JAQUICELE APARECIDA DA COSTA

PREDIÇÃO GENÔMICA VIA REDUÇÃO DE DIMENSIONALIDADE EM MODELOS ADITIVO-DOMINANTE

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
2018

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

T
C837p
2018

Costa, Jaquicele Aparecida da, 1990-
Predição genômica via redução de dimensionalidade em modelos aditivo-dominante / Jaquicele Aparecida da Costa. – Viçosa, MG, 2018.
xiii, 107 f. : il. ; 29 cm.

Orientador: Camila Ferreira Azevedo.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Modelos multinivéis (Estatísticas). 2. Análise de componentes principais. 3. Genômica. 4. Melhoramento genético. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.5

JAQUICELE APARECIDA DA COSTA

PREDIÇÃO GENÔMICA VIA REDUÇÃO DE DIMENSIONALIDADE EM MODELOS ADITIVO-DOMINANTE

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 2018.

Fabyano Fonseca e Silva

Moysés Nascimento
(Coorientador)

Camila Ferreira Azevedo
(Orientadora)

Apesar da distância, eles sempre estiveram comigo.

Aos meus amados pais e irmão.

AGRADECIMENTOS

Agradeço imensamente a Deus, o autor de todas as obras, em especial deste trabalho. Obrigada pelo Teu amor infinito, pela Tua misericórdia, pela Tua fidelidade, pela Tua força e pela Tua proteção. Tu és a fonte de alegria que conduziu toda a pesquisa. A Virgem Maria, mestra e rainha, pelo cuidado e intercessão na realização de mais um sonho.

Agradeço aos meus amados pais, Maria Helena e José Raimundo e meu amado irmão Josimar pela dedicação total, apoio e pelo amor incondicional. Obrigada por estarem comigo nos momentos difíceis e me ajudarem através das orações. Sou grata também por se alegrarem comigo nos momentos em que as dificuldades foram superadas. Esta vitória é de vocês!

Aos meus amigos/irmãos da RCC Viçosa, Grupo de Oração Cenáculo do Senhor, Ministério Universidades Renovadas e Missão Pentecostes por serem presentes de Deus que me ajudaram a sustentar e fortalecer a minha fé. Obrigada pelas orações!

Aos amigos do PPESTBIO de forma especial: a Denilson pelo incentivo e apoio para iniciar o curso de mestrado. Obrigada por acreditar em mim! A Roberta, Gabriela, Leísa, Ana Carolina, Lucas e Ithalo pelas brincadeiras de imagem e ação, pelas risadas, pelos cafés, lanches e acima de tudo por todo incentivo. Aos amigos do LICAE pelas sugestões sempre valiosas e a minha amiga Daniela por sempre estar do meu lado. Obrigada pelos momentos inesquecíveis compartilhados.

Agradeço carinhosamente a Doutora e orientadora Camila Ferreira Azevedo pelos conselhos sempre tão valiosos, pelas críticas construtivas, pela disponibilidade, pelos ensinamentos, pela paciência, confiança, apoio e dedicação a pesquisa. Obrigada por me ensinar com o seu jeito de ser a zelar por tudo que me é confiado. Sou muito grata a Deus por me dar a oportunidade de trilhar o caminho do conhecimento do seu lado. Muito obrigada!

A Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, pela oportunidade de ampliar o conhecimento.

Aos Doutores e coorientadores Moysés Nascimento e Marcos Deon Vilela de Resende, pela disponibilidade, confiança, incentivo e pelos saberes transmitidos.

Aos membros da banca examinadora, Professor Fabyano Fonseca e Silva e Professor Moysés Nascimento, pela disponibilidade e pelas sugestões e críticas que tanto contribuíram para o presente trabalho.

Aos professores do Programa de Pós-Graduação em Estatística Aplicada e Biometria, pelos saberes transmitidos.

Aos funcionários do Departamento de Estatística, de forma especial Junior e Anita, pela prontidão e por alegrar o dia com o sorriso de sempre.

A CAPES, pela concessão da bolsa de estudos.

Enfim, a todos a minha eterna gratidão. Obrigada!

BIOGRAFIA

JAQUICELE APARECIDA DA COSTA, filha de Maria Helena da Costa e de José Raimundo da Costa, nasceu em Dom Silvério, Minas Gerais, em 11 de novembro de 1990.

Em março de 2009, ingressou no curso de Licenciatura em Matemática na Universidade Federal de Viçosa, Viçosa-MG, graduando-se em setembro de 2013.

Em março de 2016, iniciou o curso de Mestrado no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa de dissertação em 26 de fevereiro de 2018.

SUMÁRIO

RESUMO

COSTA, Jaquicele Aparecida da, M.Sc., Universidade Federal de Viçosa, fevereiro de 2018, **Predição Genômica via Redução de Dimensionalidade em Modelos Aditivo-Dominante.** Orientadora: Camila Ferreira Azevedo. Coorientadores: Marcos Deon Vilela de Resende e Moysés Nascimento.

Grandes avanços no melhoramento animal e vegetal têm sido propiciados utilizando-se informações da genética molecular. Nessa perspectiva, idealizaram a Seleção Genômica Ampla (Genome Wide Selection – GWS) cuja abordagem envolve a cobertura completa do genoma utilizando milhares de marcadores SNPs (Single Nucleotide Polymorphisms). O objetivo é estimar o mérito genético dos indivíduos e para tal, as pesquisas realizadas na GWS se baseiam na busca e na aplicação de metodologias estatísticas que visam resolver os problemas enfrentados no processo de estimação, como a alta dimensionalidade e a alta colinearidade entre os marcadores. Dentre elas, se destacam os métodos de redução de dimensionalidade: Regressão via Componentes Principais (PCR), Quadrados Mínimos Parciais (PLS) e Regressão via Componentes Independentes (ICR) e o tradicional método de regularização/shrinkage, G-BLUP (Genomic Best Linear Unbiased Predictor). Assim, o primeiro capítulo contempla as ideias centrais e a importância da GWS para o melhoramento genético, a definição de efeitos aditivos e de efeitos devido à dominância, os problemas estatísticos enfrentados na estimação dos efeitos de marcadores nos fenótipos pelo método usual baseado em quadrados mínimos ordinários, bem como as metodologias estatísticas baseadas em redução dimensional para resolver tais problemas e os procedimentos de validação que tem por finalidade comparar as metodologias estatísticas da GWS. Já o segundo capítulo refere-se a proposição e aplicação de sete critérios para a escolha do número ótimo de componentes independentes a serem utilizados na ICR, considerando apenas os efeitos aditivos. Os critérios consistem em determinar que o número de componentes independentes seja igual ao número de componentes que conduz: (i) os valores genômicos estimados via PCR a um maior valor de acurácia; (ii) os valores genômicos estimados via PCR a um menor valor de viés; (iii) a PCR a 80% de explicação da variação total de X; (iv) a PCR a 80% de explicação da variação total de Y; (v) a ICR a 80% de explicação da variação total de X; além dos critérios que consistem no número de componentes independentes igual

ao número de variáveis determinadas pelos procedimentos (vi) Forward Selection e (vii) Backward Selection. O conjunto de dados simulados era composto por 2.000 marcadores SNPs e as populações simuladas totalizaram 1.000 indivíduos de 20 famílias de irmãos completos que tiveram os fenótipos e os genótipos avaliados. Além disso, os cenários simulados são baseados em dois níveis de herdabilidade e duas arquiteturas genéticas com ausência de dominância, constituindo assim, em quatro cenários, os quais foram simulados dez vezes cada. Com o intuito de demonstrar a aplicabilidade do estudo no melhoramento genético, foram avaliadas seis características de produtividade de um conjunto de dados reais de arroz asiático Oryza sativa (Número de panículas por planta, altura da planta, comprimento da panícula, número de panículas no perfilho primário, número de sementes por panícula e espiguetas por panícula) correspondente a 370 acessos de arroz, os quais foram genotipados para 44.100 marcadores SNPs. Em ambos os casos (dados simulados e reais) foi utilizada a validação independente e calculada as medidas de eficiência para comparar os critérios. De modo geral, as análises indicaram que o primeiro critério (número de componentes independentes igual ao número de componentes principais cujos os valores genômicos estimados via PCR apresentava maior valor de acurácia) se mostrou mais eficiente para os dois conjuntos de dados e apresentou as medidas de eficiência mais próximas do método exaustivo, com a vantagem de exigir menos tempo e esforço computacional. Para complementar o estudo, o terceiro capítulo consiste na aplicação dos três critérios mais eficientes do capítulo 2, os quais consistem no número de componentes independentes igual ao número de componentes que conduz os valores genômicos estimados via PCR a um maior valor de acurácia; a um menor valor de viés e a PCR a 80% de explicação da variação total de X considerando o modelo aditivo-dominante. Ainda no contexto deste modelo, foi aplicado os três métodos de redução de dimensionalidade (PCR, PLS e ICR) levando em consideração a escolha do número ótimo de componentes que conduz os valores genômicos aditivos, valores genômicos devido à dominância ou os valores genômicos totais (aditivo + dominância) a uma maior acurácia. Todos os métodos de redução de dimensionalidade foram comparados com o G-BLUP em termos de eficiência na estimação dos valores genômicos. As populações simuladas foram constituídas por 1.000 indivíduos de 20 famílias de irmãos completos, sendo genotipados para 2000 marcadores SNPs e as análises correspondentes a quatro cenários (dois níveis de herdabilidade × duas

arquiteturas genéticas) sendo assumido dominância completa. Os resultados do capítulo 3 assinalaram que se manteve a superioridade do critério 1 nos modelos aditivo-dominante. Além disso, para a estimação dos efeitos aditivos e devido a dominância concomitantemente por meio dos métodos de redução de dimensionalidade, é recomendável utilizar o número de componentes que conduz o valor genômico devido à dominância a uma maior acurácia. Ademais, ao confrontar as metodologias de redução dimensional (ICR, PCR e PLS) com o G-BLUP, verifica-se que a PCR é superior em termos de acurácia e o método vantajosamente apresenta um dos menores tempos computacionais na execução das análises. Ademais, nenhum dos métodos considerados capturaram adequadamente as herdabilidades simuladas e apresentaram viés.

ABSTRACT

COSTA, Jaquicele Aparecida da, M.Sc., Universidade Federal de Viçosa, February, 2018. **Genomic Prediction by Reduction of Dimensionality in Additive-Dominant Models.** Adviser: Camila Ferreira Azevedo. Co-advisers: Marcos Deon Vilela de Resende and Moysés Nascimento.

Great advances in animal and plant breeding have been provided using molecular genetic information. In this perspective, they proposed Genome Wide Selection (GWS), whose approach involves complete coverage of the genome using thousands of single nucleotide polymorphisms (SNPs). The objective is to estimate the genetic merit of the individuals and to that end, the researches carried out in GWS are based on the search and application of methodologies that aim to solve the problems faced in the estimation process, such as high dimensionality and high colinearity between the markers. Among them, we highlight the dimensionality reduction methods: Principal Component Regression (PCR), Partial Least Squares (PLS) and Independent Regression Component (ICR) and the traditional method of regularization / shrinkage, G-BLUP (Genomic Best Linear Unbiased Predictor). Thus, the first chapter considers the central ideas and importance of GWS for genetic improvement, definition of additive effects and effects due to dominance, the statistical problems faced in estimating the effects of markers on phenotypes by the usual method based on ordinary least squares, as well as the alternative statistical methodologies to solve such problems and validation procedures that aim to compare GWS methodologies. The second chapter refers to the proposition and application of seven criteria for choose the optimal number of independent components to be used in the ICR, considering only the additive effects. The criteria that consist of the number of independent components equal to the number of components that leads: (i) the estimated genomic values by PCR to a higher accuracy; (ii) estimated genomic values by PCR at a lower bias value; (iii) the PCR at 80% of the explanation of the total variation of X; (iv) PCR at 80% of the total variation of Y; (v) the ICR at 80% of explanation of the total variation of X; in addition to the criteria that consist of the number of independent components equal to the number of variables determined by the procedures (vi) Forward Selection and (vii) Backward Selection. The simulated data set consisted of 2.000 SNPs and the simulated populations totaled 1.000 individuals from 20 families of complete siblings that had the phenotypes and genotypes evaluated. In addition, the

simulated scenarios are based on two levels of heritability and two genetic architectures, constituting in four scenarios, which were simulated ten times each assuming absence of dominance. In order to demonstrate the applicability of the study to genetic improvement, were evaluated six characteristics of productivity of a real data set Asian rice Oryza sativa (Number of panicles per plant, plant height, panicle length, number of panicles in the tiller primary, number of seeds per panicle and spikelets per panicle) corresponding to 370 accessions of rice, which were genotyped for 44.100 markers SNPs. In both cases (simulated and real data) the independent validation was used and the efficiency measures were calculated to compare the criteria. In general, the analyzes indicated that the first criterion (number of independent components equal to the number of principal components whose genomic values estimated by PCR showed highest accuracy) proved to be more efficient for both sets of data and presented the measures of efficiencies closer to the exhaustive method, with the advantage of requiring less computational time and effort. To complement the study, the third chapter consists of the application of the three most efficient criteria of chapter 2, which consist of the number of independent components equal to the number of components that leads the estimated genomic values via PCR to a highest accuracy value; to a lower value of bias and the PCR to 80% of explanation of the total variation of X considering the additive-dominant model. In the context of this model, the three dimensionality reduction methods (PCR, PLS and ICR) were applied taking into account the choice of the optimal number of components that leads to the additive genomic values, genomic values due to dominance or total genomic values (additive + dominance) to greater accuracy. All dimensionality reduction methods were compared with G-BLUP in terms of efficiency in the estimation of genomic values. Simulated populations were composed of 1.000 individuals from 20 families of complete siblings, with genotyped 2000 SNPs markers and analyzes corresponding to four scenarios (two levels of heritability × two genetic architectures). The simulations assumed complete dominance. The results of chapter 3 pointed out that the superiority of criterion 1 was maintained in the additive-dominant models. In addition, for the estimation of the additive effects and due to the dominance concomitantly by means of dimensionality reduction methods, it is recommended to use the number of components that drives the genomic value due to the dominance to a greater accuracy. In addition, when comparing the methodologies of dimensional

reduction (ICR, PCR and PLS) with G-BLUP, it is verified that the PCR is superior in terms of accuracy and the method advantageously presents one of the smallest computational times in the execution of the analyzes. In addition, none of the methods considered adequately captured the simulated heritabilities and showed bias.

INTRODUÇÃO GERAL

A utilização de informações provenientes do DNA tem revolucionado as pesquisas na área de melhoramento animal e vegetal. Sob esse enfoque, cada vez mais os marcadores moleculares têm se destacado e há uma crescente busca por marcadores cujo processo de genotipagem seja fácil e de baixo custo operacional (Vignal et al., 2002). Nesta perspectiva, a Seleção Genômica Ampla (Genome Wide Selection – GWS) foi proposta por Meuwissen et al. (2001), sendo caracterizada por utilizar milhares de marcadores SNPs (Single Nucleotide Polymorphisms) que cobrem amplamente o genoma. Os marcadores SNPs têm sido intensamente utilizados nos últimos tempos como forma de capturar as informações genotípicas que possam influenciar na variabilidade fenotípica dos indivíduos (Goddard e Hayes, 2007).

Segundo Meuwissen et al. (2001), a utilização de marcadores na predição genômica, possibilita um aumento do percentual do ganho genético em relação ao ciclo de seleção, um aumento na acurácia da predição dos valores genômicos e a redução no intervalo entre gerações. Os marcadores SNPs apresentam a vantagem de possuir baixo percentual de variação e apesar de serem marcadores condominantes bialélicos, diferencia-se dos outros marcadores, como por exemplo microssatélites multialélicos, por apresentar uma ampla gama de técnicas de genotipagem disponíveis, enquanto os outros têm um processo de genotipagem padronizado (Vignal et al., 2002).

No entanto, a implementação da GWS a fim de estimar os valores genômicos dos indivíduos (Genomic Estimated Breeding Values - GEBVs) enfrenta alguns desafios estatísticos, como a multicolinearidade e a alta dimensionalidade. Gianola et al. (2003) e Vignal et al. (2002) reportaram que a alta densidade do número de marcadores diminui a probabilidade de uma mudança de base ocorrer independentemente de outra na mesma posição do genoma e além disso, o alto custo

das técnicas de genotipagem por indivíduo faz com que o número de observações individuais seja muito inferior em relação ao número de marcadores, isto é, o número de observações é menor que o número de variáveis explicativas ou parâmetros.

A presença de alta dimensionalidade inviabiliza a utilização de metodologias fundamentadas em quadrados mínimos ordinários (Ordinary Least Squares – OLS), dado que por meio deste método não é possível a estimação conjunta de todos os efeitos de marcadores no fenótipo. Uma alternativa para se utilizar os quadrados mínimos ordinários seria estimar cada efeito de marcador isoladamente e realizar testes para verificar se este efeito é significativo ou não. No entanto, esta prática é ineficiente, pois influencia na superestimação dos efeitos de marcadores, apresenta baixo valor de acurácia e possibilita identificar apenas os QTLs de maiores efeitos (Resende et al., 2012). Devido à alta dimensionalidade e ao desiquilíbrio de ligação (linkage disequilibrium - LD), os marcadores são altamente correlacionados.

Nesse sentido, a GWS vem ganhando destaque ao propor metodologias que visam resolver tais problemas (multicolinearidade e alta dimensionalidade). Resende et al. (2012) retrata que as metodologias estatísticas aplicadas a GWS podem ser divididas em três grupos: métodos baseados em Regressão explícita se destacando os métodos RR-BLUP, reparametrização do BLUP genômico (Genomic Best Linear Unbiased Predictor - G-BLUP) e os métodos bayesianos (BayesA, BayesB, BLASSO, entre outros); métodos baseados em Regressão implícita se destacando a Regressão Kernel, reproducing kernel Hilbert space (RKHS) e Redes neurais; e os métodos de redução de dimensionalidade, tendo destaque a Regressão via Componentes Principais (PCR), Quadrados Mínimos Parciais (PLS) e Regressão via Componentes Independentes (ICR). Dentre estes, os métodos de redução de dimensionalidade se

destacam por apresentar grande aplicabilidade e teoria relativamente simples quando comparados aos demais métodos aplicados a GWS.

A aplicação dos métodos de redução de dimensionalidade envolve a escolha ideal do número de componentes, os quais são combinações lineares das variáveis originais. Sob o contexto de regressão, o número de componentes deve ser menor ou igual ao mínimo entre o número de observações e o número de variáveis explicativas menos 1 (devido a constante da regressão). Apesar da redução do número de variáveis ao longo do processo de estimação, a eficiência na estimação dos valores genômicos deve ser mantida.

No desenvolvimento da análise de componentes principais e dos quadrados mínimos parciais, a variação total contida nos dados é igual a variação total dos componentes, sendo que os primeiros componentes explicam grande parte da variabilidade dos dados (Garthwaite, 1994; Roggo et al.; 2007). Assim, um critério de decisão para o PLS e para a PCR é adotar uma percentagem desejada da variação total explicada pelos componentes e determinar o número de componentes que atingisse esse valor. No entanto, para a ICR, Cadavid et al. (2008) sugere utilizar o mesmo número de componentes da PCR, no entanto, ainda não foram definidos os critérios de escolha formalmente.

Em virtude do caráter herdável dos efeitos aditivos, apenas estes eram considerados nas predições genômicas, todavia, vários autores (Hill et al., 2008; Goddard et al., 2009; Bennewitz e Meuwissen, 2010; Toro e Varona, 2010; Su et al., 2012; Wellmann e Bennewitz, 2012; Denis e Bouvet, 2013; Zeng et al., 2013; Wang e Da, 2014; Muñoz et al., 2014; Azevedo et al., 2015; Costa et al., 2015; Almeida Filho et. al, 2016) têm reportado a importância da inclusão dos efeitos devido à dominância, mostrando que pode haver melhorias e aumento na acurácia na predição dos valores

genômicos. No entanto, as metodologias de redução dimensional ainda não foram aplicadas no contexto dos modelos aditivo-dominante.

Diante do exposto, o Capítulo 1 apresenta as principais ideias sob a qual a seleção genômica ampla está fundamentada, bem como os principais desafios estatísticos enfrentados na estimação dos efeitos dos marcadores no fenótipo, tais como a alta dimensionalidade e a multicolinearidade. Além disso, descreve o tradicional método de estimação de parâmetros denominado quadrados mínimos ordinários e retrata detalhadamente as metodologias alternativas aplicadas a GWS como os métodos de redução de dimensionalidade. Ademais, são apresentados os métodos de seleção de variáveis Forward e Backward e os processos de validação, que são utilizados a fim de comparar as metodologias estatísticas.

O Capítulo 2 parte do pressuposto que um passo importante na aplicação dos métodos de redução dimensional é a escolha do número ótimo de componentes a serem utilizados no modelo e que não há uma formalização quanto aos critérios de escolha do número ótimo de componentes independentes. Assim, este capítulo consiste na proposição de sete critérios para escolha do número ótimo de componentes independentes a serem utilizados na Regressão via Componentes Independentes e que expliquem satisfatoriamente a variabilidade dos dados em um modelo aditivo. Os critérios consistem em determinar que o número de componentes independentes seja igual ao número de componentes que conduz: (i) os valores genômicos estimados via PCR a um maior valor de acurácia; (ii) os valores genômicos estimados via PCR a um menor valor de viés; (iii) a PCR a 80% de explicação da variação total de X; (iv) a PCR a 80% de explicação da variação total de Y; (v) a ICR a 80% de explicação da variação total de X; além dos critérios que consistem no número de componentes independentes igual ao número de variáveis determinadas pelos procedimentos (vi)

Forward Selection e (vii) Backward Selection. A análise de componentes independentes assume que os dados provêm de distribuições não gaussianas, visto que segundo Papoulis (1991), os casos em que se é verificada a normalidade dos dados a correlação igual a zero entre as variáveis já implica em independência das mesmas. Desta forma, bastaria uma análise de componentes principais para determinar o número ótimo de componentes independentes que eficientemente contribuiriam para estimar os valores genômicos dos indivíduos.

O Capítulo 3 avalia os três critérios mais eficientes obtidos no Capítulo 2 para determinar o número ótimo de componentes independentes a serem utilizados na ICR e propõe a aplicação dos métodos de redução de dimensionalidade considerando o modelo aditivo-dominante, além disso é feita a comparação com o tradicional método utilizado na GWS, o G-BLUP. No capítulo 2, verificou-se que os critérios que apresentaram maiores valores de acurácia foram aqueles em que se determina que o número de componentes independentes seja igual ao número de componentes que conduz: os valores genômicos estimados via PCR a um maior valor de acurácia; os valores genômicos estimados via PCR a um menor valor de viés e a PCR a 80% de explicação da variação total de X. Dentre estes, é analisado qual o melhor critério que determina o número ótimo de componentes independentes a serem utilizados na ICR, tendo em vista o modelo aditivo-dominante. Além disso, foram computadas as informações considerando o valor máximo de acurácia atingida considerando o valor genômico aditivo, o valor genômico devido à dominância e o valor genômico total considerando-se cada método de redução de dimensionalidade (ICR, PCR, PLS) com o objetivo de determinar qual a melhor forma de se estimar os valores genômicos dos indivíduos. Em seguida, é feita a comparação entre os métodos de redução de dimensionalidade e o G-BLUP para verificar a eficiência dos mesmos.

CAPÍTULO 1

REVISÃO DE LITERATURA

1. Seleção Genômica Ampla

1.1.Definição e Importância

A principal contribuição da genética molecular para o melhoramento genético é a utilização de informações provenientes do DNA, os marcadores moleculares. O uso dessas informações permite o aumento da eficiência na predição de valores genéticos e possibilita que os indivíduos geneticamente superiores sejam identificados rapidamente, visto que não é necessário esperar que os indivíduos expressem a característica desejada.

Meuwissen et al. (2001) propuseram a seleção genômica ampla (Genome Wide Selection – GWS), que é caracterizada pelo uso de um grande número de marcadores SNPs (Single Nucleotide Polymorphisms) presentes no genoma e os fenótipos de interesse, capturando os genes que afetam esse caráter quantitativo.

Os principais desafios estatísticos encontrados ao se aplicar a GWS são a alta dimensionalidade, o número de marcadores é muito maior que o número de indivíduos genotipados e fenotipados, e a multicolinearidade, os marcadores são altamente correlacionados. A alta dimensionalidade impossibilita a obtenção das estimativas do efeito dos marcadores no fenótipo por meio de métodos fundamentados em quadrados mínimos ordinários. Ademais, caso não haja a alta dimensionalidade, a multicolinearidade impossibilitaria a obtenção de estimativas estáveis dos efeitos dos marcadores sobre os fenótipos por meio dos quadrados mínimos ordinários.

As estimativas dos efeitos genéticos podem ser divididas em duas principais fontes: os efeitos aditivos e os efeitos devido à dominância, negligenciando os efeitos causados pela epistasia, devido a superparametrização do modelo aplicado a GWS. O

valor aditivo é o componente de maior interesse na seleção de indivíduos geneticamente superiores, visto que consiste na parte herdável da variação genética, sendo, portanto, os mais explorados. No entanto, no âmbito da predição genômica, o efeito devido à dominância vêm sendo alvo de vários estudos, em que reportam a sua importância (Hill et al., 2008; Bennewitz e Meuwissen, 2010; Wellmann e Bennewitz, 2012).

A estimativa de dominância é essencial em espécies de propagação vegetativa (Denis e Bouvet, 2013) e em populações cruzadas. Assim, incluir efeitos aditivos e devido à dominância no direcionamento de cruzamentos é uma forma de aumentar o ganho genético capitalizando a heterose (Wellmann e Bennewitz, 2012; Toro e Varona, 2010). A relevância de modelos aditivo-dominante aplicados a Seleção Genômica vêm sido discutidos por Hill et al. (2008), Bennewitz e Meuwissen (2010) e Wellmann e Bennewitz (2012). Embora, os estudos na GWS incluindo efeitos de dominância serem escassos, os modelos aditivo-dominante já foram empregados no BLUP genômico (Wang e Da, 2014; Muñoz et al., 2014; Su et al., 2012), na regressão ridge (RR-BLUP - Goddard et al., 2009), nos métodos bayesianos (Toro e Varona, 2010; Zeng et al., 2013; Azevedo et al., 2015; Almeida Filho et al., 2016). No entanto, os modelos aditivo-dominante até o momento não foram empregados nos métodos de redução de dimensionalidade.

1.2.Efeito Genético aditivo e Efeito Genético devido à dominância

O valor fenotípico (phenotype – p) de um indivíduo, valor que é observado e/ou mensurado, é resultado do valor genético (genotype - g) e do valor ambiental (environment - e), sendo que tais componentes se relacionam da seguinte forma:

$$p = g + e. \quad (1)$$

Temos ainda que g pode ser decomposto em outros dois principais componentes:

$$g = a + d \quad (2)$$

em que que a representa o valor genético aditivo e d representa o valor genético devido à dominância.

