FERNANDO DE OLIVEIRA BUSSIMAN

**Avaliação genética de cavalos Campolina: do *pedigree* à possibilidade de
seleção genômica**

Pirassununga
2021

FERNANDO DE OLIVEIRA BUSSIMAN

Avaliação genética de cavalos Campolina: do pedigree à possibilidade de seleção genômica

Tese apresentada ao Programa de Pós-Graduação em Nutrição e Produção Animal da Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo para a obtenção do título de Doutor em Ciências.

Departamento:

Nutrição e Produção Animal

Área de concentração:

Nutrição e Produção Animal

Orientador:

Prof. Dr. Júlio Cesar de Carvalho Balieiro (VNP-FMVZ/USP)

De acordo: _____

Orientador

Coorientadores:

Prof. Dr. Joanir Pereira Eler (ZMV-FZEA/USP)

Prof. Dr. Fabyano Fonseca e Silva (*in memoriam*) (DZO/UFV)

Autorizo a reprodução parcial ou total desta obra, para fins acadêmicos, desde que citada a fonte.

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO

(Biblioteca Virginie Buff D'Ápice da Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo)

T. 4115
FMVZ

Bussiman, Fernando de Oliveira

Avaliação genética de cavalos Campolina: do *pedigree* à possibilidade de seleção genômica / Fernando de Oliveira Bussiman. – 2021.
233 f. : il.

Tese (Doutorado) – Universidade de São Paulo. Faculdade de Medicina Veterinária e Zootecnia. Departamento de Nutrição e Produção Animal, Pirassununga, 2021.

Programa de Pós-Graduação: Nutrição e Produção Animal.

Área de concentração: Nutrição e Produção Animal.

Orientador: Prof. Dr. Júlio Cesar de Carvalho Balieiro.

Coorientadores: Prof. Dr. Joanir Pereira Eler; Prof. Dr. Fabyano Fonseca e Silva.

1. Linhagens. 2. Redução de posto. 3. Análise confirmatória de fatores. 4. Modelos de equações estruturais. 5. Estratégias de genotipagem. I. Título.

São Paulo, 02 de junho de 2020
CEUAX N [9782030320](#)

Ilmo(a). Sr(a).

Responsável: Júlio Cesar De Carvalho Balieiro

Área: Nutrição E Produção Animal

Equipe envolvida: Fernando De Oliveira Bussiman - [executor](#) (faculdade De Medicina Veterinária E Zootecnia);

Título do projeto: "USO DE SIMULAÇÃO E MODELAGEM COMO FERRAMENTAS DE IDENTIFICAÇÃO DE ANIMAIS PARA GENOTIPAGEM".

Parecer Consubstanciado da CEUA FMVZ

A Comissão de Ética no Uso de Animais da Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo, na reunião de 29/04/2020, **ANALISOU** e **APROVOU** o protocolo de estudo acima referenciado. A partir desta data, é dever do pesquisador:

1. Comunicar toda e qualquer alteração do protocolo.
2. Comunicar imediatamente ao Comitê qualquer evento adverso ocorrido durante o desenvolvimento do protocolo.
3. Os dados individuais de todas as etapas da pesquisa devem ser mantidos em local seguro por 5 anos para possível auditoria dos órgãos competentes.
4. **Relatórios parciais** de andamento deverão ser enviados **anualmente** à CEUA até a conclusão do protocolo.

Prof. Dr. Marcelo Bahia Labruna

Coordenador da Comissão de Ética no Uso de Animais

Faculdade de Medicina Veterinária e Zootecnia da Universidade
de São Paulo

Camilla Mota Mendes

Vice-Coordenador

Faculdade de Medicina Veterinária e Zootecnia da Universidade
de São Paulo

FOLHA DE AVALIAÇÃO

Autor: BUSSIMAN, Fernando de Oliveira

Título: **Avaliação genética de cavalos Campolina:** do *pedigree* à possibilidade de seleção genômica

Tese apresentada ao Programa de Pós-Graduação em Nome do Programa da Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo para obtenção do título de Doutor em Ciências.

Data: 17/11/2021

Banca Examinadora

Prof. Dr. Júlio César de Carvalho Balieiro

Instituição: Faculdade de Medicina Veterinária e Zootecnia (USP) Julgamento: Aprovado

Prof. Dr. Ricardo Vieira Ventura

Instituição: Faculdade de Medicina Veterinária e Zootecnia (USP) Julgamento: Aprovado

Prof. Dr. José Bento Sterman Ferraz

Instituição: Faculdade de Zootecnia e Engenharia de Alimentos (USP) Julgamento: Aprovado

Prof. Dr. Roberto Carneiro

Instituição: Faculdade de Ciências Agrárias e Veterinárias (UNESP) Julgamento: Aprovado

Prof. Dr. Fernando Sebastião Baldi Rey

Instituição: Faculdade de Ciências Agrárias e Veterinárias (UNESP) Julgamento: Aprovado

DEDICATÓRIA

À minha “pãe”, Claudete Cândida de Oliveira, exemplo de mulher de fibra, que sempre me deu força, apoio e incentivo, sendo meu exemplo pessoal de honestidade e caráter.

Ao meu avô, Cláudio Cândido de Oliveira (in memoriam), que sempre me apoiou, mesmo quando meus caminhos me afastaram dele, mostrando-me os valores morais e éticos que levo comigo.

À minha avó, Neusa Felisberto de Oliveira, matriarca que une nossa família, um dos maiores exemplos de força e determinação que tenho.

Ao Prof. Dr. Fabyano Fonseca e Silva (in memoriam), que apesar do pouco tempo que tivemos de convivência serviu de inspiração e fonte de imensuráveis conselhos seja para minha formação acadêmica, seja para esta tese.

AGRADECIMENTOS

A Deus, pois, sem Ele nada do que fiz teria sido possível. A caminhada não foi fácil até aqui, mas pude sempre perceber a mão poderosa de Deus agindo em minha vida, mudando a minha história, me dando condições de trilhar o caminho que escolhi!

Agradeço à minha família, pois se frutifiquei foi porque tive boas raízes. Em especial agradeço ao meu padrasto Nedinaldo Gonçalves, que mesmo tendo entrado em nossas vidas recentemente contribuiu muito para minha formação pessoal, me apoiando e dando forças.

À minha mãe Claudete Cândida de Oliveira, por todo o apoio que sempre me deu, pelos excelentes conselhos, pelas orações fora de hora e por ter segurado as pontas quando eu mesmo já havia pensado em desistir.

À minha avó Neusa Felisberto de Oliveira, pelo apoio e incentivo, mesmo quando o caminho que escolhi me leva a ficar longe.

Ao Prof. Dr. Júlio Cesar de Carvalho Balieiro, meu orientador, a quem tenho muito orgulho de chamar amigo; por todo o apoio e incentivo, por acreditar no meu trabalho e por sempre me deixar livre para trilhar o caminho acadêmico por mim mesmo, mas, sem nunca estar indisponível para me orientar!

Ao Prof. Dr. Joanir Pereira Eler, meu co-orientador, a quem tenho a pretensão de chamar de “papis Jow”; por todos os ensinamentos, profissionais e pessoais ao longo destes anos de convivência, por todas as trocas de experiências, pelos livros e materiais, por toda a ajuda e pelas riquíssimas conversas; por ser o meu exemplo de bom professor/pesquisador.

Ao Prof. Dr. Fabyano Fonseca e Silva (in memoriam), meu co-orientador, por ter sempre me ajudado nas análises mais “cabeludas”, por sempre estar disponível para discutir algum por menor estatístico e por sempre me dar apoio.

À Profa. Dra. Daniela Lino Lourenço, por ter me ajudado imensamente na condução dos estudos de simulação, por todos os ensinamentos e por ter me permitido a inserção em seu grupo de pesquisa mesmo à distância.

Ao Prof. Dr. Ricardo Vieira Ventura, pela amizade e ajuda, por todo o apoio que sempre me deu, pelo auxílio na discussão dos resultados e pelas excelentes contribuições em todos os capítulos deste estudo.

