

Dada a função acima, uma interpretação bayesiana do LASSO conhecida por LASSO bayesiano, ou simplesmente BLASSO, foi proposta por DE LOS CAMPOS et al. (2009), que estabelece como função de verossimilhança dos dados uma Normal, dada por $\prod_{i=1}^{n} N(y_i|\boldsymbol{x_i'\beta}, \sigma_\varepsilon^2)$ e os efeitos dos marcadores β_i são considerados um produto de exponenciais duplas com média zero, resultando na função $\prod_{j=1}^{p}(\lambda/2)\exp(-\lambda|\beta_j|)$. O parâmetro de suavização λ tem a função de aproximar os efeitos dos marcadores de zero, o que usualmente ocorre, pois a maioria dos SNP's tem efeito nulo. No BLASSO, λ controla a priori assumida por $\boldsymbol{\beta}$, e quanto maior for o valor assumido por este parâmetro, maior será a densidade nos valores próximos de zero.

2.5. Estimação do mérito genético, capacidade preditiva, herdabilidade e acurácia

Os cálculos da capacidade preditiva, herdabilidade e acurácia são baseados no mérito genético dos indivíduos (EGBV's). Tais valores são estimativas do valor real fenotípico, baseados no genótipo e nos efeitos de marcadores estimados pelo modelo, dada pela seguinte expressão:

$$\text{EGBV} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

em que **X** é a matriz de genótipos, composta pelos valores numéricos assumidos pelos SNP's e $\widehat{\boldsymbol{\beta}}$ é o vetor que contém as estimativas dos efeitos dos marcadores, de acordo com a metodologia utilizada.

A capacidade preditiva pode ser definida como o coeficiente de correlação linear entre o valor real observado da variável Y e o valor estimado pelo modelo, denotado por $r_{y,\hat{y}}$. A herdabilidade mede o grau de correspondência entre o fenótipo e o genótipo. Esse coeficiente é expresso pela relação entre a variância genotípica (v_g) e a variância fenotípica (v_{fen}), dado por $h^2 = \frac{v_g}{v_{fen}}$. E por fim, a acurácia é uma das medidas utilizadas para avaliar a qualidade do ajuste do modelo. Essa medida depende da capacidade preditiva e da herdabilidade e é denotada por (RESENDE, 2010):

$$r_{q,\hat{q}} = \frac{r_{y,\hat{y}}}{\sqrt{h^2}}$$

Esse índice varia entre 0 e 1, e quanto maior valor assumir, melhor será o ajuste.

2.6.Validação cruzada

Na prática, em estudos de seleção genômica ampla, três populações podem ser definidas, são elas: população de estimação, validação e seleção (RESENDE, 2012). A população de estimação, conhecida também como população de treinamento ou de descoberta é utilizada para verificar a associação entre os valores assumidos pelos marcadores e as variáveis fenotípicas, onde se obtém as equações de predição dos valores genéticos genômicos.

A população de validação pode ou não ser distinta da população de estimação. Quando são distintas, esse subgrupo constitui uma parcela menor que a população de descoberta, e é utilizado para testar as equações estimadas no procedimento anterior, onde a partir dos resultados verificam-se a capacidade preditiva, herdabilidade e a acurácia. Aplicar a validação e estimar medidas que avaliam o modelo na população de treinamento implica na superestimação dessas medidas, pois neste caso o modelo está sendo estimado e validado com base nos mesmos indivíduos. Assim, a validação cruzada torna-se de grande importância para contornar esse problema.

2.7. Coeficiente Cohen's Kappa

Com o objetivo de verificar a concordância entre duas avaliações independentes, o coeficiente Cohen's Kappa (k) foi proposto por Cohen (1960). Essa metodologia foi utilizada por SANTOS et al. (2012) para avaliar a concordância entre indivíduos selecionados por meio de duas metodologias diferentes na Seleção Genômica Ampla. Diferentemente do coeficiente de concordância simples, este índice leva em conta a probabilidade de a concordância ter ocorrido ao acaso, o que o torna uma medida mais precisa. O cálculo desse coeficiente é feito da seguinte maneira:

$$\hat{k} = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

em que $\Pr(a) - \Pr(e)$ representa a proporção de observações em que a concordância ocorreu além do que se esperava aleatoriamente e $1 - \Pr(e)$, a proporção de observações que não ocorreu concordância. Essa medida varia de 0 a 1 e quanto maior o índice, mais os grupos estão em concordância. A classificação do índice Cohen's Kappa segundo Landis e Koch (1977) é dada a seguir (Tabela 2).

Tabela 2. Classificação do índice Cohen's Kappa.

Índice de Kappa	Classificação
$\hat{k} \leq 0{,}2$	Ruim
$0{,}2 < \hat{k} \leq 0{,}4$	Razoável
$0{,}4 < \hat{k} \leq 0{,}6$	Bom
$0{,}6 < \hat{k} \leq 0{,}8$	Muito bom
$0{,}8 < \hat{k} \leq 1{,}0$	Excelente

Pela classificação de Landis e Koch (1977), o índice de Kappa é considerado satisfatório quando assume valores acima de 0,4.

REFERÊNCIAS BIBLIOGRÁFICAS

ABPA: Associação Brasileira de Proteína Animal. Disponível em: <http://www.abipecs.org.br/pt/estatisticas/mundial/producao-2.html>. Acesso em Nov. 2014.

AZEVEDO, C. F.; RESENDE, M. D. V.; SILVA, F. F.; Lopes, P. S.; GUIMARÃES, S. E. F. Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. **Pesquisa Agropecuária Brasileira**, v. 48, p. 619-626, 2013.

COHEN, J. A coeficient of agreement for nominal scales. **Educational and Phychological Measurement**, v. 20, p. 37-46, 1960.

CRUZ, C. D.; SALGADO, C. C.; BHERING, L. L. **Genômica Aplicada**. Visconde do Rio Branco, MG: Ed. Suprema, 424p., 2013.

DE LOS CAMPOS, G.; NAYA, h.; GIANOLA, D.; CROSSA, J.; LEGARRA, A.; MANFREDI, E.; WEIGEL, K.; COTES, J. M. Predicting Quantitative traits with regression models for dense molecular markers. **Genetics**, Austin, v. 182, p. 375-385, 2009.

FAN, B.; ONTERU, S. K.; DU, Z. Q.; GARRICK, D. J.; STALDER, K. J.; ROTHSCHILD, M. F. Genome-wide association study identifies Loci for composition and sctructural soundness traits in pigs. **PlosOne**. v. 6. p. 1-11. 2011.

FERREIRA, D. F. **Estatística Multivariada**. 2.Ed. Lavras: Ed. UFLA. 675p., 2011.

GAMERMAN, D.; LOPES, H. L. **Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)**. 2. Ed. Hardocover, 2006.

HENDERSON, C. R. Aplications of linear models in animal breeding. University of Guelph, Guelph, 462p., 1984.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. New Jersey: Prentice Hall, 2002. 773p.

KACHIGAN, S. K. Multivariate statistical analysis. New York: Radius Press, 1991.

KAISER, H. F. The Varimax criterion for analytic rotation in factor analysis. **Psychometrika**, Williamsburg, v. 23, n. 3, p. 187-200, 1958.

LANDE, R.; THOMPSON, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. **Genetics**, v. 124, p. 743-756, 1990.

LANDIS, J.; KOCH, G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p.159-174. 1977.

MAPA. Ministério da Agricultura, Pecuária e Abastecimento. 2014. Disponível em: <http://www.agricultura.gov.br/animal/especies/suinos>. Acesso em: Nov. 2014.

MEUWISSEN, T. H. E. Genomic selection: marker assisted selection on genome-wide scale. **Journal of Animal Breeding and Genetics**, v. 124, p. 321-322, 2007.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction os total genetic value using Genome-Wide dense marker maps. **Genetics Society of America**. V. 157. p. 1819-1829, 2001.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: Uma abordagem aplicada**. 1ª reimpressão. Belo Horizonte – MG. Ed. UFMG. 295p., 2007.

PAIXÃO, D. M.; CARNEIRO, P. L. S.; PAIVA, S. R.; SOUSA, K. R. S.; VERARDO, L. L.; BRACCINI NETO, J. ; PINTO, A. P. G.; HIDALGO, A. M.; NASCIMENTO, C. S.; PÉRISSÉ, I. V.; LOPES, P. S.; GUIMARÃES, S. E. F. Mapeamento de QTL nos cromossomos 1, 2, 3, 12, 14, 15 e X em suínos: características de carcaça e qualidade de carne. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 64, p. 974-982, 2012.

R DEVELOPMENT CORE TEAM. **R**: A language and environment for statistical computing. Vienna, Austria: R Fundation for Statistical Computing, 2009. Disponível em: <http://r-project .org>. Acesso Dez. 2014.

RENCHER, A. C. **Methods of multivariate analysis**. New York: John Wiley, 2002.

RENCHER, A. C.; SCHALLJE, G. B. **Linear models in statistics**. Department of Statistics, Brigham Young University, Provo-UT, 2008.

RESENDE, M.D.V.; SILVA, F. F.; VIANA, J. M. S.; PETTERNELLI, L. A.; RESENDE JR, M. F. R; VALLE, P. M. **Métodos estatísticos na seleção genômica ampla**. Colombo: Embrapa Florestas, 2011. 107p.

RESENDE, M. D.; SILVA, F. F.; LOPES, P. S.; AZEVEDO, C. F. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial.** Viçosa: Universidade Federal de Viçosa/Departamento de Estatística. 2012. 291p. Disponível em: <http://www.det.ufv.br/ppestbio/corpo_docente.php>. Acesso em Fev. 2015.

SANTOS, V. S. et al. **Seleção Genômica Ampla em suínos usando o modelo de sobrevivência de Cox**. Dissertação de estatística aplicada e biometria – Universidade Federal de Viçosa, Viçosa, MG. p.31, 2011.

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

SILVA, N. C. N.; FERREIRA, W. L.; CIRILLO, M. A.; SCALON, J. D. Uso da análise fatorial na descrição e identificação dos perfis característicos de municípios de Minas Gerais. **Revista Brasileira de Biometria**, São Paulo, v. 32, n.2, p.201-215, 2014.

SPEARMAN, C. General intelligence objectively determined and measured. **American Journal of Psychology**, Champaign, v. 15, p. 201-293, 1904.

TIBSHIRANI, R. Regression shrinkage and selection via the LASSO.**Journal of the Royal Statistics Society**. v. 58, p. 267-288, 1996.

CAPÍTULO 1

Determinação de fatores em características de suínos

Resumo: Este trabalho teve como principal objetivo utilizar análise de fatores para descrever a estrutura de variabilidade de características consideradas comercialmente importantes em suínos, visando resumir a informação contida em tais variáveis em um número menor de variáveis latentes ou fatores. Os dados utilizados neste estudo são provenientes da Granja de Melhoramento de Suínos do Departamento de Zootecnia da Universidade Federal de Viçosa (UFV), Viçosa, Minas Gerais e se referem a 41 variáveis fenotípicas, consideradas comercialmente importantes, mensuradas em uma população F2 de 345 suínos obtida pelo cruzamento de animais da raça Piau com animais Comerciais. Dos 10 fatores criados, 4 apresentaram interpretação prática, agrupando um total de 28 variáveis em fatores relacionados ao peso (14 variáveis), gordura (7 variáveis), lombo (3 variáveis) e desempenho (4 variáveis).

Palavras-chave: Melhoramento animal, estatística multivariada, redução da dimensionalidade.

1. Introdução

De acordo com o United States Departament of Agriculture (USDA), a produção de carne suína chegou a 107.514 mil toneladas no ano de 2013. O Brasil aparece como quarto maior produtor, sendo responsável por 3% da produção global (cerca de 3370 mil toneladas).

A posição de destaque alcançada pelo Brasil deve-se, dentre outros fatores, aos investimentos em pesquisa e na evolução genética da espécie, que têm sido realizados nos últimos 20 anos (MAPA, 2014).

Devido à importância econômica da suinocultura para o mercado e o contínuo crescimento da demanda, o melhoramento genético tem-se tornado cada vez mais importante para o aumento da produção e melhoria da carne. Desta forma, estudos visando o melhoramento genético de suínos têm sido cada vez mais frequentes, podendo-se citar como exemplo o estudo desenvolvido por Paixão et al. (2012) que utilizaram marcadores microssatélites na identificação de locos de características quantitativas (QTL's) associados a características de carcaça e qualidade de carne. Pinheiro et al. (2013) utilizaram modelos de regressão aleatória (MRA) para detectar QTL's para características de crescimento. Já Azevedo et al. (2013) utilizaram regressão via componentes independentes para estimação de valores genéticos genômicos e dos efeitos de marcadores SNP's para características de carcaça de uma população F2 de suínos.

Apesar de interessantes, as conclusões obtidas nestes estudos são direcionadas para uma única variável, ou seja, o pesquisador define a característica de crescimento ou qualidade de interesse e a investigação se dá a respeito dessa característica. Assim, estudos que permitissem a conclusão para mais de uma dessas características de interesse poderiam ser úteis, fornecendo resultados válidos para um conjunto de variáveis representadas por uma única variável latente.

Nesse sentido, uma abordagem que estuda a estrutura de variabilidade de variáveis visando reduzir a informação contida em tais variáveis em variáveis latentes (fatores) é a chamada análise de fatores.

Esta metodologia procura agrupar variáveis correlacionadas em fatores (ou variáveis latentes) de modo que se tenha uma redução significativa da dimensão a ser analisada. Ademais, as variáveis latentes apresentam um padrão de relacionamento devido a um fator comum que pode ser denominado de acordo com o conhecimento do

pesquisador. Após a identificação e interpretação dos fatores, as variáveis latentes podem ser preditas e seus valores utilizados em análises posteriores, como por exemplo, em seleção genômica ampla na predição de efeitos de marcadores moleculares.