Os efeitos aditivos são resultantes da ação direta de cada alelo, nos cromossomos homólogos. Assim, os efeitos aditivos podem ser vistos como o desvio do valor genético dos indivíduos que possuem o alelo A em relação à média da população. A soma dos efeitos aditivos dos genes de um mesmo loco determina o valor genético aditivo. Somente estes efeitos são herdáveis, ou seja, passam dos pais para os filhos e consequentemente os efeitos mais explorados pela literatura. O efeito devido à dominância ocorre quando os efeitos dos alelos de um determinado loco não são somente aditivos, mas interagem entre si de modo que o valor genético do heterozigoto se desvia da média dos valores genéticos dos homozigotos.

Geralmente, o efeito aditivo é o componente de maior interesse na seleção de indivíduos geneticamente superiores, pois é o efeito que se enquadra no contexto de ser um efeito herdável. No entanto, a literatura (Wang e Da, 2014; Muñoz et al., 2014; Su et al., 2012; Toro e Varona, 2010; Zeng et al., 2013; Azevedo et al., 2015; Almeida Filho et al., 2016) relata que o estudo em relação a influência do efeito devido à dominância sobre o valor fenótipo pode acarretar em um desenvolvimento relevante na genômica, ao propiciar uma análise mais acurada na seleção dos indivíduos de interesse. Os modelos aditivo-dominante são capazes de capturar ambos os efeitos, permitindo a efetiva seleção dos pais para a próxima geração, cruzamentos e clones (Azevedo et al., 2015).

1.3.Regressão Linear Múltipla, Multicolinearidade e Alta dimensionalidade sob o enfoque de quadrados mínimos

Considere o seguinte modelo linear:

$$y = 1\mu + Wm_a + Sm_d + e \quad (3)$$

em que:

y é o vetor de observações fenotípicas com dimensão I × 1, sendo I o número de indivíduos genotipados e fenotipados;

μ é a média geral da característica;

W é a matriz de incidência com valores 0, 1 e 2 para o número de alelos do marcador com dimensão I × J, sendo J o número de marcadores e m_a são efeitos aditivos dos marcadores;

S é a matriz de incidência com valores 0,1 e 0 com dimensão I × J e m_d são efeitos devido à dominância dos marcadores;

e é o vetor de erros aleatórios com estrutura de variância dada por $e \sim N(0, I\sigma_e^2)$ sendo I a matriz identidade e σ_e^2 a variância residual.

Considere a justaposição das matrizes W e S definindo a matriz X como sendo $X = [W|S]$ e os efeitos de marcadores como sendo $m = [m_a|m_d]'$.

Com o propósito de encontrar as estimativas do vetor m_a e m_d na perspectiva do método dos quadrados mínimos ordinários, deve-se minimizar a soma de quadrados dos resíduos. Para isto, considere a abordagem matricial do modelo dada por:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_I \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1J} & \dots & x_{1\,2J} \\ 1 & x_{21} & \dots & x_{2J} & \dots & x_{2\,2J} \\ 1 & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{I1} & \dots & x_{IJ} & \dots & x_{I\,2J} \end{bmatrix} \begin{bmatrix} \mu \\ m_1 \\ \vdots \\ m_{2J} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_I \end{bmatrix} \quad (4)$$

em que: $y_{I\times1}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_I \end{bmatrix}$ $X_{I\times(2J+1)}=\begin{bmatrix} 1 & x_{11} & \dots & x_{1J} & \dots & x_{1\,2J} \\ 1 & x_{21} & \dots & x_{2J} & \dots & x_{2\,2J} \\ 1 & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{I1} & \dots & x_{IJ} & \dots & x_{I\,2J} \end{bmatrix}$; $m_{(2J+1)\times1}=\begin{bmatrix} \mu \\ m_1 \\ \vdots \\ m_{2J} \end{bmatrix}$

e $e_{I\times1}=\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_I \end{bmatrix}$.

Assim, por definição a soma de quadrados dos desvios é dada conforme a expressão (5) e ao utilizar a igualdade $S\,(m) = (y - X\,m)'\,(y - Xm\,)$, essa expressão pode ser escrita de forma equivalente a expressão (6), isto é:

$$S\,(m) = e'e \tag{5}$$

$$S\,(m) = y'\,y \;-\; y'X\,m \;- m'\,X'y \;+\; m'\,X'\,X\;\,m. \tag{6}$$

Observe que na equação (6) os termos $y'_{I\times1}\,X_{I\times(J+1)}m_{(J+1)\times1}$ e $m'_{1\times(J+1)}X'_{(J+1)\times I}y_{I\times1}$ resultam em escalares e além disso:

$$m'\,Xy \;=\; (\,y'X'm)\,'. \tag{7}$$

Deste modo, conclui-se que $m'\,Xy \;=\; y'X'm$ e a equação (6) é reescrita a seguir:

$$S\,(m) = y'\,y \;-\; 2m'\,Xy \;+\; m'\,X'\,X\,m. \tag{8}$$

Para encontrar as estimativas de m, basta minimizar a soma de quadrados, sendo necessários os seguintes passos:

$$\frac{\partial S\,(m)}{\partial\,m} = -2Xy + \;2\,X'Xm. \tag{9}$$

Igualando a expressão (9) a zero, tem-se:

$$-2Xy + \;2\,X'X\,\hat{m} = \;0\;\rightarrow\;\; X'X\hat{m}\;\; = Xy. \tag{10}$$

Portanto, para estimar o vetor m pelo método dos quadrados mínimos ordinários, é imprescindível obter a inversa da matriz X′X ,para que ao multiplicar o fator $(X'X)^{-1}$ em ambos os lados da equação (10), obter:

$$\widehat{m} = (X'X)^{-1}X'y \quad (11)$$

A multicolinearidade é identificada quando há uma grande dependência entre as variáveis explicativas X. Em casos que a matriz X apresente dependência linear perfeita entre as variáveis, o determinante da matriz X′X se torna nulo e não é possível encontrar a inversa dessa matriz, consequentemente, impossibilitando a obtenção das estimativas $\widehat{m}$. No entanto, se a dependência das variáveis é aproximadamente linear, estamos a frente de um problema da multicolinearidade, e consequentemente, as estimativas de quadrados mínimos ordinários são instáveis apresentando grandes estimativas de variâncias.

Sob o enfoque da GWS, a matriz X além de apresentar multicolinearidade, apresenta também alta dimensionalidade, ou seja, o número de marcadores (J) é muito superior em relação ao número de observações fenotípicas e genotípicas (I). Além disso, quando se considera o modelo aditivo-dominante o número de variáveis explicativas é duplicado, isto é, $2J$. Sendo assim, seja $2J + 1$ o número de parâmetros (número de variáveis explicativas + a média geral) a serem estimados e se dispomos apenas de I observações ($I << 2J + 1$), tem-se um problema matemático de resolução de um sistema, o qual teremos I equações com $2J + 1$ incógnitas, não sendo possível a obtenção de estimativas para os parâmetros.

Deste modo, a presença de alta dimensionalidade impede a obtenção de estimativas dos efeitos dos marcadores sobre o fenótipo pelo tradicional método dos quadrados mínimos ordinários, sendo necessário a utilização de outras metodologias que contemplem estes desafios da análise estatística aplicada à seleção genômica.

2. Métodos de redução de dimensionalidade

Os métodos de redução de dimensionalidade são métodos que possibilitam estimar os efeitos de marcadores mesmo na presença de alta dimensionalidade e garantem que no processo de estimação os efeitos de marcadores estão livres do impacto da multicolinearidade entre os marcadores. Além disso, apresentam grande aplicabilidade e teoria relativamente simples quando comparados aos demais métodos aplicados a GWS. Os métodos de redução de dimensionalidade são baseados em variáveis latentes (variáveis não observáveis), os denominados componentes. Os componentes são combinações lineares das variáveis explicativas do modelo, e sob o contexto da GWS, são combinações lineares da incidência dos marcadores. O objetivo de tais métodos é reduzir a dimensão e garantir a ausência de multicolinearidade entre as variáveis latentes.

Os métodos de redução de dimensionalidade diferem entre si na forma como constroem os componentes. A Regressão via Componentes Principais (PCR) constroem os componentes de modo a maximizar a variância dos mesmos. O Quadrados Mínimos Parciais (PLS) leva em consideração as variáveis X e a variável Y na construção dos componentes, de modo a maximizar a covariância entre os componentes e a variável Y. Enquanto, que a Regressão via Componentes Independentes (ICR) constroem os componentes de modo a maximizar a independência entre eles.

2.1.Regressão via Componentes Principais

A regressão via componentes principais (Principal Components Regression – PCR) foi introduzida por Kendall (1957) e Hotelling (1957). Para isto, define-se o j-ésimo componente principal z_j da seguinte forma:

$$z_j = p_{j1}x_1 + p_{j2}x_2 + \cdots + p_{j\,2J}x_{2J} = p_j'X \tag{12}$$

em que $x_j's$ são as colunas da matriz X e p_j' é um vetor desconhecido que estabelece a j-ésima combinação linear, para $j = 1,2, \ldots ,2J$.

A PCR visa solucionar os problemas de alta dimensionalidade e multicolinearidade encontrados pela regressão linear múltipla via quadrados mínimos ordinários e conforme Otto (1999), a redução de dimensionalidade não deve conduzir a perda de informações relevantes das variáveis explicativas.

O objetivo principal da PCA é encontrar os componentes $Z_j's$ por meio das variáveis X $(X_1, X_2, \ldots, X_{2J})$ que maximizam a variabilidade dos componentes principais. Para tal, a variância de Z_j e covariância de Z_j e Z_k $(j \neq k)$ são dadas respectivamente por:

$$Var(Z_j) = Var\left(p_j'X\right) = p_j'Var(X)p_j = p_j'\Sigma p_j \tag{13}$$

$$Cov(Z_j, Z_k) = Cov\left(p_j'X, p_k'X\right) = p_j'\,Var(X)p_k = p_j'\Sigma p_k \tag{14}$$

em que $Var(X) = \Sigma$, ou seja, a matriz de variância e covariâncias das variáveis X.

Diante do exposto, é notório que a expressão (13) corresponde a uma forma quadrática e para maximizar a variância de Z_j será utilizado o Teorema da Maximização de Formas Quadráticas que está apresentado a seguir.

Teorema da Maximização de Formas Quadráticas: Se Σ é uma matriz simétrica $(2J \times 2J)$, então o máximo de $p_j'\Sigma p_j$ sob a restrição $p_j'p_j = 1$ é dado pelo maior dos

autovalores λ_j de Σ, $j = 1, 2, \ldots, 2J$, e pelo autovetor correspondente p_j que são soluções do sistema de equações homogêneas $(\Sigma - \lambda_j I)p_j = 0$.

Assim, definimos as variáveis latentes Z_j $(j = 1,2, \ldots, 2J)$ como sendo combinações lineares das variáveis explicativas X ($X_1, \ldots, X_{2J}$),

$$Z = XP, \tag{15}$$

em que $P = [p_1 \quad p_2 \quad \cdots \quad p_j]$ $(2J \times 2J)$ é a matriz de autovetores da matriz de covariância de X $(I \times 2J)$ e Z $(I \times 2J)$ a matriz cujas colunas são os componentes principais $Z_j's$.

Além disso, utilizando o sistema de equações homogêneas, tem-se:

$$(\Sigma - \lambda_j I)p_j = 0 \Rightarrow \Sigma p_j = \lambda_j p_j. \tag{16}$$

Substituindo a expressão (16) nas expressões de variância e covariância e utilizando as informações de que os autovetores de uma matriz simétrica são ortogonais ($p_j' p_k = 0$) e a restrição ($p_j' p_j = 1$), conclui-se que:

$$Var(Z_j) = p_j' \Sigma p_j = p_j' p_j \lambda_j = \lambda_j \tag{17}$$

$$Cov(Z_j, Z_k) = p_j' \Sigma p_k = p_j' p_k \lambda_j = 0. \tag{18}$$

A partir dos resultados obtidos nas expressões (17) e (18), verifica-se que a correlação entre os componentes Z_j e Z_k $(j \neq k)$ é dada por:

$$Cor(Z_j, Z_k) = \frac{Cov(Z_j, Z_k)}{\sqrt{Var(Z_j) Var(Z_k)}} = 0. \tag{19}$$

Assim, os componentes principais têm como característica serem componentes ortogonais $z_j's$ $(j = 1,2, \ldots, 2J)$, ou seja, não correlacionados.

A metodologia do método PCR consiste, basicamente, em eliminar componentes que não contribuam consideravelmente na explicação da variância total

presente nos dados, o que reduz a dimensionalidade das variáveis explicativas originais e por este motivo, $n_{PCR} \leq min(I, 2J) - 1$. Após a escolha do número ótimo de componentes a serem incluídos no modelo e com o objetivo de estabelecer a relação entre a variável dependente (Y) e os componentes principais $Z_m's$ ($m = 1,2, \dots, n_{PCR}$ sendo $n_{PCR} \leq min(I, 2J) - 1$), utiliza-se a regressão linear múltipla, obtendo a equação de predição a seguir:

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 z_1 + \hat{\alpha}_2 z_2 + \cdots + \hat{\alpha}_{n_{PCR}} z_{n_{PCR}} \quad (20)$$

em que $z_m's$ são os vetores colunas dos primeiros n_{PCR} componentes principais e $\hat{\alpha}_m's$ são estimados por meio do método dos quadrados mínimos ordinários.

Ademais, os coeficientes obtidos na equação de predição (20) não estão relacionados com as variáveis originais. Assim, ao estimar esses coeficientes utiliza-se as equações (15) e (20) para estimar os coeficientes associados as variáveis originais, denotados por $\hat{m}_{PCR} = [\hat{m}_a \quad \hat{m}_d]'$, basta:

$$\hat{m}_{PCR} = P_{n_{PCR}} \hat{\alpha}, \quad (21)$$

em que $P_{n_{PCR}}$ é a matriz dos primeiros n_{PCR} autovetores da matriz de covariância de X, $\hat{\alpha}$ é o vetor das estimativas dos coeficientes provenientes da regressão entre a variável Y e os primeiros n_{PCR} componentes principais.

2.1.1. Número de Componentes

Como dito anteriormente, a utilização do método PCR requer a escolha do número de componentes que serão utilizados no modelo, uma vez que é de extrema importância para a redução de dimensionalidade, sem que haja perda de informações relevantes relacionadas aos dados originais.

Assim, para determinar o número ótimo de variáveis latentes na PCR existem alguns critérios na literatura que auxiliam nesta decisão. Dentre estes, o critério comumente utilizado é baseado na variabilidade presente dos dados. Para isto,

fazendo-se a decomposição espectral (Spectral decomposition - SD) da matriz de variâncias Σ, tem-se:

$$\Sigma = P\Lambda P' \tag{22}$$

em que P é composta pelos autovetores de Σ em suas colunas e Λ é uma matriz diagonal de autovalores de Σ. Desta forma, segue que o traço da matriz Σ é dado conforme a expressão (23) ou de forma equivalente utilizando a expressão (24):

$$tr(\Sigma) = \sum_{j=1}^{2J} diag(\Sigma) = \sum_{j=1}^{2J} \sigma_j^2, \tag{23}$$

$$tr(\Sigma) = tr(P\Lambda P') = tr(\Lambda PP') = tr(\Lambda) = \sum_{j=1}^{2J} \lambda_j. \tag{24}$$

Desse modo, conclui-se que:

$$\sum_{j=1}^{2J} \lambda_j = \sum_{j=1}^{2J} \sigma_j^2. \tag{25}$$

Portanto, sob o contexto de regressão, a variabilidade total presente nas variáveis originais somente é alcançada ao se utilizar o número máximo de componentes construídos ($n_{PCR} = min(I, 2J) - 1$). No entanto, frequentemente, a maior parte dessa variabilidade pode ser explicada por um número reduzido de componentes principais ($n_{PCR} < min(I, 2J) - 1$), uma vez que os primeiros componentes principais explicam partes maiores da variabilidade total dos dados. Segundo Ferreira (2012), a escolha do número ótimo de componentes principais pode ser fundamentada na determinação de uma fração desejada da variação total, que geralmente varia entre 70% e 80%.

A construção da matriz *P* exige a abordagem de alguns temas, por isso, a seguir, são expostos conceitos relacionados à matriz de variância e covariância, autovetores e autovalores, conforme retratado por Marcoulides e Hershberger (1997), e a decomposição espectral descrita por Rencher e Christensen (2002).

2.1.2. Matriz de Variância e Covariância

A matriz de (co)variância relacionada às variáveis aleatórias $X_1, X_2, \dots, X_{2J}$ é uma matriz quadrada de ordem $(2J \times 2J)$. Em sua diagonal (posição jj) contém as respectivas variâncias $\sigma_1^2, \sigma_2^2, \dots, \sigma_{2J}^2$ e nas demais posições (jj' para $j \neq j'$) as covariâncias entre as variáveis aleatórias X_j e $X_{j'}$ $(j, j' = 1,2, \dots, 2J$ sendo $j \neq j')$ denotadas por $\sigma_{12}, \sigma_{13}, \dots, \sigma_{(2J-1)2J}$. Assim, a matriz de (co)variância pode ser expressa por:

$$Var(X) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1\,2J} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2\,2J} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{2J\,2} & \sigma_{2J\,2} & \dots & \sigma_{2J}^2 \end{bmatrix}$$

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

Dessa forma, a matriz $Var(X)$ é simétrica e matricialmente também pode ter a seguinte notação:

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

2.1.3. Autovetores e Autovalores

Seja $T: V \to V$ uma transformação linear, isto é, T é uma função linear sobre um espaço vetorial V que preserva as operações de soma e multiplicação por um escalar. Um vetor $p \in V$ ($p \neq 0$) é definido como autovetor de T, se existe um escalar real λ tal que:

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

em que o escalar λ é denominado autovalor de T associado a p.

Considere A uma matriz quadrada da transformação linear T, ou seja, $T(p) = Ap$ e como visto anteriormente, tem-se que $T(p) = \lambda p = \lambda Ip$ em que I é a matriz identidade. Assim,

$$Ap = \lambda Ip.$$

Consequentemente,

$$\lambda Ip - Ap = 0 \ \rightarrow \ (\lambda I - A)p = 0.$$

As possíveis soluções não nulas desse sistema linear homogêneo podem ser obtidas se o determinante for nulo, ou seja, $det(\lambda I - A) = 0$. Em seguida, obtêm-se os valores do escalar λ e substitui na expressão $Ap = \lambda p$ para determinar os autovetores. Um fato relevante é que os autovetores p_i e p_j associados a autovalores distintos $\lambda_i \neq \lambda_j$ são ortogonais.

2.1.4. Decomposição Espectral

Seja A uma matriz simétrica $(2J \times 2J)$ com autovetores $[\, p_1 , p_2 , \dots , p_{2J} \,]$ linearmente independentes ($\sum_{j=1}^{2J} l_j \, p_j = 0$ então $l_j = 0$ para todo $j = 1, 2, \dots, 2J$). A decomposição espectral da matriz A é feita da seguinte forma:

$$A = P\Lambda P'.$$

em que P é composta pelos autovetores de A em suas colunas e Λ é uma matriz diagonal de autovalores de A.

2.2.Quadrados Mínimos Parciais

A regressão via quadrados mínimos parciais (Partial Least Squares- PLS) idealizada por Wold (1975), busca resolver problemas em que a matriz de dados X apresenta alta dimensionalidade e multicolinearidade entre as covariáveis, a exemplo disso, as matrizes de incidência dos marcadores moleculares. Dessa forma, o objetivo,

é estimar o efeito das variáveis X sobre a variável dependente Y, cuja equação de predição é dada matricialmente pela expressão abaixo:

$$\hat{y} = X\hat{m}_{PLS}.$$

Segundo Garthwaite (1994), o método PLS apresenta similaridades com o método PCR, sendo a maior diferença entre eles dada pelo fato da PCR levar em consideração apenas as variáveis explicativas X na construção dos componentes, enquanto que o PLS também leva em consideração a variável resposta Y. Estatisticamente, o método PCR extrai componentes que maximizam a variância dos mesmos e o PLS extrai componentes que têm maior covariância com a variável Y. Desta forma, com o intuito de estimar a variável dependente Y, define-se os componentes associados a X denotados por t_m $(m \times 1,$ em que $m = 1,2,\dots,n_{PLS}$ sendo $n_{PLS} \leq min(2J, I) - 1)$ e os componentes associados a Y denotados por z_m.

Para determinar os primeiros componentes t_1 e z_1, as variáveis Y e $X_j's$ são centradas na média, definindo as variáveis U_1 e V_{1j}, como a seguir:

$$u_1 = y - \bar{y} \quad \text{e} \quad v_{1(j)} = x_j - \bar{x_j}, \tag{26}$$

para $j = 1,\dots,2J$. Define-se a variável S_1, como sendo $s_1 = V_1{}'u_1$ $(2J \times 1)$ em que $V_1 = [v_{11} \quad v_{12} \quad \cdots \quad v_{12J}]$, e aplica-se a decomposição em valores singulares (Singular value decomposition - SVD) no vetor s_1, como a seguir:

$$s_1 = L_1 k_1 q_1' \tag{27}$$

em que L_1 é uma matriz unitária $(2J \times 2J)$ com o primeiro vetor coluna igual a $\frac{s_1}{||s_1||}$ (vetor s_1 normalizado), k_1 é um vetor $(2J \times 1)$ com o primeiro valor igual a $||s_1||$ (norma do vetor s_1) e q_1 é um escalar igual a 1. Os componentes T_1 e Z_1 são, então, definidos por:

$$t_1 = V_1 L_1 \tag{28}$$

$$z_1 = u_1 q_1 \frac{Var(t_1)}{Cov(t_1, u_1)}. \tag{29}$$

No entanto, nem todas as informações contidas nas variáveis X_j $(j = 1,2,\ldots,2J)$ e na variável Y estão contidas no componente T_1, definido acima. Logo, a informação ausente em T_1 pode ser estimada por meio dos resíduos da regressão entre as variáveis X_j e T_1 ou, equivalentemente, da regressão entre as variáveis latentes V_{1j} e T_1, visto que os resíduos de ambas são idênticos (Garthwaite, 1994). Do mesmo modo, a variabilidade de Y que não está sendo explicada por T_1 pode ser estimada por meio dos resíduos da regressão ente U_1 e T_1. Dessa forma, são definidas as variáveis U_2 e $V_{2(j)}$, respectivamente,

$$\hat{v}_{2(j)} = v_{1(j)} - t_1 \hat{r}_1'$$

$$\hat{u}_2 = u_1 - t_1 \hat{p}_1'$$

ou seja, U_2 e V_{2j} são os resíduos, r_1 e p_1 são os coeficientes obtidos da regressão entre U_1e $\hat{T}_1$ e V_{1j} e $\hat{T}_1$, nesta ordem. Define-se uma nova variável S_2, como sendo $s_2 = V_2{}' u_2$ e aplica-se novamente a decomposição em valores singulares, como a seguir:

$$S_2 = L_2 k_2 q_2' \tag{30}$$

em que L_2 é uma matriz unitária $(2J \times 2J)$ com o primeiro vetor coluna igual a $\frac{s_2}{||s_2||}$ (vetor s_2 normalizado), k_2 é um vetor $(2J \times 1)$ com o primeiro valor igual a $||s_2||$ (norma do vetor s_2) e q_2 é um escalar igual a 1. Os componentes T_2 e Z_2 são, então, definidos por:

$$t_2 = \hat{V}_2 L_2 \tag{31}$$

$$z_2 = \hat{u}_2 q_2 \frac{Var(t_2)}{Cov(t_2, \hat{u}_2)}. \tag{32}$$

Os componentes $\hat{T}_3, \ldots, \hat{T}_{n_{PLS}}$ são determinados sucessivamente e de modo análogo aos anteriores.

Para determinar o número ótimo de componentes (n_{PLS}) pode-se utilizar os mesmos procedimentos utilizados na regressão via componentes principais (PCR), como por exemplo, o critério baseado na variabilidade presente nos dados. Vale ressaltar que a metodologia do PLS também garante que os componentes são não correlacionados. De fato, a correlação entre $V_{2(j)}$ e T_1 é igual a 0, visto que são, respectivamente, resíduo e regressor. Deste modo, como cada componente $T_2, T_3, \ldots, T_{n_{PLS}}$ são combinações lineares de $V_{2(j)}$ então, estes também não são correlacionados com T_1.

Após os procedimentos de decomposição acima e da definição do número de componentes, determina-se os coeficientes da regressão ($\hat{\beta}_m's, m = 1,2, \ldots, n_{PLS}$) entre os componentes $T_1, T_2, T_3, \ldots, T_{n_{PLS}}$ e a variável Y por meio do método dos quadrados mínimos ordinários, obtendo a equação de predição a seguir:

$$\hat{y} = \hat{T}\hat{\beta}, \tag{33}$$

em que $\hat{\beta}$ é o vetor das estimativas dos coeficientes da regressão estimados via OLS.

Note que os coeficientes $\hat{\beta}_m's$ também não estão associados as variáveis originais, não tendo assim, uma interpretação biológica. As estimativas dos coeficientes originais não são obtidas trivialmente, uma vez que as colunas da matriz V $(V_{1(j)}, V_{2(j)}, \ldots, V_{2J(j)})$ não são comparadas diretamente com X como na PCR, pois são deflacionadas sucessivamente. No entanto, segundo Wold (1975) os coeficientes associados as variáveis originais podem ser estimados via:

$$\hat{m}_{PLS} = L(R'L)^{-1}\hat{\beta}, \tag{34}$$

em que L ($2J \times n_{PLS}$) é denominada matriz de carregamento de X cujas colunas são

as primeiras colunas das matrizes $L_1, L_2, \dots, L_{n_{PLS}}$ e R ($2J \times n_{PLS}$) é a matriz cujas colunas contém os coeficientes $r_1, r_2, \dots, r_{n_{PLS}}$.

A seguir, o método SVD utilizado no processo de construção dos componentes do PLS, é descrito por Härdle e Hlávka (2007).

2.2.1. Decomposição em Valores Singulares

Seja X uma matriz de dimensão $m \times n$, cujas colunas estão centradas em zero, então a decomposição em valores singulares (Singular value Decomposition - SVD) de X é dada por:

$$X = USV'$$

em que:

U é uma matriz com colunas ortonormais ($m \times r$), os quais são denominados vetores singulares à esquerda;

S é uma matriz diagonal, $S = diag(\sqrt{\lambda_i})$, sendo λ_i os autovalores não-nulos e positivo das matrizes XX' ou $X'X$, denominados valores singulares;

V é uma matriz com colunas ortonormais ($n \times r$), os quais são denominados vetores singulares à direita.