À Profa. Dra. Rachel dos Santos Bueno, a quem considero como uma mãe na FZEA, obrigado por toda a ajuda, mesmo que a senhora não lembre, foi parte fundamental da minha formação. Obrigado pelas conversas e pelos cafezinhos no GMAB.

À Profa. Dra. Lúcia Galvão de Albuquerque, quem “abriu meus olhos” para a avaliação genética, obrigado pela excelente disciplina, obrigado pelos puxões de orelha nas resoluções das listas, obrigado por tido a chance de aprender com a senhora.

À Profa. Dra. Clarice Garcia Borges Demétrio, obrigado por ter me permitido acompanhar suas disciplinas. Obrigado por me mostrar um universo todo novo dentro da estatística, solidificando meus conhecimentos e dando base para toda a minha atuação no futuro. Mais que isso, obrigado por ter sido um exemplo profissional para mim nos semestres em que convivemos.

Ao Prof. Dr. Roberto Carvalheiro, obrigado por sempre compartilhar seu conhecimento de forma tão primorosa com seus alunos, nos inspirando a seguir o caminho acadêmico.

Obrigado pelas muitas conversas na FCAV e nos churrascos pós disciplina. Mas, sobretudo, obrigado por todo o ensinamento ao longo destes anos de convivência.

Ao Prof. Dr. Henrique Nunes de Oliveira, por ter me apresentado por primeiro a Inferência Bayesiana, pelas aulas, conversas e, principalmente, por toda sua ajuda e disponibilidade para auxiliar quem quer que fosse. Muito obrigado por todas dicas e sugestões professor!

Ao Prof. Dr. José Bento Sterman Ferraz, quem me colocou no melhoramento genético, muito obrigado por ter me despertado para este mundo lá nos idos de 2012/2013. Obrigado por toda a parceria, por todas as conversas e trocas de experiências.

Ao Prof. Dr. Fernando Sebastião Baldi Rey, por seus conselhos valioso quanto ao single-step e a seleção genômica. Muito obrigado pelas várias disciplinas ministradas ou organizadas que me proporcionaram sólido conhecimento ao longo desta caminhada.

Ao Prof. Dr. Mehdi Sargolzaei por ter desenvolvido o QMSim (“que programinha ‘salva vidas’”) e por ter sugerido algumas das formas que usei na simulação viabilizando a condução deste estudo.

À Profª. Dra. Arioni Augusti Boligon, por toda a ajuda na comparação das matrizes de (co)variância genética aditiva durante a execução das análises de redução de ranque.

Ao Prof. Dr. Guilherme Jordão de Magalhães Rosa por ter me apresentado os modelos gráficos e a inferência causal; pelas nossas conversas sempre que possível; e pelo Ganbei!

Ao Prof. Dr. Francisco Peñagaricano pela imensa inspiração nos modelos de variáveis latentes e por ter me apresentado a análise confirmatória de fatores.

Ao Dr. Rafael Espigolan, por todas as conversas e ensinamentos trocados durante toda nossa convivência e, em especial por toda sua ajuda na construção das matrizes de Kernel para o capítulo final desta tese.

À MSc. Alexandra Asanovna Elbakyan, por ter programado o tão controverso “SciHub”; quando a Universidade de São Paulo não cobria o paywall de diversas publicações necessárias para o desenvolvimento deste trabalho, foi esse site que disponibilizou os documentos (sim, este agradecimento tem um quê de protesto, ainda que velado, ao modus operandi das publicações).

À MSc. Elisângela Chicaroni Mattos, a quem considero uma irmã, meu muito obrigado por toda a sua ajuda, pelas nossas conversas no meio do dia quando eu “surtava com alguma coisa”; obrigado por sempre estar aberta a me ouvir e me apoiar, obrigado pelos cafezinhos também e por toda a sua consultoria informática.

À minha irmã MAIS VELHA, Dra. Bárbara da Conceição Abreu Silva (ela é “dotôra” ela), o que falar dessa pessoa incrível? Sem muito saber o que dizer aqui, quero te agradecer por ter me dado moradia quando precisei, sem nem pensar duas vezes. Obrigado por nossas conversas tarde da noite dialogando sobre os problemas da vida. Obrigado por me apoiar sempre, sem nunca deixar de me colocar com os pés no chão. Obrigado por ser esse ser humano incrível que, contribuiu e continua contribuindo para o meu crescimento, pessoal e profissional.

Ao meu irmãozinho, Felipe Eguti de Carvalho a quem considero as vezes como um filho também, obrigado por toda a parceria, pelo incentivo e até mesmo pelos desafios que você me impunha quando pedia ajuda. Você contribuiu muito para minha formação!

Ao meu “pai” Bruno Casali Sarzi, por todas as noites tomando café e discutindo sobre a qualidade duvidosa dos meus scripts, por toda a ajuda e apoio sempre, por suportar meus constantes descontentamentos com a pós-graduação como um todo e por ser essa pessoa maravilhosa.

Às minhas “filhas” Alisson Stefany Acero Valderrama (Chiquita), Fernanda Schneberger dos Santos (Fernande) e Caroline Assis Almeida (Caror), que me permitiram descobrir nas pequenas orientações a minha vocação pessoal, muito obrigado por tudo, obrigado por nunca ficarem bravas com minhas críticas e exigências constantes; obrigado por todo o apoio e carinho que vocês me deram/dão.

Ao meu “coorientado”, Gabriel Costa Medeiros, por sempre considerar meus comentários, mesmo quando exigi demais; pela amizade e parceria, pelo bom trabalho desenvolvido e por ter paciência com minha primeira experiência de orientação mais formal.

Aos amigos Ferdinando Sarzi Júnior e Simone Josely Casali Sarzi, por toda nossa amizade, por todos os conselhos e por toda a ajuda.

Aos amigos do BioMa: Lucas Tassoni Andrietta, Wecksley Leonardo de Souza, Diógenes Lodi Pinto e Lígia Garcia Mesquita, por toda a amizade e apoio.

Aos demais amigos do GMAB-FZEA/USP, por todo nosso companheirismo e pelas trocas de experiências.

A todos que direta, ou indiretamente, contribuíram para a conclusão deste doutorado, bem como na minha formação pessoal/profissional, meu muito obrigado!

À Faculdade de Medicina Veterinária e Zootecnia (FMVZ/USP) e à Faculdade de Zootecnia e Engenharia de Alimentos (FZEA/USP) por terem amparado, estruturalmente, minha pesquisa.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pela bolsa (Processo Nº: 2018/26468-3) que viabilizou o desenvolvimento deste estudo. O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“On closer examination, however, the hope of finding a “first” comes to grief because of the historically dynamic character of ideas. If we describe a result with sufficient vagueness, there seems to be an endless sequence of those who had something within the vague specifications. Even plagiarists usually introduce innovations! If we specify the idea or result precisely, it turns out that exact duplications seldom occur, so that every mathematical event is a “first”, and the priority question becomes trivial”
(MAY, 1975, p. 316)¹.

“Às vezes você tem que ir além do medo para ver a beleza que há do outro lado”.
(Papai Henry)².

¹ MAY, K.O. Historiographic vices II. Priority chasing. **Historia Mathematica**, vol. 2, no. 3, p. 315–317, 1975.

² O BOM DINOSSAURO; Direção: Peter Sohn. Produção: Denise Ream. Estados Unidos: Walt Disney Motion Pictures, 2016. 1 DVD.

RESUMO

BUSSIMAN, F. O. **Avaliação genética de cavalos Campolina: do pedigree à possibilidade de seleção genômica** [Genetic evaluation of Campolina horses: from pedigree to the genomic selection possibility]. 2021. 233 f. Tese (Doutorado em Ciências) – Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, 2021.