A análise de fatores foi utilizada no estudo apresentado por Silva et. al (2011). Nesse trabalho os autores objetivaram criar fatores para características de carcaça (tais como comprimento e rendimento de carcaça, comprimento do lombo, etc.) visando à posterior detecção de QTL's. Porém, além das características de carcaça, outras características associadas ao peso (pesos ao abate, da paleta e do pernil), à gordura (espessuras de toucinho e bacon) e à qualidade da carne (índice de saturação, maciez) também apresentam grande importância econômica e geralmente são mensuradas em programas de melhoramento de suínos, porém não foram utilizadas no estudo.

Diante do exposto, este trabalho teve por objetivo verificar a estrutura empírica de diversas características de uma população de suínos, de modo que variáveis correlacionadas sejam agrupadas em um número menor de variáveis latentes (interpretáveis), reduzindo a dimensionalidade do conjunto de dados.

2. Material e Métodos

Os dados utilizados neste estudo são provenientes da Granja de Melhoramento de Suínos do Departamento de Zootecnia da Universidade Federal de Viçosa (UFV), Viçosa, Minas Gerais e se referem a 41 variáveis fenotípicas, consideradas comercialmente importantes, mensuradas em uma população F2 de 345 suínos obtida pelo cruzamento de animais da raça Piau com animais comerciais. As 41 variáveis analisadas estão descritas na Tabela a seguir:

Tabela 1. Descrição das 41 variáveis fenotípicas analisadas

Variável	**Descrição**	**Variável**	**Descrição**
PCARC	Peso de carcaça (kg)	PC	Peso do carré (kg)
PCD	Peso da carcaça direita (kg)	PL	Peso do lombo (kg)
TLD	Tamanho da leitegada ao desmame	PB	Peso do bacon (kg)
TLN	Tamanho da leitegada ao nascimento	PCOST	Peso da costela (kg)
IDA	Idade de abate (dias)	PF	Peso do filezinho (kg)

RCARC	Rendimento de carcaça (%)	PBR	Peso da banha rama (kg)
MBCC	Comprimento de carcaça pelo método de classificação brasileiro (cm)	CR	Consumo de Ração (kg)
MLC	Comprimento de carcaça pelo método de classificação americano (cm)	GPD	Ganho de peso médio diário (kg)
ETSH	Maior espessura de toucinho na região da copa, na linha dorso-lombar (mm)	CA	Conversão alimentar (kg/kg)
ETUC	Espessura de toucinho imediatamente após a última costela na linha dorso-lombar (mm)	NT	Número de tetos
ETUL	Espessura de toucinho entre a última e a penúltima vértebra lombar, na linha dorso-lombar (mm)	PA	Peso ao abate (kg)
ETL	Espessura de toucinho medida na região acima da última vértebra lombar, na linha dorso-lombar (mm)	PN	Peso ao nascer (kg)
ETO	Espessura de toucinho (mm)	pH45	pH medido 45 minutos post-mortem
EBACON	Espessura do bacon (mm)	pH24	pH medido 24 horas post-mortem
PROLOM	Profundidade do lombo (mm)	L	Luminosidade
AOL	Área de olho de lombo (cm^2)	GOINTR	Gordura intramuscular (%)
CORAC	Peso do coração (kg)	PGOTEJ	Perda por gotejamento (%)
PP	Peso do pernil (kg)	PCOZ	Perda por cozimento (%)

PPL	Peso do pernil sem pele e sem gordura (kg)	MACIEZ	Maciez objetiva (força de cisalhamento)
PCOPA	Peso da copa (kg)		
PPA	Peso da paleta (kg)	C	Índice de saturação

As variáveis foram corrigidas para efeito fixo de sexo, lote e a presença ou ausência do gene halotano.

Visando agrupar as variáveis correlacionadas aplicou-se ao conjunto de dados corrigido a análise de fatores. Tal metodologia tem como objetivos reduzir a dimensionalidade e descrever a variabilidade dos dados por meio de variáveis latentes (fatores), de modo que essas novas variáveis (interpretáveis e não observáveis) sejam capazes de explicar a maior parte da variação total.

O modelo fatorial adotado para uma variável X_i observável, com média μ_i pode ser representado da seguinte forma (JOHNSON & WICHERN, 2007, SILVA et al., 2014):

$$X_i - \mu_i = l_{i1}F_1 + l_{i2}F_2 + \cdots + l_{im}F_m + \varepsilon_i,$$

em que $i = 1,2, \ldots, p$ e $m \leq p$, sendo p o número de variáveis originais observáveis; o coeficiente l_{ij} é chamado de carga fatorial da i-ésima variável sobre o j-ésimo fator comum, sendo $j = 1,2, \ldots, m; F_1, F_2, \ldots, F_m$ são denominados fatores comuns, variáveis aleatórias inobserváveis e ε_i são os erros aleatórios que estão associados somente a i-ésima variável corrigida X_i, respectivamente.

Para medir a adequabilidade dos dados utilizou-se o critério de Kaiser-Meyer-Olkin (KMO) e o teste de esfericidade de Bartlett (FERREIRA, 2011). Segundo Hair et. al (2005) valores de KMO acima de 0,5 são aceitáveis, por outro lado, Pallant (2007), sugere 0,6 como um valor razoável.

O número de fatores foi definido considerando um percentual de explicação de 70% da variabilidade total, que segundo Ferreira (2011) é suficiente para a redução dos dados de maneira satisfatória.

A alocação das variáveis em cada fator foi feita através dos loadings (l_{ij}), ou cargas fatoriais, que consistem na correlação entre cada variável e os respectivos fatores. Esses valores, assim como a correlação simples, variam entre -1 e 1 e, quanto maior carga fatorial (em módulo) mais correlacionada a variável será com o respectivo fator. Logo, as variáveis farão parte do fator ao qual estiverem mais correlacionadas.

As comunalidades foram utilizadas para avaliar a proporção de cada variável explicada pelo fator a qual ela pertence e a proporção explicada pelo erro aleatório. Segundo Figueiredo Filho (2010) tais valores devem ser superiores a 0,5. Finalmente, visando uma melhor interpretação da distribuição das variáveis nos respectivos fatores, utilizou-se a rotação Varimax.

3. Resultados e discussão

Após a análise de fatores, as variáveis latentes interpretáveis foram estimadas e seus valores foram apresentados em gráficos de duas dimensões. Este procedimento permite ao pesquisador verificar quais animais estão mais relacionados aos fatores encontrados.

De acordo com o índice de KMO (0,75), considerado satisfatório pelos critérios de Pallant (2007) e Hair et al. (2005), e com o teste de esfericidade de Bartlett, que apresentou significância estatística ($p < 0{,}01$), verificou-se que há adequabilidade dos dados para análise de fatores.

Com base no critério de escolher um número de fatores tal que a explicação da variação total fosse superior a 70%, observou-se a formação de 10 fatores (Tabela 2). Destes, apenas 4 apresentaram interpretação prática e são descritos abaixo, por ordem de importância.

O primeiro fator (F1) foi formado por variáveis relacionadas ao peso de diversos caracteres do animal e características de carcaça (Tabela 2). São elas: PCARC, PCD, MBCC, MLC, CORAC, PP, PPL, PCOPA, PPA, PC, PB, PCOST, PF e PA. Este resultado indica que as variáveis relacionadas ao peso e carcaça estão altamente correlacionadas entre si, o que possibilita a criação de um fator que pode ser denotado como "peso". Esses resultados são corroborados pelo trabalho realizado por Silva et al. (2011), em que as variáveis MBCC e MLC se agruparam no mesmo fator. Além disso, observa-se que todas as variáveis apresentam valores da correlação positivos, ou seja, quanto maior o valor dessas variáveis, maior será o valor dos escores da nova variável "peso". Dessa maneira, ao analisar os escores dessa nova variável, estaremos analisando todas as variáveis pertencentes a este, conjuntamente.

Tabela 2. Loadings para cada variável em relação a todos os fatores, a variação explicada de cada um dos fatores e as comunalidades (h2)

Variáveis	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	h2
PCARC	**0.87**	0.40	0.11	0.03	-0.07	0.06	0.00	-0.10	0.01	0.04	0.96
PCD	**0.87**	0.41	0.11	0.03	-0.06	0.06	0.00	-0.09	-0.01	0.06	0.96
TLD	-0.08	0.02	0.01	0.05	0.03	-0.07	**0.90**	0.08	-0.04	-0.06	0.84
TLN	0.05	0.03	-0.07	-0.02	0.01	-0.02	**0.92**	-0.10	0.00	-0.02	0.86
IDA	0.26	0.18	0.01	**-0.78**	-0.08	0.01	-0.08	0.03	-0.11	-0.01	0.73
RCARC	0.15	0.32	0.26	-0.09	0.08	0.10	-0.12	-0.28	0.07	**0.33**	0.42
MBCC	**0.82**	-0.25	-0.08	-0.03	0.09	-0.06	0.02	0.02	0.12	0.05	0.78
MLC	**0.84**	-0.22	-0.06	0.00	0.14	-0.10	0.09	0.03	0.16	-0.04	0.82
ETSH	0.13	**0.74**	0.00	-0.09	-0.01	-0.02	0.07	0.02	0.01	0.00	0.57
ETUC	0.06	**0.82**	0.12	0.08	-0.05	0.07	0.02	0.10	0.02	0.05	0.72
ETUL	0.05	**0.86**	0.05	0.07	-0.05	-0.02	0.08	0.00	-0.10	0.00	0.76
ETL	0.02	**0.88**	0.00	0.01	-0.05	-0.01	0.06	0.01	-0.10	0.01	0.78
ETO	0.04	**0.87**	-0.08	0.09	-0.08	-0.01	-0.06	0.05	0.02	-0.02	0.79
EBACON	0.05	**0.79**	-0.16	-0.03	-0.12	0.01	-0.07	-0.02	0.06	0.05	0.68
PROLOM	0.01	-0.07	**0.83**	0.02	0.01	-0.03	-0.04	0.05	0.02	-0.11	0.72
AOL	0.27	0.05	**0.82**	-0.03	0.03	0.09	-0.05	-0.08	0.00	0.09	0.77
CORAC	**0.52**	-0.15	0.15	-0.03	-0.02	0.00	-0.14	0.23	0.02	-0.09	0.39
PP	**0.77**	0.30	0.25	0.00	-0.09	0.13	0.00	-0.08	-0.11	-0.02	0.79
PPL	**0.78**	-0.07	0.37	-0.03	-0.06	0.10	0.03	-0.09	-0.10	0.06	0.79
PCOPA	**0.71**	0.25	0.13	0.01	-0.03	0.01	-0.02	-0.16	-0.07	0.04	0.62
PPA	**0.86**	0.10	0.02	0.04	-0.06	0.07	-0.05	-0.10	-0.10	-0.01	0.78
PC	**0.58**	0.48	0.28	0.03	-0.06	0.05	-0.01	-0.06	0.18	0.15	0.71
PL	0.51	-0.13	**0.63**	0.08	-0.02	0.08	0.07	-0.05	0.12	0.05	0.72
PB	**0.56**	0.54	-0.14	-0.05	-0.14	0.01	0.03	-0.13	0.14	-0.09	0.69
PCOST	**0.59**	0.15	-0.05	0.15	0.09	-0.09	0.02	0.16	-0.08	0.22	0.50
PF	**0.50**	0.03	0.37	0.11	-0.18	0.16	0.00	0.05	-0.31	0.09	0.57
PBR	0.26	**0.75**	-0.03	-0.06	-0.11	0.08	-0.07	-0.03	0.03	-0.11	0.67
CR	0.15	0.18	-0.04	**0.83**	-0.12	0.03	-0.01	0.02	0.07	0.08	0.78
GPD	0.28	0.02	0.00	**0.71**	0.01	0.07	0.05	-0.54	-0.07	0.01	0.89
CA	-0.17	0.15	-0.07	-0.06	-0.11	-0.03	-0.04	**0.75**	0.15	0.07	0.67
NT	-0.01	0.00	0.08	0.15	0.00	-0.01	-0.01	0.16	**0.81**	0.10	0.72
PA	**0.85**	0.31	0.05	0.02	-0.08	0.04	-0.01	-0.02	-0.04	-0.01	0.83

PN	0.11	0.05	0.19	**0.53**	0.10	-0.10	-0.08	0.39	0.09	-0.09	0.54
pH45	0.08	-0.07	-0.07	0.03	0.03	**-0.85**	0.03	0.02	-0.03	-0.10	0.75
pH24	0.23	-0.01	-0.18	0.09	**-0.60**	-0.04	0.05	0.07	-0.16	-0.15	0.51
L	0.04	-0.19	-0.06	0.00	**0.84**	0.07	0.05	-0.03	-0.15	-0.04	0.78
GOINTR	0.09	-0.07	-0.03	0.06	0.02	0.10	-0.06	0.05	0.05	**0.84**	0.75
PGOTEJ	0.17	0.06	0.07	-0.04	0.11	**0.86**	-0.07	-0.03	-0.03	0.05	0.80
PCOZ	0.18	-0.18	0.00	0.14	0.16	**0.40**	0.20	0.39	-0.13	0.03	0.48
MACIEZ	0.08	0.03	0.09	-0.16	-0.29	-0.45	0.15	0.13	**-0.41**	0.29	0.61
C	0.01	-0.20	-0.07	0.08	**0.85**	0.05	0.04	0.04	0.06	-0.03	0.79
Variância Explicada Acumulada	0.20	0.36	0.42	0.47	0.53	0.58	0.62	0.66	0.69	0.71	-

O segundo fator (F2) foi composto por variáveis relacionadas à gordura (todas as variáveis relacionadas à espessura de toucinho, peso do bacon e peso da banha rama). Portanto, teremos um fator que representa "gordura". Assim como no fator anterior, todas as variáveis pertencentes a este apresentaram loadings positivos, e assim, o valor do escore desse fator aumentará de acordo com o aumento das variáveis pertencentes a ele. Novamente, tais resultados são semelhantes aqueles encontrados por Silva et. al (2011), em que variáveis de espessura de toucinho formaram um fator.