Os autovalores λ_i, são ordenados de forma decrescente, isto é: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq \lambda_{r+1} \geq \cdots \geq \lambda_n$ com $\lambda_i > 0$ para $1 \leq i \leq r$ e $\lambda_i = 0$ caso contrário. Além disso, os vetores $v_i's$ da matriz V correspondem aos r autovetores da matriz XX' ou $X'X$ e os vetores $u_i's$ da matriz U são definidos por: $u_i = \frac{1}{\sigma_i} X v_i$.

A matriz de variância e covariância de X quando as colunas são centradas em zero se resume a $\Sigma = \frac{1}{n} X'X$. Escrevendo a matriz $X'X$ por meio da SVD e aplicando propriedades de transposição de matriz e que as colunas da matriz U é composta por vetores ortogonais, tem-se:

$$X'X = (USV')'USV' \rightarrow X'X = VS'SV'$$

Assim, utilizando a decomposição espectral (SD) e a decomposição em valores singulares (SVD), tem-se que a matriz Σ é definida por:

$$\text{SD: } \Sigma = P\Lambda P' \quad \text{e} \quad \text{SVD: } \Sigma = \frac{1}{n} VS'SV'$$

Desse modo, conclui-se que:

$$P = V \text{ e } \Lambda = \frac{1}{n} S'S \ .$$

Para um caso mais simples, em que x é um vetor centrado em zero de dimensão $m \times 1$, então a decomposição em valores singulares (Singular value Decomposition-SVD) de x é dada por:

$$x = Usv$$

em que:

U é uma matriz unitária ($m \times m$) com o primeiro vetor coluna igual a $\frac{x_1}{||x_1||}$, ou seja, o vetor x_1 normalizado;

s é um vetor ($m \times 1$) com o primeiro valor igual a norma $||x_1||$.

v é um escalar igual a 1.

2.3.Regressão via Componentes Independentes

A regressão via Componentes Independentes (Independent Component Regression - ICR) proposta por Jutten e Hérault (1991) e Comon (1994) e apresentada sob o contexto da GWS por Azevedo et al. (2013), visa solucionar os problemas da alta dimensionalidade e da multicolinearidade da matriz de dados X (que contém valores centrados na média e que provem de uma distribuição não gaussiana). Sob o contexto de regressão, a ICR decompõe a matriz de dados X como a seguir:

$$X = SA' \tag{35}$$

em que a matriz X $(I \times 2J)$ é decomposta em uma matriz de componentes independentes S $(I \times min(2J, I) - 1)$ e em uma matriz A $(2J \times min(2J, I) - 1)$ denominada matriz de misturas. Caso, a matriz A fosse conhecida e quadrada facilmente obteríamos os componentes independentes. No entanto, como o intuito, geralmente, é reduzir a dimensionalidade, a matriz A não é quadrada, além de ser desconhecida. Sem perda de generalidade, podemos definir a matriz de misturas como o produto de uma matriz de branqueamento K $(2J \times min(2J, I) - 1)$ e a matriz R $(min(2J, I) - 1 \times min(2J, I) - 1)$ que garante a independência entre os componentes.

O primeiro processo para se estimar a matriz A é o processo de branqueamento (whitening) que torna as variáveis originais $(X_1, X_2, \dots, X_{2J})$ em não correlacionadas e com variância unitária, ou seja, a matriz de covariância dos dados branqueados é a matriz identidade. Para o branqueamento, a decomposição ortogonal é aplicada na matriz de covariância amostral de X, denotada por Σ $(2J \times 2J)$, obtendo:

$$\Sigma = P\Lambda^{-\frac{1}{2}}P'$$

em que P é composta pelos autovetores em suas colunas e Λ é uma matriz diagonal de autovalores da matriz de covariância amostral de X. Sob o contexto de regressão, a matriz K $(2J \times min(2J, I) - 1)$ é então definida como $P_r\Lambda_r^{-\frac{1}{2}}$, sendo P_r a matriz com as $min(2J, I) - 1$ primeiras colunas da matriz P com dimensão $2J \times min(2J, I) - 1$ ($min(2J, I) - 1$ primeiros autovetores) e Λ_r $(min(2J, I) - 1 \times min(2J, I) - 1)$ uma matriz com as $min(2J, I) - 1$ primeiras linhas e colunas da matriz Λ (autovalores associados a esses primeiros autovetores).

Assim, a matriz de dados branqueados será obtida por meio de XK $(I \times 2J)$. Pelo fato dos dados proverem de uma distribuição não gaussiana, a não correlação

entre as variáveis não implica em independência estatística. A independência somente é alcançada por meio do processo descrito a seguir.

O segundo processo para garantir a independência entre as colunas de S é feita com base na maximização da não gaussianidade. Esse processo se baseia no teorema central do limite, o qual afirma que a soma de N variáveis aleatórias independentes e identicamente distribuídas, para N suficientemente grande e que satisfaçam certas condições gerais, terá distribuição aproximadamente gaussiana. Consequentemente, a soma de duas variáveis aleatórias é mais gaussiana do que as próprias variáveis. Sob o contexto da ICA, como as variáveis X são uma combinação linear dos componentes, pode-se concluir que os componentes possuem distribuição mais distante da Gaussiana. Portanto, podemos obter os componentes independentes ao maximizar a não-gaussianidade da matriz de dados branqueados.

Os principais algoritmos de maximização da não gaussianidade são baseados em curtose e em negentropia. O algoritmo mais utilizado foi proposto por Hyvärinen (1998), denominado FastICA, e é baseado em negentropia. A negentropia é definida como:

$$J(R) = H\left(R_{gaussiana}\right) - H(R), \tag{36}$$

em que $H(R) = -\int_R f_R(r)\, ln f_R(r) dr$ é a entropia de uma variável aleatória T com função densidade de probabilidade $f_R(.)$ e $H\left(R_{gaussiana}\right)$ é a entropia de uma variável aleatória T com distribuição gaussiana.

A entropia no contexto estatístico é uma medida da incerteza média associada à observação de uma variável aleatória, sendo assim, quanto maior a entropia mais imprevisível é a observação da variável. Quando a variável possui distribuição gaussiana, a entropia e a variância são coincidentes. Considerando as variáveis R e $R_{gaussiana}$ com mesma variância, $H\left(R_{gaussiana}\right)$ é o valor máximo de entropia

encontrado, ou seja, uma variável aleatória gaussiana tem maior entropia que qualquer outra variável aleatória de mesma variância (Leite, 2013). Sendo assim, a negentropia pode quantificar o grau de não gaussianidade de uma variável aleatória e sua medida sempre será um valor não negativo.

Segundo Hyvärinen (1998), a maximização da negentropia conduz a estimação dos componentes independentes. No entanto, a dificuldade do algoritmo baseado em negentropia está no cálculo da entropia. Dessa forma, é necessário o uso de aproximações para a expressão (36), como por exemplo:

$$J(R) \propto \left\{E[G(R)] - E\left[G\left(R_{gaussiana}\right)\right]\right\}^2,$$

em que G é uma função não quadrática e a escolha da função G influencia na aproximação da negentropia (Hyvärinen, 1999), assim, as funções G mais empregadas com esse intuito são:

$$G_1(r) = \frac{1}{a}\log\cosh(ar) \quad \text{e} \quad G_2(r) = -exp\left(-\frac{r^2}{2}\right)$$

em que a é uma constante $(1 \leq a \leq 2)$ e deve-se escolher uma função G que não cresça muito rapidamente.

O FastICA é um algoritmo iterativo e após sua convergência é possível encontrar uma matriz R $(min(2J, I) - 1 \times min(2J, I) - 1)$ que torne as colunas da matriz XK independentes e consequentemente as colunas de S, uma vez que os componentes independentes podem ser obtidos via:

$$S = XKR. \tag{37}$$

Após a obtenção dos componentes independentes, utilizamos a regressão linear múltipla para obter a equação de predição a seguir:

$$\hat{y} = \hat{\gamma}_0 + \hat{\gamma}_1 s_1 + \ldots + \hat{\gamma}_{n_{icr}} s_{n_{icr}} \tag{38}$$

em que as estimativas dos coeficientes $\hat{\gamma}_m$ ($m = 0,1,\ldots,n_{ICR} \leq min(2J, I) - 1$) da regressão são estimados via método dos quadrados mínimos ordinários. Note que na equação (38) os coeficientes também não estão relacionados as variáveis originais. Para estimação dos coeficientes com interpretação biológica ($\hat{m}_{ICR}$), faz-se:

$$\hat{m}_{ICR} = KR\,\hat{\gamma} \tag{39}$$

sendo $\hat{\gamma}$ o vetor das estimativas dos coeficientes provenientes da regressão entre a variável Y e os componentes independentes S ($I \times n_{ICR}$).

A princípio, na análise de componentes independentes, não é possível determinar a variância dos componentes independentes, no entanto, isso se torna possível quando se assume que os componentes independentes tenham variância unitária e média igual a 0 (Hyvärinen, 1999). Diferentemente dos componentes principais, todos os componentes independentes explicam partes pequenas da variância total dos dados. Além disso, também não é possível determinar a ordem com que os componentes independentes são extraídos.

3. Métodos de regularização ou shrinkage

Os métodos de regularização fazem uso de estimadores do tipo shrinkage e assumem que os efeitos dos marcadores são variáveis de efeito aleatório. Tais metodologias, têm o objetivo de limitar a amplitude das estimativas dos coeficientes a fim de reduzir a variância. Por consequência, produzem estimativas estáveis visto que considera o tamanho da amostra e a variabilidade devido aos efeitos aleatórios ou residuais.

3.1.G-BLUP

O método G- BLUP (Genomic Best Linear Unbiased Predictor) consiste no método BLUP fenotípico tradicional utilizando as matrizes de parentesco genômico

no lugar das matrizes de parentesco baseadas em pedigree. Dessa forma, tem-se que o modelo linear misto a nível de indivíduos pode ser expresso por:

$$y = 1\mu + Za + Zd + e, \tag{40}$$

em que:

y é o vetor de observações fenotípicas com dimensão $I \times 1$, sendo I o número de indivíduos genotipados e fenotipados;

μ é a média geral da característica e 1 é o vetor com dimensão $I \times 1$ cujos seus elementos são iguais a 1;

a é o vetor de efeitos genômicos aditivos dos indivíduos ($I \times 1$) com matriz de incidência Z ($I \times I$), sendo a estrutura de variância dada por $a \sim N(0, G_a\sigma_a^2)$ em que σ_a^2 é a variância aditiva e G_a ($I \times I$) é a matriz de parentesco genômica para os efeitos aditivos;

d é o vetor de efeitos genômicos devido à dominância dos indivíduos com matriz de incidência Z (I × I), sendo a estrutura de variância dada por $d \sim N(0, G_d\sigma_d^2)$ em que σ_d^2 é a variância devido à dominância e G_d (I × I) é a matriz de parentesco genômica para os efeitos devido à dominância;

e é o vetor de efeitos residuais aleatórios com $e \sim N(0, I\sigma_e^2)$ e σ_e^2 é a variância residual.

Utilizando as equações de modelos mistos, podemos predizer os valores genômicos aditivos e devido à dominância dos indivíduos por meio do método G-BLUP dado matricialmente por:

$$\begin{bmatrix} 1'1 & 1'Z & 1'Z \\ Z'1 & Z'Z + G_a^{-1}\dfrac{\sigma_e^2}{\sigma_a^2} & Z'Z \\ Z'1 & Z'Z & Z'Z + G_d^{-1}\dfrac{\sigma_e^2}{\sigma_d^2} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a} \\ \hat{d} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z'y \end{bmatrix},$$

em que $G_a = \frac{WW'}{\sum_{j=1}^{J}(2p_jq_j)}$ (matriz de parentesco genômico dos efeitos aditivos) e $G_d = \frac{SS'}{\sum_{j=1}^{J}(2p_jq_j)^2}$ (matriz de parentesco genômico dos efeitos de dominância) sendo p_j e q_j as frequências alélicas do j-ésimo marcador (Vitezica et al., 2013) e os componentes de variância (σ_e^2, σ_a^2 e σ_d^2) são estimados via REML (Restricted Maximum Likelihood).

Destaca-se ainda que o modelo a nível de indivíduos é equivalente ao modelo a nível de marcas, sendo:

$$a = Wm_a$$

$$d = Sm_d$$

em que m_a são os efeitos aditivos dos marcadores e m_d são os efeitos devido à dominância dos marcadores.

4. Seleção de Variáveis Latentes

Segundo James et al. (2013), a abordagem do método dos quadrados mínimos ordinários é mais eficiente na estimação dos coeficientes da regressão, quando o número de observações é muito maior que o número de parâmetros. Quando existe a alta dimensionalidade, número de observações é menor que o número de parâmetros, para se fazer uso do método dos quadrados mínimos ordinários é necessário aplicar as metodologias de seleção de variáveis até que o número de observação seja menor que o número de parâmetros a serem estimados. Os procedimentos de seleção de variáveis consistem no descarte de variáveis explicativas que não estão diretamente relacionadas com a variável resposta. No entanto, sob contexto de seleção genômica, milhares de variáveis estão disponíveis e esse procedimento se torna inviável.

Desta forma, os métodos de redução de dimensionalidade se fazem necessários, além de possuírem teoria relativamente simples. Na aplicação dos métodos de redução de dimensionalidade é de extrema importância a escolha do número de componentes

a serem inseridos no modelo, consequentemente, a seleção de variáveis latentes/componentes é imprescindível. Os procedimentos de seleção de variáveis são viáveis sob este contexto, uma vez que o número de variáveis latentes deve ser igual a $min(2J, I) - 1$. Dentre estes procedimentos existentes, é apresentado o Forward Selection e Backward Selection.

4.1.Forward Selection

O método do Forward Selection inicia-se a partir do modelo básico, sem nenhuma variável explicativa, e a cada passo as variáveis podem ser incorporadas ao modelo ou não. O algoritmo é descrito por James et al. (2013) da seguinte forma:

i) Considere M_0 o modelo básico (sem nenhuma variável explicativa);

ii) Seja p o número de parâmetros a serem estimados pelo modelo e seja k a iteração em que o algoritmo se encontra, $k = 0, 1, 2, \ldots, p - 1$. Tem-se que $p - k$ modelos adicionam um parâmetro (uma variável explicativa) no modelo M_k. Dentre estes $p - k$ modelos, defina como M_{k+1} o modelo que possui o maior coeficiente de determinação (R^2) ou a menor soma de quadrados dos resíduos. Estas medidas estão associadas ao erro de estimação da população de treinamento;

iii) Após explicitar os modelos $M_0, M_1, \ldots, M_p$, escolha dentre estes o único melhor modelo utilizando alguns critérios, como por exemplo, o coeficiente de determinação ajustado (R^2_{aju}), o critério de Informação de Akaike (AIC), o critério de Informação Bayesiano (BIC), o critério de C_p de Mallow, entre outros. O melhor modelo é aquele que possui o maior R^2_{aju} e os menores AIC, BIC e C_p. Estas medidas estão associadas ao erro de estimação da população de validação.

4.2.Backward Selection

O método do Backward Selection é similar ao Forward Selection, no entanto inicia-se com o modelo completo (todas as variáveis explicativas incluídas no modelo) e a cada passo as variáveis são retiradas ou não do modelo. O algoritmo também é descrito por James et al. (2013) conforme a seguir:

i) Considere M_p o modelo completo (com todas as variáveis explicativas);

ii) Para $k = p, p-1, \ldots, 1$, tem-se k modelos que removem apenas um parâmetro do modelo M_k. Dentre estes k modelos, defina M_{k-1} como o modelo que possui maior coeficiente de determinação (R^2) ou a menor soma de quadrado dos resíduos;

iii) Após explicitar os modelos $M_p, M_{p-1}, \ldots, M_1$, escolha dentre estes o único melhor modelo baseando-se por exemplo, no coeficiente de determinação ajustado (R^2_{aju}), no critério de Informação de Akaike (AIC), no critério de Informação Bayesiano (BIC), no critério de C_p de Mallow, entre outros.

Segundo James et al. (2013), os critérios AIC e BIC são definidos para uma grande classe de modelos ajustados pelo método da máxima verossimilhança. No entanto, para os modelos cujo erros seguem uma distribuição gaussiana, maximizar a função de verossimilhança equivale a minimizar a soma de quadrados dos erros, ou seja, o método da máxima verossimilhança equivale o método quadrados mínimos ordinários. Dessa forma, sob este contexto, os critérios AIC e BIC são definidos, respectivamente, como:

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}(SQR + 2d\hat{\sigma}^2)$$

$$\text{BIC} = \frac{1}{n}(SQR + \log(n)d\hat{\sigma}^2),$$

em que SQR é a soma de quadrado dos resíduos, $\hat{\sigma}^2$ é o quadrado médio do resíduo, d é o número de parâmetros e n é o número de observações.

5. Validação em Seleção Genômica

Segundo Resende et. al (2012), a população em estudo ou o banco de dados em estudo pode ser dividido em dois subconjuntos, o primeiro denominado população de estimação e o segundo denominado população de validação. A seguir são apresentados detalhes de cada uma delas.

i) População de estimação/treinamento: abrange indivíduos que possuem disponível as informações fenotípicas e as genotípicas. Esta população é utilizada com intuito de estimar os efeitos dos marcadores sobre a característica de interesse;

ii) População de validação: abrange indivíduos que possuem disponível também as informações fenotípicas e as genotípicas. Baseada nas estimativas dos efeitos de marcadores encontradas na população de estimação, é possível predizer os valores genéticos genômicos dos indivíduos desta população. Isso possibilita calcular medidas de eficiência tais como acurácia e viés.

Também de acordo com Resende et al. (2012), considerando as duas populações definidas anteriormente, existem três formas de validação a fim de avaliar a eficiência associada as estimativas, tendo os efeitos da escolha da população de estimação e validação minimizados. São destacadas três formas de validação:

i) Apenas um conjunto de dados é utilizado para a população de estimação e para a população validação, ou seja, todos os indivíduos participam da população de estimação e da população de validação simultaneamente;

ii) O conjunto de dados é dividido em dois subconjuntos (estimação e validação) e estes subconjuntos são distintos, ou seja, nenhum indivíduo do conjunto de estimação participa do conjunto de validação e vice-versa. Esta metodologia é conhecida por validação independente;

iii) O processo de validação é divido em ciclos, a cada ciclo o conjunto de dados é dividido em dois subconjuntos (estimação e validação) e estes subconjuntos são distintos. Este processo é repetido por k ciclos até que todos os indivíduos tenham participado da população de estimação em $k-1$ ciclos e da população de validação em um único ciclo.

Partindo do pressuposto que os métodos mais utilizados são os procedimentos dos itens ii) e iii) (Resende et al. 2012), estes serão abordados detalhadamente. Os processos de validação cruzada consistem na divisão aleatória de um conjunto de dados com n elementos em dois subconjuntos (população de estimação e população de validação) (James et al. 2013).

A metodologia de validação cruzada descrita por James et al. (2013) leave-on-out consiste em utilizar o par correspondente (x_1, y_1) como população de validação e os restantes $(x_i\,, y_i)$ com $i = \{2, .., n\}$ como população de estimação. O mesmo procedimento é utilizado considerando $(x_2\,, y_2)$ como população de validação e $(x_i\,, y_i)$ com $i = \{1,3, .., n\}$ como população de estimação e assim sucessivamente este procedimento é repetido n vezes. No final é obtido o vetor de valores estimados e então pode ser calculada as medidas de eficiência, como capacidade preditiva, acurácia e viés.

Conforme retrata James et al. (2013), o processo de validação cruzada leave-on-out é um caso particular do processo k-fold. Na metodologia k-fold o conjunto de dados composto por n elementos é dividido em k grupos com g elementos cada, assim

teríamos $n = gk$. Em seguida, um grupo é utilizado como população de validação e os $g - 1$ restantes como população de estimação. Mas é imprescindível ressaltar que todos os grupos em uma das etapas será utilizado como população de validação. A cada ciclo pode ser calculada as medidas de eficiência e após o fim do processo é possível obter a média aritmética e o desvio-padrão dos k valores.

A validação independente também é um caso particular do método k-fold sendo $k = 2$. Assim os n elementos são divididos em dois grupos distintos (k_0 e k_1) sendo que o grupo k_0 é utilizado como população de validação e o grupo k_1 forma a população de estimação.

6. Referências Bibliográficas

ALMEIDA FILHO J. E.; GUIMARÃES J. F.; E SILVA F. F.; DE RESENDE M. D.; MUÑOZ P.; KIRST M.; RESENDE M. F J. R. The contribution of dominance to phenotype prediction in a pine breeding and simulated population. **Heredity**, v. 117, p. 33-41, 2016.

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.

AZEVEDO, C. F.; RESENDE, M. D. V.; SILVA, F. F.; VIANA, J. M. S.; VALENTE, M.S. F. RESENDE JUNIOR, M. F. R.; MUÑOZ, P. Ridge, Lasso and Bayesian additive-dominance genomic models. **BMC Genetics**, v. 16, p. 105, 2015.

BENNEWITZ J.; MEUWISSEN T. H. E. The distribution of QTL additive and dominance effects in porcine F2 crosses. **Journal of Animal Breeding and Genetics**, v. 127, p. 171-179, 2010.

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

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

COSTA, E. V.; DINIZ, D.B.; VERONEZE, R.; RESENDE, M. D. V.; AZEVEDO, C. F.; GUIMARÃES, S. E. F.; SILVA, F. F.; LOPES, P. S. Estimating additive and dominance variances for complex traits in pigs combining genomic and pedigree information. **Genetics and Molecular Research**, v. 14, p. 6303-6311, 2015.

DENIS M.; BOUVET J. M. Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. **Tree Genetics Genomes**, v. 9, p. 37-51, 2013.

FERREIRA, D. F. **Estatística Multivariada**. 2.ed.rev e amp, p. 411-419, 2012.

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

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

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

GODDARD M. E.; WRAY N. R.; VERBYLA K.; VISSCHER P. M. Estimating effects and making predictions from genome-wide marker data. **Statistical Science**, v. 24, p. 517-529, 2009.

HÄRDLE, W.; HLÁVKA, Z. **Multivariate Statistics: Exercises and Soluctions**. Springer, p. 145-150, 2007.

HILL W. G.; GODDARD M. E.; VISSCHER P. M. Data and theory point to mainly additive genetic variance for complex traits. **PLoS Genet**, v. 4, p. e1000008, 2008.

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

HYVÄRINEN, A. Fast and robust fixed-point algorithms for independent component analysis. **IEEE transactions on Neural Networks**, v. 10, n. 3, p. 626-634, 1999.

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

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning Garther**, Springer p. 205-212, 2013.

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

LEITE, I. C. C. Análise de Componentes Independentes aplicada a avaliação de imagem radiográfica de sementes. Tese (Doutorado em Estatística e Experimentação Agropecuária, Lavras, 2013.

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

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

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

MUÑOZ, P. R.; RESENDE, J. R M. F. R.; GEZAN, S. A.; RESENDE, M. D. V.; DE LOS CAMPOS G.; KIRST, M.; HUBER, D.; PETER, G. F. Unraveling additive from nonadditive effects using genomic relationship matrices. **Genetics**, v. 198, p. 1759-1768, 2014.

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

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

RENCHER, A. C; Christensen, W.F. **Methods of multivariate analysis**. John Wiley & Sons, 2rd ed, 2002.

RESENDE, M. D. V.; 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, 2012. 291p. Disponível em: http://www.ppestbio.ufv.br/?page_id=448.

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

SU,G.; CHRISTENSEN, O. F.; OSTERSEN, T.; HENRYON, M.; LUND, M. S. Estimating Additive and Non-Additive Genetic Variances and Predicting

Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers, **PLoS One**, v. 7, p. e45293, 2012.

TORO M. A, VARONA L. A note on mate allocation for dominance handling in genomic selection. **Genetics Selection Evolution**, v. 42, p. 33, 2010.

VITEZICA, Z. G.; VARONA, L.; LEGARRA, A. On the additive and dominance variance and covariance of individuals within the genomic selection scope. **Genetics**, Austin, v. 195, n. 4, p. 1223-1230, 2013.

VIGNAL, A.; MILAN, D.; SANCRISTOBAl, M.; EGGEN, A. A review on SNP and other types of molecular markers and their use in animal genetics. **Genetics Selection Evolution**, v. 34, p. 275, 2002.

WANG, C.; DA, Y. Quantitative genetics model as the unifying model for defining genomic relationship and inbreeding coefficient. **PLoS One**, v. 9, p. e114484, 2014.

WOLD, H. **Soft modelling by latent variables: the non-linear iterative partial least squares approach**. Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett, J. Gani, ed., Academic Press, London, 1975.

WELLMANN R.; BENNEWITZ J. Bayesian models with dominance effects for genomic evaluation of quantitative traits. **Genetics Research**, v. 94, p. 21-37, 2012.

ZENG J.; TOOSI A.; FERNANDO R. L, DEKKERS J. C. M.; GARRICK D. J. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. **Genetics Selection Evolution**, v.45, p.11, 2013.

CAPÍTULO 2

CRITÉRIOS PARA ESCOLHA DO NÚMERO ÓTIMO DE COMPONENTES INDEPENDENTES EM MODELOS ADITIVOS NO CONTEXTO DE SELEÇÃO GENÔMICA

Resumo

A Seleção Genômica Ampla (Genome-Wide Selection- GWS) consiste na análise de milhares de marcadores SNPs (Single Nucleotide Polymorphisms) amplamente distribuídos no genoma a fim de estimar o mérito genético dos indivíduos. Dessa forma, os indivíduos geneticamente superiores podem ser identificados rapidamente, o que tem contribuído para grandes avanços no melhoramento animal e vegetal. No entanto, devido a alguns desafios estatísticos tais como a multicolinearidade (marcadores altamente correlacionados) e a alta dimensionalidade, as metodologias de redução dimensional, Regressão via Componentes Principais (PCR) e Regressão via Componentes Independentes (ICR), vêm ganhando destaque na GWS. A aplicação desses métodos exige a escolha do número ótimo de componentes a serem utilizados no modelo e que possam explicar de forma satisfatória a variabilidade total dos dados. Para a PCR já foram definidos critérios para a escolha do número ótimo de componentes. Todavia, embora a ICR tenha se mostrado superior em relação a predição de valores genômicos aditivos, quando comparado aos demais métodos de redução de dimensionalidade, ainda não existem critérios formais para a seleção do número de componentes independentes. O critério exaustivo que escolhe o número de componentes independentes que está associado a uma maior acurácia exige elevada demanda computacional. Em virtude disso, a relevância do presente trabalho está em propor sete critérios de escolha do número de componentes independentes baseando-se no número de componentes principais e em métodos de seleção de variáveis. Os critérios propostos definem o número de componentes independentes

igual ao número de componentes que conduz a: (i) PCR a um maior valor de acurácia; (ii) PCR a um menor valor de viés; (iii) 80% da variação total de X explicada pelos componentes principais; (iv) 80% da variação total de Y explicada pelos componentes principais; (v) 80% da variação total de X explicada pelos componentes independentes; e o número de componentes independentes igual ao número de variáveis determinadas pelos procedimentos (vi) Forward Selection; (vii) Backward Selection. O conjunto de dados simulados constituiu-se de 1.000 indivíduos de 20 famílias de irmãos completos que foram genotipados e fenotipados para 2.000 marcadores SNPs e contemplando quatro cenários (dois níveis de herdabilidade × duas arquiteturas genéticas). Para a comparação entre os critérios, foram computadas as medidas de eficiência (herdabilidade molecular, acurácia e viés de predição). A fim de mostrar a aplicabilidade dos critérios no melhoramento vegetal, foi utilizado um conjunto de dados reais em que foram avaliadas seis características de produtividade do arroz asiático Oryza sativa (Número de panículas por planta, altura da planta, comprimento da panícula, número de panículas no perfilho primário, número de sementes por panícula e espiguetas por panícula) referente a 370 acessos de arroz e 44.100 marcadores de SNPs. Os resultados para dados simulados e reais mostraram que o critério no qual o número de componentes independentes é igual ao número de componentes principais que conduz a um maior valor de acurácia se mostrou mais eficiente além de reduzir drasticamente o tempo computacional no processo de predição. Os critérios 2 e 3 também se destacaram. Dentre todos critérios avaliados, nenhum conseguiu recuperar os valores de herdabilidade que foram simulados e além disso, em geral, foram viesados.