Embora os custos da genotipagem tenham diminuído ao longo dos últimos anos, em cavalos, chips comerciais ainda apresentam um preço elevado e, atualmente, há no mercado apenas duas alternativas disponíveis. Com efeito, o número de cavalos genotipados nos estudos publicados é baixo e, em geral, os pesquisadores encontram problemas quanto a representatividade das populações por parte dos animais genotipados. Assim, o objetivo do presente trabalho foi propor estratégias de genotipagem em populações de cavalos não genotipadas, prospectando, antes de investimentos financeiros, quais os melhores métodos para seleção genômica nesta espécie. Foi utilizada como material de estudo, a população de cavalos Campolina. Inicialmente um estudo sobre os fundadores, bem como a formação de linhagens no pedigree foi conduzido. Para tanto, foi utilizada análise de componentes principais e conseguiu-se determinar a ocorrência de três linhagens fundadoras na população recente desta raça. Além disso foi conduzido um estudo de redução de posto, no qual testaram-se 19 modelos sendo nove modelos de componentes principais; nove modelos fatores analíticos e um modelo multicaracterística padrão. Em abordagem frequentista, via máxima verossimilhança restrita em algoritmo de informação média, pôde-se determinar que o melhor modelo é um modelo fator analítico com cinco fatores. Foi ainda empregado um estudo utilizando análise confirmatória de fatores para determinar como as características mensuradas na população Campolina se combinam e quais os efeitos causais entre as características latentes que, de fato, causam os fenótipos observáveis. Como resultados, a qualidade morfológica tem um efeito causal sobre a morfologia que, por sua vez, tem um efeito causal sobre o andamento. A partir destes resultados, um estudo com modelos recursivos bem como estruturais permitiu concluir como os atributos de andamento se interconectam causando uns aos outros para criar a pontuação total de andamento. Por fim, um estudo com dados simulados foi feito. Para tanto foram simulados três cenários com diferentes estruturas populacionais e quatro características. Foram então testadas seis diferentes estratégias de genotipagem juntamente quatro diferentes modelos estatísticos para predições genômicas. Houve efeito significativo da interação entre característica e cenário, bem como da interação entre característica e estratégia de genotipagem sobre a acurácia de predição do valor genômico. Sobre a porcentagem de viés, houve efeito significativo apenas da estratégia de genotipagem. No tocante a modelos, os métodos em passo único foram mais acurados e menos viesados tanto para características contínuas quanto para binárias. Futuras implementações de seleção genômica em cavalos deveriam considerar a amostragem de animais com o maior número de progênies para genotipagem, implementando modelos de passo único semi-paramétricos para as predições genômicas.

Palavras-chave: linhagens; redução de posto; análise confirmatória de fatores; modelos de equações estruturais; estratégias de genotipagem.

ABSTRACT

BUSSIMAN, F. O. **Genetic evaluation of Campolina horses:** from pedigree to the genomic selection possibility [Avaliação genética de cavalos Campolina: do pedigree à possibilidade de seleção genômica]. 2021. 233 f. PhD. Thesis (Doctoral) – College of Veterinary Medicine and Animal Science, University of Sao Paulo, 2021.

Although genotyping costs have decreased over the last few years, in horses, commercial chips still carry a high price and, currently, there are only two alternatives available on the market. In fact, the number of genotyped horses in the published studies is low and, in general, researchers find problems regarding the representativeness of populations by the genotyped animals. Thus, the aim of the present work was to propose genotyping strategies in non-genotyped horse populations, prospecting, before financial investments, which are the best methods for genomic selection in this species. The Campolina horse population was used as study material. Initially a study on the founders as well as the formation of pedigree lines was conducted. Therefore, principal component analysis was used, and it was possible to determine the occurrence of three founder lines in the recent population of this breed. In addition, a study of rank reduction was conducted, in which 19 models were tested, being nine models of principal components; nine factor analytic models and one standard multi-trait model. In a frequentist approach, through restricted maximum likelihood in an average information algorithm, was found best model is a factor-analytic model with five common factors. A study using confirmatory factor analysis was also employed to determine how measured traits in the Campolina population are combined and what causal effects among latent traits causes the observable phenotypes. As a result, morphological quality has a causal effect on morphology which, in turn, has a causal effect on gait. From these results, a study with recursive as well as structural equation models allowed to conclude how the gait attributes interconnect causing each other to create the total gait score. Finally, a study with simulated data was carried out. For this, three scenarios with different population structures and four traits were simulated. Six different genotyping strategies were then tested along with four different statistical models for genomic predictions. There was a significant effect of the interaction between trait and scenario, as well as the interaction between trait and genotyping strategy on prediction accuracy. On the percentage of bias, there was only a significant effect of the genotyping strategy. About models, the single-step methods were more accurate and less biased for both continuous and binary traits. Future implementations of genomic selection in horses should consider sampling animals with the largest number of progenies for genotyping, implementing semi-parametric single-step models for genomic predictions.

Keywords: pedigree lines; rank reduction; confirmatory factor analysis; structural equation models; genotyping strategy.

SUMÁRIO

1. INTRODUÇÃO	15
1.1. Hipótese	19
1.2. Objetivos gerais.....	19
1.3. Objetivos específicos	19
2. REVISÃO DE LITERATURA.....	20
2.1. Um pouco de história.....	20
2.1.1. O cavalo Campolina.....	20
2.1.2. Melhoramento genético	30
2.2. Alguns conceitos estatísticos	45
2.2.1. Métodos multivariados.....	45
a) <i>Análise de componentes principais</i>	45
b) <i>Análise de fatores</i>	52
2.2.2. Modelos mistos	59
2.3. “Um pouquinho de inspiração”	67
2.4. Referências	69
3. FOUNDER ANALYSIS AND FAMILY IDENTIFICATION IN THE CURRENT CAMPOLINA HORSE POPULATION	84
3.1. Introduction.....	85
3.2. Material and Methods	86
3.2.1. Pedigree file	86
3.2.2. Founders and contributions.....	86
3.2.3. Lineage discovery	87
3.2.4. Phenotypes and data editing.....	88
3.2.5. Lineage effect.....	90
3.3. Results.....	91
3.3.1. Founders and contributions.....	91
3.3.2. Lineage discovery and its effect.....	92
3.4. Discussion	97
3.5. Conclusion	100
3.6. Implications	101
3.7. References.....	101
4. REDUCED RANK ANALYSIS OF MORPHOMETRIC AND FUNCTIONAL TRAITS IN CAMPOLINA HORSES: PRINCIPAL COMPONENTS AND FACTOR ANALYTICAL MODELS	104
4.1. Introduction.....	105
4.2. Material and Methods	108
4.2.1. Pedigree and phenotypic data	108
4.2.2. Statistical models	111
4.2.3. Multitrait mixed model (MTM)	112
4.2.4. Rationale for rank reduction	113
4.2.5. Principal components model (PCM).....	114
4.2.6. Factor analytical model (FAM).....	114

4.2.7.	Recall of breeding value prediction	115
4.2.8.	Model comparison	116
4.3.	Results.....	117
4.3.1.	Model comparisons	117
4.3.2.	Genetic parameters.....	120
4.3.3.	Sampling errors and rank correlations	125
4.4.	Discussion	128
4.5.	Conclusion	134
4.6.	References.....	134
5.	CONFIRMATORY FACTOR ANALYSIS AND STRUCTURAL EQUATION MODELS TO DISSECT THE RELATIONSHIP BETWEEN GAIT AND MORPHOLOGY IN HORSES	139
5.1.	Introduction.....	140
5.2.	Material and Methods	142
5.2.1.	Dataset and phenotype adjustment.....	142
5.2.2.	Confirmatory factor analysis.....	146
5.2.3.	Multitrait mixed model	148
5.2.4.	Structural equation model	149
5.3.	Results.....	149
5.3.1.	Measurement model.....	149
5.3.2.	Multitrait mixed model	152
5.3.3.	Structural equation model	154
5.4.	Discussion	157
5.5.	Conclusion	161
5.6.	References.....	161
6.	BAYESIAN RECURSIVE AND STRUCTURAL EQUATION MODELS TO INFER CAUSAL LINKS AMONG GAIT VISUAL SCORES IN CAMPOLINA HORSES	166
6.1.	Introduction.....	167
6.2.	Material and Methods	169
6.2.1.	Phenotypic and pedigree data	169
6.2.2.	Statistical analyses	171
a)	<i>Multitrait model.....</i>	<i>171</i>
b)	<i>Inductive causation algorithm.....</i>	<i>172</i>
c)	<i>Structural equation models</i>	<i>173</i>
6.3.	Results.....	174
6.3.1.	Convergence diagnoses.....	174
6.3.2.	Multitrait model	175
6.3.3.	Inductive causation algorithm.....	176
6.3.4.	Structural equation model	178
6.4.	Discussion	181
6.5.	Conclusion	185
6.6.	References.....	186