Já o terceiro fator (F3) pode ser nomeado como "lombo" visto que o mesmo agrupou três características relacionadas ao lombo (peso, profundidade e área). Novamente as variáveis que constituem o fator apresentam correlações positivas, indicando que altos valores dessas três variáveis estão associados a altos escores para o fator "lombo". As características de lombo presentes no estudo de Silva et. al (2011) (área de olho do lombo e profundidade do lombo) formaram seu segundo fator.

Podemos observar que o quarto fator (F4) foi formado pelas variáveis relacionadas ao desempenho do animal (idade ao abate, consumo de ração, ganho de peso diário e peso ao nascer). Podemos então intitular esse fator como "desempenho".

Praticamente todas as variáveis pertencentes aos fatores que possuem interpretação prática citados acima possuem um valor aceitável para a comunalidades ($h2 > 0{,}50$), com exceção da variável "peso do coração", que apresentou $h2 = 0{,}39$. As demais variáveis apresentaram em maioria valores acima de 0,70.

De acordo com a Figura 1, verifica-se, conforme esperado, a falta de associação entre os fatores definidos dada a hipótese para a construção da análise de fatores (Ferreira, 2011).

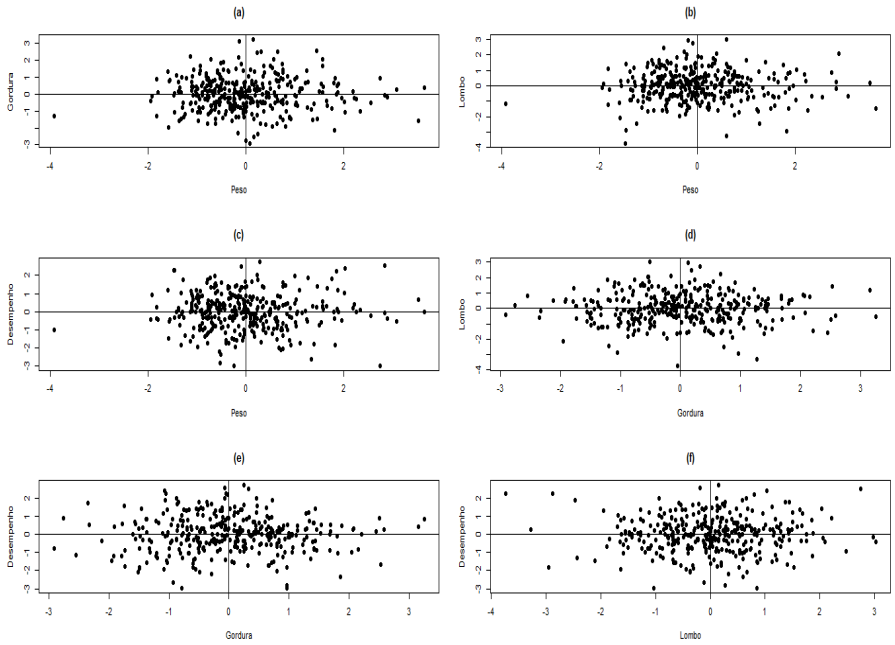

Figura 1: Gráficos de dispersão entre os escores dos fatores interpretáveis encontrados.

Como a grande maioria das variáveis apresentaram loadings positivos, podemos classificar os animais de acordo com os quadrantes. Desta forma, no primeiro quadrante, estão os indivíduos que se destacam positivamente em relação aos dois fatores avaliados no gráfico, ao contrário do terceiro quadrante, onde os animais não se destacam em nenhuma das variáveis (Figura 1). No segundo quadrante estão os indivíduos que possuem valores elevados da característica do eixo Y e valores baixos no eixo X, como por exemplo, no segundo quadrante da Figura 1(a) estão os indivíduos que apresentam alto valor de gordura e baixo valor do peso. Oposto do que acontece no quarto quadrante, em que os indivíduos possuem altos escores para a variável do eixo X e baixos escores para a variável do eixo Y.

Após a identificação e estimação dos escores para cada fator, as variáveis latentes podem ser utilizados em análises posteriores, como por exemplo, na identificação de QTL's (Silva et al., 2011). Outra análise interessante, que pode fazer uso das variáveis latentes (interpretáveis) obtidas neste estudo é a aplicação em seleção genômica ampla (SGA). Tal análise, que será realizada posteriormente, possibilita a inserção de informações genômicas na predição do mérito genético do animal visando a

posterior seleção dos mesmos. Nessa abordagem a seleção dos animais seria realizada levando em conta um grupo de variáveis e não apenas em uma única variável, como é rotineiramente apresentado na literatura. Para exemplificar a utilização de variáveis latentes na SGA, podemos citar o estudo realizado por AZEVEDO et al. (2013), em que foram estimados individualmente os méritos genéticos de sete características fenotípicas, são elas: espessuras de toucinho medidas imediatamente após a última costela na linha dorso-lombar (ETUC); a 6,5 cm da linha dorso-lombar (ETO); entre a última e a penúltima vértebra lombar (ETUL); menor espessura na região acima da última vértebra lombar, na linha dorso-lombar (ETL); espessura de bacon (EBACON); comprimento de carcaça pelo método de classificação americano (MLC) e rendimento de carcaça (RCARC). No contexto da análise de fatores, e com os resultados aqui obtidos, poderíamos estimar seis dessas sete variáveis por meio apenas do segundo fator ("gordura"), ou seja, o mérito genético obtido pata tal fator é altamente correlacionado com aqueles obtidos para seis variáveis fenotípicas de interesse, de modo a simplificar a interpretação dos resultados e possibilitar a seleção destas variáveis simultaneamente.

4. Conclusões

A análise de fatores conseguiu reduzir as 41 características inicialmente avaliadas para apenas 10 fatores, sendo estes com um percentual satisfatório de variabilidade explicada. Dos 10 fatores criados, 4 apresentaram interpretação prática, agrupando um total de 28 variáveis nos fatores relacionados ao peso (14 variáveis), gordura (7 variáveis), lombo (3 variáveis) e desempenho (4 variáveis).

REFERÊNCIAS BIBLIOGRÁFICAS

AZEVEDO, C. F.; RESENDE, M. D. V.; SILVA, F. F.; Lopes, P. S.; GUIMARÃES, S. E. F. Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. **Pesquisa Agropecuária Brasileira**, v. 48, p. 619-626, 2013.

FERREIRA, D. F. **Estatística Multivariada**. 2.Ed. Lavras: Ed. UFLA, 2011.

FIGUEIREDO FILHO, D. B; JÚNIOR, J. A. S. Visão além do alcance: uma introdução à análise fatorial. **Opinião Pública**, Campinas, v. 16, n. 1, p. 160-185, 2010.

HAIR, Jr; BLACK, W. C; BABIN, B. J; ANDERSON, R. E; TATHAM, R. L. **Multivariate Data Analysis**. 6ª edição. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

MAPA. **Ministério da Agricultura, Pecuária e Abastecimento**. 2014. Disponível em: <http://www.agricultura.gov.br/animal/especies/suinos >. Acesso em: Nov. 2014.

PAIXÃO, D. M.; CARNEIRO, P. L. S.; PAIVA, S. R.; SOUSA, K. R. S.; VERARDO, L. L.; BRACCINI NETO, J. ; PINTO, A. P. G.; Hidalgo, A. M.; NASCIMENTO, C. S.; PÉRISSÉ, I. V.; LOPES, P. S.; GUIMARÃES, S. E. F. Mapeamento de QTL nos cromossomos 1, 2, 3, 12, 14, 15 e X em suínos: características de carcaça e qualidade de carne. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 64, p. 974-982, 2012.

PALLANT, J. **SPSS Survival Manual**. Open University Press, 2007.

PINHEIRO, V. R.; SILVA, F. F.; GUIMARÃES, S. E. F; RESENDE, M. D. V.; LOPES, P. S.; CRUZ, COSME DAMIÃO; AZEVEDO, C. F. Mapeamento de QTL para características de crescimento de suínos por meio de modelos de regressão aleatória. **Pesquisa Agropecuária Brasileira**, v. 48, p. 190-196, 2013.

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

SILVA, N. C. N.; FERREIRA, W. L.; CIRILLO, M. A.; SCALON, J. D. Uso da análise fatorial na descrição e identificação dos perfis característicos de municípios de Minas Gerais. **Revista Brasileira de Biometria**, São Paulo, v. 32, n.2, p.201-215, 2014.

USDA. **United States Department of Agriculture**. 2014. Disponível em: <http://www.abipecs.org.br/pt/estatisticas/mundial/producao-2.html>. Acesso em: Nov. 2014.

CAPÍTULO 2

Análise de fatores aplicada à seleção genômica em suínos

Resumo: O objetivo deste trabalho foi propor e avaliar a utilização de análise de fatores para construção de variáveis latentes que representam um conjunto de variáveis de suínos para posterior uso em seleção genômica ampla. Nessa abordagem, a seleção é feita em função das variáveis latentes ao invés de considerar as variáveis originais. Para tanto, foram utilizados dados fenotípicos e genotípicos provenientes de 345 suínos obtidos pelo cruzamento das raças Piau e Comercial, oriundos da Granja de Melhoramento de Suínos do Departamento de Zootecnia da Universidade Federal de Viçosa (UFV), no período de novembro de 1998 a julho de 2001, referentes a 237 marcadores SNP's e 41 variáveis fenotípicas. Após a aplicação da análise de fatores foram obtidas quatro variáveis latentes com interpretação prática (F1 – "Peso"; F2 – "Gordura"; F3 – "Lombo"; F4 – "Desempenho"). Posteriormente, tais variáveis latentes foram submetidas aos procedimentos Bayes A, Bayes B, RR-BLUP Bayes e LASSO bayesiano, em que o último apresentou melhores resultados em termos de acurácia, principalmente para o segundo fator, denominado "Gordura", utilizado em análises posteriores. A utilização de variáveis latentes em estudos de seleção genômica é uma abordagem interessante e promissora, visto que o valor da acurácia referente à variável latente "Gordura" foi semelhante àqueles obtidos quando as variáveis foram analisadas separadamente e os valores de concordância entre os 10% melhores indivíduos selecionados por meio do fator ("Gordura") e pelas variáveis fenotípicas analisadas individualmente foram satisfatórios. Além disso, o padrão dos efeitos de marcadores encontrados para as variáveis foi semelhante ao que foi encontrado para o fator.

Palavras-chave: melhoramento genético, análise multivariada, inferência bayesiana.

1. Introdução

De acordo com a Associação Brasileira de Proteína Animal (ABPA), a produção mundial de carne suína chegou a 107.514 mil toneladas no ano de 2013. O Brasil é o quarto maior produtor, sendo responsável por cerca de 3% da produção (3370 mil toneladas). De acordo com o MAPA (2014), o país se destaca também em relação à exportação da carne de porco, sendo responsável por 10% do volume total, o que acarreta em um lucro de US$1 bilhão por ano.

Devido à importância econômica da produção e exportação de carne suína para o mercado, têm-se investido bastante no melhoramento genético, tendo como finalidade o aumento da produção e a melhoria de qualidade da carne consumida e exportada (MAPA, 2014).

Dentre algumas publicações resultantes de pesquisas direcionadas ao melhoramento genético de suínos, podemos citar o estudo desenvolvido por Fan et al. (2011), em que foram aplicadas metodologias bayesianas a fim de estimar de efeitos de marcadores SNP's em análise de associação genômica para características de lombo e carcaça de uma população da raça Large White. Paixão et al. (2012) utilizaram marcadores microssatélites para identificação de locos de características quantitativas (QTL's) associados a características de carcaça e qualidade de carne. Azevedo et al. (2013) aplicaram regressão via componentes independentes para estimação de valores genéticos genômicos (seleção genômica ampla) para características de carcaça de uma população F2 de suínos.

Dentre as diversas metodologias para obtenção de ganhos em programas de melhoramento, a Seleção Genômica Ampla (SGA) desenvolvida por Meuwissen et al. (2001) apresenta grande destaque, visto que tal metodologia possibilita incorporar informações do genoma diretamente na predição do mérito genético dos indivíduos, permitindo alta eficiência seletiva, rapidez na obtenção dos ganhos genéticos com a seleção e baixo custo, em comparação com a tradicional seleção baseada em dados fenotípicos ou mesmo pela seleção assistida por marcadores moleculares (RESENDE, 2012).

Em geral, estudos de seleção genômica ampla analisam cada característica individualmente, ou seja, os resultados obtidos são válidos apenas para uma única variável. Entretanto, em programas de melhoramento o interesse recai em ganhos para mais de uma característica conjuntamente. Dessa forma, desenvolver uma abordagem

que trabalhe com análises que considerem várias características simultaneamente pode ser interessante, visto que poderíamos estudar um conjunto de caracteres importantes conjuntamente. Uma metodologia possível de ser utilizada para este fim é a análise fatorial (ou análise de fatores - AF). Tal metodologia permite a obtenção de variáveis latentes (fatores comuns) que representam um conjunto das variáveis originais. A partir de então, análises posteriores podem ser realizadas utilizando as variáveis latentes criadas. A AF foi utilizada com sucesso em Silva et al. (2011) para a detecção de QTL's em uma população de suínos, em que os fatores postulados representavam algumas características de carcaça.

Outra abordagem possível seria a utilização de variáveis latentes em seleção genômica, visto que a sua utilização além de possibilitar a seleção de indivíduos para um conjunto de caracteres simultaneamente, reduziria o tempo computacional para obtenção dos resultados, já que em algumas situações as metodologias bayesianas utilizadas rotineiramente na SGA demandam bastante tempo computacional.