Palavras-chave: Predição Genômica, melhoramento vegetal, análise de componentes independentes, análise de componentes principais.

1. INTRODUÇÃO

A genética molecular tem contribuído consideravelmente para o melhoramento genético ao utilizar informações provenientes diretamente do DNA e assim, propiciar com maior rapidez e eficiência a identificação e a seleção de indivíduos geneticamente superiores. Neste contexto, a Seleção Genômica Ampla (Genome-Wide Selection- GWS) desenvolvida por Meuwissen et al. (2001) consiste no uso de milhares de marcadores SNPs (Single Nucleotide Polymorphisms), que cobrem amplamente o genoma, juntamente com as informações fenotípicas buscando predizer os valores genéticos genômicos (Genomic estimated breeding values - GEBVs) dos indivíduos com base nas estimativas dos efeitos desses marcadores no fenótipo.

No entanto, o avanço das técnicas de sequenciamento e o alto custo de genotipagem por amostra influencia para a redução do número de observações fenotípicas em relação ao número de informações genotípicas. Sob esta perspectiva, são detectados problemas relativos a alta dimensionalidade (número de marcadores maior que o número de observações fenotípicas individuais) e multicolinearidade (marcadores altamente correlacionados). Tais problemas, impossibilitam o uso de métodos baseados em quadrados mínimos ordinários (Ordinary Least Squares – OLS) (Gianola et al., 2003).

Em virtude disso, a GWS tem obtido êxito ao buscar superar estes desafios estatísticos baseando-se em métodos de regressão explícita, regressão implícita e em redução de dimensionalidade (Resende et al., 2012). Dentre estes, os métodos de redução de dimensionalidade se destacam pela facilidade de sua aplicação e de entendimento da teoria que envolve tais metodologias, como a Regressão via

Componentes Principais (Principal Component Regression – PCR) (Kendall, 1957; Hotelling, 1957) e a Regressão via Componentes Independentes (Independent Component Regression – ICR) (Jutten E Héraut, 1991; Comon, 1994). A metodologia da ICR foi proposta na predição genômica considerando modelos aditivos por Azevedo et al. (2013) e se mostrou superior em relação a PCR por estar associado a elevados valores de acurácia e a ausência de viés de predição, sendo assim mais eficiente para predizer o mérito genético aditivo de cada indivíduo. Por outro lado, o método ICR exige maior demanda de tempo computacional para executar as análises se comparado a PCR.

A aplicação dos métodos de redução de dimensionalidade requer a escolha do número ótimo de variáveis latentes (combinações lineares dos marcadores) que serão inseridas no modelo de regressão linear múltipla, sendo este passo de extrema importância para que não haja perda de informações relevantes para a variação dos dados. Para a PCR já existem critérios formais de escolha, como por exemplo, o critério baseado na variabilidade total dos dados visto que os primeiros componentes explicam grande parte desta variabilidade (Ferreira, 2012). No entanto, para a ICR ainda não existem critérios formais e a dificuldade é gerada primordialmente por que os componentes independentes explicam partes pequenas e em diferentes proporções da variabilidade total dos dados.

Sendo assim, baseado na importância da ICR na seleção genômica e na elevada demanda computacional para se determinar o número ideal de componentes independentes, este estudo tem por objetivo principal propor e avaliar sete critérios para a escolha do número ótimo de componentes independentes inseridos no modelo, sem que haja redução na eficiência da predição dos valores genômicos aditivos. Por fim, para avaliação dos critérios, é utilizado um banco de dados simulados, com quatro

cenários considerando duas herdabilidades e duas arquiteturas genéticas. A comparação dos critérios foi realizada considerando medidas de eficiência tais como acurácia, herdabilidade molecular e viés. A predição genômica em arroz vêm sendo abordada por alguns estudos como Spindel et al. (2015), Grenier et al. (2015), Spindel et al. (2016) e Hassen et al. (2018). Assim, os critérios foram avaliados em dados reais, na predição genômica de seis características de produtividade em arroz, visando elucidar a importância do método nos programas de melhoramento.

2. MATERIAIS E MÉTODOS

2.1.Dados Simulados

A descrição da geração do conjunto de dados simulados foi apresentada por Azevedo et al. (2015). Um total de 2.000 marcadores SNPs equidistantes separados por 0,1 centiMorgan (cM) entre os dez cromossomos foram simulados. Os QTLs foram distribuídos nas regiões abrangidas pelo SNPs. Um montante de 1.000 indivíduos de 20 famílias de irmãos completos foram genotipados e fenotipados.

Características com duas arquiteturas genéticas foram simuladas, uma seguindo um modelo infinitesimal (locos não ligados, com efeitos iguais – herança poligênica) e outra com cinco genes de efeitos maiores, responsável por 50% da variabilidade genética (herança mista). No primeiro caso, para cada um dos 100 QTLs foi atribuído um efeito aditivo de pequena magnitude no fenótipo (sob a definição de distribuição Normal). Para o segundo caso, pequenos efeitos aditivos foram designados para os restantes 95 locus. Os efeitos foram normalmente distribuídos com média zero e variância genética permitindo o nível de herdabilidade desejado. O valor fenotípico foi obtido adicionando ao valor genotípico um efeito ambiental proveniente de uma distribuição normal $N(0, {\sigma^2}_e)$, em que a variação ${\sigma^2}_e$ foi definida de acordo

com dois níveis de herdabilidade no sentido restrito de 0,20 e 0,30, respectivamente. Os níveis de herdabilidade foram escolhidos para representar uma característica com baixa herdabilidade e outra com herdabilidade moderada, casos em que se espera que a seleção genômica seja superior à seleção fenotípica (Azevedo et al., 2015).

As simulações assumiram ausência de dominância e os efeitos aditivos foram normalmente distribuídos com média zero e variância dependendo da arquitetura genética (herança mista e poligência). Marcadores com MAF (Minor Allele Frequency – Frequência do Menor Alelo) menor do que 5% foram excluídos das análises. Quatro cenários diferentes foram utilizados nas análises: dois níveis de herdabilidades em sentido restrito (cerca de 0,20 e 0,30) × duas arquiteturas genéticas (herança poligênica e mista). Os cenários são descritos conforme Tabela 1 e eles foram analisados considerando os métodos de redução de dimensionalidade, ICR e PCR. Cada tipo de população (ou cenários) foi simulada 10 vezes.

Tabela 1 – Cenários com as respectivas médias das herdabilidades aditivas (h_a^2) e arquiteturas genéticas (características controladas por genes de pequeno efeito – herança poligênica e características controladas por genes de pequeno e maior efeito - herança mista).

Cenários	Arquitetura Genética	h_a^2
Cenário 1	Herança poligênica	0,20
Cenário 2	Herança poligênica	0,30
Cenário 3	Herança mista	0,20
Cenário 4	Herança mista	0,30

2.2.Dados Reais

O arroz asiático Oryza sativa é um dos alimentos mais consumidos em grande parte do mundo. Assim, o crescimento populacional justifica o interesse dos pesquisadores em tornar as variedades deste arroz altamente produtivas. O banco de

dados utilizado neste estudo é composto por sete características de produtividade, referentes a 370 acessos de arroz, o qual foram genotipados para 44.100 marcadores SNPs. Este conjunto de dados é público e faz parte de dois projetos, o Projeto OryzaSNP e o Projeto OMAP (Ammiraju et al., 2006; Zhao et al., 2011) e está disponível no site https://ricediversity.org/data/.

As plantações foram supervisionadas ao longo da fase de crescimento dos acessos, no período de maio a outubro dos anos 2006 e 2007. Foi utilizado um delineamento em blocos completos com duas repetições, em que as linhas de plantações tinham comprimento igual a 5m. As plantas estavam espaçadas em 25cm entre si e 0,50m entre as fileiras. Maiores detalhes podem ser encontrados em Zhao et al. (2011).

Os marcadores SNPs com baixa taxa de atendimento na genotipagem (call rate < 70%) e baixa frequência do alelo mais raro (MAF < 1% - Minor Allele Frequency) foram removidos da análise, uma vez que são pouco informativos e não apresentam relevância genética na população. Após o controle de qualidade do banco de dados genômicos foi totalizado 36.901 marcadores SNPs.

As seis características de produtividade de arroz que serão utilizadas neste estudo são: (i) número de panículas por planta, (ii) altura da planta, (iii) comprimento da panícula, (iv) número de panículas no perfilho primário, (v) número de sementes por panícula e (vi) espiguetas por panícula.

2.3.Modelo linear básico

Considere o seguinte modelo linear:

$$y = 1\mu + Xm_a + e \quad (1)$$

em que:

y é o vetor de observações fenotípicas com dimensão I × 1, sendo I o número de indivíduos genotipados e fenotipados; μ é a média geral da característica; m_a é o vetor de efeitos aditivos de marcadores com matriz de incidência X composta por valores 0, 1 e 2 cuja dimensão é I × J, sendo J o número de marcadores; e é o vetor de erros aleatórios com estrutura de variância dada por $e \sim N(0, I\sigma_e^2)$ sendo I a matriz identidade e σ_e^2 a variância residual.

2.4.Regressão via Componentes Principais

A regressão via componentes principais (Principal Components Regression - PCR) tem como objetivo resolver problemas de alta dimensionalidade e multicolinearidade. Para isto, os componentes principais são escritos como combinações lineares da variável X da seguinte forma:

$$Z = XP \tag{2}$$

em que as colunas de Z representam os primeiros n_{PCR} $(1 \leq n_{PCR} \leq min(I, 2J) - 1)$ componentes principais, X é a matriz de incidência dos marcadores e P é a matriz dos primeiros n_{PCR} autovetores da matriz de covariância de X. Os componentes principais Z_m $(m = 1, \ldots, n_{PCR})$ são chamados também componentes ortogonais, visto que não são correlacionados, isto é, $Cor\ (Z_m\ , Z_{m'}) = 0$ para todo $m \neq m'$.

Com o intuito de relacionar a variável resposta Y e as variáveis latentes $Z_1, Z_2, \ldots, Z_{n_{PCR}}$ realiza-se uma regressão linear múltipla obtendo a seguinte equação de predição:

$$\hat{y} = Z\hat{\alpha} \tag{3}$$

em que $\hat{\alpha}_m$ $(m = 1, \ldots, n_{pcr})$ são as estimativas dos coeficientes da regressão entre Y e Z que podem ser obtidas pelo método dos quadrados mínimos ordinários (Ordinary Least Squares – OLS).

No entanto, estes coeficientes não estão relacionados as variáveis originais, para isso basta combinar as equações (2) e (3) e as estimativas dos efeitos associados as variáveis originais (marcadores) são dadas por:

$$\hat{m}_{PCR} = P\hat{\alpha}.$$

2.5.Regressão via Componentes Independentes

O método ICR (Independent component regression) proposto sob o contexto de predição genômica aditiva por Azevedo et al. (2013), consiste em decompor a matriz de covariáveis X em combinações de componentes independentes, o que garante a retirada da multicolinearidade dos dados, além de reduzir a dimensionalidade. Por esse método, existe o pressuposto de que os dados sejam provenientes de uma distribuição não normal. Portanto, a ICR pode ser aplicada de forma eficiente à GWS, em que a matriz de incidência de marcadores X parametrizada com os valores 0, 1 e 2 (distribuição não normal).

Assim, tem-se a decomposição dada por:

$$X = SA' \tag{4}$$

em que a matriz X ($I \times 2J$) é decomposta em uma matriz de componentes independentes S ($I \times min(2J, I) - 1$) e em uma matriz A ($2J \times min(2J, I) - 1$) denominada matriz de misturas e que geralmente é desconhecida.

Para estimar a matriz A, o primeiro passo é obter uma matriz K (conhecida como matriz de branqueamento – whitening matrix) por meio da decomposição ortogonal da matriz de covariância de X, fazendo com que a matriz de covariância de XK seja igual a matriz identidade, sendo assim, a correlação entre as colunas de XK é igual a 0 e variância igual a 1. Para isso, a decomposição ortogonal é aplicada a matriz de covariância amostral de X, denotada por Σ ($2J \times 2J$), obtendo:

$$\Sigma = P\Lambda^{-\frac{1}{2}}P'$$

em que P é composta pelos autovetores em suas colunas e Λ é uma matriz diagonal de autovalores da matriz de covariância amostral de X. Sob o contexto de regressão, a matriz K $(2J \times min(2J, I) - 1)$ é então definida como $P_r\Lambda_r^{-\frac{1}{2}}$, sendo P_r é matriz com as $min(2J, I) - 1$ primeiras colunas da matriz P ($min(2J, I) - 1$ primeiros autovetores) e Λ_r $(min(2J, I) - 1 \times min(2J, I) - 1)$ é uma matriz com as $min(2J, I) - 1$ primeiras linhas e colunas da matriz Λ (autovalores associados a esses primeiros autovetores).

Com o intuito de atingir a independência entre os componentes, é utilizado o algoritmo proposto por Hyvärinen (1998), que é fundamentado no princípio de máxima entropia ($J(R)$), obtendo uma estimativa para a matriz R ($min(2J, I) - 1 \times min(2J, I) - 1$) pela seguinte expressão:

$$J(R) \propto \left\{E[G(R)] - E\left[G\left(R_{gaussiana}\right)\right]\right\}^2{}_2 \quad (5)$$

em que G é uma função não quadrática e a escolha da função G influencia na aproximação da negentropia (Hyvärinen, 1999). Após o algoritmo, tem-se que os componentes independentes podem ser expressos por:

$$S = XKR. \quad (6)$$

Posteriormente, a equação de predição entre a variável resposta Y e os componentes independentes $S_1, S_2, \ldots, S_{n_{icr}}$ é dada por:

$$\hat{y} = \hat{\gamma}_0 + \hat{\gamma}_1 s_1 + \ldots. + \hat{\gamma}_{n_{icr}} s_{n_{icr}} \quad (7)$$

em que as estimativas dos coeficientes $\hat{\gamma}_m$ $(m = 0,1,\ldots, n_{ICR}$, sendo $n_{ICR} \leq min(2J, I) - 1)$ são obtidas via método dos quadrados mínimos ordinários.

Os coeficientes da equação (7) também não estão relacionados às variáveis originais e para estimar os efeitos associados as variáveis originais (marcadores) basta combinar as equações (6) e (7) como apresentado a seguir:

$$\widehat{m}_{ICR} = KR\,\hat{\gamma} \tag{8}$$

sendo $\hat{\gamma}$ o vetor das estimativas dos coeficientes provenientes da regressão entre a variável Y e os componentes independentes S ($I \times n_{ICR}$).

2.6. Critérios para Escolha do Número Ótimo de Componentes Independentes

A Regressão via Componentes Principais (PCR) e a Regressão via componentes Independentes (ICR) exigem a escolha do número de componentes adequado que expliquem de forma satisfatória a variação dos dados, ou seja, que a redução de dimensionalidade não conduza a perda de informações relevantes. Critérios de decisão quanto ao número ótimo de componentes considerando o método PCR já são amplamente utilizados, como por exemplo, adotar uma porcentagem de interesse (entre 70% e 80%) baseada na fração da variabilidade total dos dados (Ferreira, 2012). Contudo, a ICR não possui um critério já formalizado, principalmente, sob o contexto de grandes bancos de dados, como na seleção genômica.

Na predição genômica as principais medidas para avaliação de eficiência são a acurácia (correlação entre o valor genômico estimado e valor genético) e a capacidade preditiva (correlação entre o valor genômico estimado e valor fenotípico), e por este motivo, Azevedo et al. (2014) e Azevedo et al. (2015) escolheram o número de componentes independentes que estava associado a uma maior acurácia e a uma maior capacidade preditiva. No entanto, quando se trata de bancos de dados de alta dimensionalidade (bancos genômicos), os algoritmos e as ferramentas computacionais

da ICR tornam inviáveis as análises exaustivas que variam o número de componentes para a decisão do número ótimo de componentes independentes, como feito por estes autores.

Segundo alguns autores (Cadavid et al., 2008; Azevedo et al., 2013) é possível recorrer as ferramentas computacionais da PCR e utilizar o número de componentes independentes igual ao número de componentes principais. Assim, os critérios analisados neste trabalho visam determinar o número ótimo de componentes independentes a partir de critérios descritos conforme a seguir:

i) **Critério 1**: Para cada número de componentes principais ($m = 1, \dots, min(I, 2J) - 1$), são estimados os efeitos de marcadores na população de estimação via PCR e estes são utilizados na população de validação para estimar os valores genômicos aditivos dos indivíduos desta população. Em seguida, calcula-se para os dados simulados a acurácia ($r_{a\hat{a}}$), correlação entre o valor genômico estimado e o valor genômico real $\left(r_{a\hat{a}} = Cor\ (\hat{a}, a)\right)$, e para os dados reais a capacidade preditiva ($r_{y\hat{a}}$) , correlação entre o valor genômico estimado e o valor fenotípico $\left(r_{y\hat{a}} = Cor\ (\hat{a}, y)\right)$. Escolhe-se o número de componentes principais cujo valor genômico aditivo conduz a uma maior acurácia e a uma maior capacidade preditiva. Dessa forma, o presente critério determina que o número de componentes independentes deve estar associado a esse número de componentes principais.

ii) **Critério 2**: Para cada número de componentes principais ($m = 1, \dots, min(I, 2J) - 1$), são estimados os efeitos de marcadores na população de estimação via PCR e estes são utilizados na população de validação para estimar os valores genômicos aditivos dos indivíduos desta população. Em seguida, calcula-se o coeficiente de regressão ($b_{y\hat{a}}$), coeficiente obtido da

regressão entre o fenótipo e o valor genômico aditivo estimado e posteriormente o viés de predição dado por $(1 - b_{y\hat{a}})$. Escolhe-se o número de componentes principais cujo valor genômico conduz a um menor viés de predição. Dessa forma, o presente critério determina que o número de componentes independentes deve estar associado a esse número de componentes principais.

iii) **Critério 3**: A porcentagem de explicação da variação total de X ao utilizar m componentes principais é dada por: $p_m(\%) = \frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{2J} \sigma_j^2}$ em que λ_j é o autovalor correspondente ao j-ésimo autovetor da matriz de covariância de X e σ_j^2 é a variância da j-ésima variável X. O presente critério determina que o número de componentes independentes a ser utilizado no modelo deve ser igual ao número de componentes principais que expliquem 80% da variação total de X.

iv) **Critério 4**: Utilizando o coeficiente de determinação $(R^2 = Cor(y, \hat{a})^2 \times 100\%)$, escolhe-se o número de componentes principais que explicam um determinado percentual da variação total de Y. O presente critério determina que o número de componentes independentes a ser inserido no modelo deve ser igual ao número de componentes principais que expliquem 80% da variação total de Y.

v) **Critério 5:** Assumindo que os componentes independentes possuem média 0 e variância igual a 1, tem-se que a variação contabilizada pelo k-ésimo componente é dada por $\frac{I \sum_{j=1}^{2J} a_{jk}^2}{\sum_{i=1}^{I} \sum_{j=1}^{2J} x_{ij}^2}$, em que a_{jk} é o elemento da j-ésima linha e k-ésima coluna da matriz de misturas ($j = 1,2, \dots, 2J$ e $k = 1, \dots, min(I, 2J) - 1$), x_{ij} é o elemento da i-ésima linha e j-ésima coluna da matriz centrada de variáveis explicativas X ($i = 1,2, \dots, I$) e I é o número de observações

(Bingham e Hyvärinen, 2000; Helwig e Hong, 2013). Escolhe-se então, o número de componentes independentes que expliquem 80% da variação total de X.

vi) **Critério 6**: Baseando-se no algoritmo do Forward Selection, considere o modelo M_0 sem nenhum componente independente. Em seguida, para $n_{ICR} = min(I, 2J) - 1$ (o número máximo de componentes independentes possíveis) e iteração k ($k = 0, 1, \dots, min(I, 2J) - 1$), determine como M_{k+1} o modelo que apresentar maior R^2. Escolha dentre estes $M_{k+1}{}'$s modelos, o único modelo que apresenta menor BIC (Critério de Informação Bayesiano). O presente critério determina quais e quantos componentes independentes devem estar no modelo escolhido.

vii) **Critério 7**: Baseando-se no algoritmo do método Bacward Selection, considere o modelo completo $M_{min(I,2J)-1}$, isto é, o modelo com o número máximo de componentes construídos. Posteriormente, defina como M_{k-1} ($k = min(I, 2J) - 1\ , min(I, 2J) - 2, \dots, 1$) o modelo que apresentar maior R^2 e escolha dentre estes $M_{k-1}{}'$s modelos, o único modelo que apresenta menor BIC. O presente critério determina quais e quantos componentes independentes devem estar no modelo escolhido.

2.7.Comparação das metodologias

Nos dados simulados, os métodos de redução de dimensionalidade serão comparados por meio de uma validação independente em que as nove primeiras replicatas serão assumidas como populações de estimação e utilizadas para estimar os efeitos dos marcadores SNPs no fenótipo. A décima replicata será assumida como população de validação e utilizada para predizer os valores genéticos genômicos

aditivos via estimativas dos efeitos dos marcadores obtidos na população de estimação. Assim, será viável calcular medidas de eficiência para a predição genômica, tais como acurácia ($r_{\hat{a}a}$), viés de predição ($1 - b_{y\hat{a}}$) e herdabilidade genômica (h^2_{aM}), das estimativas baseadas em cada um dos nove cenários.

As medidas utilizadas são descritas a seguir: (i) a acurácia, que é dada pela correlação entre os GEBVs e os valores paramétricos; (ii) o viés de predição, o qual é definido como sendo um menos o coeficiente da regressão entre o fenótipo e o GEBV ($1 - b_{y\hat{a}}$). Para os coeficientes de regressão abaixo de 1 ($b_{y\hat{a}} < 1$) entende-se que os GEBVs foram superestimados, para os coeficientes de regressão acima de 1 ($b_{y\hat{a}} > 1$) conclui-se que os GEBVs foram subestimados e para os coeficientes iguais a 1 ($b_{y\hat{a}} = 1$) conclui-se que os GEBVs são não viesados; (iii) a herdabilidade molecular aditiva é dada por $h^2_{aM} = \frac{\sigma^2_{a_M}}{\sigma^2_{a_M} + \sigma^2_e}$, em que $\sigma^2_{a_M} = \sum_{j=1}^{J} 2p_j q_j m^{2}_{aj}$ é a variância genômica aditiva, σ^2_e é a variância residual, p_j e q_j são as frequências alélicas do j-ésimo marcador. Após a obtenção das medidas de acurácia, coeficientes de regressão e herdabilidade para cada repetição em cada cenário será obtida a média desses valores.

A eficiência do método ICR considerando cada critério de escolha foi comparado utilizando seis características de produtividade avaliadas em arroz. As medidas de eficiência utilizadas foram a capacidade preditiva $r_{\hat{a}y}$, que consiste na correlação entre os valores genômicos aditivos estimados e os valores fenotípicos da população de validação e o viés de predição. Os dados reais foram avaliados sob um processo de validação *k-fold* com $k = 10$.

2.8.Recursos Computacionais

As configurações do computador utilizado nas análises estatísticas foram: processador Intel(R) Core(TM) i7-6500 (CPU 2.50 GHz) com 16 Gb de memória RAM. Todas as implementações dos métodos utilizados foram realizadas no software R (R Development Core Team, 2018). Os pacotes (e funções) utilizadas nas análises foram pls (pcr), caret (icr) e leaps (regsubsets) para as metodologias PCR, ICR, forward e backward, conforme (Mevik et al., 2016; Kuhn, 2017; Lumley, 2017) respectivamente.

3. RESULTADOS E DISCUSSÃO

Os resultados médios e os desvios das simulações referentes ao número de componentes, herdabilidade molecular aditiva, acurácia e viés considerando a ICR e cada critério de escolha do número ótimo de componentes independentes são apresentados na Tabela 2 e na Tabela 3. Além disso, é apresentado também os resultados das análises referentes ao cálculo do número de componentes necessários para atingir o valor máximo de acurácia via ICR pelo modo exaustivo.

Dentre os sete critérios analisados, os critérios 1, 6 e 7 apresentaram os valores de acurácia mais próximos do valor máximo de acurácia obtida pelo método exaustivo, considerando os quatros cenários. Em contrapartida, os valores mais baixos de acurácia estão associados ao critério 2, exceto no cenário 3. Apesar dos critérios 6 e 7 apresentarem elevados valores de acurácia, ambos os critérios foram os que mais superestimaram os valores genômicos, revelando que as estimativas encontradas apresentam variabilidade além da simulada.

Além disso, em um primeiro momento, quando é analisado o coeficiente de regressão pelos critérios 1, 2 e 3 nos cenários 3 e 4 (Tabela 3), nota-se que o critério 2

apresenta ausência de viés (valores próximos de 1) e os outros superestimam os valores genômicos. No entanto, os critérios 1 e 3 apresentam valores mais próximos da unidade que o obtido pelo método exaustivo, destacando-se ainda mais o critério 1. A relevância da propriedade do não vício está quando a seleção envolve indivíduos de muitas gerações usando efeitos dos marcadores estimados em uma só geração e quando deseja-se não apenas selecionar os indivíduos, mas também determinar os méritos genômicos dos indivíduos (Resende et al., 2014).