7. EVALUATING HOW TO IMPLEMENT GENOMIC SELECTION IN A NON-GENOTYPED POPULATION USING SINGLE-STEP METHOD COMBINED WITH KERNEL REGRESSIONS: A SIMULATION STUDY	192
7.1. Introduction.....	193
7.2. Material and Methods	195
7.2.1. Simulated scenarios	196
7.2.2. Genotyping strategies.....	202
7.2.3. Breeding value prediction	203
7.2.4. Statistical analyses	205
7.2.5. Model comparison	205
7.2.6. Validation scheme.....	208
7.3. Results and Discussion	209
7.3.1. Linkage disequilibrium and population structure	209
7.3.2. Genotyping strategy	211
7.3.3. Model comparison	214
7.4. Conclusion	218
7.5. References.....	219
8. CONSIDERAÇÕES FINAIS.....	227
APÊNDICE	230
APÊNDICE	232

1. INTRODUÇÃO

*“Disadvantage can be turned into an advantage. Four ounces can move 1,000 pounds”
(Sergeant Qiang)³.*

A humanidade vem, ao longo dos séculos, modificando os animais; acasalando-os de acordo com padrões desejados, e fazendo com que as diversas espécies domésticas se adequem às suas necessidades. O melhoramento animal toma por base a reprodução controlada dos animais domésticos, aqui principalmente os de interesse zootécnico, objetivando melhorar atributos, ou características, desejáveis nos rebanhos. A este processo dá-se o nome de seleção, e, este é um conceito que combina diversas áreas da ciência de acordo com a espécie e característica sendo selecionadas.

Podemos entender o Melhoramento Genético como sendo o ramo da ciência que surge da combinação da genética, estatística, fisiologia reprodutiva, ciência da computação e genética molecular. Do ponto de vista prático, o objetivo central do melhoramento genético animal é, portanto, identificar corretamente os pais da futura geração de indivíduos, no sentido de que esta geração seja superior em termos genético-fenotípicos a seus pais.

Devido a natureza multidisciplinar do melhoramento, o processo de identificação dos indivíduos geneticamente superiores, tem como consequência o entendimento da arquitetura genética das diversas características de interesse; a compreensão da dinâmica populacional dos rebanhos sob seleção; o aprimoramento de métodos estatístico-computacionais para melhorar as predições; e o aumento da rentabilidade da atividade pecuária.

Por muito tempo, a seleção dos indivíduos baseou-se em critérios subjetivos, produzindo pouco, ou nenhum, progresso genético nas populações. Entretanto com o advento da genética e seu alinhamento com a estatística, métodos mais precisos foram aplicados para prever o valor dos indivíduos do ponto de vista genético (chamado usualmente de valor genético aditivo). Inicialmente o método dos mínimos quadrados foi empregado para correção das observações para causas de variação conhecidas e, posteriormente, os índices de seleção proporcionaram a predição dos valores genéticos aditivos.

Foi somente com o surgimento das equações de modelos mistos, combinando a versatilidade dos mínimos quadrados, com o poder dos índices de seleção, que o melhoramento genético foi então impulsionado e, desde então, diversos métodos foram desenvolvidos e

³ MULAN; Direção: Niki Caro. Produção: Chris Bender; Tendo Nagenda; Jason T. Reed e Elizabeth Martin. Estados Unidos: Walt Disney Motion Pictures, 2020. 1 DVD.

aplicados, tanto para estimação de parâmetros genéticos, quanto para a predição dos valores genéticos aditivos.

De maneira simplificada, a metodologia dos modelos mistos pressupõe um modelo aditivo, no qual há efeitos considerados fixos bem como efeitos aleatórios. No caso do melhoramento genético animal, os efeitos aleatórios são, em geral, assumidos como tendo uma distribuição normal; com média zero e matriz de (co)variância proporcional ao parentesco entre todos os indivíduos em avaliação. Por muito tempo a forma de se mensurar o parentesco entre indivíduos considerou dados de pedigree. Não obstante, com o surgimento de técnicas moleculares os pesquisadores vislumbravam formas de se mensurar o parentesco entre os animais por meio da genotipagem.

Foi somente com a viabilização da genotipagem dos indivíduos, com painéis de marcadores moleculares do tipo SNP (do inglês, *single-nucleotide polymorphism* – polimorfismo de nucleotídeo único), que foi possível estimar de forma apropriada o parentesco, agora genômico, entre os animais. Surgiram então os primeiros estudos de seleção assistida por marcadores e, mais tarde, seleção genômica, bem como, os estudos de associação genômica ampla (buscando entender melhor as bases genéticas das características).

No início, a genotipagem com painéis incluindo milhares de SNPs era bastante onerosa, mas à medida que foi se popularizando teve seu custo barateado e, como resultado, o número de animais genotipados nos rebanhos aumentou. Assim como, aumentou também a quantidade de publicações tratando do uso de informações genômicas para seleção. Como o número de animais genotipados era, em geral, inferior ao número de animais fenotipados, inicialmente os estudos focavam na predição de valores genômicos aditivos, somente para os animais genotipados.

A metodologia mais difundida à época era o que chamamos de GBLUP (do inglês, *genomic best linear unbiased Predictor* – melhor preditor linear genômico não viesada). Ou algum outro método BLUP, com penalização para o grande número de marcadores. Surgiram então os primeiros trabalhos utilizando inferência Bayesiana para acomodar melhor o grande número de marcadores nas equações de modelos mistos, bem como, proporcionar maior poder preditivo para os efeitos dos marcadores.

Entretanto, em todos esses métodos era necessário primeiro “treinar” os modelos preditivos em uma população genotipada, com alta acurácia do valor genético (chamada de população de referência) e, posteriormente, validar as predições em uma população, relacionada geneticamente a população de referência (o que ficou conhecido em inglês como *multi step methods* – métodos de vários passos).

Recentemente, um grande avanço foi feito no sentido de se combinar as informações genômicas, fenotípicas e de pedigree, de modo que, a partir do parentesco entre os animais não-genotipados e os genotipados, são imputados relacionamentos genômicos para todos os animais em avaliação. Desta forma, é possível, em um passo único (*single-step*), obter-se o valor genético genômico aditivo dos indivíduos, combinando de forma ótima todas as fontes de informação. Este método é conhecido como ssGBLUP (do inglês, *single-step genomic best linear Predictor* – melhor preditor linear não viesado genômico de passo único).

Seja sob uma abordagem frequentista (escola clássica da estatística), seja em uma abordagem Bayesiana (escola probabilística da estatística), o ssGBLUP vem sendo apontado como um método bastante robusto para predições, diretas e indiretas, dos valores genéticos genômicos dos animais. Entretanto, quando a população genotipada não representa de forma satisfatória a população controlada pelo pedigree, problemas de estimabilidade são encontrados tanto para os componentes de variância, quanto para as predições dos GEBVs (do inglês, *genomic estimated breeding values* – valores genéticos genômicos estimados).

Embora os custos da genotipagem em larga escala venham diminuindo ao longo dos últimos anos, e a seleção genômica tenha sido empregada, com sucesso, aumentando a acurácia de predição dos GEBVs, e diminuindo o intervalo de gerações das populações sob seleção, existem diversas populações que nunca foram genotipadas e, portanto, ainda não foram desenvolvidas estratégias corretas de genotipagem dos animais. Um exemplo disso são as populações equinas que, geralmente, quando possuem animais genotipados, estes representam uma parcela muito pequena da população e altamente pré-selecionada.

O custo da genotipagem varia de acordo com a espécie, sendo que em equinos os painéis de SNPs ainda são bastante onerosos aos criadores, levando a uma certa relutância quando da implementação de seleção genômica, bem como limitações nos tamanhos amostrais dos estudos genômicos em cavalos. Desta forma, a pergunta “quem são os animais a serem genotipados que maximizariam a acurácia de predição dos GEBVs?” parece ser crucial para esta espécie, bem como para as demais, pois, tal resposta conduziria a uma correta aplicação de recursos, bem como uma utilização ótima dos recursos dispendidos.