Diante do exposto, este trabalho teve como objetivo propor e avaliar a utilização de análise de fatores para construção de variáveis latentes que representam um conjunto de caracteres de suínos para posterior uso em seleção genômica ampla.

2. Material e Métodos

Os dados genéticos e fenotípicos utilizados neste estudo são provenientes da Granja de Melhoramento de Suínos do Departamento de Zootecnia da Universidade Federal de Viçosa (UFV), em Viçosa, Minas Gerais no período de novembro de 1998 a julho de 2001 e se referem a 41 variáveis fenotípicas (Tabela 1), consideradas comercialmente importantes, mensuradas em uma população F2 de 345 suínos obtida pelo cruzamento de animais da raça Piau com animais Comerciais.

Tabela 1. Descrição das 41 variáveis fenotípicas analisadas

Variável	Descrição	Variável	Descrição
PCARC	Peso de carcaça (kg)	PC	Peso do carré (kg)
PCD	Peso da carcaça direita (kg)	PL	Peso do lombo (kg)
TLD	Tamanho da leitegada ao desmame	PB	Peso do bacon (kg)
TLN	Tamanho da leitegada ao	PCOST	Peso da costela (kg)

	nascimento		
IDA	Idade de abate (dias)	PF	Peso do filezinho (kg)
RCARC	Rendimento de carcaça (%)	PBR	Peso da banha rama (kg)
MBCC	Comprimento de carcaça pelo método de classificação brasileiro (cm)	CR	Consumo de Ração (kg)
MLC	Comprimento de carcaça pelo método de classificação americano (cm)	GPD	Ganho de peso médio diário (kg)
ETSH	Maior espessura de toucinho na região da copa, na linha dorso-lombar (mm)	CA	Conversão alimentar (kg/kg)
ETUC	Espessura de toucinho imediatamente após a última costela na linha dorso-lombar (mm)	NT	Número de tetos
ETUL	Espessura de toucinho entre a última e a penúltima vértebra lombar, na linha dorso-lombar (mm)	PA	Peso ao abate (kg)
ETL	Espessura de toucinho medida na região acima da última vértebra lombar, na linha dorso-lombar (mm)	PN	Peso ao nascer (kg)
ETO	Espessura de toucinho (mm)	pH45	pH medido 45 minutos post-mortem
EBACON	Espessura do bacon (mm)	pH24	pH medido 24 horas post-mortem
PROLOM	Profundidade do lombo (mm)	L	Luminosidade
AOL	Área de olho de lombo (cm^2)	GOINTR	Gordura intramuscular (%)
CORAC	Peso do coração (kg)	PGOTEJ	Perda por gotejamento (%)

PP	Peso do pernil (kg)	PCOZ	Perda por cozimento (%)
PPL	Peso do pernil sem pele e sem gordura (kg)	MACIEZ	Maciez objetiva (força de cisalhamento)
PCOPA	Peso da copa (kg)		
PPA	Peso da paleta (kg)	C	Índice de saturação

Os marcadores SNP's estão distribuídos da seguinte forma nos cromossomos de Susscrofa: SSC1 (56), SSC4 (54), SSC7 (59), SSC8 (31), SSC17 (25) e SSCX (12) totalizando assim 237 marcadores identificados para os animais F_2. Esses marcadores foram obtidos apenas em regiões nas quais se observaram a presença QTL em estudos prévios (HIDALGO et al., 2011) nessa mesma população, caracterizando assim um mapeamento fino apenas em regiões cromossômicas de interesse, o que explica o número reduzido de marcadores utilizados (AZEVEDO et al., 2013).

Visando estudar várias características fenotípicas em termos de um número menor de variáveis latentes, que representem subconjuntos das variáveis originais, utilizou-se análise de fatores. Tal metodologia possibilita o agrupamento das variáveis originais em subconjuntos de variáveis mutuamente não correlacionadas (denominadas variáveis latentes ou fatores), de modo que essas novas variáveis possam apresentar uma interpretação prática (definida a critério do pesquisador). Sob essa abordagem, a variabilidade de cada variável fenotípica X é dividida naquela atribuída aos fatores comuns (comunalidade) e na variabilidade devido ao erro aleatório (unicidade).

A obtenção de tais variáveis é realizada por meio do modelo fatorial, adotado para uma variável fenotípica X_i observável, com média μ_i, representado da seguinte forma (SILVA et al., 2014):

$$X_i - \mu_i = l_{i1}F_1 + l_{i2}F_2 + \cdots + l_{im}F_m + \varepsilon_i,$$

em que $i = 1,2, \ldots, 41$ e $m \leq 41$, sendo 41 o número de variáveis fenotípicas originais observáveis; o coeficiente l_{ij} é chamado de carga fatorial (ou loadings) da i-ésima variável fenotípica sobre o j-ésimo fator comum, sendo $j = 1,2, \ldots, m$; $F_1, F_2, \ldots, F_m$ são denominados fatores comuns (variáveis aleatórias não observáveis) e ε_ié o vetor de erros aleatórios ou de fatores específicos que estão associados somente a i-ésima variável fenotípica X_i, respectivamente.

As variáveis fenotípicas foram corrigidas para efeito fixo de sexo, lote e a presença ou ausência do gene halotano.

Para avaliar a adequabilidade do modelo utilizou-se o critério de Kaiser-Meyer-Olkin (KMO) (MINGOTI, 2007) e o teste de esfericidade de Bartlett (FERREIRA, 2011). O número de fatores foi determinado considerando um percentual de explicação de 70% da variabilidade total, que, segundo Ferreira (2011) é suficiente para a redução dos dados de maneira satisfatória. A alocação das variáveis em cada fator foi feita através dos loadings (l_{ij}) após rotação por meio do critério Varimax. Assim, quanto maior o valor do loading (em módulo) mais relacionada à variável fenotípica é com o respectivo fator.

Posteriormente, visando à obtenção dos valores referentes às variáveis latentes que serão utilizadas na seleção genômica, estimaram-se, por meio do método da regressão, os escores fatoriais (FERREIRA, 2011):

$$\hat{F}_j = \hat{\Gamma}^T\left(\hat{\Gamma}\hat{\Gamma}^T + \hat{\psi}\right)^{-1}(Y_j - \bar{Y}),$$

em que $\hat{\Gamma}_{\text{pxm}}$ é a matriz dos loadings, $\hat{\psi}$ a matriz das unicidades, Y_j é o vetor referentes a j-ésima unidade amostral ($j = 1,2,\ldots,345$) e $\overline{Y}$ o vetor de médias referente as 41 variáveis fenotípicas avaliadas. A obtenção das matrizes de loadings e unicidades foi realizada por meio do método dos componentes principais (MINGOTI, 2007).

De posse dos escores fatoriais, aplicaram-se quatro metodologias bayesianas (Bayes A, Bayes B, RR-BLUP Bayes e Lasso Bayesiano) objetivando estimar os seus respectivos valores genéticos genômicos a serem utilizados na predição e seleção de indivíduos, utilizando os 237 marcadores SNP's.

No contexto da SGA, o modelo proposto por Meuwissen et al. (2001) é definido abaixo, em notação matricial:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X\beta} + \boldsymbol{e},$$

em que $\boldsymbol{y}_{345x1}$ é o vetor coluna que contém os valores da variável $\boldsymbol{Y}$ para cada indivíduo; $\mathbf{1}$ é o vetor coluna composto por 1's de dimensão $345x1$; $\boldsymbol{\mu}_{345x1}$ é a média da variável $\boldsymbol{Y}$; $\boldsymbol{\beta}_{237x1}$ é o vetor que contém os efeitos dos marcadores SNP's; $\boldsymbol{X}_{345x237}$ é a matriz de incidência que relaciona efeitos de SNP's aos valores da característica $\boldsymbol{Y}$; e $\boldsymbol{e}$ é o vetor dos erros aleatórios.

No método Bayes A (MEUWISSEN et al., 2001) é considerada uma variância diferente para cada marcador ($\sigma^2_{\beta i}$), e os efeitos de marcadores (β_i) são estimados considerando a informação combinada da função de verossimilhança (obtida através dos

dados) e das distribuições à priori para as variâncias. As distribuições à priori para β_i e $\sigma^2_{\beta i}$ seguem, respectivamente: $\beta_i|\sigma^2_{\beta i} \sim N(0, \sigma^2_{\beta i})$, $\sigma^2_{\beta i} \sim \chi^{-2}(v_\beta, S^2_\beta)$ e $\sigma^2_u = 2\sum_{i=1}^{m} p_i(1-p_i)\sigma^2_{\beta i}$ é a variância aditiva, em que v_β representa os graus de liberdade, S^2_β é o parâmetro da escala de distribuição e p_i representa as frequências alélicas. Meuwissen et al. (2001) consideram os valores 4,012 ou 4,2 para v_β e 0,002 e 0,0429 para S^2_β.

O método Bayes B assume que um número de marcadores (com proporção π) tem efeito zero, e o restante dos marcadores, com proporção $1-\pi$, é amostrado com uma variância individual para cada marcador, considerando as mesmas prioris usadas no método Bayes A (CRUZ, 2013). A determinação de π pode ser feita de acordo com o conhecimento do pesquisador ou estimado por alguma metodologia específica. Este método é equivalente ao Bayes A quando $\pi = 0$. Neste método, consideram-se as seguintes distribuições dos parâmetros: $\beta_i|\pi, \sigma^2_{i1} \sim N(0, \sigma^2_{\beta i})$; $\sigma^2_{\beta i} = 0$ com probabilidade π; $\sigma^2_{\beta i} \sim \chi^{-2}(v_\beta, S^2_\beta)$ com probabilidade $1-\pi$ e a variância aditiva é dada por $\sigma^2_u = 2\sum_{i=1}^{m} p_i(1-p_i)\sigma^2_i$. Os hiperparâmetros v_β e S^2_β são os mesmos considerados por Meuwissen et al. (2001) da mesma maneira do Bayes A e p_i e $(1-p_i)$ denotam as frequências alélicas.

O método RR-BLUP Bayes (Bayesian Ridge Regression) é semelhante ao Bayes A. A diferença é que aqui se faz a pressuposição de homogeneidade das variâncias dos SNP's, ao contrário do Bayes A, que supõe uma variância para cada marcador. Temos apenas um valor assumido para σ^2. Os parâmetros de efeito de SNP's (β_i), variância dos marcadores (σ^2) e variância aditiva (σ^2_u) seguem respectivamente: $\beta_i|\sigma^2 \sim N(0, \sigma^2)$, $\sigma^2 \sim \chi^{-2}(v, S^2)$ e $\sigma^2_u = 2\sigma^2\sum_{i=1}^{m} p_i(1-p_i)$, em que p_i e $(1-p_i)$ denotam as frequências alélicas e v e S^2 se equivalem a v_β e S^2_β, citados anteriormente.

O LASSO bayesiano, ou BLASSO, foi uma proposta por DE LOS CAMPOS et al. (2009), de uma interpretação bayesiana baseada no LASSO (TIBSHIRANI, 1996) que considera os efeitos dos marcadores β_i um produto de exponenciais duplas com média zero, que indica que cada marcador possui essa distribuição, de parâmetro λ. A priori conjunta é dada por $\prod_{j=1}^{p}(\lambda/2)\exp(-\lambda|\beta_j|)$. O parâmetro de suavização λ tem a função de aproximar os efeitos dos marcadores de zero, o que usualmente ocorre, pois a maioria dos SNP's tem efeito nulo. No BLASSO, λ controla a priori assumida por $\boldsymbol{\beta}$, e quanto maior for o valor assumido por este parâmetro, maior será a densidade nos valores próximos de zero.

Todas as metodologias acima citadas tem como pressuposto a função de verossimilhança dos dados uma Normal, dada por $\prod_{i=1}^{n} N(y_i|\boldsymbol{x_i'\beta}, \sigma_\varepsilon^2)$.

Após a estimação dos efeitos de marcadores para cada fator interpretável por cada metodologia apresentada anteriormente, as mesmas foram comparadas por meio da acurácia, encontrada a partir da seguinte expressão (RESENDE et al., 2010):

$$r_{q,\hat{q}} = \frac{r_{y,\hat{y}}}{\sqrt{h^2}},$$

em que $r_{y,\hat{y}}$ é a capacidade preditiva do modelo, dada por $r_{y,\hat{y}} = \frac{Cov(y,\hat{y})}{\sqrt{Var(y).Var(\hat{y})}}$ e h^2 é conhecida como herdabilidade do carácter, definida por $h^2 = \frac{VG}{VF}$, sendo VG a variância genética e VF a variância fenotípica.

Os cálculos da capacidade preditiva, herdabilidade e acurácia foram baseados no mérito genético (EGBV's). Tais valores são estimativas do valor real fenotípico, baseados no genótipo e nos efeitos de marcadores, dada pela seguinte expressão:

$$\text{EGBV} = \mathbf{Z\hat{b}},$$

em que $\mathbf{Z}$ é a matriz de genótipos, composta pelos valores numéricos assumidos pelos SNP's e $\widehat{\boldsymbol{b}}$ é o vetor que contém as estimativas dos efeitos dos marcadores, de acordo com as metodologias utilizadas.

Visando avaliar a qualidade do ajuste, e para que os efeitos dos marcadores não sejam superestimados devido à estimação e validação da mesma amostra (CRUZ, 2013), uma técnica de validação cruzada foi adotada. A população F2 dos suínos foi dividida em três populações distintas, onde se utilizou duas destas para estimação dos efeitos de marcadores e a outra para validação, onde se verificou a partir da capacidade preditiva e acurácia, a semelhança com os valores reais fenotípicos. As três combinações possíveis para essa situação foram utilizadas e tomaram-se os valores médios para acurácia como referência. Essa abordagem foi também utilizada por Azevedo et al. (2013).