O critério 1, assim como os outros critérios, não são adequados para a estimação da herdabilidade nos cenários 2, 3 e 4, visto que os valores obtidos são distantes do valor paramétrico. No entanto, observa-se que estes valores são próximos do valor da herdabilidade que o critério exaustivo atinge considerando o valor máximo de acurácia via ICR. Nos critérios 4, 5, 6, e 7 são encontradas estimativas de herdabilidade iguais a 1 e pode se perceber que estes critérios são associados aos maiores números de componentes. Dessa forma, avaliamos que o número de componentes influencia na estimativa da herdabilidade e a que medida em que se inclui componentes no modelo a herdabilidade tende a 1. Isto pode ser explicado, pelo método ICR assumir os SNPs como efeitos fixos, uma vez que segundo Resende et al. (2014) quando se assume os marcadores como efeitos fixos, implicitamente se assume que a herdabilidade é igual 1.

A associação existente entre o elevado número de componentes e os critérios 4, 5, 6 e 7 considerando todos os cenários é justificável. Em relação ao critério 4, isso pode ser explicado pelo fato da PCR levar em consideração apenas as variáveis explicativas X, e não a variável resposta Y, na construção dos componentes, necessitando de um número maior de componentes para explicar Y do que para explicar X (critério 3). Já no critério 5, o elevado número de componentes justifica-se

pois ao contrário da PCR, em que os primeiros componentes explicam a maior parte da variância total contida no conjunto de dados (X), a ICR divide em todos os componentes independentes partes pequenas da explicação da variação total dos dados (Figura 1).

No que se refere aos critérios Forward Selection e Backward Selection (critérios 6 e 7, respectivamente), os métodos de seleção de variáveis visam retirar do modelo variáveis que não são relevantes ou que não estão muito relacionadas com a variável dependente (James et al., 2013). Assim, em um modelo de regressão cujo conjunto de dados das variáveis explicativas são altamente correlacionadas, pode se tornar o modelo menos complexo reduzindo o número dessas variáveis para predizer a variável resposta. No caso da ICR os componentes são independentes (além de serem não correlacionados, os componentes não possuem qualquer relação funcional entre si), e dessa forma são necessários mais variáveis, isto é, mais componentes para explicar a variável resposta, nos critérios 6 e 7. Destaca-se que para a predição dos valores genômicos utilizando estes critérios de escolha não é adequada visto que foram os critérios associados aos maiores vieses (valores dos coeficientes próximos de 0).

Vale ressaltar que outros critérios foram propostos e analisados como por exemplo, Critério de Informação de Akaike (AIC); Critério de Informação Bayesiano (BIC); Coeficiente de determinação (R^2); quadrado médio dos resíduos (QME); Coeficiente de determinação ajustado (R_a^2). No entanto, a aplicação destes critérios sugeridos não se tornou viável, visto que o tempo computacional seria o mesmo que no método exaustivo (que seria o valor máximo de acurácia alcançada). Da mesma forma, utilizando o método Stepwise Selection o número de variáveis selecionadas resultou no modelo completo (considerando 999 componentes) que esteve associado a baixos valores de acurácia.

O número de componentes, herdabilidade molecular aditiva, capacidade preditiva e viés de predição referentes as seis características de arroz são apresentados na Tabela 4, considerando cada critério de escolha do número ótimo de componentes independentes. Da mesma forma, o número de componentes independentes necessários pelo modo exaustivo também pode ser visto na Tabela 4.

O critério 1 apresentou valores de capacidade preditiva mais próximos do máximo para as características número de panículas por planta, altura e comprimento da panícula. Já o critério 2 apresentou valores de capacidade preditiva mais próximos do máximo para as características número de panículas por planta e comprimento da panícula. O critério 3 para as características número de panículas por planta e comprimento da panícula, o critério 5 para número de sementes por panícula e o critério 4 não se destacou para nenhuma característica, no entanto os três critérios (3,4 e 5) superestimaram os valores genéticos genômicos nessas características.

Ao analisar o coeficiente de regressão pelos critérios 1, 2, 3 e 4 observa-se que o critério 2 em comparação com os demais apresenta o viés mais próximo de 1 para quatro características (número de panículas por planta, altura da planta, número de panículas no perfilho primário e número de sementes por panícula), sendo que para altura da planta este critério subestima o valor genômico. Já o critério 1 apresentou ausência de viés também para número de panículas por planta e comprimento da panícula. Observa-se ainda que os critérios 6 e 7 superestimaram consideravelmente os valores genômicos para todas as características.

Na estimativa da herdabilidade, os critérios que estavam associados a valores mais próximos de herdabilidade atingida pelo método exaustivo foram: critério 1 para características altura da planta e número de sementes por panícula, critério 2 para comprimento da panícula, critério 3 para número de panículas por planta e número de

panículas no perfilho primário e critério 4 para espiguetas por panícula. Além disso, os critérios 5, 6 e 7 apresentam elevado número de componentes e estão relacionados aos valores mais altos de herdabilidade, próximos ou iguais a 1, sendo assim, o número de componentes influencia na estimativa da herdabilidade, o que também foi verificado na análise utilizando dados simulados e estes critérios devem ser descartados.

Bisne et al. (2009) reportaram que valores de herdabilidade ocilantes entre alto e médio, indicam sucesso de seleção. Assim, considerando o conjunto de dados reais, destaca-se as características que se enquadram neste contexto. Dentre elas, o número de panículas por planta atingiu valores de alta herdabilidade ao apresentar valores no intervalo de 0,69 à 0,74 para o método exaustivo, critério 1, critério 2 ou critério 3, respectivamente. Já os estudos de Akinwale et al. (2011) e Seyoum et al. (2012) apresentaram valores inferiores para essa característica, cerca de 0,59 e 0,50 respectivamente.

As características altura da planta e comprimento de panícula apresentaram valores médios de herdabilidade no critério 1, respectivamente, iguais a 0,47 e 0,42 (valor próximo atingido pelo critério 2). Para altura da planta Grenier et al. (2015), Spindel et al. (2015) e Spindel et al. (2016) reportaram valores similares, no entanto, para comprimento de panícula Spindel et al. (2016) alcançaram uma herdabilidade superior ao reportado neste estudo.

Para a característica espiguetas por panícula, os critérios 1 e 2 apresentaram herdabilidade de 0,69 sendo que estes critérios obtiveram valor mais próximo atingido nos trabalhos desenvolvidos por Seyoum et al. (2012) e Akinwale et al. (2011) que alcançaram valores de 0,60 e 0,61, respectivamente. Observa-se ainda que as herdabilidades apresentadas por Akinwale et al. (2011) e Seyoum et al. (2012) são

atingidas via pedigree e no contexto do presente trabalho foi considerada a herdabilidade genômica.

Os tempos computacionais associados aos dados simulados e reais em segundos e horas são apresentados na Tabela 5. O tempo computacional referente ao método exaustivo do conjunto de dados simulados, considerando uma replicata de cada cenário, é de 163 horas. Todavia, pelos critérios 2 e 3 o tempo é reduzido drasticamente para cerca de 0,05 horas (aproximadamente 3 minutos); mas ambos os critérios apresentam valores baixos de acurácia e distantes da acurácia atingida pelo modo exaustivo. Em contrapartida, pelo critério 1 a redução do tempo também é relevante sendo necessários cerca de 0,06 horas (aproximadamente 3,7 minutos) e o critério mostra-se superior quando comparado aos outros métodos visto que, apresenta os valores de acurácia, coeficientes de regressão e herdabilidade próximos dos valores obtidos pelo método exaustivo.

Considerando os dados reais, observa-se também na Tabela 5, que o critério exaustivo exige elevado tempo computacional, cerca de 886 horas, utilizando um elevado número de marcadores moleculares. Este tempo é radicalmente maior ao atentar-se para o fato de que na genotipagem de bovinos e ovinos são identificados cerca de 500.000 e 600.000 marcadores SNPs (Wilkinson et al., 2017; Brito et al., 2017), ou seja, são centenas de milhares de efeitos de marcadores a serem estimados considerando apenas o modelo aditivo.

Verifica-se ainda que o tempo computacional é drasticamente reduzido levando em consideração os sete critérios que foram propostos. Os critérios 3, 4 e 5, possuem nesta ordem, tempo de execução da análise dado em horas por 0,61; 0,70 e 5,51; contudo superestimaram os valores genéticos genômicos para as características nas quais apresentaram elevado valor de acurácia. Da mesma forma, os critérios 6 e 7

apresentaram tempo de 0,91 horas para ambos, no entanto os GEBVs foram drasticamente superestimados. Dessa forma, destaca-se ainda os critérios 1 e 2 que apresentaram tempo computacional em horas de 1,65 e 1,40, utilizando estes critérios o método ICR foi pouco viesado para as características em que foram obtidos maiores valores de acurácia.

De forma geral, o critério 1, número de componentes independentes igual ao número de componentes principais que conduz a um maior valor de acurácia, mostra-se como uma alternativa eficaz e computacionalmente viável comparada os critérios utilizados por Azevedo et al. (2013, 2014, 2015) tanto para dados simulados quanto para as características de dados reais.

4. CONCLUSÕES

O critério que define o número de componentes independentes como sendo o número de componentes principais que conduz PCR a um maior valor de acurácia, mostrou-se como uma alternativa viável na determinação do número ótimo de componentes independentes a serem utilizados na predição dos valores genômicos tanto para dados simulados quanto para dados de arroz. A principal vantagem apresentada por este critério é a redução drástica no tempo computacional para executar as análises. Com isso, a ICR torna-se viável para a predição genômica considerando um número elevado de marcadores moleculares.

Tabela 2: A herdabilidade aditiva paramétrica ($h^2_{Ma_{par}}$), o número de componentes (Nc), herdabilidade aditiva (h^2_{aM}), acurácia ($r_{a\hat{a}}$) e coeficiente de regressão ($\hat{b}_{y\hat{a}}$) considerando cada critério de escolha do número de componentes independentes e os cenários de herança poligênica.

Cenário	$\boldsymbol{h^2_{Ma_{par}}}$	**Critério**	Nc	$\boldsymbol{h^2_{aM}}$	$r_{a\hat{a}}$	$\hat{b}_{y\hat{a}}$
Cenário 1	0.20	**Exaustivo**	**36 ± 0.00**	**0.20±0.03**	**0.71±0.02**	**0.90±0.05**
		Critério 1	66 ± 78	0.22 ± 0.07	0.70 ± 0.01	0.89± 0.11
		Critério 2	5 ± 4	0.04 ± 0.03	0.34 ± 0.15	1.10± 0.10
		Critério 3	130 ± 0.00	0.27 ± 0.03	0.70 ± 0.02	0.75± 0.04
		Critério 4	730 ± 0.00	1.00 ± 0.00	0.48± 0.04	0.25± 0.03
		Critério 5	780 ± 0.00	1.00± 0.00	0.46± 0.04	0.23± 0.02
		Critério 6	630± 470	1.00± 0.00	0.58± 0.17	0.07± 0.06
		Critério 7	630± 470	1.00± 0.00	0.58± 0.17	0.07± 0.06
Cenário 2	0.30	**Exaustivo**	**44 ± 0.00**	**0.22±0.04**	**0.75±0.02**	**0.92±0.07**
		Critério 1	40±25	0.21± 0.04	0.74± 0.02	0.94± 0.06
		Critério 2	7± 8	0.08± 0.04	0.48± 0.13	0.98± 0.06
		Critério 3	130± 0.00	0.28± 0.04	0.72± 0.02	0.78± 0.05
		Critério 4	730± 0.00	1.00± 0.00	0.50± 0.04	0.25± 0.03
		Critério 5	780± 0.00	1.00± 0.00	0.48± 0.04	0.23± 0.03
		Critério 6	950± 33	1.00± 0.00	0.71± 0.01	0.04± 0.01
		Critério 7	950± 33	1.00± 0.00	0.71± 0.01	0.04± 0.01

Número de componentes independentes que conduz a: ICR a um maior valor de acurácia (Exaustivo); PCR a um maior valor de acurácia (Critério 1); PCR a um menor valor de viés (Critério 2); 80% da variação total de X explicada pelos componentes principais (Critério 3); 80% da variação total de Y explicada pelos componentes principais (Critério 4); 80% da variação total de X explicada pelos componentes independentes (Critério 5); Forward Selection (Critério 6); Backward Selection (Critério 7).

Tabela 3: A herdabilidade aditiva paramétrica ($h^2_{Ma_{par}}$), o número de componentes (Nc), herdabilidade aditiva (h^2_{aM}), acurácia ($r_{a\hat{a}}$) e coeficiente de regressão ($\hat{b}_{y\hat{a}}$) considerando cada critério de escolha do número de componentes independentes e os cenários de herança mista.

Cenários	$\boldsymbol{h^2_{Ma_{par}}}$	**Critérios**	Nc	$\boldsymbol{h^2_{aM}}$	$r_{a\hat{a}}$	$\hat{b}_{y\hat{a}}$
Cenário 3	0.20	**Exaustivo**	**277 ± 0.00**	**0.53±0.06**	**0.77±0.02**	**0.77±0.05**
		Critério 1	260± 96	0.51± 0.14	0.77± 0.02	0.80± 0.10
		Critério 2	16± 17	0.15± 0.09	0.53± 0.19	1.00± 0.03
		Critério 3	130± 0.00	0.36± 0.04	0.75± 0.01	0.89± 0.06
		Critério 4	730± 0.00	1.00± 0.00	0.66± 0.02	0.45± 0.02
		Critério 5	780± 0.00	1.00± 0.00	0.63± 0.03	0.42± 0.03
		Critério 6	960± 16	1.00± 0.00	0.72± 0.02	0.04± 0.01
		Critério 7	960± 16	1.00± 0.00	0.72± 0.02	0.04± 0.01
Cenário 4	0.30	**Exaustivo**	**189 ± 0.00**	**0.48±0.03**	**0.80±0.02**	**0.83±0.03**
		Critério 1	200± 110	0.50± 0.14	0.80± 0.02	0.83± 0.09
		Critério 2	4± 3	0.11± 0.05	0.45± 0.11	1.00± 0.04
		Critério 3	130± 0.00	0.42± 0.03	0.79± 0.02	0.87± 0.04
		Critério 4	730± 0.00	1.00± 0.00	0.66± 0.03	0.45± 0.02
		Critério 5	780± 0.00	1.00± 0.00	0.63± 0.03	0.42± 0.03
		Critério 6	960± 30	1.00± 0.00	0.73± 0.01	0.04± 0.01
		Critério 7	960± 30	1.00± 0.00	0.73± 0.01	0.04± 0.01

Número de componentes independentes que conduz a: ICR a um maior valor de acurácia (Exaustivo); PCR a um maior valor de acurácia (Critério 1); PCR a um menor valor de viés (Critério 2); 80% da variação total de X explicada pelos componentes principais (Critério 3); 80% da variação total de Y explicada pelos componentes principais (Critério 4); 80% da variação total de X explicada pelos componentes independentes (Critério 5); Forward Selection (Critério 6); Backward Selection (Critério 7).

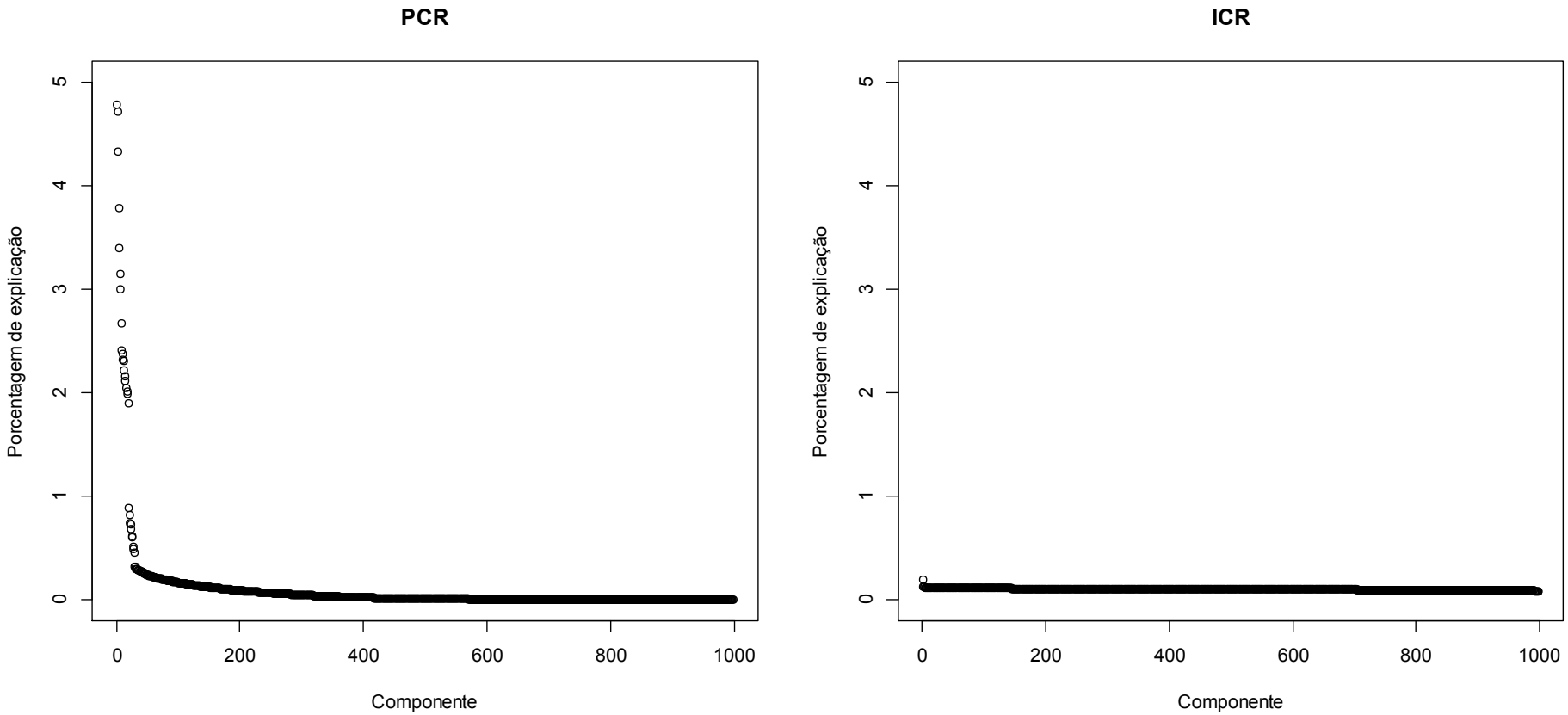

Figura 1: Porcentagem de explicação das variáveis explicativas em relação a cada componente considerando a análise de componentes principais (PCR) e a análise componentes independentes (ICR) considerando as matrizes de incidência de marcadores.

Tabela 4: O número de componentes (Nc), herdabilidade aditiva (h^2_{aM}), capacidade preditiva ($r_{y\hat{a}}$) e viés de predição ($\hat{b}_{y\hat{a}}$) considerando o método exaustivo e cada critério de escolha do número de componentes independentes.

Características	Critérios	Nc	h^2_{aM}	$r_{y\hat{a}}$	$\hat{b}_{y\hat{a}}$
Número de panículas por planta	**Exaustivo**	**117**	**0,69±0,05**	**0,82±0,06**	**0,98±0,11**
	Critério 1	125	0,73±0,03	0,82±0,01	0,97±0,03
	Critério 2	105	0,74±0,01	0,82±0,01	0,97±0,02
	Critério 3	44	0,69±0,03	0,82±0,01	1,02±0,03
	Critério 4	55	0,80±0,03	0,71±0,02	0,79±0,01
	Critério 5	263	1,00±0,00	0,70±0,07	0,66±0,10
	Critério 6	295	1,00±0,00	0,08±0,13	0,00±0,01
	Critério 7	295	1,00±0,00	0,08±0,13	0,00±0,01
Altura da planta	**Exaustivo**	**175**	**0,65±0,07**	**0,81±0,05**	**1,00±0,17**
	Critério 1	213	0,47±0,01	0,78±0,01	1,17±0,01
	Critério 2	5	0,24±0,01	0,56±0,01	1,14±0,03
	Critério 3	44	0,35±0,01	0,72±0,01	1,20±0,02
	Critério 4	55	0,38±0,02	0,72±0,01	1,18±0,02
	Critério 5	263	0,94±0,06	0,71±0,06	0,74±0,07
	Critério 6	295	1,00±0,00	0,04±0,09	0,00±0,01
	Critério 7	295	1,00±0,00	0,04±0,09	0,00±0,01
Comprimento da panícula	**Exaustivo**	**157**	**0,52±0,08**	**0,69±0,06**	**0,92±0,18**
	Critério 1	140	0,42±0,02	0,68±0,03	1,04±0,05
	Critério 2	152	0,44±0,03	0,69±0,02	1,05±0,06
	Critério 3	44	0,38±0,02	0,66±0,03	1,06±0,06
	Critério 4	55	0,39±0,02	0,65±0,03	1,04±0,04
	Critério 5	263	0,94±0,06	0,51±0,12	0,56±0,19
	Critério 6	295	1,00±0,00	0,04±0,12	0,00±0,01

	Critério 7	295	1,00±0,00	0,04±0,12	0,00±0,01
Número de panículas no perfilho primário	**Exaustivo**	**123**	**0,46±0,08**	**0,64±0,08**	**0,90±0,25**
	Critério 1	154	0,78±0,06	0,51±0,03	0,57±0,04
	Critério 2	3	0,19±0,03	0,43±0,06	0,97±0,07
	Critério 3	44	0,52±0,05	0,54±0,04	0,75±0,08
	Critério 4	55	0,60±0,03	0,55±0,02	0,70±0,05
	Critério 5	263	0,74±0,26	0,51±0,13	0,64±0,26
	Critério 6	295	1,00±0,00	0,01±0,07	0,00±0,01
	Critério 7	295	1,00±0,00	0,01±0,07	0,00±0,01
Número de sementes por panícula	**Exaustivo**	**48**	**0,31±0,08**	**0,56±0,08**	**1,02±0,13**
	Critério 1	27	0,31±0,04	0,46±0,03	0,82±0,05
	Critério 2	22	0,28±0,04	0,47±0,02	0,89±0,07
	Critério 3	44	0,37±0,01	0,45±0,04	0,74±0,07
	Critério 4	55	0,37±0,02	0,46±0,05	0,75±0,08
	Critério 5	263	0,89±0,11	0,54±0,09	0,60±0,12
	Critério 6	295	1,00±0,00	0,01±0,17	0,00±0,01
	Critério 7	295	1,00±0,00	0,04±0,12	0,00±0,01
Espiguetas por panícula	**Exaustivo**	**52**	**0,42±0,09**	**0,66±0,08**	**1,03±0,13**
	Critério 1	140	0,69±0,06	0,50±0,05	0,72±0,07
	Critério 2	152	0,69±0,02	0,60±0,02	0,73±0,02
	Critério 3	44	0,31±0,03	0,56±0,02	1,00±0,05
	Critério 4	55	0,33±0,02	0,56±0,02	0,98±0,06
	Critério 5	263	0,73±0,23	0,50±0,08	0,60±0,13
	Critério 6	295	1,00±0,00	0,01±0,16	0,00±0,01
	Critério 7	295	1,00±0,00	0,01±0,16	0,00±0,01

Número de componentes independentes que conduz a: ICR a um maior valor de acurácia (Exaustivo); PCR a um maior valor de acurácia (Critério 1); PCR a um menor valor de viés (Critério 2); 80% da variação total de X explicada pelos componentes principais (Critério 3); 80% da variação total de Y explicada pelos componentes principais (Critério 4); 80% da variação total de X explicada pelos componentes independentes (Critério 5); Forward Selection (Critério 6); Backward Selection (Critério 7).

Tabela 5: Tempo computacional em segundos (horas) considerando os dados simulados e dados reais e cada critério de escolha do número de componentes independentes.

Natureza dos dados	Critérios	Tempo Computacional
Simulados	**Exaustivo**	**587776,39 (163,2712)**
	Critério 1	224,80 (0,0624)
	Critério 2	197,67 (0,0549)
	Critério 3	197,22 (0,0548)
	Critério 4	888,99 (0,2469)
	Critério 5	2235,96 (0,6211)
	Critério 6	1351,34 (0,3753)
	Critério 7	1350,24 (0,3750)
Reais	**Exaustivo**	**3189757,20 (886,04)**
	Critério 1	5945,9 (1,65)
	Critério 2	5029,09 (1,40)
	Critério 3	2179,76 (0,61)
	Critério 4	2524,28 (0,70)
	Critério 5	19827,66 (5,51)
	Critério 6	3260,44 (0,91)
	Critério 7	3262,01 (0,91)

Número de componentes independentes que conduz a: ICR a um maior valor de acurácia (Exaustivo); PCR a um maior valor de acurácia (Critério 1); PCR a um menor valor de viés (Critério 2); 80% da variação total de X explicada pelos componentes principais (Critério 3); 80% da variação total de Y explicada pelos componentes principais (Critério 4); 80% da variação total de X explicada pelos componentes independentes (Critério 5); Forward Selection (Critério 6); Backward Selection (Critério 7).

5. REFERÊNCIAS

AKINWALE, M. G., GREGORIO, G., NWILENE, F., AKINYELE, B. O., OGUNBAYO, S. A., ODIYI, A. C. Heritability and correlation coefficient analysis for yield and its components in rice (Oryza sativa L.). **African Journal of Plant Science**, v. 5, p. 207-212, 2011.

AMMIRAJU, J.S.S; LUO, M.; GOICOECHEA, J. L; WANG, W.; KUDRNA, D.; MUELLER, C., TALAG, J.; KIM, H.; SISNEROS N.B.; BLACKMON, B.; FANG, E.; TOMKINS, J.B.; BRAR, D.; MACKILL, D.; MACCOUCH, S.; KURATA, N.; LAMBERT, G.; GALBRAITH, D.W.; ARUMUGANATHAN, K.; RAO, K.; WALLING, J.G.; GILL, N.; YU, Y.; SANMIGUEL, P.; SODERLUND, C.; JACKSON, S.; WING, R.A. The Oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. **Genome Research**, 16:140-147, 2006.

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.

AZEVEDO, C. F.; RESENDE, M. D. V.; SILVA, F. F.; VIANA, J. M. S.; VALENTE, M. S. F., RESENDE JUNIOR, M. F. R.; MUÑOZ, P. Ridge, LASSO And Bayesian Additive-Dominance Genomic Models. **BMC Genetics**, v. 16, p. 105, 2015.

AZEVEDO, C. F.; SILVA, F. F.; RESENDE, M. D.; LOPES, M. S.; DUIJVESTEIJN, N.; GUIMARÃES, S. E. F.; LOPES, P. S.; KELLY, M. J.; VIANA, J. M. S.; KNOL, E. F. Supervised independent component analysis as an alternative

method for genomic selection in pigs. **Journal of Animal Breeding and Genetics**, v. 131, p. 452-461, 2014.

BINGHAM, E.; HYVÄRINEN, A. A fast fixed-point algorithm for independent component analysis of complex valued signals. **International journal of neural systems**, v. 10, p. 1-8, 2000.