Atualmente, técnicas como simulação e modelagem são alternativas custo-efetivas para estudos de estratégia de genotipagem, uma vez que possibilitam mimetizar populações reais, mas em situações mais controladas, testar diferentes formas de amostragem dos candidatos à genotipagem, bem como, avaliar diferentes modelos quanto a sua habilidade preditiva de acordo com características de interesse na população estudada. Uma das grandes vantagens da simulação é possibilidade de se conhecer o valor genético verdadeiro: TBV (do inglês, *true*

breeding value – valor genético verdadeiro), possibilitando a aplicação de diversas técnicas estatísticas para a correta definição de modelos e estratégias de genotipagem.

Embora bastante efetiva, a simulação requer cuidado, pois a população simulada precisa, minimamente, representar a população real, para que as conclusões sejam válidas. A correta simulação do desequilíbrio de ligação (associação não aleatória entre pares de marcadores) depende da formulação adequada da estrutura da população, do esquema de seleção, da natureza da característica sendo simulada, e da correta formulação do genoma a ser simulado. Portanto, é preciso que se conheça a fundo, genética e fenotipicamente, a população em estudo.

Estudos de pedigree, como por exemplo estrutura populacional e probabilidades gene origem podem auxiliar na correta simulação da estrutura populacional desejada. Estudos de estimação de componentes de variância, aliados a técnicas multivariadas e inferência causal podem auxiliar na compreensão de quais características devem, de fato, ser simuladas e, por fim, a aplicação de diferentes estratégias de genotipagem, bem como modelagem, pode auxiliar na implementação futura de seleção genômica nessas populações não-genotipadas, uma vez que seria possível a correta definição da ótima estratégia de amostragem dos candidatos à genotipagem.

Neste sentido, este estudo está organizado sobre o raciocínio proposto acima, sob o qual uma série de estudos separados, mas complementares, foi conduzida a fim de se conhecer profundamente o pedigree da população em estudo; a natureza genético-quantitativa das características que de fato necessitam ser estudadas, levando-se em conta todas as características de interesse na população estudada. Por fim, diversas estratégias de genotipagem, de acordo com a dinâmica de seleção da espécie estudada foram testadas.

Para este estudo a população escolhida foi a de cavalos Campolina, uma raça de cavalos marchadores brasileira, de importância nacional e internacional. O cavalo Campolina representa atualmente o sexto maior rebanho de cavalos de raça pura no Brasil, tem importância histórico-cultural, bem como socioeconômica no estado de Minas Gerais e vem atraindo olhares do mercado equestre internacional, por tratar-se de um animal de grande porte, com andamento confortável à montada, adequando-se as preferências médias dos criadores internacionais.

Por fim, quebrando o protocolo, essa introdução objetivou apenas despertar a curiosidade para o tema, uma vez que conhecimentos histórico-estatísticos, necessários ao entendimento deste estudo, são abordados na revisão de literatura, bem como nas introduções individuais de cada capítulo. Assim, fica a cargo do leitor o estudo aprofundado dos principais

conceitos abordados na revisão de literatura que, pode ser “pulada”, sem prejuízo, por aqueles mais bem preparados.

1.1. Hipótese

Estudos de pedigree, aliados a estudos de redução de posto, e inferência causal auxiliam na correta simulação para definição da melhor estratégia de genotipagem em populações não-genotipadas.

1.2. Objetivos gerais

Diante da escassez de informações genóticas acerca da arquitetura genética das características de interesse em equinos, este trabalho visa utilizar técnicas de simulação para avaliar diferentes estratégias de genotipagem, viabilizando a correta determinação dos animais a serem genotipados, antes dos recursos serem movimentados para a genotipagem.

1.3. Objetivos específicos

- a) Entender a formação de linhagens no pedigree real da população de cavalos Campolina;
- b) Utilizar inferência causal para entender a relação entre os fenótipos de interesse nesta população de cavalos;
- c) Utilizar as metodologias de Componentes Principais e Análise de Fatores para redução de posto em modelos multicaracterística;
- d) Simular uma estrutura populacional que represente a estrutura da população de cavalos Campolina;
- e) Simular características contínuas e binárias, sendo as binárias com a mesma distribuição do tipo de marcha existente na raça;
- f) Avaliar diferentes estratégias de genotipagem para maximizar a acurácia em estudos de seleção genômica utilizando a metodologia *single-step*; e
- g) Comparar estratégias de modelagem paramétricas e semi-paramétricas em características de interesse em equinos.

2. REVISÃO DE LITERATURA

*“If I have seen further it is by standing on the shoulders of Giants”
(Isaac Newton, 1675)⁴.*

Antes de serem abordados os conceitos fundamentais a cerca deste trabalho, uma revisão do ponto de vista histórico é apresentada, tanto para a formação da raça de cavalos Campolina, quanto para a formação do Melhoramento Genético enquanto ciência. A primeira se justifica pela curiosidade, uma vez que este contexto histórico pode, no futuro, ajudar a elucidar resultados de pesquisas como, por exemplo, a origem genética destes cavalos. A segunda se justifica em uma afirmação feita por Fisher em 1959:

Mais atenção deveria ser dada a história da ciência, tanto por cientistas, quanto por historiadores, em especial por biólogos [em sentido mais amplo – profissionais das áreas biológicas em geral], isso deveria ser uma tentativa consciente de entender os pensamentos dos grandes mestres do passado, para ver em que circunstâncias ou meio intelectual suas ideias foram formadas, onde que eles pegaram o caminho errado ou pararam de maneira precoce no caminho certo. Um senso de continuidade e o caráter progressivo e cumulativo de uma ciência em avanço é o melhor profilático que eu posso sugerir contra as alternâncias maniaco-depressivas do culto à moda e ao impulso, que ameaçam sufocar os esforços científicos de uma grande nação, por mais gigantes que eles sejam (FISHER, 1959, p. 17, tradução nossa).

Esta revisão de literatura trata de se aprofundar nas bases estatísticas fundamentais dos três principais métodos utilizados ao longo do desenvolvimento deste trabalho. Conceitos específicos são abordados de forma oportuna nos respectivos capítulos. Assim serão tratados nesta revisão os conceitos gerais relacionados à: 1) Análise de Componentes Principais; 2) Análise de Fatores; e 3) Modelos Mistos.

2.1. Um pouco de história

2.1.1. O cavalo Campolina

A raça de cavalos Campolina tem sua origem na derrota de Cassiano Antônio da Silva Campolina (1836-1904) em uma cavallhada no ano de 1870 (ARAÚJO, 2013). De acordo com Martins da Silva (2010), as cavallhadas são um teatro equestre que representam as batalhas entre

⁴ NEWTON, I. [Carta enviada a Sir Robert Hooke]. Destinatário: Robert Hooke. Cambridge, 5 fev. 1675. 1 carta. Disponível em: <<https://www.newtonproject.ox.ac.uk/view/texts/normalized/OTHE00101>>. Acesso em: 04 out. 2021.

cristãos e mouros na Idade Média. Introduzidas no Brasil pelos colonizadores, caíram no gosto popular sob duas formas: as encamisadas e as cavalarias (DEL PRIORE, 1994). As encamisadas tratavam de representar táticas de guerra e, as cavalarias constituíam-se de justas e diversas outras competições equestres (DEL PRIORE, 1994).

Por ocasião da derrota, Cassiano decidiu criar uma raça de cavalos, para que viesse a ganhar a competição no futuro, tarefa na qual foi ajudado pelo Imperador Dom Pedro II, que lhe presenteou, no mesmo ano (1870), com uma égua, de nome Medéia, a qual estava prenhe de um garanhão Andaluz (ARAÚJO, 2013; ABCCCAMPOLINA, 2017a). Medéia deu à luz um potro, de nome Monarca, o qual foi garanhão na Fazenda do Tanque (de propriedade de Cassiano), este potro é considerado o fundador da raça Campolina (ABCCCAMPOLINA, 2017a).