Após a seleção dos melhores indivíduos, correspondente a 10% da população de validação, por variável e fator (variável latente obtida) de acordo com a estimação dos seus méritos genéticos, verificou-se a concordância entre cada variável e o seu respectivo fator, ou seja, a concordância entre os indivíduos selecionados por fator e por cada variável a qual ele pertence. A medida utilizada com esse intuito foi o coeficiente Cohen's Kappa (COHEN, 1960). Esse índice pode ser medido a partir da seguinte expressão:

$$\hat{k} = \frac{Pr(a) - Pr(e)}{1 - Pr(e)},$$

em que $Pr(a) - Pr(e)$ representa a proporção de observações em que a concordância ocorreu além do que se esperava aleatoriamente e $1 - Pr(e)$, a proporção de observações que não ocorreu concordância. Essa medida varia de 0 a 1 e quanto maior o índice, mais os grupos estão em concordância.

Para observar os efeitos de cada um dos marcadores nas variáveis e os seus devidos cromossomos de interesse, utilizou-se o Manhattan Plot, que corresponde ao gráfico de dispersão que contém os cromossomos correspondentes no eixo X e os efeitos estimados dos marcadores no eixo Y. A partir desse gráfico, podemos ter uma visão clara do padrão de comportamento dos SNP's nos respectivos cromossomos.

Para a análise de fatores foi utilizada a função factanal, através dos pacotes psych (REVELLE, 2008) e GPArotation (BERNAARDS & JENNRICH, 2005). Os modelos bayesianos foram estimados pela função BGLR (com 100.000 interações, 20.000 de burn-in e thin assumindo o valor 10), implementada no pacote BGLR (CAMPOS & RODRIGUEZ, 2009). Os Manhattan plots foram obtidos com a utilização da função mhtp, do pacote gap (ZHAO, 2007). Todas as funções foram aplicadas no software R (Development Core Team, 2014) e os códigos e algoritmos utilizados estão apresentados no apêndice A.

3. Resultados e discussão

De acordo com o índice de KMO (0,85) e do teste de esfericidade de Bartlett, que apresentou significância estatística ($p < 0{,}01$), verificou-se que há adequabilidade dos dados para análise de fatores. Considerando como critério o número de fatores cuja explicação da variação total seja superior a 70%, observou-se a formação de dez fatores (Tabela 2 – Apêndice B). Destes, quatro apresentaram interpretação prática (Tabela 3).

Tabela 3. Fatores interpretáveis, respectivas variáveis associadas e loadings (parênteses).

Fator	**Variáveis associadas**
Peso	PCARC (0,87), PCD (0,87), MBCC (0,82), MLC (0,84), CORAC (0,52), PP (0,77), PPL (0,78), PCOPA (0,71), PPA (0,86), PC (0,58), PB (0,56), PCOST (0,59), PF (0,50) e PA

	(0,85)
Gordura	ETSH (0,74), ETUC (0,82), ETUL (0,86), ETL (0,88), ETO (0,87), EBACON (0,79) e PBR (0,75)
Lombo	PROLOM (0,83), AOL (0,82) e PL (0,63)
Desempenho	IDA (-0,78), CR (0,83), GPD (0,71) e PN (0,53)

Legenda: PBR – peso da banha rama; EBACON – espessura do bacon; ETO – espessura de toucinho; ETL - espessura de toucinho medida na região acima da última vértebra lombar, na linha dorso-lombar; ETUL – espessura de toucinho entre a última e a penúltima vértebra lombar, na linha dorso-lombar; ETUC – espessura de toucinho imediatamente após a última costela na linha dorso-lombar; ETSH – maior espessura de toucinho na região da copa, na linha dorso-lombar.

O primeiro fator agrupou variáveis relacionadas ao peso dos indivíduos (Tabela 3). Sob a abordagem da análise de fatores, temos que essas variáveis são altamente correlacionadas entre si, e a um fator comum (variável latente), que pode ser identificado e nomeado como "peso". O segundo fator abrangeu todas as 5 variáveis associadas a espessura de toucinho, além de espessura do bacon (EBACON) e peso da banha rama (PBR), sendo as 7 variáveis associadas a "gordura", segunda variável latente (Tabela 3). Já as três variáveis fenotípicas relacionadas a características do lombo (AOL, PROLOM e PL) formaram o terceiro fator, denotado como "lombo". Finalmente, o quarto fator, último com interpretação prática, agrupou as variáveis idade ao abate (IDA), consumo de ração (CR), ganho de peso diário (GPD) e PN (peso ao nascer), relacionadas ao desempenho, nome dado a essa nova variável. As variáveis CR, GPD e PN se correlacionaram positivamente com esse fator, porém a idade ao abate apresentou correlação negativa (Tabela 3). Tal resultado já era esperado, visto que quanto maior for o consumo de ração, o ganho de peso diário e o peso ao nascer, maior será o seu desempenho e menor será a quantidade de dias para o abate desse animal.

Para os 4 fatores obtidos, estimou-se o modelo com base nas metodologias Bayes A, Bayes B, RR-BLUP Bayes e BLASSO e os valores de acurácia para cada abordagem foram obtidos. Utilizando essa medida como critério, podemos observar que a melhor metodologia foi o LASSO bayesiano (Figura 1). Esse resultado é corroborado pelo estudo de DE LOS CAMPOS et al. (2009) em que o BLASSO obteve resultados superiores ao Bayes A e Bayes B. Observa-se também que o segundo fator, denotado por "gordura", foi o que apresentou maior valor de acurácia (0,56) comparando-se com os demais fatores (Figura 1).

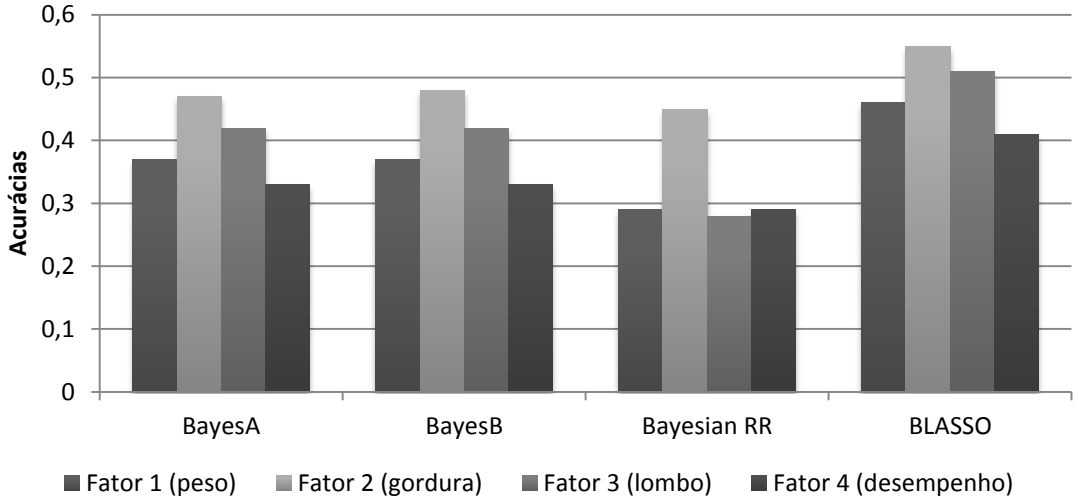

Figura 1. Acurácias das metodologias utilizadas, de acordo com o fator interpretável.

Diante da maior acurácia e da disponibilidade de trabalhos na literatura que utilizam as variáveis que compõem o segundo fator (Paixão et al., 2012; Silva et al., 2011; Azevedo et al., 2013; Azevedo et al., 2015), apresentou-se apenas os resultados e discussões para o mesmo. Os resultados para as demais variáveis são apresentados no apêndice C.

Observa-se que o valor da acurácia referente à variável latente "gordura" (0,56) é semelhante àqueles obtidos quando as variáveis PBR (0,49), EBACON (0,63), ETO (0,58), ETL (0,59) e ETUC (0,55) que foram analisadas individualmente pelo método BLASSO (Figura 2). Apenas as acurácias referentes às variáveis fenotípicas ETUL (0,32) e ETSH (0,28) apresentaram valores discrepantes, inferiores a 0,40 (Figura 2). A proximidade destes valores sugere que a construção da variável latente produza resultados semelhantes em termos de acurácia àqueles obtidos via análises individuais das variáveis fenotípicas. Ademais, a acurácia para as variáveis ETUC, ETUL, ETL e EBACON apresentaram valores superiores aos obtidos a partir de metodologias de seleção genômica baseadas em métodos de redução de dimensionalidade apresentadas em Azevedo et al. (2013).

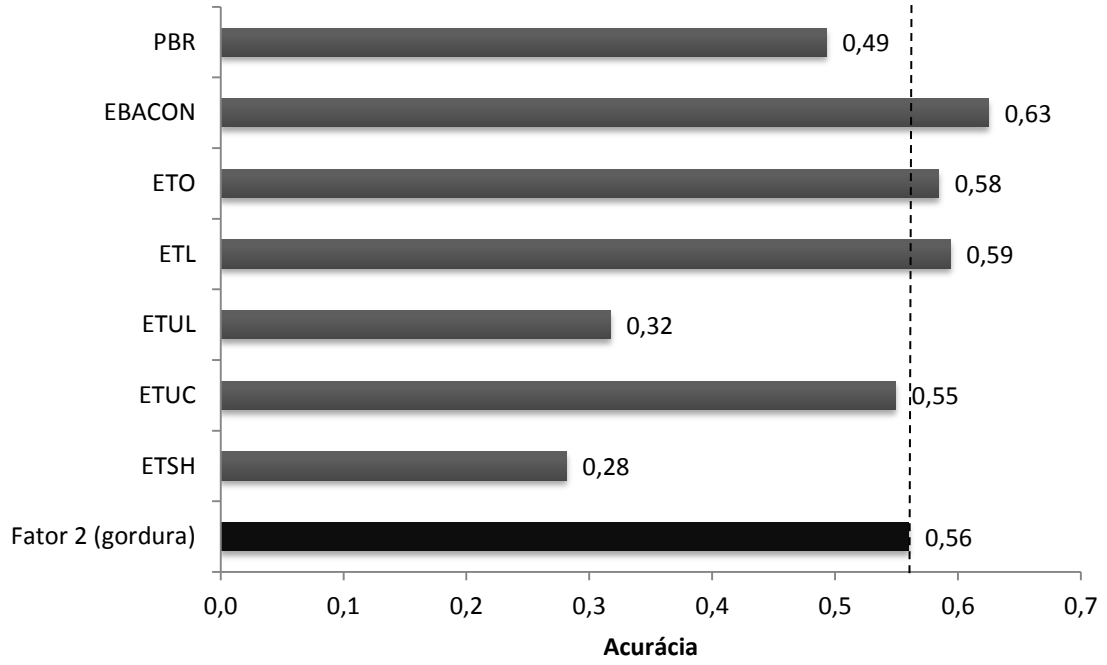

Figura 2. Acurácia relativa às variáveis e ao fator gordura.

PBR – peso da banha rama; EBACON – espessura do bacon; ETO – espessura de toucinho; ETL - espessura de toucinho medida na região acima da última vértebra lombar, na linha dorso-lombar; ETUL – espessura de toucinho entre a última e a penúltima vértebra lombar, na linha dorso-lombar; ETUC – espessura de toucinho imediatamente após a última costela na linha dorso-lombar; ETSH – maior espessura de toucinho na região da copa, na linha dorso-lombar.

Com relação à concordância entre os 10% melhores indivíduos selecionados por meio do fator (gordura) e pelas variáveis fenotípicas analisadas individualmente (ETSH, ETUC, ETUL, ETL, ETO, EBACON e PBR), observa-se resultados satisfatórios, visto que o menor coeficiente encontrado foi 0,50 para espessura do bacon (EBACON) e peso da banha rama (PBR) (Tabela 4). Pode-se notar também que de acordo com os loadings obtidos na análise fatorial (Tabela 2) essas variáveis são as que menos se relacionam com o fator, com valores de 0,79 e 0,75, respectivamente. Segundo Landis & Koch (1977) valores iguais ou superiores a 0,5 podem ser considerados bons, visto que essa medida varia entre 0 e 1. A variável que apresentou maior concordância foi à espessura de toucinho imediatamente após a última costela (ETUC), apresentando o índice Kappa de 0,72 (Tabela 4). A tabela 4 apresenta os valores obtidos para o índice de concordância e sua classificação de acordo com Landis & Koch (1977). Tais resultados sugerem novamente que a seleção de indivíduos por meio de um fator comum altamente correlacionado com um grupo de variáveis pode ser uma estratégia interessante quando objetiva-se selecionar indivíduos para várias variáveis simultaneamente.

Tabela 3. Coeficientes de concordância relacionados a cada variável e o fator gordura.

Variáveis	Cohen's Kappa	Classificação
ETSH	0,60	Bom
ETUC	0,72	Muito bom
ETUL	0,60	Bom
ETL	0,57	Bom
ETO	0,60	Bom
EBACON	0,50	Bom
PBR	0,50	Bom

PBR – peso da banha rama; EBACON – espessura do bacon; ETO – espessura de toucinho; ETL - espessura de toucinho medida na região acima da última vértebra lombar, na linha dorso-lombar; ETUL – espessura de toucinho entre a última e a penúltima vértebra lombar, na linha dorso-lombar; ETUC – espessura de toucinho imediatamente após a última costela na linha dorso-lombar; ETSH – maior espessura de toucinho na região da copa, na linha dorso-lombar.