BISNE, R; SARAWGI, A. K.; VERULKAR, S. B. Study of heritability, genetic advance and variability for yield contributing characters in rice. Bangladesh Journal of Agricultural Research, v. 34, p. 175-179, 2009.

BRITO, L. F., MCEWAN, J. C., MILLER, S. P., PICKERING, N. K., BAIN, W. E., DODDS, K. G., SCHENKEL, F. S., CLARKE, S. M.. Genetic diversity of a New Zealand multi-breed sheep population and composite breeds' history revealed by a high-density SNP chip**. BMC genetics**, v. 18, n. 1, p. 25, 2017.

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

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

FERREIRA, D. F. **Estatística Multivariada**. 2.ed.rev e amp, p. 411‑419, 2012.

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

GRENIER, C.; CAO, T. V.; OSPINA, Y.; QUINTERO, C.; CHÂTEL, M. H.; TOHME, J.; COURTOIS, B.; AHMADI, N. Accuracy of Genomic Selection in a Rice Synthetic Population Developed for Recurrent Selection Breeding. **PLOS One**, v. 10, p. e0136594, 2015.

HASSEN, M. B.; CAO, T. V.; BARTHOLOMÉ, J.; ORASEN, G.; COLOMBI, C.; RAKOTOMALALA, J.; RAZAFINIMPIASA, L.; BERTONE, C.; BISELLI, C.; VOLANTE, A.; DESIDERIO, F.; JACQUIN, L.; VALÈ, G.; AHMADI, N. Rice diversity panel provides accurate genomic predictions for complex traits in the progenies of biparental crosses involving members of the panel. Theoretical and Applied Genetics, **Theoretical and Applied Genetics**, v. 131, p. 417-435, 2018.

HELWIG, N. E.; HONG, S. A critique of tensor probabilistic independent component analysis: implications and recommendations for multi-subject fMRI data analysis. **Journal of neuroscience methods**, v. 213, p. 263-273, 2013.

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

HYVÄRINEN, A. Fast and robust fixed-point algorithms for independent component analysis. **IEEE transactions on Neural Networks**, v. 10, n. 3, p. 626-634, 1999.

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

JAMES, G.; WITTEN, D.; HASTIE, Trevor; TIBSHIRANI, R. **An Introduction to Statistical Learning**, p. 205-212, 2013.

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

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

LUMLEY, T. based on Fortran code by MILLER, A. **Leaps: Regression Subset Selection.** R package version 3.0. https://CRAN.R-project.org/package=leaps, 2017.

KUHN, M. Contributions from Jed WING, J.; WESTON, S.; WILLIAMS, A.; KEEFER, C.; ENGELHARDT, A.; COOPER, T.; MAYER, Z.; KENKEL, B.; the R Core Team, BENESTY, M.; LESCARBEAU, R.; SCRUCCA, A. Z. L.; TANG, Y.; CANDAN, C.; HUNT, T**. Caret: Classification and Regression Training.** R package version 6.0-77. https://CRAN.R-project.org/package=caret, 2017.

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

MEVIK, B. H; WEHRENS, R.; LILAND, K. H. **Pls: Partial Least Squares and Principal Component Regression.** R package version 2.6-0. https://CRAN.R-project.org/package=pls, 2016.

R Development Core Team. **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria, Available: http://www.R-project.org, 2016.

RESENDE, M. D. V.; SILVA, F. F; AZEVEDO, C. F. **Estatística Matemática, Biométrica e Computacional**: **Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência.** Visconde do Rio Branco: Suprema, v. 1, p. 881, 2014.

RESENDE, M. D. V.; 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, 2012. 291p. Disponível em: http://www.ppestbio.ufv.br/?page_id=448.

SEYOUM, M.; ALAMEREW, S.; BANTTE, K. Genetic variability, heritability, correlation coefficient and path analysis for yield and yield related traits in upland rice (Oryza sativa L.). **Journal of plant sciences**, v. 7, p. 13, 2012.

SPINDEL, J. E.; BEGUM, H.; AKDEMIR, D.; COLLARD, B.; REDOÑA, E.; JANNINK, J. L.; MCCOUCH, S. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. **Heredity,** v. 116, p. 395, 2016.

SPINDEL, J.; BEGUM, H.; AKDEMIR, D.; VIRK, P.; COLLARD, B.; REDOÑA, E. ATLIN, G.; JANNINK, J.L.; MCCOUCH, S.R. Genomic Selection and Association Mapping in Rice (Oryza sativa): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. **PLOS Genetics**, v. 11, p. e1004982, 2015.

WILKINSON, S., BISHOP, S. C., ALLEN, A. R., MCBRIDE, S. H., SKUCE, R. A., BERMINGHAM, M., WOOLLIAMS, J. A., GLASS, E. J. Fine-mapping host genetic variation underlying outcomes to Mycobacterium bovis infection in dairy cows. **BMC genomics**, v. 18, p. 477, 2017.

ZHAO, K., TUNG, C. W., EIZENGA, G. C., WRIGHT, M. H., ALI, M. L., PRICE, A. H., NORTON, J. G., ISLAM, A.R., REYNOLDS, A., MEZEY, J., MCCLUNG, A. M., BUSTAMANTE, C. D., MCCLUNG, A. M. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. **Nature communications**, v. 2, p. 467, 2011.

CAPÍTULO 3

EFICIÊNCIA NA PREDIÇÃO GENÔMICA EM MODELOS ADITIVO-DOMINANTE

Resumo

A Seleção Genômica Ampla (Genome Wide Selection – GWS) utiliza uma alta densidade de marcadores SNPs (Single Nucleotide Polymorphisms) que cobrem amplamente o genoma a fim de estimar o efeito dos marcadores no fenótipo dos indivíduos. Dessa forma, viabiliza a estimação do mérito genético dos indivíduos e otimiza o processo de identificação dos indivíduos geneticamente superiores. No entanto, devido ao elevado custo no processo de genotipagem por amostra e a alta correlação entre os marcadores, a GWS enfrenta alguns desafios na estimação dos efeitos dos marcadores, dentre eles: a alta dimensionalidade e a multicolinearidade. Para resolver tais problemas, a GWS busca metodologias baseadas em redução de dimensionalidade tais como a Regressão via Componentes Principais (PCR), a Regressão via Componentes Independentes (ICR) e os Quadrados Mínimos Parciais (PLS) e baseadas em regularização como o BLUP genômico (G-BLUP). Os métodos de redução de dimensionalidade são fundamentados em variáveis latentes (componentes) e se destacam por apresentar teoria simples e de fácil aplicação, porém até o momento não foram aplicados a GWS considerando o modelo aditivo-dominante. A escolha do número de componentes a serem incluídos no modelo pode ser feita de maneira exaustiva, variando o número de componentes e escolhendo aquele número que está associado a uma maior acurácia. No entanto, sob o contexto aditivo-dominante, tem-se as acurácias associadas ao valor genômico aditivo, valor genômico devido à dominância e ao valor genômico total. Além disso, a ICR enfrenta desafios na determinação do número ótimo de componentes uma vez que o procedimento exaustivo exige elevada demanda

computacional. Diante do exposto, a relevância do presente trabalho está em identificar qual o melhor critério para determinar o número ótimo de componentes independentes a serem utilizados no modelo aditivo-dominante. Os critérios definidos foram o número de componentes independentes igual ao número de componentes que conduz a PCR a um maior valor de acurácia (critério 1), a PCR a um menor viés (critério 2) e a PCR a 80% da explicação de X (critério 3). Também objetiva-se comparar os métodos ICR, PCR e PLS sob as três informações genômicas (aditiva, dominante e total) além de compara-los ao G-BLUP. As comparações serão realizadas por meio de populações simuladas compostas por 1.000 indivíduos de 20 famílias de irmãos completos e 2.000 marcadores SNPs. Os fenótipos estão associados a dois níveis de herdabilidade e duas arquiteturas genéticas, constituindo assim quatro cenários que assumiram dominância completa. Os resultados mostraram que para todos os efeitos, o critério no qual determina o número de componentes independentes igual o número de componentes que conduz a PCR a um maior valor de acurácia, produziu estimativas mais acuradas. Para a estimação simultânea dos efeitos de marcadores aditivos e devido a dominância, a melhor alternativa é a escolha do número de componentes que conduz o valor genômico devido a dominância a uma maior acurácia. De modo geral, a PCR mostrou-se mais eficiente que os demais métodos e nenhuma das metodologias conseguiram capturar de forma eficaz as herdabilidades simuladas e apresentaram estimativas viesadas.

Palavras-chave: Predição Genômica, análise de componentes independentes, modelos aditivo-dominante, métodos de redução dimensional, G-BLUP.

1. INTRODUÇÃO

A genética molecular tem contribuído para grandes avanços no melhoramento animal e vegetal ao utilizar as informações provenientes diretamente do DNA, o que permite identificar com maior rapidez os indivíduos que são geneticamente superiores. Nesta perspectiva, Meuwissen et al. (2001) idealizaram a seleção genômica ampla (Genome Wide Selection – GWS) cujo enfoque se baseia na utilização de milhares de marcadores SNPs (Single Nucleotide Polymorphisms) amplamente distribuídos no genoma. O objetivo principal da GWS é estimar o efeito desses marcadores sobre o fenótipo dos indivíduos, uma vez que se supõe a existência de desiquilíbrio de ligação entre marcadores SNPs e os loci de características quantitativas (Quantitative Trait loci – QTL), e posteriormente estimar os valores genéticos genômicos (GEBVs) dos indivíduos.

No entanto, o processo de estimação dos efeitos de marcadores na GWS enfrenta alguns desafios estatísticos, tais como a multicolinearidade (marcadores altamente correlacionados) e a alta dimensionalidade, uma vez que o número de marcadores é muito maior que o número de observações individuais, devido ao alto custo de genotipagem por amostra. Tais desafios impossibilitam a obtenção de estimativas por meio do método dos quadrados mínimos ordinários, sendo necessário outras metodologias estatísticas que contemplem estas adversidades.

Dentre as principais metodologias estatísticas aplicadas a GWS, existem as metodologias baseadas em modelos mistos como BLUP genômico (G-BLUP- Genomic Best Linear Unbiased Predictor), as metodologias bayesianas como o LASSO bayesiano (BLASSO) e existem metodologias baseadas em redução dimensional. Na última classe destacam-se a Regressão via Componentes Principais (PCR), o Quadrados Mínimos

Parciais (PLS) e a Regressão via Componentes Independentes (ICR), e estes assumem que os efeitos dos marcadores são efeitos fixos (Resende et al., 2012).

Os métodos de redução de dimensionalidade se destacam pela simplicidade de sua aplicação em relação as outras metodologias e podem ser amplamente utilizados para predizer os valores genômicos dos indivíduos. Além disso, dentre esses, a ICR tem se mostrado eficiente para estimar os valores genômicos aditivos dos indivíduos conforme Azevedo et al. (2013), no entanto o método exaustivo para modelos superparametrizados, como ocorre com a seleção genômica devido à alta densidade do número de marcadores, exigem grande demanda computacional para a execução das análises.

Ademais, os estudos de predição genômica devem apresentar uma abordagem completa usando modelos genéticos adequados, incluindo efeitos aditivos e efeitos devido à dominância, que são essenciais para a seleção de cruzamentos e clones e trazem melhorias para as estimativas de efeitos aditivos para a seleção (Denis e Bouvet, 2013; Wellmann e Bennewitz, 2012; Toro e Varona, 2010). Apesar de nos primeiros estudos de GWS a dominância ter sido negligenciada, devido ao fato dos efeitos não ser de característica herdável e além disso, contribuir para a duplicação do número de parâmetros a serem estimados, pesquisas vêm sendo conduzidas a respeito da relevância dos modelos aditivo-dominante na predição genômica (Hill et al., 2008; Bennewitz e Meuwissen, 2010 e Wellmann e Bennewitz, 2012). Os modelos aditivo-dominante já foram empregados no BLUP genômico (Wang e Da, 2014; Muñoz et al., 2014; Su et al., 2012), na regressão ridge (RR-BLUP - Goddard et al., 2009), nos métodos bayesianos (Toro e Varona, 2010; Zeng et al., 2013; Azevedo et al., 2015b; Almeida Filho et al., 2016). No entanto, os modelos aditivo-dominante ainda não foram empregados nas metodologias fundamentadas em redução dimensional.

Diante do exposto, a relevância do presente trabalho está em propor a aplicação dos métodos de redução de dimensionalidade em modelos aditivo-dominante e avaliar neste contexto os três principais critérios propostos por Costa (2018) para escolher o número ótimo de componentes independentes. Em seguida, avaliar qual o melhor procedimento de escolha do número de componentes dos métodos de redução de dimensionalidade em modelos aditivos-dominante, seja com base nos valores genômicos aditivos, devido a dominância ou nos valores totais. Ademais, a eficiência na predição dos valores genômicos (aditivo, devido à dominância e total) dos mesmos será comparada ao tradicional aplicado a GWS, o BLUP genômico.

2. MATERIAIS E MÉTODOS

2.1.Dados Simulados

A constituição do conjunto de dados simulados foi descrita por Azevedo et al. (2015b). Um total de 2.000 marcadores SNPs equidistantes separados por 0,1 centiMorgan (cM) entre os dez cromossomos foram simulados. Os QTLs foram distribuídos nas regiões abrangidas pelo SNPs. Os indivíduos fenotipados e genotipados totalizam 1.000 indivíduos de 20 famílias de irmãos completos.

Foram simuladas características com duas arquiteturas genéticas, uma seguindo um modelo infinitesimal (locos não ligados, com efeitos iguais – herança poligênica) e outra com cinco genes de efeitos maiores, responsável por 50% da variabilidade genética (herança mista). Na primeira arquitetura genética, efeitos de pequena magnitude no fenótipo foram assumidos para cada um dos 100 QTLs e na segunda arquitertura genética grandes efeitos foram designados para 5 QTLs (representam 50% da variabilidade genética total) e pequenos efeitos foram designados para os demais 95 QTLs.

Os efeitos aditivos e de dominância foram normalmente distribuídos com média zero e variância genética permitindo o nível de herdabilidade desejado. As simulações

assumiram independência entre efeitos aditivos e devido à dominância, com efeitos de dominância tendo a mesma distribuição que os efeitos aditivos. Visando obter o valor fenotípico, foi adicionado ao valor genético um efeito ambiental proveniente de uma distribuição normal $N(0, {\sigma^2}_e)$, em que a variância ${\sigma^2}_e$ foi definida de acordo com dois níveis de herdabilidade no sentido restrito (0,20 e 0,30, respectivamente) e com dois níveis de herdabilidade no sentido amplo (0,30 e 0,50, respectivamente).

Os níveis de herdabilidade foram escolhidos para representar uma característica com baixa herdabilidade e outra com herdabilidade moderada, casos em que se espera que a seleção genômica seja superior à seleção fenotípica (Azevedo et al., 2015b). As magnitudes das herdabilidades no sentido restrito e no sentido amplo estão associadas com um grau médio de nível de dominância (d/a) de aproximadamente 1 (domínio completo) em uma população com frequências alélicas intermediárias.

2.2.Cenários

As análises abrangeram quatro cenários diferentes sendo estes constituídos por: dois níveis de herdabilidade no sentido amplo (0,30 e 0,50) × duas arquiteturas genéticas (herança poligênica e mista). A descrição destes quatro cenários pode ser vista na Tabela 1 e cada tipo de cenário foi simulado dez vezes.

Tabela 1 – Cenários com as respectivas médias das herdabilidades aditivas (h_a^2), devido a dominância (h_d^2) e total (h_g^2), arquiteturas genéticas (características controladas por genes de pequeno efeito – herança poligênica e características controladas por genes de pequeno e maior efeito - herança mista).

Cenário	Arquitetura Genética	h_a^2	h_d^2	h_g^2
Cenário 1	Herança poligênica	0,20	0,10	0,30
Cenário 2	Herança poligênica	0,30	0,20	0,50
Cenário 3	Herança mista	0,20	0,10	0,30
Cenário 4	Herança mista	0,30	0,20	0,50

2.3.Modelo a nível de marcadores

Considere o seguinte modelo linear:

$$y = 1\mu + Wm_a + Sm_d + e$$

em que:

y é o vetor de observações fenotípicas com dimensão I × 1, sendo I o número de indivíduos genotipados e fenotipados;

μ é a média geral da característica e 1 é um vetor de mesma dimensão de y em que todos os elementos são iguais a um;

W é a matriz de incidência com valores 0, 1 e 2 para o número de alelos do marcador com dimensão I × J, sendo J o número de marcadores e m_a são efeitos aditivos dos marcadores;

S é a matriz de incidência com valores 0,1 e 0 com dimensão I × J e m_d são efeitos devido à dominância dos marcadores;

e é o vetor de erros aleatórios com estrutura de variância dada por $e \sim N(0, I\sigma_e^2)$ sendo I a matriz identidade e σ_e^2 a variância residual.

Considere a justaposição das matrizes W e S definindo a matriz X da seguinte forma $X = [W|S]$ e os efeitos de marcadores como sendo $m = [m_a|m_d]'$.

2.4.Regressão via Componentes Principais

O método regressão via componentes principais (Principal Components Regression -PCR) idealizado por Kendall (1957) e Hotelling (1957), consiste na decomposição da matriz X a fim de maximizar a variância dos componentes. Os componentes principais são dados por:

$$Z = XP \quad (1)$$

em que Z é a matriz com os n_{PCR} primeiros componentes principais e P é a matriz com os n_{PCR} primeiros autovetores da matriz de variância de X.

Sob o contexto de regressão, o número de componentes principais deve ser definido como $1 \leq n_{PCR} \leq \min(I, 2J) - 1$. Os componentes principais Z_m ($m = 1, \ldots, n_{PCR}$) são caracterizados como sendo componentes ortogonais, visto que não são correlacionados, isto é, $Cor\ (Z_m\ , Z_{m'}) = 0$ para todo $m \neq m'$. Posteriormente, realiza-se uma regressão linear múltipla entre a variável Y e os componentes Z, obtendo a seguinte predição:

$$\hat{y} = Z\hat{\alpha} \tag{2}$$

em que $\hat{\alpha}_m$ ($m = 1, \ldots, n_{PCR}$) são as estimativas dos coeficientes da regressão entre Y e Z obtidos pelo método dos quadrados mínimos ordinários (Ordinary Least Squares – OLS).

Visto que estes coeficientes $\hat{\alpha}_m$ não estão associados as variáveis originais, é necessário combinar as equações (1) e (2) para se obter as estimativas dos efeitos associados as variáveis originais (marcadores) e que são dadas por:

$$\widehat{m}_{PCR} = P\hat{\alpha}. \tag{3}$$

2.5.Quadrados Mínimos Parciais

O método regressão via quadrados mínimos parciais (Partial Least Squares- PLS) proposto por Wold (1975) decompõe a matriz X e o vetor y com o objetivo de maximizar a covariância entre os componentes e a variável Y. Para determinar o primeiro componente T_1, as variáveis Y e $X_j's$ são centradas na média, definindo as variáveis U_1 e V_{1j}, como a seguir:

$$u_1 = y - \bar{y} \quad \text{e} \quad v_{1(j)} = x_j - \overline{x_j}, \tag{4}$$

para $j = 1, \ldots, 2J$. Define-se a variável S_1 como sendo $s_1 = V_1{}'u_1$ ($2J \times 1$) em que $V_1 = [v_{11} \quad v_{12} \quad \cdots \quad v_{12J}]$, e aplica a decomposição em valores singulares (Singular value decomposition - SVD) no vetor s_1 conforme a seguir:

$$s_1 = L_1 k q_1' \quad (5)$$

em que L_1 é uma matriz unitária $(2J \times 2J)$ com o primeiro vetor coluna igual a $\frac{s_1}{||s_1||}$ (vetor s_1 normalizado), k_1 é um vetor $(2J \times 1)$ com o primeiro valor igual a $||s_1||$ (norma do vetor s_1) e q_1 é um escalar igual a 1. Os componentes T_1 e Z_1 são, então, definidos por:

$$t_1 = V_1 L_1 \quad (6)$$

$$z_1 = u_1 q_1 \frac{Var(t_1)}{Cov(t_1, u_1)}. \quad (7)$$

No entanto, nem todas as informações contidas nas variáveis X_j $(j = 1,2,\ldots,2J)$ e na variável Y estão contidas no componente T_1, definido acima. Logo, a informação ausente em T_1 pode ser estimada por meio dos resíduos da regressão entre as variáveis X_j e T_1 ou, equivalentemente, da regressão entre as variáveis latentes V_{1j} e T_1, visto que os resíduos de ambas são idênticos (Garthwaite, 1994). Do mesmo modo, a variabilidade de Y que não está sendo explicada por $\hat{T}_1$ pode ser estimada por meio dos resíduos da regressão ente U_1 e $\hat{T}_1$. Dessa forma, são definidas as variáveis U_2 e $V_{2(j)}$, respectivamente,

$$\hat{v}_{2(j)} = v_{1(j)} - t_1 \hat{r}_1'$$

$$\hat{u}_2 = u_1 - t_1 \hat{p}_1'.$$

ou seja, u_2 e v_{2j} são os resíduos, r_1 e p_1 são os coeficientes obtidos da regressão entre u_1e t_1 e v_{1j} e t_1, nesta ordem. Define-se uma nova variável S_2, como sendo $s_2 = V_2{}' u_2$ e aplica-se novamente a decomposição em valores singulares, como a seguir:

$$s_2 = L_2 k_2 q_2' \quad (8)$$

em que L_2 é uma matriz unitária $(2J \times 2J)$ com o primeiro vetor coluna igual a $\frac{s_2}{||s_2||}$ (vetor s_2 normalizado), k_2 é um vetor $(2J \times 1)$ com o primeiro valor igual a $||s_2||$ (norma do vetor s_2) e q_2 é um escalar igual a 1. Os componentes T_2 e Z_2 são, então, definidos por:

$$t_2 = \hat{V}_2 L_2 \tag{9}$$

$$z_2 = \hat{u}_2 q_2 \frac{Var(t_2)}{Cov(t_2, \hat{u}_2)}. \tag{10}$$

Os componentes $t_3,\ldots,t_{n_{PLS}}$ $(1 \leq n_{PLS} \leq \min(I, 2J) - 1)$ são determinados sucessivamente e de modo análogo aos anteriores. Além disso, pode-se garantir que todos os componentes são ortogonais (Garthwaite, 1994). Em seguida, por meio da regressão linear múltipla entre os componentes associados a X e a variável Y, obtêm-se a equação de predição a seguir:

$$\hat{y} = T\widehat{\beta}, \tag{11}$$

em que $\hat{\beta}$ $(\hat{\beta}_m's, m = 1,2, \ldots, n_{PLS})$ é o vetor das estimativas dos coeficientes da regressão estimados via OLS. Os coeficientes $\hat{\beta}_m's$ não têm uma interpretação biológica e as estimativas dos coeficientes originais não são obtidas trivialmente, uma vez que as colunas da matriz V $(V_{1(j)}, V_{2(j)}, \ldots, V_{2J(j)})$ não são comparadas diretamente com X como na PCR, pois são deflacionadas sucessivamente. No entanto, segundo Wold (1975) os coeficientes associados as variáveis originais podem ser estimados via:

$$\widehat{m}_{PLS} = L(R'L)^{-1}\hat{\beta}, \tag{12}$$

em que L é a matriz cujas colunas são $L_1, L_2, \ldots, L_{n_{PLS}}$, denominada matriz de carregamento de X, e R é a matriz cujas colunas contém os coeficientes $r_1, r_2, \ldots, r_{n_{PLS}}$.

2.6.Regressão via Componentes Independentes

A regressão via Componentes Independentes (Independent Component Regression - ICR) foi idealizado por Jutten e Hérault (1991) e Comon (1994) e proposta sob o contexto da GWS em modelos aditivos por Azevedo et al. (2013). A ICR apresenta o propósito de maximizar a independência dos componentes decompondo a matriz de dados X (cujos valores são centrados na média) da seguinte forma:

$$X = SA' \quad (13)$$

em que S ($I \times min(2J, I) - 1$) é a matriz de componentes independentes e A ($2J \times min(2J, I) - 1$) é denominada matriz de misturas. A matriz A, geralmente desconhecida, é uma função de duas matrizes K e R, a primeira obtida por meio do processo de branqueamento (whitening) e a segunda garante a independência entre os componentes.

A matriz K ($2J \times min(2J, I) - 1$) é definida como $P_r \Lambda_r^{-\frac{1}{2}}$, sendo P_r a matriz com as $min(2J, I) - 1$ primeiras colunas da matriz de autovetores da matriz de covariância de X ($min(2J, I) - 1$ primeiros autovetores) e Λ_r uma matriz com as $min(2J, I) - 1$ primeiras linhas e colunas da matriz de autovalores associados a esses primeiros autovetores. A matriz R pode ser obtida por meio de um algoritmo desenvolvido por Hyvärinen (1998) baseado no princípio da máxima entropia (J(R)) cuja aproximação pode ser dada por:

$$J(R) \propto \left\{E[G(R)] - E\left[G\left(R_{gaussiana}\right)\right]\right\}^2.$$

em que G é uma função quadrática e a escolha da função G influencia na aproximação da negentropia (Hyvarinen, 1999), assim, as funções G mais empregadas são:

$$G_1(r) = \frac{1}{a}\log\cosh(ar) \quad \text{e} \quad G_2(r) = -exp\left(-\frac{r^2}{2}\right)$$

em que a é uma constante ($1 \leq a \leq 2$) e deve-se escolher uma função G que não cresça rapidamente, o que conduz a estimadores mais robustos, computacionalmente mais simples e com propriedades estatísticas mais interessantes.

Após a convergência deste algoritmo iterativo, obtém-se a matriz R que maximiza a independência das colunas de S. Os componentes independentes são definidos como:

$$S = XKR. \quad (14)$$

Assim, a equação de predição que relaciona a variável Y e os componentes independentes S é expressa por:

$$\hat{y} = S\hat{\gamma} \quad (15)$$

em que as estimativas dos coeficientes $\hat{\gamma}_m$ ($m = 0,1,\ldots,n_{ICR}$ sendo $1 \leq n_{ICR} \leq min(2J, I) - 1$) da regressão são estimados via método dos quadrados mínimos ordinários.