De acordo com Procópio, Bergmann e Costa (2003), Monarca foi utilizado em um rebanho de éguas descendentes de animais ibéricos, trazidos pelos conquistadores europeus, das raças Andaluz, Lusitano, Bérbere e Sorraia. Posteriormente, visando os objetivos iniciais (originalmente, porte elevado e, mais tarde, andamento) foram introduzidos animais Anglo-Normando, Clydesdale, Holsteiner, American Saddle Horse e Mangalarga (PROCÓPIO; BERGMANN; COSTA, 2003), bem como, Puro-Sangue Inglês e Percheron (EDWARDS, 1994).

É digno de nota que, nos tempos do Brasil Colônia, quando os navegadores portugueses e espanhóis tentavam trazer animais de raças “nobres” como o Lusitano e o Andaluz, a maioria dos animais eram roubados ainda no porto, então os navegadores passaram a trazer animais da raça Garrano, além do Sorraia que, para a época, eram considerados menos nobres; é plausível, portanto, que as raças nacionais, como o Campolina, possuam em sua composição animais Garrano também (*informação verbal*)⁵.

De acordo com Araújo (2013), Cassiano faleceu antes que fosse à forra na cavallhada; falecendo em 1904, deixou toda sua fortuna para que fosse construída uma obra em benefício da região de Entre Rios de Minas-MG (cidade em que nasceu e faleceu), nascia ali o Hospital Cassiano Campolina (ABCCCAMPOLINA, 2017b). A tropa que formara, Cassiano a deixou para o Coronel Joaquim Pacheco de Resende (ABCCCAMPOLINA, 2017a), responsável pela fundação das linhagens Gas e Passatempo, esta última em 2013, pertencia a Gastão Resende, tataraneto do testamenteiro de Cassiano, quem esclareceu que a famosa marca “CC” nos animais Campolina significa Constância e Caráter (ARAÚJO, 2013).

⁵ Informação fornecida por Gobesso, A.A.O. na palestra “Panorama das raças brasileiras de equídeos”, durante o V Simpósio de Produção Animal – O Potencial dos Equídeos Marchadores Brasileiros, Pirassununga, 2016.

Na década de 30 um grupo de criadores fundou o Consórcio Profissional Cooperativo dos Criadores da Raça Campolina, objetivando unificar o trabalho de seleção que cada criador empregava; após cerca de 70 anos de desenvolvimento da raça Campolina, em 1951 surgiu a Associação Brasileira dos Criadores do Cavalo Campolina (ABCCCampolina), definindo o primeiro padrão racial para direcionar os esforços em aperfeiçoar a raça de acordo com critérios oficiais da ABCCCampolina (ABCCCAMPOLINA, 2017a).

Historicamente, o cavalo Campolina está ligado ao desenvolvimento socioeconômico-cultural do estado de Minas Gerais (REZENDE; MOURA, 2004), estado que detém o maior número de cabeças de animais da raça. Além disso, os animais Campolina vem sendo exportados para diversos países como: México, Venezuela, Estados Unidos e Alemanha (DE OLIVEIRA BUSSIMAN et al., 2018c). Em termos gerais, o cavalo Campolina figura na 6ª posição em número de animais de raça pura no Brasil (BRASIL, 2016) e o *stud book* da ABCCCampolina conta com mais de 100.000 animais registrados.

Uma vez que a raça de cavalos Campolina surgiu em MG, todos os rebanhos, em cada estado, guardam algum relacionamento genético com o rebanho de MG (Figura 1). Por outro lado, recentemente o parentesco médio entre estados aumentou nos animais nascidos a partir de 1990, enquanto o parentesco dentro de estados diminuiu (DE OLIVEIRA BUSSIMAN et al., 2018b). Isso se deve a uma maior comercialização de animais e novas tecnologias reprodutivas sendo adotadas pelos criadores, entretanto o parentesco dentro de estados permanece mais elevado que o parentesco entre estados (na ordem de 3:1). Além disso, essa distribuição de rebanhos aparentados em diferentes níveis no território nacional evidencia a estratificação da população (TEEGEN; EDEL; THALLER, 2009).

Figura 1 - Parentesco médio entre cavalos Campolina de acordo com os estados de nascimento e regiões geográficas para toda a população (A) e para os animais nascidos entre 1990 e 2016 (B)

Fonte: de Oliveira Bussiman et al. (2018b). Abreviações: N = Norte; NE = Nordeste; CO = Centro-Oeste; SE = Sudeste; S = Sul; PA = Pará; TO = Tocantins; AL = Alagoas; BA = Bahia; CE = Ceará; MA = Maranhão; PE = Pernambuco; SE = Sergipe; DF = Distrito Federal; GO = Goiás; MT = Mato Grosso; ES = Espírito Santo; MG = Minas Gerais; RJ = Rio de Janeiro; SP = São Paulo e PR = Paraná.

De acordo com o regulamento do serviço do registro genealógico da ABCCCampolina de 2018, o cavalo Campolina é um “equino de tipo sela, proporcional, equilibrado em sua aparência, nobre, atento e dócil, apresentando linhas harmoniosas e bem definidas” (ABCCCAMPOLINA, 2018, Cap. XXV, p. 25). O mesmo padrão racial preconiza, ainda, que os animais devam possuir “constituição forte e vigorosa, estrutura proporcional e harmoniosa, com musculatura definida e bem distribuída, pele fina, pelos finos e macios [...] temperamento dócil e ativo, de fácil manejo e equitação” (ABCCCAMPOLINA, 2018, Cap. XXV, p. 25).

A altura mínima para registro (aos 36 meses) de machos inteiros é de 1,52 m; de 1,45 m para fêmeas e, para castrados é de 1,50 m; sendo considerado ideal que machos possuam altura de 1,60 m e fêmeas de 1,54 m. Para concessão do registro é feita uma pontuação baseada em critérios morfológicos e de andamento variando de 0 a 500 pontos, dos quais 200 são concedidos por aspectos de andamento e, os outros 300 ficam distribuídos da seguinte forma: 35 pontos para Aparência geral; 70 pontos para Cabeça; 30 pontos para Pescoço; 85 pontos para Tronco e, 80 pontos para membros (ABCCCAMPOLINA, 2018, Cap. XXV, p. 25-26).

São aceitas, para fins de registro, todas as pelagens. Ao todo, sem considerar nomenclaturas para pelagens compostas, existem 20 diferentes pelagens na raça Campolina, a pelagem predominante é a Baia, seguida da Alazã e, posteriormente, da Castanha (Figura 2). Considerando-se a ocorrência de pelagem composta (e incluindo neste grupo a pelagem

Pampa), 21.673 animais apresentam esta característica, o que representa 22,54% do total de animais com informação para pelagem no banco de dados da ABCCCampolina.

Figura 2 - Distribuição do número de animais por pelagem na população de cavalos Campolina

Fonte: Bussiman (2021).

Uma das características mais marcantes do cavalo Campolina é a cabeça, sendo esta, parte fundamental do julgamento morfológico nas competições da raça, o que pode ser notado pela ênfase dada às características avaliadas na cabeça no momento da concessão de registro (23,33% - equiparável a ênfase dada aos membros 26,67%). De acordo com o padrão racial do cavalo Campolina, uma boa cabeça deve ser:

Harmoniosa e proporcional no seu comprimento, largura e altura. Com contornos ósseos nítidos e suaves. Deve apresentar características sexuais definidas [...] [, ter forma] trapezoidal tendendo a retangular [...] [, perfil] retilíneo na região da fronte e suavemente convexilíneo na região do chanfro [...] [e as orelhas,] com pavilhão de abertura mediana e terminadas preferencialmente em forma de ponta de lança [(Figura 3)] (ABCCCAMPOLINA, 2018, Cap. XXV, p. 25).

Figura 3 - Campeões do julgamento de “Melhor Cabeça” da 39ª Semana Nacional do Cavalo Campolina Marchador de 2019. Rei das Marias, reservado grande campeão melhor cabeça (A); Gata do Anaití, reservada grande campeã melhor cabeça (B)

Fonte: A: Túlio (2019); B: Mariotti (2019).