Por fim, podemos verificar o padrão de comportamento dos efeitos dos marcadores nos respectivos cromossomos de acordo com as variáveis e o respectivo fator (Figura 3). De maneira geral, o padrão dos efeitos de marcadores encontrados para as variáveis foi semelhante ao que foi encontrado para o fator. Especificamente, pode-se notar que o fator "gordura" apresentou maiores efeitos nos cromossomos 1, 4, 7 e 17, os mesmos que se destacaram para a variável ETSH (Figura 3). Esse resultado corrobora com o que foi encontrado por Guo et al. (2008), que localizaram QTL's para a mesma característica nos cromossomos 1 e 7 em duas populações de suínos Meishan x Large White e com os resultados obtidos Yin et al. (2012), que identificaram a presença de QTL para essa mesma variável no cromossomo 17. A espessura de toucinho imediatamente após a última costela (ETUC), variável que apresentou maior destaque em relação à concordância na seleção apresentou maiores valores para os cromossomos 1, 4, 7 e 17, assim como o fator 2. Resultado concordante com o estudo realizado por Fan et al. (2011), que identificaram QTL's nos mesmos cromossomos em uma população de fêmeas da raça Large White e com HIDALGO et al. (2013) que localizaram um QTL no cromossomo 4 ligado a essa mesma característica. O cromossomo 1 apresentou maior relevância também nas variáveis ETUL, ETL e ETO, semelhante ao que foi encontrado por BEECKMANN et al.(2003), ROHRER et al.(1998) e por Azevedo et al. (2013).

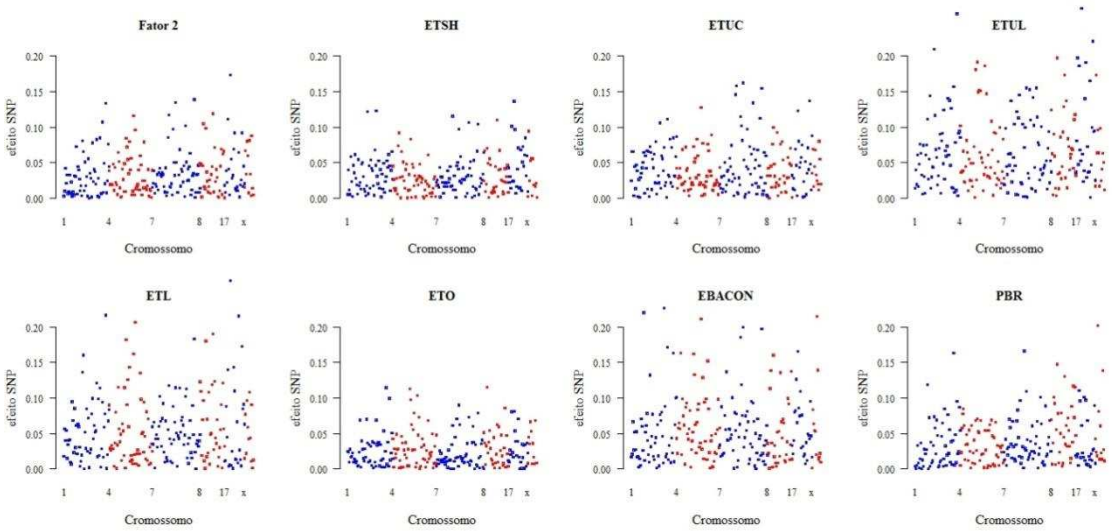

Figura 3. Acurácia relativa às variáveis e ao fator gordura.

PBR – peso da banha rama; EBACON – espessura do bacon; ETO – espessura de toucinho; ETL - espessura de toucinho medida na região acima da última vértebra lombar, na linha dorso-lombar; ETUL – espessura de toucinho entre a última e a penúltima vértebra lombar, na linha dorso-lombar; ETUC – espessura de toucinho imediatamente após a última costela na linha dorso-lombar; ETSH – maior espessura de toucinho na região da copa, na linha dorso-lombar.

4. Conclusões

A utilização da análise de fatores na construção de variáveis latentes para posterior uso em estudos de seleção genômica é uma abordagem interessante e promissora, visto que o valor da acurácia referente à variável latente foi semelhante àqueles obtidos quando as variáveis foram analisadas separadamente. Além disso, os valores de concordância entre os 10% melhores indivíduos selecionados por meio do fator e pelas variáveis fenotípicas analisadas individualmente foram satisfatórios e o padrão dos efeitos de marcadores encontrados para as variáveis foi semelhante ao que foi encontrado para o fator.

REFERÊNCIAS BIBLIOGRÁFICAS

ABPA: Associação Brasileira de Proteína Animal. Disponível em: <http://www.abipecs.org.br/pt/estatisticas/mundial/producao-2.html>. Acesso em Nov. 2014.

AZEVEDO, C. F.; NASCIMENTO, M; SILVA, F. F e; RESENDE, M. D. V. de; LOPES, P. S.; GUIMARÃES, S. E. F. A comparison of dimensionality reduction methods to predict genomic breeding values for carcass traits in pigs. **Genetics and Molecular Research**, 2015 (no prelo).

AZEVEDO, C. F.; RESENDE, M. D. V.; SILVA, F. F.; Lopes, P. S.; GUIMARÃES, S. E. F. Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. **Pesquisa Agropecuária Brasileira**, v. 48, p. 619-626, 2013.

BEECKMANN, P. et al. Linkage and QTL mapping for Susscrofa chromosome 1. **Journal of Animal Breeding and Genetics**, v.120, p.1-10, 2003.

BERNAARDS & JENNRICH, GPA Factor Rotation.
URL: http://cran.r-project.org/web/packages/GPArotation/index.html, 2014.

DE LOS CAMPOS, G.; NAYA, h.; GIANOLA, D.; CROSSA, J.; LEGARRA, A.; MANFREDI, E.; WEIGEL, K.; COTES, J. M. Predicting quantitative traits with regression models for dense molecular markers. **Genetics**, Austin, v. 182, p. 375-385, 2009.

CAMPOS & RODRIGUEZ, Bayesian Generalized Linear Regression.
URL: http://cran.r-project.org/web/packages/BGLR/index.html, 2014.

COHEN, J. A coeficient of agreement for nominal scales. **Educational and Psychologial Measurement.**, v. 20, p. 37-46, 1960.

CRUZ, C. D.; SALGADO, C. C.; BHERING, L. L. **Genômica Aplicada**. Visconde do Rio Branco, MG: Ed. Suprema, 424p., 2013.

FAN, B.; ONTERU, S. K.; DU, Z. Q.; GARRICK, D. J.; STALDER, K. J.; ROTHSCHILD, M. F. Genome-wide association study identifies Loci for composition and sctructural soundness traits in pigs. **PlosOne**. v. 6. p. 1-11. 2011.

FERREIRA, D. F. **Estatística Multivariada**. 2.Ed. Lavras: Ed. UFLA. 675p., 2011.

GUO, Y-M.; LEE, G. J.; ARCHIBALD, A. L.; HALEY, C. S.; Quantitative trait loci for production traits in pigs: a combined analysis of two Meishan x Large White populations. **Animal genetics**, v. 39.p. 486-495, 2008.

HIDALGO, A.M. et al. **Fine mapping and single nucleotide polymorphism effects estimation on pig chromossomes 1,4,7,8,17 and x.**. Dissertation of Genetics and Breeding - University Federal of Viçosa, Viçossa, MG.p.31, 2011.

LANDIS, J.; KOCH, G.The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p.159-174. 1977.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction os total genetic value using Genome-Wide dense marker maps. **Genetics Society of America**.V. 157. p. 1819-1829, 2001.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: Uma abordagem aplicada**. 1ª reimpressão. Belo Horizonte – MG. Ed. UFMG. 295p., 2007.

PAIXÃO, D. M.; CARNEIRO, P. L. S.; PAIVA, S. R.; SOUSA, K. R. S.; VERARDO, L. L.; BRACCINI NETO, J. ; PINTO, A. P. G.; HIDALGO, A. M.; NASCIMENTO, C. S.; PÉRISSÉ, I. V.; LOPES, P. S.; GUIMARÃES, S. E. F. Mapeamento de QTL nos cromossomos 1, 2, 3, 12, 14, 15 e X em suínos: características de carcaça e qualidade de carne. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 64, p. 974-982, 2012.

R DEVELOPMENT CORE TEAM. **R**: A language and environment for statistical computing. Vienna, Austria: R Fundation for Statistical Computing, 2009. Disponível em: <http://r-project .org>. Acesso em: Dez. 2015.

RESENDE, M. D. V.; SILVA, F. F.; VIANA, J. M.; PETTERNELLI, L. A.; RESENDE JR, M. F. R.; VALLE, P. **Computação da Seleção Genômica Ampla**. Colombo: EMBRAPA Florestas, 79p. 2010.

RESENDE, M.D.V.; SILVA, F. F.; VIANA, J. M. S.; PETTERNELLI, L. A.; RESENDE JR, M. F. R; VALLE, P. M. **Métodos estatísticos na seleção genômica ampla**. Colombo: Embrapa Florestas, 2011. 107p.

RESENDE, M. D.; SILVA, F. F.; LOPES, P. S.; AZEVEDO, C. F. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial.** Viçosa: Universidade Federal de Viçosa/Departamento de Estatística. 2012. 291p. Disponível em: http://www.det.ufv.br/ppestbio/corpo_docente.php. Acesso em Fev. 2015.

REVELLE, Procedures for Psychological, Psychometric, and Personality Research.
URL: http://cran.r-project.org/web/packages/psych/index.html, 2015.

ROHRER, G.A. et al. Identification of quantitative trait loci affecting carcass composition in swine: I. Fat deposition traits. **Journal of Animal Science**, v.76, p.47-54, 1998.

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

SILVA, N. C. N.; FERREIRA, W. L.; CIRILLO, M. A.; SCALON, J. D. Uso da análise fatorial na descrição e identificação dos perfis característicos de municípios de Minas Gerais. **Revista Brasileira de Biometria**, São Paulo, v. 32, n.2, p.201-215, 2014.

TIBSHIRANI, R. Regression shrinkage and selection via the LASSO. **Journal of the Royal Statistics Society**. v. 58, p. 267-288, 1996.

YIN, Q.; YANG, H-W.; HAN, X-L.; FAN, B.; LIU, B. Isolation, mapping, SNP detection and association with backfat traits of the porcine CTNNBL1 and DGAT2 genes. **Molecular Biology Reports**, v. 39, p. 4485-4490. 2012.

ZHAO, Genetic analysis package.
URL: http://cran.r-project.org/web/packages/gap/index.html, 2014.

CONSIDERAÇÕES FINAIS

Este trabalho foi uma proposta da implementação de técnica multivariada da análise de fatores no contexto da seleção genômica ampla.

Com a aplicação da análise de fatores, houve uma redução de 41 variáveis fenotípicas para dez variáveis latentes, sendo que quatro apresentaram interpretação prática (F1 – “Peso”; F2 – “Gordura”; F3 – “Lombo”; F4 – “Desempenho”), agrupando um total de 28 caracteres fenotípicos, o que é uma redução considerável da quantidade de variáveis.

Após a construção dos fatores, foi feita uma aplicação de metodologias bayesianas no contexto de seleção genômica ampla para estimar o mérito genético dos indivíduos através dessas novas variáveis, as quais representam um conjunto de características correlacionadas. Os resultados foram satisfatórios em termos de acurácia, concordância entre os 10% melhores indivíduos selecionados por meio do fator (“Gordura”) e pelas variáveis fenotípicas analisadas individualmente. Além disso, o padrão dos efeitos de marcadores encontrados para as variáveis foi semelhante ao que foi encontrado para o fator.

Logo, a utilização de variáveis latentes em estudos de seleção genômica é uma abordagem interessante e promissora. Portanto, novos trabalhos neste contexto serão realizados, com a utilização de um conjunto de dados mais onde todo o genoma esteja representado.

APÊNDICES

Apêndice A: Algoritmos utilizados para análise

As rotinas computacionais dos métodos descritos neste trabalho foram implementadas no software livre R (R Development Core Team, 2012) e estão descritas a seguir.

A.1.Leitura do conjunto de dados e correção das variáveis.

```
dados<-read.table("bdartigo2.txt",h=T)
mapa<-read.table("mapa2.txt",h=T)
attach(dados)
head(dados)
library("psych")
library("GPArotation")
library("BGLR")
library("gap")
# Separar as variáveis:
v<-matrix(NA,345,41)
for (i in 5:45){
  v[,i-4]<-dados[,i]
}
head(v)
v
# Correção da Matriz v:
sexo<-dados[,2]
lote<-dados[,3]
hal<-dados[,4]
vcor<-matrix(NA,345,41)
for (i in 1:41){
vcor[,i]<-lm(v[,i]~sexo+lote+hal)$residuals
}
# Fator 2
ETSH<-vcor[,9]
ETUC<-vcor[,10]
ETUL<-vcor[,11]
ETL<-vcor[,12]
ETO<-vcor[,13]
```

```
EBACON<-vcor[,14]
PBR<-vcor[,27]
# Demais variáveis
PCARC<-vcor[,1]
PCD<-vcor[,2]
TLD<-vcor[,3]
TLN<-vcor[,4]
IDA<-vcor[,5]
RCARC<-vcor[,6]
MBCC<-vcor[,7]
MLC<-vcor[,8]
PROLOM<-vcor[,15]
AOL<-vcor[,16]
CORAC<-vcor[,17]
PP<-vcor[,18]
PPL<-vcor[,19]
PCOPA<-vcor[,20]
PPA<-vcor[,21]
PC<-vcor[,22]
PL<-vcor[,23]
PB<-vcor[,24]
PCOST<-vcor[,25]
PF<-vcor[,26]
CR<-vcor[,28]
GPD<-vcor[,29]
CA<-vcor[,30]
NT<-vcor[,31]
PA<-vcor[,32]
PN<-vcor[,33]
ph45<-vcor[,34]
ph24<-vcor[,35]
L<-vcor[,36]
GOINTR<-vcor[,37]
PGOTEJ<-vcor[,38]
PCOZ<-vcor[,39]
MACIEZ<-vcor[,40]
C<-vcor[,41]
```