Dessa forma, as estimativas dos efeitos dos marcadores ($\hat{m}_{ICR}$), ou seja, as estimativas dos coeficientes relacionadas as variáveis originais, são dados pela equação a seguir:

$$\hat{m}_{ICR} = KR\,\hat{\gamma}. \quad (16)$$

2.7. Método G-BLUP

O método G-BLUP (Genomic Best Linear Unbiased Predictor) é baseado em um modelo a nível de indivíduos dado por:

$$y = 1\mu + Za + Zd + e,$$

em que:

y é o vetor de observações fenotípicas com dimensão $I \times 1$, sendo I o número de indivíduos genotipados e fenotipados;

μ é a média geral da característica e 1 é o vetor com dimensão $I \times 1$ cujos seus elementos são iguais a 1;

a é o vetor de efeitos genômicos aditivos dos indivíduos ($I \times 1$) com matriz de incidência Z ($I \times I$) referente aos indivíduos, sendo a estrutura de variância dada por $a \sim N(0, G_a\sigma_a^2)$ em que σ_a^2 é a variância aditiva e G_a ($I \times I$) é a matriz de parentesco genômica para os efeitos aditivos;

d é o vetor de efeitos genômicos devido à dominância dos indivíduos com matriz de incidência Z (I × I) referente aos indivíduos, sendo a estrutura de variância dada por

$d \sim N(0, G_d\sigma_d^2)$ em que σ_d^2 é a variância devido à dominância e G_d (I × I) é a matriz de parentesco genômica para os efeitos devido à dominância;

e é o vetor de efeitos residuais aleatórios com $e \sim N(0, I\sigma_e^2)$ e σ_e^2 é a variância residual.

Os valores genômicos aditivos e devido à dominância podem ser estimados via equações de modelo misto dadas por:

$$\begin{bmatrix} 1'1 & 1'Z & 1'Z \\ Z'1 & Z'Z + G_a^{-1}\frac{\sigma_e^2}{\sigma_a^2} & Z'Z \\ Z'1 & Z'Z & Z'Z + G_d^{-1}\frac{\sigma_e^2}{\sigma_d^2} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a} \\ \hat{d} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z'y \end{bmatrix},$$

em que σ_e^2, σ_a^2 e σ_d^2 correspondem a variância residual, aditiva e devido à dominância respectivamente, e obtidas via REML (Restricted Maximum Likelihood), $G_a = \frac{WW'}{\sum_{j=1}^{J}(2p_jq_j)}$ (matriz de parentesco genômico dos efeitos aditivos) e $G_d = \frac{SS'}{\sum_{j=1}^{J}(2p_jq_j)^2}$ (matriz de parentesco genômico dos efeitos de dominância) sendo p_j e q_j as frequências alélicas do j-ésimo locos conforme Vitezica et al. (2013).

2.8.Escolha do número ótimo de componentes

A aplicação dos métodos de redução de dimensionalidade requer a escolha do número ótimo de componentes a serem utilizados no modelo, a fim de que os componentes extraídos expliquem satisfatoriamente a variabilidade dos dados. Azevedo et. al (2014, 2015a) utilizam o critério exaustivo para a escolha do número de componentes principais, parciais e independentes no modelo aditivo. Estes autores variaram o número de componentes de 1 ao número máximo de componentes construídos $min(2J, I) - 1$ e adotaram aquele número que está associado ao maior valor de acurácia na predição dos valores genômicos aditivos. Sob o contexto aditivo dominante, o

procedimento exaustivo será aplicado a ICR, PCR e PLS sob quantidades referentes a acurácia aditiva, devido à dominância e total.

No entanto, para a ICR este procedimento se torna inviável quando o banco de dados é composto por um elevado número de indivíduos e de marcadores e principalmente, sob o contexto de modelo aditivo-dominante. O estudo desenvolvido por Costa (2018) consistiu na proposição de sete critérios para determinação do número ótimo de componentes independentes em modelos aditivos, porém apenas três deles apresentaram resultados satisfatórios. Os três critérios de escolha do número de componentes também serão avaliados sob as três quantidades referente aos valores genômicos, aditivos (a), devido a dominância (d) e total (g). Os critérios que mais se destacaram foram:

i) **Critério 1**: Para cada número de componentes principais ($m = 1, \dots, min(I, 2J) - 1$), são estimados os efeitos de marcadores na população de estimação via PCR e estes são utilizados na população de validação para estimar os valores genômicos aditivos dos indivíduos desta população. Em seguida, calcula-se para os dados simulados as acurácias ($r_{a\hat{a}}, r_{d\hat{d}}, r_{g\hat{g}}$), correlação entre o valor genômico estimado e o valor genômico real. Escolhe-se o número de componentes principais cujo valor genômico aditivo conduz a uma maior acurácia sob as três quantidades. Dessa forma, o presente critério determina que o número de componentes independentes deve estar associado a esse número de componentes principais.

ii) **Critério 2**: Para cada número de componentes principais ($m = 1, \dots, min(I, 2J) - 1$), são estimados os efeitos de marcadores na população de estimação via PCR e estes são utilizados na população de validação para estimar os valores genômicos aditivos dos indivíduos desta população. Em seguida,

calcula-se os coeficientes de regressão ($b_{y\hat{a}}, b_{y\hat{d}}, b_{y\hat{g}}$), coeficiente obtido da regressão entre o fenótipo e o valor genômico aditivo estimado e posteriormente os vieses de predição dado por ($1 - b_{y\hat{a}}$,$1 - b_{y\hat{d}}$, $1 - b_{y\hat{g}}$). Escolhe-se o número de componentes principais cujo valor genômico conduz a um menor viés de predição. Dessa forma, o presente critério determina que o número de componentes independentes deve estar associado a esse número de componentes principais.

iii) **Critério 3**: A porcentagem de explicação da variação total de X ao utilizar m componentes principais é dada por: $p_m(\%) = \frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{2J} \sigma_j^2}$ em que λ_j é o autovalor correspondente ao j-ésimo autovetor da matriz de covariância de X e σ_j^2 é a variância da j-ésima variável X. O presente critério determina que o número de componentes independentes a ser utilizado no modelo deve ser igual ao número de componentes principais que expliquem 80% da variação total de X.

2.9.Comparação das metodologias de seleção genômica ampla

Os métodos de redução de dimensionalidade e o G-BLUP serão comparados por meio de uma validação independente em que as nove primeiras replicatas serão assumidas como populações de estimação e utilizadas para estimar o valor genético genômico dos indivíduos e a décima replicata será assumida como população de validação para calcular as medidas de eficiência descritas na Tabela 2.

As medidas de eficiência são interpretadas da seguinte forma: i) o método é considerado superior quando apresenta maior valor de acurácia ($r_{a\hat{a}}$ e $r_{d\hat{a}}$),; ii) o viés de predição ($1 - b_{y\hat{a}}$ e $1 - b_{y\hat{d}}$) próximo ou igual 0, ou equivalentemente $b_{y\hat{a}} = 1$ (ou $b_{y\hat{d}} = 1$), os valores genômicos aditivos (ou devido a dominância) são não viesados; iii)

a herdabilidade mais próxima da simulada. Após a obtenção das medidas de acurácia, viés e herdabilidade para cada replicata em cada cenário, será obtida a média desses valores.

Tabela 2. Expressões para o cálculo das medidas de eficiência: acurácias ($r_{\hat{a}a}$ *e* $r_{\hat{d}d}$), coeficientes de regressão ($b_{y\hat{a}}$ *e* $b_{y\hat{d}}$), herdabilidades moleculares (h^2_{aM} *e* h^2_{dM}) e eficiências relativas (ER_a *e* ER_d).

Medida de Eficiência	**Expressões**
Acurácia	$r_{\hat{a}a} = Cor(a, \hat{a})$
	$r_{\hat{d}d} = Cor(d, \hat{d})$
Coeficiente de Regressão	$b_{y\hat{a}} = \frac{Cov(y, \hat{a})}{Var(\hat{a})}$
	$b_{y\hat{d}} = \frac{Cov(y, \hat{d})}{Var(\hat{d})}$
Herdabilidade	$h^2_{aM} = \frac{\sigma^2_{a_M}}{\sigma^2_{a_M} + \sigma^2_{d_M} + \sigma^2_e}$
	$h^2_{dM} = \frac{\sigma^2_{d_M}}{\sigma^2_{a_M} + \sigma^2_{d_M} + \sigma^2_e}$
Eficiência Relativa	$\mathrm{ER}_a = \frac{r_{\hat{a}a\ MR}}{r_{\hat{a}a\ G-BLUP}}$
	$\mathrm{ER}_d = \frac{r_{\hat{d}d\ MR}}{r_{\hat{d}d\ G-BLUP}}$

$\sigma^2_{a_M} = \sum_{j=1}^{J} 2p_j q_j m^2_{aj}$ é a variância genômica aditiva e $\sigma^2_{d_M} = \sum_{j=1}^{J} (2p_j q_j)^2 m^2_{dj}$ é a variância genômica devido a dominância, p_j e q_j são as frequências alélicas do j-ésimo marcador, $r_{a\hat{a}\ MR}$ e $r_{\widehat{a}a\ G-BLUP}$, $r_{d\hat{d}\ MR}$ e $r_{d\hat{d}\ G-BLUP}$ são as acurácias aditivas e devido a dominância dos métodos de redução de dimensionalidade e do G-BLUP, respectivamente.

2.10. Recursos Computacionais

As configurações do computador utilizado nas análises estatísticas foram: processador Intel(R) Core(TM) i7-6500 (CPU 2.50 GHz) com 16 Gb de memória RAM. Todas as implementações dos métodos utilizados foram realizadas no software R (R Development Core Team, 2018) disponível em (http://cran.r-project.org). Na Tabela 3 é apresentado um resumo dos pacotes e das funções usadas na implementação de cada um dos métodos propostos.

Tabela 3. Métodos propostos e respectivas ferramentas de implementação no *software* R.

Metodologia	Pacote	Função	Referência
PCR	*pls*	*pcr*	(Mevik et al., 2016)
PLS	*pls*	*pls*	(Mevik et al., 2016)
ICR	*caret*	*icr*	(Kuhn et al., 2017)
G-BLUP	*sommer*	*mmer*	(Covarrubias-Pazaran, 2016)

3. RESULTADOS E DISCUSSÃO

Os resultados médios e os desvios das simulações dos valores genômicos aditivos, devido à dominância e total referentes ao número de componentes, herdabilidade molecular, acurácia, viés e eficiência relativa (aditiva e devido a dominância) considerando a ICR (critério exaustivo, 1, 2 e 3), PCR (critério exaustivo), PLS (critério exaustivo) e G-BLUP são apresentados nas Tabelas 4, 5, 6 e 7 correspondentes aos cenários 1, 2, 3 e 4, respectivamente.

i) Critério de escolha do número de componentes independentes

As acurácias aditivas e devido a dominância via ICR, considerando o valor genômico aditivo, são maiores no critério 1, para todos os cenários, sendo que em todos os cenários a acurácia aditiva é muito próxima da acurácia máxima atingida pelo método

exaustivo, o que ocorre também com a acurácia devido à dominância nos cenários 2 e 4. Em relação a acurácia, quando se considera apenas o valor genômico devido à dominância, os critérios 1 e 2 apresentam acurácias aditivas similares nos cenários 1, 2, e 4, sendo que no cenário 3 as acurácias aditivas obtidas pelos critérios 1 e 2 são superiores ao exaustivo. Em todos os cenários, a acurácia aditiva obtida pelo critério 1 é superior ou igual ao valor de acurácia aditiva obtida pelo critério 2, considerando apenas os valores genômicos aditivos ou apenas os valores devido à dominância. No que se refere a acurácia devido à dominância, temos que os valores obtidos pelos critérios 1 e 2 são similares nos cenários 1 e 4 e próximas dos valores obtidos pelo método exaustivo. No cenário 3, o critério 1 mostra-se mais acurado na estimação dos efeitos devido à dominância e próximo do método exaustivo. No cenário 2, o critério 2 mostrou-se mais eficiente, no entanto, a acurácia obtida pelo critério 1 é mais próxima do valor obtido quando se considera apenas o valor genômico aditivo. Considerando o valor genômico total, observa-se que as acurácias aditivas e devido a dominância obtidas via os critérios 1 e 3 são similares, sendo que todos os valores apresentados são superiores as acurácias obtidas via critério 2 e também são mais próximos dos valores alcançados pelo método exaustivo. Além disso, as acurácias aditivas e devido à dominância são próximas aos valores obtidos quando considerado o valor genômico devido à dominância. Desta forma, conclui-se que o critério 1 está associado a uma maior acurácia aditiva e a uma maior acurácia devido à dominância.

Nos cenários 1, 2 e 3, os critérios 1 e 2, considerando o valor genômico aditivo, subestimaram consideravelmente os valores aditivos e devido à dominância, enquanto que no cenário 4, o critério 1 superestimou o valor genômico aditivo e o critério 2 subestimou o valor genômico devido a dominância. Vale ressaltar que em todos os cenários, o método exaustivo também subestima os valores genômicos, exceto para os

valores aditivos no cenário 4. Os critérios 1 e 2, considerando o valor genômico devido a dominância, superestimaram o valor genômico aditivo nos cenários 3 e 4. Já nos cenários 1 e 2, o critério 1 subestimou os GEBVs aditivos. Apesar do critério 2 apresentar viés aditivo mais próximo do valor obtido pelo método exaustivo, observa-se que o critério 1 apresentou valores mais próximos de viés aditivo quando se considera apenas o valor genômico aditivo como critério de escolha em todos os cenários. Já em relação ao viés devido a dominância, verifica-se que em todos os cenários, os valores genômicos devido a dominância foram subestimados pelos critérios 1 e 2. No entanto, o viés obtido via critério 2 é menor que o proveniente do critério 1, sendo que estes valores foram similares aos valores obtidos pelo método exaustivo nos cenários 2 e 4. Da mesma forma, observa-se em todos cenários que o critério 1 obteve vieses devido a dominância mais próximos das análises que consideram apenas o valor genômico aditivo. Em relação ao viés considerando o valor genético total, percebe-se que o critério 1 obteve valores mais próximos do método exaustivo, subestimando os valores genômicos aditivos, exceto nos cenários 3 e 4, e os valores devido a dominância. Além disso, exceto no cenário 3 para o viés devido a dominância, observa-se que os vieses são similares aos valores encontrados quando se considera apenas o valor genômico devido a dominância. Dentre todos os critérios apresentados, o critério 2 foi o que mais subestimou os valores genômicos em todos os cenários, sendo que estes valores foram também os mais distantes do método exaustivo.

Considerando o método ICR e avaliando os critérios 1 e 2, nas análises referentes apenas ao critério de escolha associado aos valores genômicos aditivos, mostram que o critério 1 nos cenários 1 e 3, ao contrário dos cenários 2 e 4, apresentaram valores de herdabilidade aditiva mais distantes do valor paramétrico, no entanto, em todos os quatro cenários estes valores são próximos da herdabilidade aditiva alcançada pelo método

exaustivo. Em relação a herdabilidade devido a dominância, tem-se que em todos os cenários, os valores obtidos pelo critério 1 também são distantes do valor paramétrico quando considerado apenas o valor genômico devido a dominância, sendo que nos cenários 2 e 4 os valores obtidos se aproximam do método exaustivo. Na ICR por meio dos 1 e 2 e levando em consideração apenas os valores genômicos devido a dominância, tem-se que os critérios 1 e 2 apresentaram valores próximos de herdabilidade aditiva, sendo que nos cenários 2 e 4 os valores de herdabilidade aditiva obtidos por ambos os métodos também são próximos dos valores obtidos do método exaustivo que considera apenas o valor genômico aditivo e valor genômico devido a dominância. Em relação a herdabilidade devido a dominância, tem-se que em todos os cenários, o critério 1 e 2 apresentaram valores próximos entre si e do modo exaustivo. Além disso, os valores de herdabilidade devido a dominância são próximos dos valores obtidos pelo critério 1 quando se considera apenas o valor genômico aditivo em todos os cenários. Os resultados das análises considerando o valor genômico total, revelam que nos cenários 1 e 2, o critério 1 e 3 apresentaram valores similares de herdabilidade aditiva, sendo que nestes cenários a herdabilidade foi mais próxima da herdabilidade aditiva obtida quando se considera apenas o valor genômico devido a dominância. Em relação a herdabilidade devida a dominância, os critérios 1 e 3 também apresentaram valores similares e próximos dos resultados obtidos pelo método exaustivo nos cenários 1,2 e 4 e em todos os cenários a herdabilidade devido a dominância foi próxima dos resultados obtidos considerando apenas o valor genômico aditivo ou devido a dominância. Em geral, o método ICR é ineficiente para a estimação da herdabilidade aditiva e devido à dominância.

Desta forma, o critério 1 está associado a uma maior acurácia aditiva e devido à dominância, em todos os cenários tendo como base o modelo aditivo dominante. Isto respalda os resultados obtidos por Costa (2018) referentes a dados simulados e reais em

que os modelos eram apenas aditivos e apresenta-se como uma alternativa mais acurada em relação aos critérios apresentados por Azevedo et al. (2013, 2014, 2015a). Assim, o critério 1 será considerado nas próximas comparações.

ii) Critério de escolha do número de componentes com base nos efeitos aditivos, devido à dominância e total

Os resultados médios de eficiência relativa em relação ao máximo valor alcançado, ou seja, razão entre a acurácia do método e critério de escolha do número de componentes em relação ao valor genômico (aditivo, devido à dominância e total) e a acurácia máxima (aditivo e devido à dominância) atingida são apesentados na Tabela 8.

Para todos os métodos, ICR considerando o método exaustivo, ICR considerando o critério 1, PCR e PLS, o critério de escolha do número de componentes com base nos valores genômicos devido a dominância apresentou, em média, maiores acurácias aditivas e devido a dominância em relação aos outros valores genômicos. Dessa forma, caso seja interesse especificamente um dos efeitos, aditivo ou devido à dominância, deve-se escolher aquele associado ao critério que lhe fornece maior acurácia do efeito de interesse, no entanto, escolhemos aquele critério que combinava melhor as duas informações. Isto corrobora com o estudo realizado por Huang e Mackay (2016) que descreve que a parametrização tradicional para os efeitos aditivos e de dominância conduz ao fato de a variância aditiva explicar a maioria das variações genéticas, mesmo sob ação gênica de dominância. Nesse caso, deve se privilegiar a dominância para que consiga capturar ambos efeitos adequadamente.

iii) Comparação entre os métodos

Considerando apenas critério de escolha do número de componentes com base nos valores genômicos devido a dominância, tem-se que a ICR com o critério 1 e com o método exaustivo apresentou, em média, menores acurácias em relação aos métodos de redução de dimensionalidade, seguido pelo PLS e por último a PCR (Tabela 4, 5, 6 e 7). No entanto, os trabalhos desenvolvidos por Azevedo et al. (2015a) mostraram que para um conjunto de dados reais, a ICR apresentou-se mais eficiente na estimação dos valores genéticos genômicos baseado no modelo apenas aditivo e também Azevedo et al. (2015a) verificaram que a ICR e PCR apresentaram similaridades e possivelmente isto se justifica pelo fato das variáveis analisadas adotarem um comportamento linear. Considerando todos os cenários, o PLS foi, em média, para os valores aditivos 1,25% mais eficiente que o G-BLUP e para os valores devido à dominância, em média, foi menos eficiente que o G-BLUP. A PCR foi, em média, para os valores aditivos 4,25% e para os valores devido à dominância 14,75% mais eficiente que o G-BLUP.

Os métodos PLS e PCR subestimam as herdabilidades aditivas e devido à dominância enquanto o G-BLUP, na maioria dos cenários, subestima as herdabilidades aditivas e superestima as herdabilidades devido à dominância. Em relação ao viés das estimativas dos valores genômicos, os métodos PLS, PCR e G-BLUP superestimaram os valores aditivos (coeficientes de regressão < 1) e subestimaram os valores devido à dominância (coeficientes de regressão > 1).

Os tempos computacionais associados aos dados simulados em segundos e horas referentes a cada método são apresentados na Tabela 9. O tempo computacional referente a ICR pelo método exaustivo considerando uma replicata de cada cenário, é de 221 horas. Entretanto, pelos critérios 1, 2 e 3 o tempo é reduzido drasticamente para cerca de 0,18

horas (aproximadamente 10,8 minutos), mas apenas o critério 1 se mostrou mais eficiente considerando os três modelos (aditivo, dominante e aditivo-dominante).

Considerando os quatro métodos (ICR, PCR, PLS e G-BLUP), observa-se que o G-BLUP apresentou menor tempo de análise computacional, cerca de 340,22 horas (aproximadamente 5,4 minutos), valores estes bem próximos do tempo de execução das análises da PCR, PLS e G-BLUP, sendo o tempo de execução das análises dado em horas por 374,46 e 444,34 (6 e 7,2 minutos) respectivamente. No entanto, a PCR mostrou-se como um método mais eficiente quando comparado ao G-BLUP para as análises que consideram o valor genômico aditivo, devido a dominância ou total.

Assim, a PCR mostrou-se como um método eficiente, em relação aos outros métodos de dimensionalidade e o G-BLUP, na estimação de valores genômicos aditivos e devido a dominância considerando análises baseadas no número de componentes que conduzem os valores genômicos devido à dominância a acurácia máxima, e esta, portanto, deve ser considerada a abordagem mais adequada. Além de apresentar baixo esforço computacional.

4. CONCLUSÕES

Dessa forma, conclui-se que o critério 1 no qual consiste no número de componentes independentes igual ao número de componentes principais que conduz PCR a um maior valor de acurácia, mostrou-se mais eficiente na estimação dos valores genômicos dos indivíduos, visto que atingiu valores de acurácia bem próximos dos valores alcançados pelo modo exaustivo.

As análises ainda revelam que para a estimação simultânea dos efeitos aditivos e devido a dominância pelos métodos de redução de dimensionalidade, o número de

componentes escolhido deve ser aquele que conduz o valor devido à dominância a uma maior acurácia.

Além disso, ao comparar os métodos de redução dimensional (ICR, PCR e PLS) com o G-BLUP observa-se que a eficiência relativa da PCR é superior em termos de acurácia, além de apresentar um dos menores tempos computacionais na execução das análises. Ademais, nenhum dos métodos considerados capturaram adequadamente as herdabilidades simuladas e apresentaram viés.

Tabela 4: As herdabilidades paramétricas aditiva, devido à dominância e total ($h^2_{M_{par}}$), o número de componentes (Nc), herdabilidade aditiva e devido a dominância (h^2_{aM} e h^2_{dM}), acurácia aditiva e devido à dominância ($r_{a\hat{a}}$ e $r_{d\hat{d}}$) e coeficientes de regressão ($\hat{b}_{y\hat{a}}$ e $\hat{b}_{y\hat{d}}$) e eficiência relativa considerando cada critério de escolha e método de redução de dimensionalidade e o método G-BLUP considerando o cenário 1 (características controladas por genes de pequenos efeitos – herança poligênica).

VG	$h^2_{M_{par}}$	Método	Critério	Nc	h^2_{aM}	$r_{a\hat{a}}$	$\hat{b}_{y\hat{a}}$	h^2_{dM}	$r_{d\hat{d}}$	$\hat{b}_{y\hat{d}}$	EF_a	EF_d
Aditivo	0,20	ICR	Exaustivo	20 ± 0,00	0,05 ± 0,01	0,58 ± 0,04	1,30± 0,12	0,10 ± 0,01	0,22 ± 0,05	1,80 ± 0,17	1,07	0,71
			Critério 1	56 ± 22	0,07 ± 0,02	0,57 ± 0,04	1,20± 0,09	0,03 ± 0,01	0,19 ± 0,08	1,70 ± 0,13	1,06	0,61
			Critério 2	16 ± 16	0,20 ± 0,05	0,45 ± 0,15	1,15± 0,08	0,02 ± 0,02	0,06 ± 0,08	1,80 ± 0,11	0,83	0,19
		PCR	Exaustivo	56 ± 22	0,12 ± 0,02	0,60 ± 0,03	1,00± 0,10	0,02 ± 0,01	0,32 ± 0,11	2,50 ± 0,70	1,11	1,03
		PLS	Exaustivo	2,80 ± 1,50	0,22 ± 0,07	0,59 ± 0,05	0,80± 0,09	0,06 ± 0,06	0,31 ± 0,08	2,10 ± 1,10	1,09	1,00
Dominância	0,10	ICR	Exaustivo	211 ± 0,00	0,10 ± 0,01	0,53 ± 0,04	0,89± 0,07	0,04 ± 0,02	0,24 ± 0,05	1,30 ± 0,09	0,98	0,77
			Critério 1	120 ± 120	0,29± 0,04	0,54 ± 0,03	1,10± 0,23	0,04± 0,01	0,23 ± 0,04	1,50 ± 0,33	1,00	0,74
			Critério 2	250 ± 45	0,27± 0,04	0,52 ± 0,04	0,82± 0,05	0,05± 0,02	0,23 ± 0,05	1,20 ± 0,07	0,96	0,74
		PCR	Exaustivo	120 ± 120	0,16 ± 0,08	0,58 ± 0,03	0,91± 0,20	0,04 ± 0,04	0,37 ± 0,04	1,90 ± 0,76	1,07	1,19
		PLS	Exaustivo	1,10 ± 0,33	0,13 ± 0,02	0,58 ± 0,04	0,88± 0,06	0,01 ± 0,01	0,33 ± 0,07	4,60 ± 0,83	1,07	1,06
Total	0,30	ICR	Exaustivo	60 ± 0,00	0,07 ± 0,01	0,56 ± 0,04	1,20± 0,09	0,03 ± 0,02	0,19 ± 0,05	1,60 ± 0,13	1,04	0,61
			Critério 1	78 ± 32	0,29± 0,02	0,54 ± 0,03	1,10± 0,23	0,03± 0,01	0,23 ± 0,04	1,50 ± 0,33	1,00	0,74
			Critério 2	9,90 ± 6,70	0,14 ± 0,02	0,38 ± 0,16	1,40 ± 0,10	0,01± 0,01	0,08 ± 0,07	1,90 ± 0,14	0,70	0,26
			Critério 3	190 ± 0,33	0,29 ± 0,02	0,54 ± 0,04	0,91± 0,06	0,05 ± 0,01	0,23 ± 0,05	1,30 ± 0,09	1,00	0,74
		PCR	Exaustivo	78 ± 32	0,14 ± 0,03	0,60 ± 0,04	0,97± 0,04	0,03 ± 0,01	0,37 ± 0,06	2,10 ± 0,40	1,11	1,19
		PLS	Exaustivo	1,60 ± 0,53	0,12 ± 0,02	0,59 ± 0,04	0,87± 0,07	0,02 ± 0,01	0,26 ± 0,07	3,50 ± 1,20	1,10	0,84
		G-BLUP	-	-	0,15 ± 0,05	0,54 ± 0,06	0,54± 0,20	0,13 ± 0,06	0,31 ± 0,08	2,90 ± 0,50	1,00	1,00

Número de componentes independentes que conduz a: ICR, PCR e PLS a um maior valor de acurácia (Exaustivo); PCR a um maior valor de acurácia (Critério 1); PCR a um menor valor de viés (Critério 2); 80% da variação total de X explicada pelos componentes principais (Critério 3).