Outra característica peculiar dos cavalos Campolina é seu andamento, que pode se apresentar sob duas modalidades distintas, ambas aceitas para fins de registro; são elas: marcha batida e marcha picada (ABCCCAMPOLINA, 2018, Cap. XXV, p. 26). Por definição o andamento chamado de marcha no Brasil é um andamento dissociado com momentos de tríplice apoio dos membros no solo e, a dissociação, significa que cada membro se encontra em uma fase diferente da trajetória descrita pelo membro durante a locomoção (DE OLIVEIRA BUSSIMAN et al., 2019). Esses andamentos podem ser entendidos como uma forma de passo acelerado, uma vez que são andamentos a quatro tempos (NICODEMUS; CLAYTON, 2003).

Fruto da dissociação, animais marchadores executam uma ampla gama de combinações de apoios durante a locomoção, e alguns deles são mais abundantes em um andamento do que em outro como, por exemplo, o tempo de apoios diagonais, que é maior na marcha batida do que na marcha picada, ou o tempo de apoios laterais que é maior na marcha picada do que na batida. Vale ressaltar ainda que, em ambos os andamentos, os tríplices apoios estão presentes intercalados a apoios bipedais (DE OLIVEIRA BUSSIMAN et al., 2019).

Há, ainda, outras diferenças metabólicas e genéticas quanto a esses andamentos, para uma revisão mais aprofundada recomenda-se a leitura da sessão 2.3.5.5 em de Oliveira Bussiman (2018). Entretanto, do ponto de vista morfológico não são esperadas grandes diferenças (Figura 4), exceto em medidas angulares, como ângulo escapulo-solo e escapulo-umeral, maiores em animais de marcha picada; comprimento de quartela e garupa (maiores nos animais de marcha batida) e largura de peito e garupa, também maiores em animais de marcha batida (SANTIAGO et al., 2014).

Figura 4 - Grandes campeões de marcha na 39ª Semana Nacional do Cavalo Campolina Marchador de 2019. Malibu do LM – Marcha Picada (A); Ferro Elétrico do Haras Ventania – Marcha Batida (B)

Fonte: A: Haras (2020); B: Duarte (2014).

Do ponto de vista de seleção genética, a raça de cavalos Campolina nunca esteve sob um programa formal de seleção com predição de valores genéticos, a seleção vem sendo feita de forma empírica, baseada em resultados de competições e como principal ferramenta os criadores se valem do uso irrestrito da endogamia (DE OLIVEIRA BUSSIMAN et al., 2018b). Em estudo recente de estimação de parâmetros genéticos para características lineares e de andamento observou-se que, de maneira geral, características morfométricas em cavalos Campolina são geneticamente positivamente correlacionadas (Figura 5), ao passo que escores de andamento apresentam correlação genética negativa com a maioria das características morfométricas (DE OLIVEIRA BUSSIMAN et al., 2018c).

Figura 5 - Correlações genéticas entre as características mensuradas ao registro nos cavalos Campolina

Fonte: de Oliveira Bussiman et al. (2018c). Abreviações: AC = Altura de Cernelha; AD = Altura de Dorso; AG = Altura de Garupa; Acos = Altura de Costados; CP = Comprimento de Pernas; ACAG = Razão entre AC e AG; LC = Largura da Cabeça; LP = Largura do Peito; LA = Largura de Ancas; CC = Comprimento da Cabeça; Cpes = Comprimento do Pescoço; CDL = Comprimento do Dorso-Lombo; CDLCcorp = Razão entre CDL e Ccorp; CG = Comprimento da Garupa; CE = Comprimento da Espádua; Ccorp = Comprimento Corporal; PT = Perímetro Torácico; PC = Perímetro da Canela; e AT = Andamento Total.

Outra particularidade do cavalo Campolina é que o efeito do técnico sobre a variabilidade da pontuação de andamento é bastante elevado (DE OLIVEIRA BUSSIMAN et al., 2018c, 2018a) e, além disso, diferentes características medidas por um mesmo técnico apresentam correlações variando de -0,23 a 0,66 (DE OLIVEIRA BUSSIMAN et al., 2018a). Desta forma além de contribuir para uma baixa herdabilidade (Figura 6) dos escores visuais de andamento, o efeito de técnico não seria o mesmo em características diferentes (Figura 7). Por

outro lado, as correlações genéticas entre cada um dos atributos do andamento e a pontuação total é positiva, facilitando a seleção apenas para pontuação total (Figura 7).

Figura 6 - Porcentagem da variância fenotípica explicada pela ação aditiva dos genes (U), Porcentagem da variância fenotípica explicada por técnico (TEC) e Porcentagem da variância fenotípica de origem ambiental (E), em características ligadas a andamento em cavalos Campolina

Fonte: de Oliveira Bussiman et al. (2018a).

Figura 7 - Correlações genéticas (acima da diagonal) e entre técnicos (abaixo da diagonal) em características de andamento em cavalos Campolina

Fonte: de Oliveira Bussiman et al. (2018a). Abreviações: AT = Andamento Total; C = Cômulo; E = Estilo; R = Regularidade; De = Desenvolvimento; e Di = Dissociação.

No tocante à seleção, do ponto de vista do melhoramento genético os métodos empregados atualmente pela ABCCCampolina são bastante empíricos e não científicos, levando a pouca (ou mesmo nenhuma) melhoria de cunho genético na população como um todo, o que fica evidente quando olhamos para as tendências genéticas publicadas por de Oliveira Bussiman et al. (2018c). Por outro lado, parece que isso beneficia a população, pois, uma vez que a seleção praticada não elevou a endogamia nos garanhões, estes ainda assegurariam níveis suficientes de variabilidade genética na população, o que podemos verificar pelas medidas de diversidade genética apresentadas por de Oliveira Bussiman et al. (2018b).

O número elevado de características mensuradas no momento do registro genealógico dos animais dificulta a seleção genética de cavalos Campolina. Uma vez que muitas das características introduzem redundância nos modelos de avaliação multicaracterística, e a utilização de modelos unicaracterística levaria a ciclos de seleção independentes muito longos, técnicas que combinem essas características tenderiam a facilitar as decisões de seleção, quando da possível instalação de um programa de melhoramento genético de cavalos Campolina.

Como veremos ao longo do presente estudo, o uso de Índices de Seleção (HAZEL, 1943; HAZEL; LUSH, 1942; SMITH, 1936), técnicas multivariadas, como: Análise Confirmatória de Fatores (BROWN, 2015; PEÑAGARICANO et al., 2015) e Análise de Componentes Principais (BUZANSKAS et al., 2013; SOLAR DIAZ et al., 2020), ou mesmo sua combinação, ajudaria a facilitar a implementação de programas de melhoramento nesta população. Se aliarmos a este conjunto de métodos o poder dos Modelos de Equações Estruturais (HAAVELMO, 1943; VALENTE et al., 2010, 2013; VALENTE; ROSA, 2013; WRIGHT, 1918, 1921), podemos identificar as características nas quais os esforços de seleção devem ser empregados, uma vez que isso levaria a melhoria global da população.

Algumas “máximas do cavalo” foram confirmadas neste trabalho, como por exemplo, a ideia de que “forma prediz função”. Por outro lado, o enfoque aqui não deveria ser interpretado como um apego irrestrito à seleção morfológica, uma vez que o registro genealógico “carimba” o animal com o rótulo “confere com o padrão”, a seleção deveria ser praticada, dentro dos animais registrados, para as características que, de fato, precificam os animais, como andamento ou tipo de andamento, na presente população. Desta forma, seria sim possível a obtenção do cavalo “bonito e bom de sela”, outra máxima para qual o presente estudo tentou dar uma solução mais refinada.

Por fim, vale acrescentar aqui, que todos os resultados levantados com este breve histórico do cavalo Campolina, bem como as hipóteses cunhadas com os resultados dos estudos apresentados nos capítulos subsequentes, não possuem o objetivo de serem definitivos. Este trabalho tem muito mais o papel de servir como um direcionamento e levantar inquietações científicas no leitor, no sentido de despertar o interesse para novas possibilidades de pesquisa, bem como respaldar criadores e amantes de cavalos a melhor selecionarem seus animais.