A.2. Análise de fatores.

```
# KMO, Scree-plot e análise de fatores (componentes principais)
C<-cor(vcor)
KMO(C)
autos<-eigen(C)
plot(seq(1:ncol(vcor)),autos$values,type="l",ylab="Autovalores",xlab="
Fatores")
AF<-principal(vcor,nfactors=10,rotate="varimax")
AF
escores<-AF$scores
head(escores)
e.peso<-escores[,1]
e.gordura<-escores[,2]
e.lombo<-escores[,3]
e.desempenho<-escores[,4]
# Gráfico de dispersão entre os fatores:
par(mfrow=c(3,2))
plot(e.peso,e.gordura,pch=16,xlab="Peso",ylab="Gordura",main="(a)")
abline(0,0,0,0)
plot(e.peso,e.lombo,pch=16,xlab="Peso",ylab="Lombo",main="(b)")
abline(0,0,0,0)
plot(e.peso,e.desempenho,pch=16,xlab="Peso",ylab="Desempenho",main="(c
)")
abline(0,0,0,0)
plot(e.gordura,e.lombo,pch=16,xlab="Gordura",ylab="Lombo",main="(d)")
abline(0,0,0,0)
plot(e.gordura,e.desempenho,pch=16,xlab="Gordura",ylab="Desempenho",ma
in="(e)")
abline(0,0,0,0)
plot(e.lombo,e.desempenho,pch=16,xlab="Lombo",ylab="Desempenho",main="
(f)")
abline(0,0,0,0)
```

A.3. Metodologias Bayesianas (fator 2).

```
######### Criando arquivo de genótipos e fenótipos (fator 2) #########
geno<-dados[,46:282]
head(geno)
dim(geno)
feno<-e.gordura
feno<-as.matrix(feno)
```

```
rownames(feno)<-c(ID)
feno
######### Validação Cruzada - Método 1 #########
##### FENÓTIPOS #####
# Populações individuais:
feno1=feno[1:115]
feno1=as.matrix(feno1)
dim(feno1)
feno2=feno[116:230]
feno2=as.matrix(feno2)
feno3=feno[231:345]
feno3=as.matrix(feno3)
# Populações agrupadas:
feno12=feno[1:230]
feno12=as.matrix(feno12)
feno13=rbind(feno1,feno3)
feno23=feno[116:345]
feno23=as.matrix(feno23)
##### GENÓTIPOS #####
### Populações individuais ###
geno1=geno[1:115,]
geno1=as.matrix(geno1)
geno2=geno[116:230,]
geno2=as.matrix(geno2)
geno3=geno[231:345,]
geno3=as.matrix(geno3)

### Populações agrupadas ###
geno12=geno[1:230,]
geno12=as.matrix(geno12)
geno13=rbind(geno1,geno3)
geno23=geno[116:345,]
geno23=as.matrix(geno23)
### Frequências alélicas ###
p2=matrix(0,ncol(geno),1)
q2=matrix(0,ncol(geno),1)
for(i in 1:ncol(geno))
{
q2[i,]=(2*length(which(geno[,i]==0))+length(which(geno[,i]==1)))/(2*le
ngth(which(geno[,i]==0))+2*length(which(geno[,i]==1))+2*length(which(g
eno[,i]==2)))
```

```
p2[i,]=(length(which(geno[,i]==1))+2*length(which(geno[,i]==2)))/(2*le
ngth(which(geno[,i]==0))+2*length(which(geno[,i]==1))+2*length(which(g
eno[,i]==2)))
}
### Títulos das linhas (para seleção) ###
nlinhas=dados[,1]
np1=nlinhas[1:115]     # população 1
np2=nlinhas[116:230]   # população 2
np3=nlinhas[231:345]   # população 3
##### Bayes B #####
### Treinamento: 1 e 2 | Validação: 3 ###
BB1=BGLR(y=feno12,ETA=list(list(X=geno12,model='BayesB',probIn=0.9)),n
Iter=100000,burnIn=20000,thin=10)
gebv_val_BB3=geno3%*%BB1$ETA[[1]]$b
cor_BB1=cor(feno3,gebv_val_BB3)
cor_BB1
## Herdabilidade ##
V=BB1$ETA[[1]]$varB      # variância do marcador
Ve=BB1$varE              # variância residual
Va=t(2*p2*q2)%*%V
Vfen=Va+Ve
h2aBB1=Va/Vfen
h2aBB1
acurBB1<-cor_BB1/sqrt(h2aBB1)
acurBB1
### Treinamento: 1 e 3 | Validação: 2 ###
BB2=BGLR(y=feno13,ETA=list(list(X=geno13,model='BayesB',probIn=0.9)),n
Iter=100000,burnIn=20000,thin=10)
gebv_val_BB2=geno2%*%BB2$ETA[[1]]$b
cor_BB2=cor(feno2,gebv_val_BB2)
cor_BB2
## Herdabilidade ##
V=BB2$ETA[[1]]$varB      # variância do marcador
Ve=BB2$varE              # variância residual
Va=t(2*p2*q2)%*%V
Vfen=Va+Ve
h2aBB2=Va/Vfen
h2aBB2
acurBB2<-cor_BB2/sqrt(h2aBB2)
acurBB2
### Treinamento: 2 e 3 | Validação: 1 ###
```

```
BB3=BGLR(y=feno23,ETA=list(list(X=geno23,model='BayesB',probIn=0.9)),n
Iter=100000,burnIn=20000,thin=10)
gebv_val_BB1=geno1%*%BB3$ETA[[1]]$b
cor_BB3=cor(feno1,gebv_val_BB1)
cor_BB3
## Herdabilidade ##
V=BB3$ETA[[1]]$varB      # variância do marcador
Ve=BB3$varE              # variância residual
Va=t(2*p2*q2)%*%V
Vfen=Va+Ve
h2aBB3=Va/Vfen
h2aBB3
acurBB3<-cor_BB3/sqrt(h2aBB3)
acurBB3
mcBB<-(cor_BB1+cor_BB2+cor_BB3)/3
mcBB # média das capacidades preditivas BayesB
mhBB<-(h2aBB1+h2aBB2+h2aBB3)/3
mhBB # média das herdabilidadesBayesB
maBB<-(acurBB1+acurBB2+acurBB3)/3
maBB
##### Bayes A #####
### Treinamento: 1 e 2 | Validação: 3 ###
BA1=BGLR(y=feno12,ETA=list(list(X=geno12,model='BayesA')),nIter=100000
,burnIn=20000,thin=10)
gebv_val_BA3=geno3%*%BA1$ETA[[1]]$b
cor_BA1=cor(feno3,gebv_val_BA3)
cor_BA1
## Herdabilidade ##
V=BA1$ETA[[1]]$varB      # variância do marcador
Ve=BA1$varE              # variância residual
Va=t(2*p2*q2)%*%V
Vfen=Va+Ve
h2aBA1=Va/Vfen
h2aBA1
acurBA1<-cor_BA1/sqrt(h2aBA1)
acurBA1
### Treinamento: 1 e 3 | Validação: 2 ###
BA2=BGLR(y=feno13,ETA=list(list(X=geno13,model='BayesA')),nIter=100000
,burnIn=20000,thin=10)
gebv_val_BA2=geno2%*%BA2$ETA[[1]]$b
cor_BA2=cor(feno2,gebv_val_BA2)
```

```
cor_BA2
## Herdabilidade ##
V=BA2$ETA[[1]]$varB      # variância do marcador
Ve=BA2$varE              # variância residual
Va=t(2*p2*q2)%*%V
Vfen=Va+Ve
h2aBA2=Va/Vfen
h2aBA2
acurBA2<-cor_BA2/sqrt(h2aBA2)
acurBA2
### Treinamento: 2 e 3 | Validação: 1 ###
BA3=BGLR(y=feno23,ETA=list(list(X=geno23,model='BayesA')),nIter=100000
,burnIn=20000,thin=10)
gebv_val_BA1=geno1%*%BA3$ETA[[1]]$b
cor_BA3=cor(feno1,gebv_val_BA1)
cor_BA3
## Herdabilidade ##
V=BA3$ETA[[1]]$varB      # variância do marcador
Ve=BA3$varE              # variância residual
Va=t(2*p2*q2)%*%V
Vfen=Va+Ve
h2aBA3=Va/Vfen
h2aBA3
acurBA3<-cor_BA3/sqrt(h2aBA3)
acurBA3
mcBA<-(cor_BA1+cor_BA2+cor_BA3)/3
mcBA # média das capacidades preditivas BayesA
mhBA<-(h2aBA1+h2aBA2+h2aBA3)/3
mhBA # média das herdabilidadesBayesA
maBA<-(acurBA1+acurBA2+acurBA3)/3
maBA
##### Bayesian RR #####
### Treinamento: 1 e 2 | Validação: 3 ###
BRR1=BGLR(y=feno12,ETA=list(list(X=geno12,model='BRR')),nIter=100000,b
urnIn=20000,thin=10)
ahat_BRR1=BRR1$ETA[[1]]$b #vetor de efeitos estimados SNPs BRR
gebv_val_BRR3=geno3%*%BRR1$ETA[[1]]$b
cor_BRR1=cor(feno3,gebv_val_BRR3)
cor_BRR1
## Herdabilidade ##
V=BRR1$ETA[[1]]$varB      # variância do marcador
```

```
Ve=BRR1$varE              # variância residual
Va=2*V*sum(p2*q2)
Vfen=Va+Ve
h2aBRR1=Va/Vfen
h2aBRR1
acurBRR1<-cor_BRR1/sqrt(h2aBRR1)
acurBRR1
### Treinamento: 1 e 3 | Validação: 2 ###
BRR2=BGLR(y=feno13,ETA=list(list(X=geno13,model='BRR')),nIter=100000,b
urnIn=20000,thin=10)
ahat_BRR2=BRR2$ETA[[1]]$b #vetor de efeitos estimados SNPs BRR
gebv_val_BRR2=geno2%*%BRR2$ETA[[1]]$b
cor_BRR2=cor(feno2,gebv_val_BRR2)
cor_BRR2
## Herdabilidade ##
V=BRR2$ETA[[1]]$varB      # variância do marcador
Ve=BRR2$varE              # variância residual
Va=2*V*sum(p2*q2)
Vfen=Va+Ve
h2aBRR2=Va/Vfen
h2aBRR2
acurBRR2<-cor_BRR2/sqrt(h2aBRR2)
acurBRR2
### Treinamento: 2 e 3 | Validação: 1 ###
BRR3=BGLR(y=feno23,ETA=list(list(X=geno23,model='BRR')),nIter=100000,b
urnIn=20000,thin=10)
ahat_BRR3=BRR3$ETA[[1]]$b #vetor de efeitos estimados SNPs BRR
gebv_val_BRR1=geno1%*%BRR3$ETA[[1]]$b
cor_BRR3=cor(feno1,gebv_val_BRR1)
cor_BRR3
## Herdabilidade ##
V=BRR3$ETA[[1]]$varB      # variância do marcador
Ve=BRR3$varE              # variância residual
Va=2*V*sum(p2*q2)
Vfen=Va+Ve
h2aBRR3=Va/Vfen
h2aBRR3
acurBRR3<-cor_BRR3/sqrt(h2aBRR3)
acurBRR3
maBRR<-(acurBRR1+acurBRR2+acurBRR3)/3
maBRR        # média das acurácias do BayesianRR
```

```
mcBRR<-(cor_BRR1+cor_BRR2+cor_BRR3)/3
mcBRR # média das capacidades BayesianRR
mhBRR<-(h2aBRR1+h2aBRR2+h2aBRR3)/3
mhBRR # média das herdabilidadesBayesianRR
##### Bayesian LASSO #####
### Treinamento: 1 e 2 | Validação: 3 ###
BL1=BGLR(y=feno12,ETA=list(list(X=geno12,model='BL')),nIter=100000,bur
nIn=20000,thin=10)
ahat_BL1=BL1$ETA[[1]]$b #vetor de efeitos estimados SNPs BLASSO
gebv_val_BL3=geno3%*%BL1$ETA[[1]]$b
cor_BL1=cor(feno3,gebv_val_BL3)
cor_BL1
## Herdabilidade ##
v=BL1$ETA[[1]]$tau2
Ve=BL1$varE
t=matrix(v)*BL1$varE
Va=sum(2*p2*q2*t)
Vfen=Va+Ve
h2aBL1=Va/Vfen
h2aBL1
acurBL1<-cor_BL1/sqrt(h2aBL1)
acurBL1
### Treinamento: 1 e 3 | Validação: 2 ###
BL2=BGLR(y=feno13,ETA=list(list(X=geno13,model='BL')),nIter=100000,bur
nIn=20000,thin=10)
ahat_BL2=BL2$ETA[[1]]$b #vetor de efeitos estimados SNPs BRR
gebv_val_BL2=geno2%*%BL2$ETA[[1]]$b
cor_BL2=cor(feno2,gebv_val_BL2)
cor_BL2
## Herdabilidade ##
v=BL2$ETA[[1]]$tau2
Ve=BL2$varE
t=matrix(v)*BL2$varE
Va=sum(2*p2*q2*t)
Vfen=Va+Ve
h2aBL2=Va/Vfen
h2aBL2
acurBL2<-cor_BL2/sqrt(h2aBL2)
acurBL2
### Treinamento: 2 e 3 | Validação: 1 ###
```

```
BL3=BGLR(y=feno23,ETA=list(list(X=geno23,model='BL')),nIter=100000,bur
nIn=20000,thin=10)
ahat_BL3=BL3$ETA[[1]]$b #vetor de efeitos estimados SNPs BRR
gebv_val_BL1=geno1%*%BL3$ETA[[1]]$b
cor_BL3=cor(feno1,gebv_val_BL1)
cor_BL3
## Herdabilidade ##
v=BL3$ETA[[1]]$tau2
Ve=BL3$varE
t=matrix(v)*BL3$varE
Va=sum(2*p2*q2*t)
Vfen=Va+Ve
h2aBL3=Va/Vfen
h2aBL3
acurBL3<-cor_BL3/sqrt(h2aBL3)
acurBL3
maBL<-(acurBL1+acurBL2+acurBL3)/3
maBL
media_BL<-(cor_BL1+cor_BL2+cor_BL3)/3
media_BL
maBL # média das acurácias BLASSO
mhBL<-(h2aBL1+h2aBL2+h2aBL3)/3
mhBL # média das herdabilidades BLASSO
mcBL<-(cor_BL1+cor_BL2+cor_BL3)/3
mcBL
### Capacidades preditivas ###
mcBA
mcBB
mcBRR
mcBL
### Herdabilidades ###
mhBA
mhBB
mhBRR
mhBB
### Acurácias ###
maBA
maBB
maBRR
maBL # melhor método BLASSO
```

Apêndice B: Tabela 2.