Tabela 5: As herdabilidades paramétricas aditiva, devido à dominância e total ($h^2_{M_{par}}$), o número de componentes (Nc), herdabilidade aditiva e devido a dominância (h^2_{aM} e h^2_{dM}), acurácia aditiva e devido à dominância ($r_{a\hat{a}}$ e $r_{d\hat{d}}$) e coeficientes de regressão ($\hat{b}_{y\hat{a}}$ e $\hat{b}_{y\hat{d}}$) e eficiência relativa considerando cada critério de escolha e método de redução de dimensionalidade e o método G-BLUP considerando o cenário 2 (características controladas por genes de pequenos efeitos – herança poligênica).

VG	$h^2_{M_{par}}$	Método	Critério	Nc	h^2_{aM}	$r_{a\hat{a}}$	$\hat{b}_{y\hat{a}}$	h^2_{dM}	$r_{d\hat{d}}$	$\hat{b}_{y\hat{d}}$	EF_a	EF_d
Aditivo	0,30	ICR	Exaustivo	158 ± 0,00	0,37 ± 0,01	0,62 ± 0,04	1,20± 0,06	0,06 ± 0,05	0,25 ± 0,04	1,70 ± 0,09	1,07	0,68
			Critério 1	270 ± 110	0,38± 0,02	0,62 ± 0,03	1,10± 0,12	0,07 ± 0,01	0,27 ± 0,05	1,50 ± 0,17	1,07	0,73
			Critério 2	50 ± 87	0,21± 0,04	0,46 ± 0,10	1,30± 0,05	0,03 ± 0,02	0,07 ± 0,11	1,70 ± 0,07	0,79	0,19
		PCR	Exaustivo	270 ± 110	0,32± 0,02	0,68 ± 0,03	0,96± 0,09	0,11 ± 0,04	0,40 ± 0,06	1,40 ± 0,38	1,17	1,08
		PLS	Exaustivo	4,80 ± 1,60	0,38 ± 0,04	0,66 ± 0,03	0,93± 0,04	0,14 ± 0,04	0,23 ± 0,05	1,30 ± 0,22	1,14	0,62
Dominância	0,10	ICR	Exaustivo	444 ± 0,00	0,35 ± 0,01	0,59 ± 0,03	0,88 ± 0,04	0,09 ± 0,04	0,31 ± 0,03	1,30 ± 0,05	1,02	0,84
			Critério 1	210 ± 120	0,36± 0,02	0,60 ± 0,03	1,10± 0,12	0,07 ± 0,01	0,26 ± 0,03	1,60 ± 0,17	1,03	0,70
			Critério 2	370 ± 27	0,36± 0,01	0,60 ± 0,03	0,95± 0,04	0,09 ± 0,04	0,30 ± 0,03	1,30 ± 0,05	1,03	0,81
		PCR	Exaustivo	210 ± 120	0,29 ± 0,07	0,66 ± 0,04	0,98± 0,09	0,09 ± 0,05	0,42 ± 0,04	1,60 ± 0,50	1,14	1,14
		PLS	Exaustivo	1,00 ± 0,00	0,19 ± 0,02	0,60 ± 0,03	1,00 ± 0,04	0,01 ± 0,01	0,33 ± 0,05	5,50 ± 0,34	1,03	0,89
Total	0,50	ICR	Exaustivo	315 ± 0,00	0,37 ± 0,01	0,61 ± 0,03	1,00± 0,06	0,08 ± 0,01	0,29 ± 0,03	1,50 ± 0,08	1,05	0,78
			Critério 1	280 ± 90	0,36 ± 0,02	0,60 ± 0,03	1,10 ± 0,12	0,07 ± 0,01	0,26 ± 0,03	1,60 ± 0,17	1,03	0,70
			Critério 2	14 ± 13	0,14 ± 0,04	0,38 ± 0,03	1,50 ± 0,17	0,02 ± 0,02	0,12 ± 0,11	2,10 ± 0,24	0,66	0,32
			Critério 3	190 ± 0,00	0,37± 0,01	0,61 ± 0,03	1,20± 0,05	0,07 ± 0,01	0,26 ± 0,04	1,70 ± 0,07	1,05	0,70
		PCR	Exaustivo	280 ± 90	0,33 ± 0,05	0,68 ± 0,03	0,96± 0,08	0,12 ± 0,04	0,41 ± 0,04	1,30 ± 0,28	1,17	1,11
		PLS	Exaustivo	2,70 ± 1,10	0,29 ± 0,09	0,65 ± 0,03	0,97± 0,05	0,06 ± 0,05	0,33 ± 0,08	2,70 ± 1,40	1,12	0,89
		G-BLUP	-	-	0,27 ± 0,03	0,58 ± 0,03	0,63 ± 0,12	0,20 ± 0,03	0,37 ± 0,06	3,20 ± 0,54	1,00	1,00

Número de componentes independentes que conduz a: ICR, PCR e PLS a um maior valor de acurácia (Exaustivo); PCR a um maior valor de acurácia (Critério 1); PCR a um menor valor de viés (Critério 2); 80% da variação total de X explicada pelos componentes principais (Critério 3).

Tabela 6: As herdabilidades paramétricas aditiva, devido à dominância e total ($h^2_{M_{par}}$), o número de componentes (Nc), herdabilidade aditiva e devido a dominância (h^2_{aM} e h^2_{dM}), acurácia aditiva e devido à dominância ($r_{a\hat{a}}$ e $r_{d\hat{d}}$) e coeficientes de regressão ($\hat{b}_{y\hat{a}}$ e $\hat{b}_{y\hat{d}}$) e eficiência relativa considerando cada critério de escolha e método de redução de dimensionalidade e o método G-BLUP considerando o cenário 3 (características controladas por genes de grandes e pequenos efeitos – herança mista).

VG	$h^2_{M_{par}}$	Método	Critério	Nc	h^2_{aM}	$r_{a\hat{a}}$	$\hat{b}_{y\hat{a}}$	h^2_{dM}	$r_{d\hat{d}}$	$\hat{b}_{y\hat{d}}$	EF_a	EF_d
Aditivo	0,20	ICR	Exaustivo	17 ± 0,00	0,26 ± 0,01	0,55 ± 0,05	1,10± 0,11	0,02 ± 0,01	0,10 ± 0,09	1,60 ± 0,15	0,92	0,29
			Critério 1	85 ± 67	0,28± 0,02	0,53 ± 0,04	1,00 ± 0,17	0,02 ± 0,02	0,16 ± 0,04	1,50 ± 0,25	0,88	0,46
			Critério 2	14 ± 16	0,20± 0,02	0,45 ± 0,04	1,20 ± 0,05	0,02 ± 0,02	0,04 ± 0,13	1,60 ± 0,17	0,75	0,11
		PCR	Exaustivo	85 ± 67	0,12 ± 0,04	0,59 ± 0,04	0,94± 0,15	0,03 ± 0,03	0,31 ± 0,10	2,00 ± 1,50	0,98	0,89
		PLS	Exaustivo	3,30 ± 1,40	0,26 ± 0,08	0,59± 0,06	0,70± 0,14	0,09 ± 0,05	0,36 ± 0,08	1,30 ± 1,30	0,98	1,03
Dominância	0,10	ICR	Exaustivo	487 ± 0,00	0,16 ± 0,01	0,40 ± 0,04	0,50± 0,05	0,06 ± 0,01	0,22 ± 0,04	0,71 ± 0,07	0,67	0,63
			Critério 1	170 ± 110	0,23± 0,02	0,48 ± 0,04	0,89 ± 0,18	0,04 ± 0,01	0,21 ± 0,04	1,30 ± 0,26	0,80	0,60
			Critério 2	130 ± 53	0,23± 0,02	0,48 ± 0,07	0,88 ± 0,07	0,03 ± 0,01	0,18 ± 0,07	1,20 ± 0,10	0,80	0,51
		PCR	Exaustivo	170 ± 110	0,18 ± 0,07	0,56± 0,05	0,81± 0,16	0,06 ± 0,05	0,40 ± 0,06	1,30 ± 0,63	0,93	1,14
		PLS	Exaustivo	1,00 ± 0,00	0,12 ± 0,01	0,59± 0,04	0,87± 0,06	0,01± 0,01	0,36 ± 0,04	3,40 ± 0,86	0,98	1,03
Total	0,30	ICR	Exaustivo	18 ± 0,00	0,26 ± 0,01	0,51 ± 0,05	1,10± 0,11	0,02 ± 0,01	0,10 ± 0,09	1,60 ± 0,15	0,85	0,29
			Critério 1	120 ± 95	0,23 ± 0,02	0,48 ± 0,04	0,89 ± 0,18	0,04 ± 0,01	0,21 ± 0,04	1,60 ± 0,26	0,80	0,60
			Critério 2	5,20 ± 3,40	0,20 ± 0,01	0,45 ± 0,03	1,30 ± 0,18	0,01 ± 0,03	0,01 ± 0,01	1,80 ± 0,12	0,75	0,03
			Critério 3	190 ± 0,00	0,20 ± 0,01	0,45 ± 0,07	0,72 ± 0,15	0,04 ± 0,01	0,20 ± 0,06	1,00 ± 0,21	0,75	0,57
		PCR	Exaustivo	120 ± 95	0,14 ± 0,06	0,58± 0,05	0,90± 0,17	0,04 ± 0,04	0,36 ± 0,11	1,60 ± 1,10	0,97	1,03
		PLS	Exaustivo	1,60 ± 1,30	0,15± 0,07	0,58± 0,05	0,84± 0,11	0,03 ± 0,05	0,29 ± 0,08	2,90 ± 1,30	0,97	0,83
		G-BLUP	-	-	0,14± 0,03	0,60± 0,04	0,71± 0,15	0,13 ± 0,04	0,35 ± 0,07	2,70 ± 0,87	1,00	1,00

Número de componentes independentes que conduz a: ICR, PCR e PLS a um maior valor de acurácia (Exaustivo); PCR a um maior valor de acurácia (Critério 1); PCR a um menor valor de viés (Critério 2); 80% da variação total de X explicada pelos componentes principais (Critério 3).

Tabela 7: As herdabilidades paramétricas aditiva, devido à dominância e total ($h^2_{M_{par}}$) do cenário 4, o número de componentes (Nc), herdabilidade aditiva e devido a dominância (h^2_{aM} e h^2_{dM}), acurácia aditiva e devido à dominância ($r_{a\hat{a}}$ e $r_{d\hat{d}}$) e coeficientes de regressão ($\hat{b}_{y\hat{a}}$ e $\hat{b}_{y\hat{d}}$) e eficiência relativa considerando cada critério de escolha e método de redução de dimensionalidade e o método G-BLUP considerando o cenário 4 (características controladas por genes de grandes e pequenos efeitos – herança mista).

VG	$h^2_{M_{par}}$	Método	Critério	Nc	h^2_{aM}	$r_{a\hat{a}}$	$\hat{b}_{y\hat{a}}$	h^2_{dM}	$r_{d\hat{d}}$	$\hat{b}_{y\hat{d}}$	EF_a	EF_d
Aditivo	0,30	ICR	Exaustivo	203 ± 0,00	0,30 ± 0,01	0,55 ± 0,05	0,96 ± 0,04	0,06 ± 0,01	0,25 ± 0,04	1,40 ± 0,05	0,90	0,61
			Critério 1	290 ± 140	0,29 ± 0,02	0,54 ± 0,05	0,89 ± 0,11	0,06 ± 0,01	0,25 ± 0,04	1,30 ± 0,15	0,89	0,61
			Critério 2	6,40 ± 4,40	0,19± 0,02	0,44 ± 0,04	1,20± 0,07	0,02 ± 0,01	0,01 ± 0,13	1,30 ± 0,09	0,72	0,02
		PCR	Exaustivo	290 ± 140	0,30 ± 0,08	0,66 ± 0,03	0,82 ± 0,12	0,13 ± 0,07	0,43 ± 0,05	1,10 ± 0,35	1,08	1,05
		PLS	Exaustivo	4,70± 1,00	0,36 ± 0,04	0,64 ± 0,04	0,78 ± 0,06	0,16 ± 0,03	0,38 ± 0,04	1,00 ± 0,20	1,05	0,93
Dominância	0,20	ICR	Exaustivo	350 ± 0,00	0,28 ± 0,01	0,53 ± 0,04	0,83 ± 0,05	0,08 ± 0,01	0,28 ± 0,04	1,20 ± 0,07	0,87	0,68
			Critério 1	270 ± 120	0,28 ± 0,02	0,53 ± 0,04	0,89 ± 0,09	0,07 ± 0,09	0,27 ± 0,04	1,30 ± 0,13	0,87	0,66
			Critério 2	280 ± 68	0,28± 0,02	0,53 ± 0,03	0,89± 0,04	0,07 ± 0,01	0,27 ± 0,04	1,20 ± 0,05	0,87	0,66
		PCR	Exaustivo	270 ± 120	0,29 ± 0,06	0,63 ± 0,03	0,81 ± 0,10	0,12 ± 0,07	0,46 ± 0,03	1,10 ± 0,32	1.03	1.12
		PLS	Exaustivo	1,10 ± 0,33	0,18 ± 0,11	0,59 ± 0,03	0,88 ± 0,05	0,01± 0,01	0,41 ± 0,06	3,90 ± 0,56	0,97	0,76
Total	0,50	ICR	Exaustivo	288 ± 0,00	0,29 ± 0,01	0,54 ± 0,04	0,89 ± 0,04	0,08 ± 0,03	0,26 ± 0,04	1,30 ± 0,06	0,89	0,63
			Critério 1	280 ± 93	0,23 ± 0,02	0,53 ± 0,05	0,89 ± 0,09	0,07 ± 0,01	0,27 ± 0,04	1,30 ± 0,13	0,87	0,66
			Critério 2	6,80 ± 4,60	0,20 ± 0,02	0,45 ± 0,06	1,30 ± 0,10	0,02 ± 0,01	0,01 ± 0,01	1,90 ± 0,14	0,74	0,02
			Critério 3	190 ± 0,00	0,29 ± 0,01	0,54 ± 0,05	0,96 ± 0,04	0,06 ± 0,01	0,25 ± 0,04	1,40 ± 0,05	0,89	0,61
		PCR	Exaustivo	280 ± 93	0,30 ± 0,06	0,65 ± 0,03	0,82 ± 0,10	0,13 ± 0,05	0,45 ± 0,03	1,10 ± 0,27	1,07	1,10
		PLS	Exaustivo	2,90 ± 1,90	0,28 ± 0,11	0,62 ± 0,04	0,85 ± 0,07	0,09 ± 0,08	0,39 ± 0,10	2,40 ± 1,50	1,02	0,95
		G-BLUP	-	-	0,25 ± 0,06	0,61 ± 0,02	0,74 ± 0,20	0,18 ± 0,04	0,41 ± 0,04	3,20 ± 0,68	1,00	1,00

Número de componentes independentes que conduz a: ICR, PCR e PLS a um maior valor de acurácia (Exaustivo); PCR a um maior valor de acurácia (Critério 1); PCR a um menor valor de viés (Critério 2); 80% da variação total de X explicada pelos componentes principais (Critério 3).

Tabela 8: Resultados médios de eficiência relativa em relação ao máximo valor alcançado, razão entre a acurácia do método e critério de escolha do número de componentes em relação ao valor genômico (aditivo, devido à dominância e total - VG) e a acurácia máxima (aditivo e devido à dominância) atingida.

Cenário	Método	Critério	VG	Eficiência Relativa dos efeitos		
				Aditivo	Dominante	Média
Cenário 1	ICR	Exaustivo	Aditivo	1,00	0,92	0,96
			Dominante	0,91	1,00	0,96
			Total	0,97	0,79	0,88
	ICR	Critério 1	Aditivo	0,98	0,79	0,89
			Dominante	0,93	0,96	0,95
			Total	0,93	0,96	0,95
	PCR	Exaustivo	Aditivo	1,00	0,86	0,93
			Dominante	0,97	1,00	0,99
			Total	1,00	1,00	1,00
	PLS	Exaustivo	Aditivo	1,00	0,94	0,97
			Dominante	0,98	1,00	0,99
			Total	1,00	0,79	0,90
Cenário 2	ICR	Exaustivo	Aditivo	1,00	0,81	0,91
			Dominante	0,95	1,00	0,98
			Total	0,98	0,94	0,96
	ICR	Critério 1	Aditivo	1,00	0,87	0,94
			Dominante	0,97	0,84	0,91
			Total	0,98	0,84	0,91
	PCR	Exaustivo	Aditivo	1,00	0,95	0,98
			Dominante	0,97	1,00	0,99
			Total	1,00	0,98	0,99
	PLS	Exaustivo	Aditivo	1,00	0,70	0,85
			Dominante	0,91	1,00	0,96
			Total	0,98	1,00	0,99
Cenário 3	ICR	Exaustivo	Aditivo	1,00	0,45	0,73
			Dominante	0,73	1,00	0,87
			Total	0,93	0,45	0,69
	ICR	Critério 1	Aditivo	0,96	0,73	0,85
			Dominante	0,87	0,95	0,91
			Total	0,87	0,95	0,91
	PCR	Exaustivo	Aditivo	1,00	0,78	0,89
			Dominante	0,95	1,00	0,98
			Total	0,98	0,90	0,94
	PLS	Exaustivo	Aditivo	1,00	1,00	1,00
			Dominante	1,00	1,00	1,00
			Total	0,98	0,81	0,90
Cenário 4	ICR	Exaustivo	Aditivo	1,00	0,89	0,95
			Dominante	0,96	1,00	0,98
			Total	0,98	0,93	0,96
	ICR	Critério 1	Aditivo	0,98	0,89	0,94
			Dominante	0,96	0,96	0,96
			Total	0,96	0,96	0,96
	PCR	Exaustivo	Aditivo	1,00	0,93	0,97

		Dominante	0,95	1,00	0,98
		Total	0,98	0,98	0,98
		Aditivo	1,00	0,93	0,97
PLS	Exaustivo	Dominante	0,92	1,00	0,96
		Total	0,97	0,95	0,96

Número de componentes independentes que conduz a: ICR, PCR e PLS a um maior valor de acurácia (Exaustivo); PCR a um maior valor de acurácia (Critério 1).

Tabela 9: Tempo computacional em segundos (horas) considerando cada método utilizado e cada critério de escolha do número de componentes independentes.

Método	Critérios	Tempo Computacional
ICR	Exaustivo	795600 (221)
	Critério 1	654,74 (0,18)
	Critério 2	633,98 (0,18)
	Critério 3	648,12 (0,18)
PCR	Exaustivo	374,46 (0,10)
PLS	Exaustivo	444,34 (0,12)
G-BLUP		340,22 (0,09)

Número de componentes independentes que conduz a: ICR, PCR e PLS a um maior valor de acurácia (Exaustivo); PCR a um maior valor de acurácia (Critério 1); PCR a um menor valor de viés (Critério 2); 80% da variação total de X explicada pelos componentes principais (Critério 3).

5. REFERÊNCIAS BIBLIOGRÁFICAS

DE ALMEIDA FILHO J. E.; GUIMARÃES J. F.; E SILVA F. F.; DE RESENDE M. D.; MUÑOZ P.; KIRST M.; RESENDE M. F J. R. The contribution of dominance to phenotype prediction in a pine breeding and simulated population. **Heredity**, v. 117, p. 33-41, 2016.

AZEVEDO, C. F.; NASCIMENTO, M.; SILVA, F. F.; RESENDE, M. D. V.; LOPES, P.S.; GUIMARÃES, S. E. F.; GLÓRIA, L. S. Comparison of dimensionality reduction methods to predict genomic breeding values for carcass traits in pigs. **Genet. Mol. Res**, v. 14, p. 12217-12227, 2015.

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.

AZEVEDO, C.F.; RESENDE, M. D.V.; SILVA, F. F.; VIANA, J. M. S.; VALENTE, M. S. F. RESENDE JUNIOR, M.F.R.; MUÑOZ, P. RIDGE, Lasso and Bayesian additive-dominance genomic models. **BMC Genetics**, v. 16, p. 105, 2015.

AZEVEDO, C. F.; SILVA, F. F.; RESENDE, M. D.; LOPES, M. S.; DUIJVESTEIJN, N.; GUIMARÃES, S. E. F.; LOPES, P.S.; KELLY, M. J.; VIANA, J. M. S.; KNOL, E. F. Supervised independent component analysis as an alternative method for genomic selection in pigs. **Journal of Animal Breeding and Genetics**, v. 131, p. 452-461, 2014.

BENNEWITZ J.; MEUWISSEN T.H.E. The distribution of QTL additive and dominance effects in porcine F2 crosses. **Journal of Animal Breeding and Genetics**, v. 127, p. 171-179, 2010.

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

COSTA, J. A.; AZEVEDO, C. F.; RESENDE, M. D. V.; NASCIMENTO, M. Predição genômica via redução de dimensionalidade em modelos aditivo-dominante. Dissertação (Mestrado em Estatística Aplicada e Biometria), Universidade Federal de Viçosa, Viçosa, 2018.

COVARRUBIAS-PAZARAN, G. Genome assisted prediction of quantitative traits using the R package sommer. **PLoS ONE**, v. 11, n. 6, p. e0156744, 2016.

DENIS M.; BOUVET J. M. Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. **Tree Genetics Genomes**, v. 9, p. 37-51, 2013.

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

GODDARD M. E.; WRAY N. R.; VERBYLA K.; VISSCHER P. M. Estimating effects and making predictions from genome-wide marker data. **Statistical Science**, v. 24, p. 517-529, 2009.

HILL W.G.; GODDARD M. E.; VISSCHER P. M. Data and theory point to mainly additive genetic variance for complex traits. **PLoS Genet**, v. 4, p. e1000008, 2008.

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

HUANG, W.; MACKAY, T. F. C. The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis. **PLoS Genetics**, v. 12, n. 11, p. e1006421, 2016.

HYVÄRINEN, A. Fast and robust fixed-point algorithms for independent component analysis. **IEEE transactions on Neural Networks**, v. 10, n. 3, p. 626-634, 1999.

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

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

KUHN, M. Contributions from Jed WING, J.; WESTON, S.; WILLIAMS, A.; KEEFER, C.; ENGELHARDT, A.; COOPER, T.; MAYER, Z.; KENKEL, B.; the R Core Team, BENESTY, M.; LESCARBEAU, R.; SCRUCCA, A. Z. L.; TANG, Y.; CANDAN, C.; HUNT, T**. Caret: Classification and Regression Training.** R package version 6.0-77. https://CRAN.R-project.org/package=caret, 2017.

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

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

MEVIK, B. H; WEHRENS, R.; LILAND, K. H. **Pls: Partial Least Squares and Principal Component Regression.** R package version 2.6-0. https://CRAN.R-project.org/package=pls, 2016.

MUÑOZ, P. R.; RESENDE, J. R M. F. R.; GEZAN, S. A.; RESENDE, M. D. V.; DE LOS CAMPOS G.; KIRST, M.; HUBER, D.; PETER, G. F. Unraveling additive from nonadditive effects using genomic relationship matrices. **Genetics**, v. 198, p. 1759-1768, 2014.

R Development Core Team. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, Available: http://www.R-project.org, 2016.

RESENDE, M. D.V.; 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, 2012. p.151-153. Disponível em: http://www.ppestbio.ufv.br/?page_id=448.

SU, G.; CHRISTENSEN, O. F.; OSTERSEN, T.; HENRYON, M.; LUND, M. S. Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. **PLoS One**, v. 7, p. e45293, 2012.

TORO M. A, VARONA L. A note on mate allocation for dominance handling in genomic selection. **Genetics Selection Evolution**, v. 42, p. 33, 2010.

VITEZICA, Z. G.; VARONA, L.; LEGARRA, A. On the additive and dominance variance and covariance of individuals within the genomic selection scope. **Genetics**, Austin, v. 195, n. 4, p. 1223-1230, 2013.

WANG, C.; DA, Y. Quantitative genetics model as the unifying model for defining genomic relationship and inbreeding coefficient. **PLoS One**, v. 9, p. e114484, 2014.

WOLD, H. **Soft modelling by latent variables: the non-linear iterative partial least squares approach**. Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett, J. Gani, ed., Academic Press, London, 1975.

WELLMANN R.; BENNEWITZ J. Bayesian models with dominance effects for genomic evaluation of quantitative traits. **Genetics Research**, v. 94, p. 21-37, 2012.

ZENG J.; TOOSI A.; FERNANDO R. L, DEKKERS J. C. M.; GARRICK D. J. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. **Genetics Selection Evolution**, v.45, p.11, 2013.

CONCLUSÕES GERAIS

Este estudo propõe no Capítulo 2 critérios de decisão para determinar o número ótimo de componentes independentes a serem inseridos no modelo aditivo a fim de utilizar a Regressão via Componentes Independentes (ICR) na estimação dos valores genômicos aditivos. Os resultados obtidos nesse capítulo indicaram que o critério 1, cujo número de componentes independentes fosse igual ao número de componentes principais que conduz a PCR a um maior valor de acurácia, se mostrou mais acurado na estimação dos valores genômicos aditivos. Adicionalmente, os valores de acurácia apresentados por este critério e pelo modo exaustivo via ICR são bem similares sendo que a superioridade do critério 1 se dá pelo fato de demandar tempo e esforço computacional muito inferior quando comparado com o método exaustivo. Assim, o critério 1 mostra-se como uma alternativa viável para a escolha do número ótimo de componentes independentes a serem utilizados na aplicação da ICR. Em geral, todos os critérios foram viesados e não capturaram adequadamente os valores de herdabilidade aditiva que foram simulados.

No Capítulo 3 foram propostos e avaliados os três critérios mais eficientes analisados no segundo capítulo para determinar o número de componentes independentes sob o contexto de modelos aditivo-dominante. Verificou-se que prevalece a eficiência e a superioridade do critério 1 ao analisar os efeitos aditivos e devido a dominância sob as três perspectivas (valor genômico aditivo, valor genômico devido a dominância e valor genômico total) e os diagnósticos posteriores referentes a ICR se deram apoiadas no critério 1. Além disso, em virtude do número de componentes que conduzem o valor genômico aditivo, devido a dominância e total a acurácia máxima, foi avaliado qual a melhor forma de determinar o número de componentes sob estas três perspectivas a fim de estimar simultaneamente os efeitos aditivos e devido a dominância. Neste contexto, identificou-se que as análises em que o número de componentes era definido de acordo

com o valor genômico devido a dominância que atinge maior valor de acurácia, se mostrou mais eficiente para estimar concomitantemente os efeitos aditivos e efeitos devido a dominância. As análises posteriores dos métodos se deram a partir do valor genômico devido a dominância que conduz a acurácia máxima do método.

A partir das análises referentes ao valor genômico devido a dominância que conduz a acurácia máxima, comparou-se os métodos de redução de dimensionalidade com o G-BLUP, por meio da eficiência relativa. Verificou-se que dentre eles, a PCR apresentou desempenho melhor que o G-BLUP na presença dos efeitos devido a dominância, isto é, eficiência relativa superior no que se refere a acurácia para estimar os efeitos aditivos e efeitos devido a dominância. Além disso, as herdabilidades capturadas pelos métodos não foram conforme as herdabilidades simuladas e todos foram viesados.