Tabela 2. Tabela 2: Loadings obtidos na análise de fatores.

Variáveis	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	h2
PCARC	**0.87**	0.40	0.11	0.03	-0.07	0.06	0.00	-0.10	0.01	0.04	0.96
PCD	**0.87**	0.41	0.11	0.03	-0.06	0.06	0.00	-0.09	-0.01	0.06	0.96
TLD	-0.08	0.02	0.01	0.05	0.03	-0.07	**0.90**	0.08	-0.04	-0.06	0.84
TLN	0.05	0.03	-0.07	-0.02	0.01	-0.02	**0.92**	-0.10	0.00	-0.02	0.86
IDA	0.26	0.18	0.01	**-0.78**	-0.08	0.01	-0.08	0.03	-0.11	-0.01	0.73
RCARC	0.15	0.32	0.26	-0.09	0.08	0.10	-0.12	-0.28	0.07	**0.33**	0.42
MBCC	**0.82**	-0.25	-0.08	-0.03	0.09	-0.06	0.02	0.02	0.12	0.05	0.78
MLC	**0.84**	-0.22	-0.06	0.00	0.14	-0.10	0.09	0.03	0.16	-0.04	0.82
ETSH	0.13	**0.74**	0.00	-0.09	-0.01	-0.02	0.07	0.02	0.01	0.00	0.57
ETUC	0.06	**0.82**	0.12	0.08	-0.05	0.07	0.02	0.10	0.02	0.05	0.72
ETUL	0.05	**0.86**	0.05	0.07	-0.05	-0.02	0.08	0.00	-0.10	0.00	0.76
ETL	0.02	**0.88**	0.00	0.01	-0.05	-0.01	0.06	0.01	-0.10	0.01	0.78
ETO	0.04	**0.87**	-0.08	0.09	-0.08	-0.01	-0.06	0.05	0.02	-0.02	0.79
EBACON	0.05	**0.79**	-0.16	-0.03	-0.12	0.01	-0.07	-0.02	0.06	0.05	0.68
PROLOM	0.01	-0.07	**0.83**	0.02	0.01	-0.03	-0.04	0.05	0.02	-0.11	0.72
AOL	0.27	0.05	**0.82**	-0.03	0.03	0.09	-0.05	-0.08	0.00	0.09	0.77
CORAC	**0.52**	-0.15	0.15	-0.03	-0.02	0.00	-0.14	0.23	0.02	-0.09	0.39
PP	**0.77**	0.30	0.25	0.00	-0.09	0.13	0.00	-0.08	-0.11	-0.02	0.79
PPL	**0.78**	-0.07	0.37	-0.03	-0.06	0.10	0.03	-0.09	-0.10	0.06	0.79
PCOPA	**0.71**	0.25	0.13	0.01	-0.03	0.01	-0.02	-0.16	-0.07	0.04	0.62
PPA	**0.86**	0.10	0.02	0.04	-0.06	0.07	-0.05	-0.10	-0.10	-0.01	0.78
PC	**0.58**	0.48	0.28	0.03	-0.06	0.05	-0.01	-0.06	0.18	0.15	0.71
PL	0.51	-0.13	**0.63**	0.08	-0.02	0.08	0.07	-0.05	0.12	0.05	0.72
PB	**0.56**	0.54	-0.14	-0.05	-0.14	0.01	0.03	-0.13	0.14	-0.09	0.69
PCOST	**0.59**	0.15	-0.05	0.15	0.09	-0.09	0.02	0.16	-0.08	0.22	0.50
PF	**0.50**	0.03	0.37	0.11	-0.18	0.16	0.00	0.05	-0.31	0.09	0.57
PBR	0.26	**0.75**	-0.03	-0.06	-0.11	0.08	-0.07	-0.03	0.03	-0.11	0.67
CR	0.15	0.18	-0.04	**0.83**	-0.12	0.03	-0.01	0.02	0.07	0.08	0.78
GPD	0.28	0.02	0.00	**0.71**	0.01	0.07	0.05	-0.54	-0.07	0.01	0.89

CA	-0.17	0.15	-0.07	-0.06	-0.11	-0.03	-0.04	**0.75**	0.15	0.07	0.67
NT	-0.01	0.00	0.08	0.15	0.00	-0.01	-0.01	0.16	**0.81**	0.10	0.72
PA	**0.85**	0.31	0.05	0.02	-0.08	0.04	-0.01	-0.02	-0.04	-0.01	0.83
PN	0.11	0.05	0.19	**0.53**	0.10	-0.10	-0.08	0.39	0.09	-0.09	0.54
pH45	0.08	-0.07	-0.07	0.03	0.03	**-0.85**	0.03	0.02	-0.03	-0.10	0.75
pH24	0.23	-0.01	-0.18	0.09	**-0.60**	-0.04	0.05	0.07	-0.16	-0.15	0.51
L	0.04	-0.19	-0.06	0.00	**0.84**	0.07	0.05	-0.03	-0.15	-0.04	0.78
GOINTR	0.09	-0.07	-0.03	0.06	0.02	0.10	-0.06	0.05	0.05	**0.84**	0.75
PGOTEJ	0.17	0.06	0.07	-0.04	0.11	**0.86**	-0.07	-0.03	-0.03	0.05	0.80
PCOZ	0.18	-0.18	0.00	0.14	0.16	**0.40**	0.20	0.39	-0.13	0.03	0.48
MACIEZ	0.08	0.03	0.09	-0.16	-0.29	-0.45	0.15	0.13	**-0.41**	0.29	0.61
C	0.01	-0.20	-0.07	0.08	**0.85**	0.05	0.04	0.04	0.06	-0.03	0.79
Variância Explicada Acumulada	0.20	0.36	0.42	0.47	0.53	0.58	0.62	0.66	0.69	0.71	-

Apêndice C: Acurácias dos demais fatores.

C.1 Fator 1 – Peso.

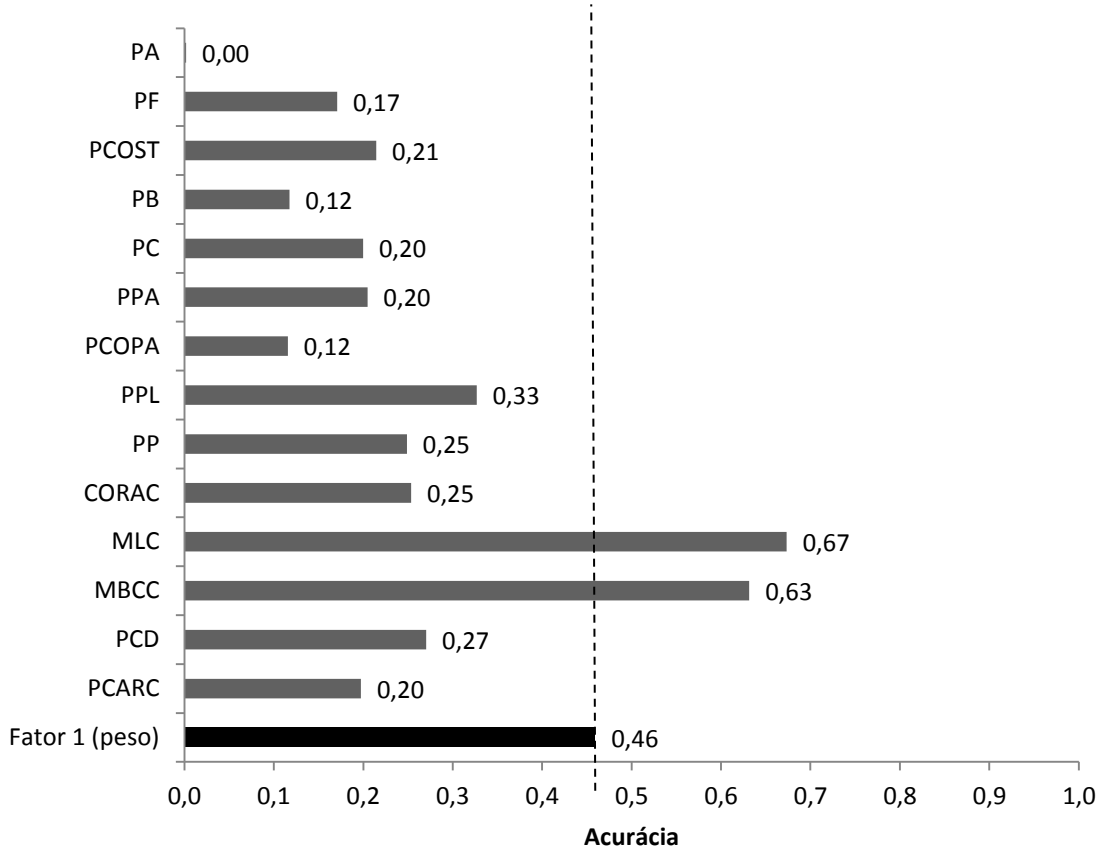

Figura 1: acurácias das variáveis relativas ao terceiro fator.

C.2 Fator 3 – Lombo.

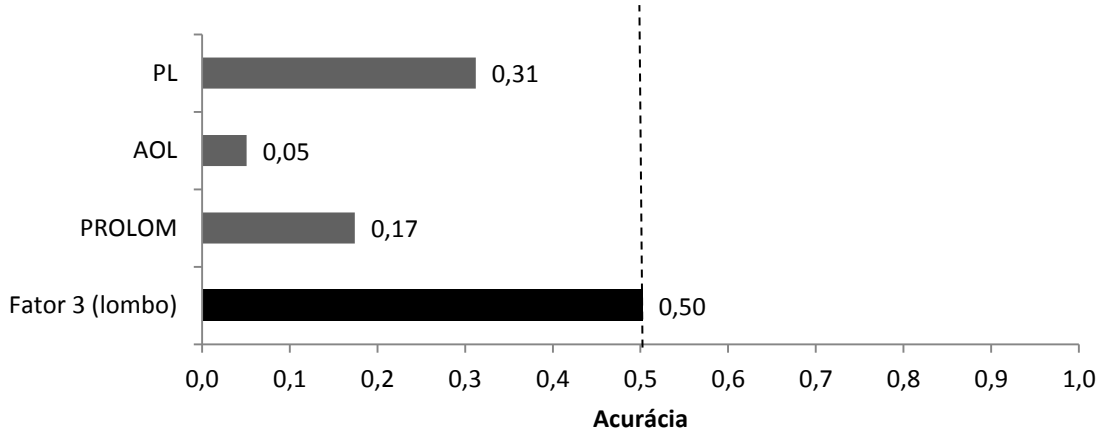

Figura 2: Acurácias das variáveis do terceiro fator.

C.3 Fator 4 – Desempenho.

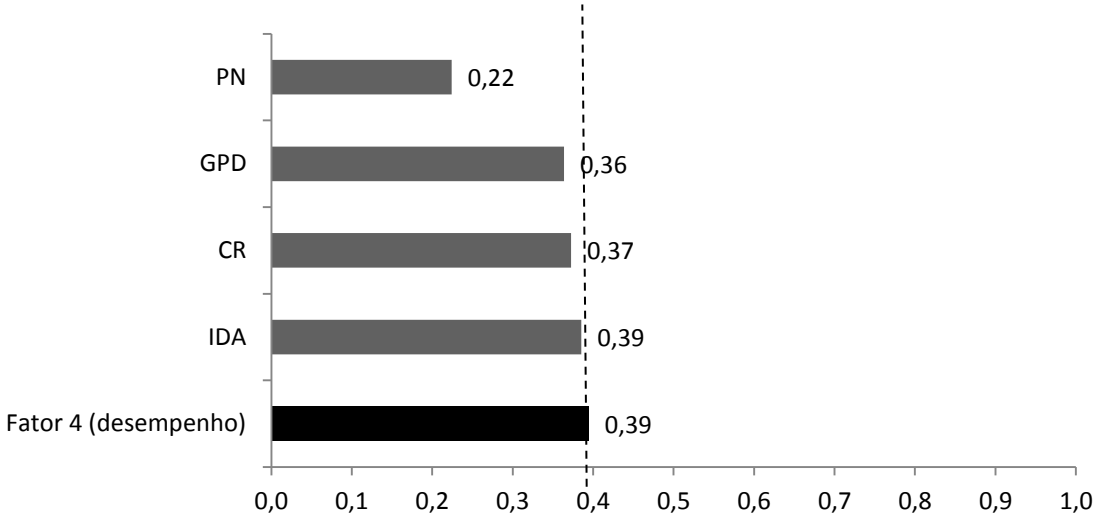

Figura 3: Acurácia das variáveis relativas ao quarto fator.

ANEXOS

ANEXO I: Teorema de Bayes

Suponha que eventos $C_1, C_2, \dots, C_k$ formem uma partição de Ω (espaço amostral) e que suas probabilidades sejam conhecidas. Suponha ainda que para um evento A, se conheçam as probabilidades $P(A|C_i)$ para todo $i = 1,2,\dots,k$. Então para qualquer j,

$$P(C_j|A) = \frac{P(A|C_j)P(C_j)}{\sum_{i=1}^{k} P(A|C_i)P(C_i)}$$