2.1.2. Melhoramento genético

A ideia de acasalar o melhor, ou mais valioso, animal não é nova. De fato, Varro (116-27 a.C. – filósofo romano) menciona que desde 2.000 a.C. o homem já tinha a ideia de

reproduzir somente certos indivíduos e, esta ideia de uma forma ou de outra, cria as condições para o desenvolvimento do melhoramento genético durante os séculos 18 e 19 (GROSU et al., 2014). Até os anos 1700, não existia nenhum meio sistemático de seleção para reprodução, então o acasalamento dos animais era feito com outros animais na vizinhança baseado em critérios pré-definidos que não mudavam, ou permaneciam similares, acasalamento após acasalamento (OLDENBROEK; VAN DER WAAIJ, 2015).

É atribuída a Sir Robert Bakewell (1725-1795) a introdução da ideia de se criar registros acurados da performance dos animais e, desta forma seria possível selecioná-los objetivamente (OLDENBROEK; VAN DER WAAIJ, 2015), esta ideia fica mais evidente quando H. Branth, um fazendeiro dinamarquês afirmou, em 1891, que “[...] a habilidade de uma vaca para produzir mais ou menos gordura no leite, a partir do alimento que ela come, depende da hereditariedade” (GROSU et al., 2014, p. 1 tradução nossa).

A afirmação de Branth trata de algo ainda bastante novo para a época: hereditariedade. Podemos pensar que à época estavam a acreditar que a resposta dos indivíduos a fatores ambientais dependia da genética, o que não está errado; mas, se nos valemos de um trabalho bastante importante publicado 32 anos antes por Charles Darwin podemos entender essa afirmação sob a ótica do que hoje chamamos evolução.

De fato, em seu trabalho *On the Origin of Species by means of Natural Selection* (Sobre a Origem das Espécies por meio de Seleção Natural em tradução livre), Darwin afirmava que, a depender de fatores ambientais, diferentes características são selecionadas, geração após geração (DARWIN, 1859). Entretanto, os conceitos darwinistas só fizeram mais sentido com o advento da genética nos trabalhos de Mendel apresentados em 1865, e publicados no ano seguinte.

Em duas apresentações orais, Gregor Mendel cunhou a base da genética moderna. Seu trabalho *Versuche über Pflanzen-Hybriden* (Experimento sobre plantas híbridas em tradução livre) (MENDEL, 1866), não teve muito impacto na época, mas foi Mendel quem propôs o mecanismo de herança dominante em organismos diploides e, agora podíamos explicar os conceitos de Darwin sob a ótica dos “fatores herdáveis” de Mendel, pois o ambiente ao inviabilizar a sobrevivência de determinados indivíduos, ou favorecer a reprodução de outros, selecionava certos “fatores” nos pais.

Mendel mostrou que o material genético é herdado dos dois pais de forma independente e, por isso, indivíduos diploides possuem duas cópias do mesmo gene, do qual apenas uma é passada à progênie (MENDEL, 1866). Mendel ainda mostrou que cada uma dessas cópias (alelos) pode ser: 1) dominante (quando apenas um alelo é necessário para a expressão do gene);

2) recessiva (quando são necessários dois alelos para que o gene seja expresso); ou 3) aditiva (quando cada alelo contribui para a expressão do gene). Esses resultados permaneceram sem reconhecimento até os anos 1900 (OLDENBROEK; VAN DER WAAIJ, 2015), quando seus trabalhos foram “redescobertos” por Correns (1900); de Vries (1900a, 1900b, 1900c) e Tschermak-Seysenegg (1900a, 1900b).

Embora as ideias de Mendel estejam intimamente ligadas ao melhoramento genético, enquanto ciência, outros conceitos (necessários) foram desenvolvidos muito antes de seu “redescobrimento” (GIANOLA; ROSA, 2015). Foi Francis Galton, em 1886, estudando a hereditariedade da altura, quem observou que:

Parecia, a partir desses experimentos, que a progênie não tendia a assemelhar-se com seus pais em tamanho, mas [tendia] a ser mais mediana que eles – ser menor que os pais, se os pais fossem altos; ser mais altos que os pais, se os pais fossem muito pequenos (GALTON, 1886, p. 246, tradução nossa).

Este fato observado por Galton é, atualmente, chamado de regressão à média e, segundo Gianola e Rosa (2015), está intimamente ligado aos conceitos de herdabilidade e resposta à seleção, pois, a diferença na média de uma dada característica em grupos de pais extremos assemelha-se ao diferencial de seleção e a diferença correspondente na média da progênie desses grupos é equivalente a resposta à seleção.

Inspirado nos trabalhos de Galton, Karl Pearson derivou uma série de fórmulas com o que seria a fundação matemática da evolução das características, e introduziu o conceito de mistura de distribuições para explicar a assimetria (PEARSON, 1894, 1903), ou desvio da normalidade, uma vez que, se o material experimental é realmente homogêneo (provém da mesma distribuição), tal desvio é fruto da evolução:

A assimetria talvez surja do fato de que as unidades agrupadas de forma conjunta, no material mensurado não são realmente homogêneas. Pode acontecer que tenhamos uma mistura de 2, 3, ... n grupos homogêneos, cada um deles desviado de sua própria média simetricamente e de uma maneira tal, que sejam representados pela curva normal com acurácia suficiente (PEARSON, 1894, p. 72, tradução nossa).

Por outro lado, eu não posso pensar que, para o problema da evolução, seja desnecessário dissecar a curva mais simétrica dada pelas medidas [encontrar as várias curvas normais e simétricas que compõem a curva assimétrica dos dados]. Sempre haverá o problema: o material é homogêneo e uma evolução real está acontecendo, ou o material é uma mistura? (PEARSON, 1894, p. 99, tradução nossa).

O primeiro caso de herança mendeliana em animais foi reportado em 1906 por Toyama Kametaro enquanto estudava a cor de casulo do bicho da seda (KAMETARO, 1906), mas até

então, havia estudos de herança mendeliana somente em características categóricas. Foi G. Udny Yule quem “reconciliou” o mendelismo com a variação contínua (GIANOLA; ROSA, 2015) em dois trabalhos: Yule (1902) e Yule, (1906), respectivamente.

A partir dos trabalhos de Yule, Godfrey Harold Hardy escreveu ao editor da revista *Science* para esclarecer sobre o ponto de equilíbrio das frequências alélicas (HARDY, 1908) na ausência de fatores que nelas interfiram (mutação, seleção, migração e deriva gênica):

Ao editor da *Science*: Eu estou relutante de me meter numa discussão de um assunto no qual não sou especialista e eu deveria esperar que o ponto muito simples que desejo fazer fosse familiar aos biólogos. Entretanto, algumas observações do Sr. Udny Yule, para as quais o Sr. R.C. Punnett chamou minha atenção, sugerem que isso ainda valha a pena ser feito. [...] Suponha que Aa é um par de caracteres mendelianos, sendo A dominante e, em uma dada geração o número de puros dominantes (AA), heterozigotos (Aa) e puros recessivos (aa) é $p : 2q : r$. Finalmente, suponha que os números são suficientemente grandes, assim os acasalamentos podem ser considerados como aleatórios, suponha que os sexos estão igualmente distribuídos entre as três variedades [AA, Aa, aa] e que eles são igualmente férteis. Um pouco de matemática, do tipo de tabela de multiplicação é suficiente para mostrar que, na próxima geração os números serão: $(p + q)^2 : 2(p + q)(q + r) : (q + r)^2$, ou $p_1 : 2q_1 : r_1$ (HARDY, 1908, p. 49, tradução nossa).

O equilíbrio das frequências alélicas permanece atribuído exclusivamente a Hardy, até que Crow (1999) mostrou que o problema havia sido abordado de forma independente e, anterior, pelo médico alemão Wilhelm Weinberg no mesmo ano em que Hardy publicou sua carta ao editor da revista *Science*. Os resultados de Weinberg (1908) eram os mesmos que os de Hardy (1908), de tal sorte que o equilíbrio das frequências alélicas passou a ser conhecido como equilíbrio de Hardy-Weinberg, dando crédito ao trabalho dos dois cientistas.

À luz dos resultados de Galton, Pearson e Yule, Ronald Fisher introduziu, em um paper de notável dificuldade (GIANOLA; ROSA, 2015), o modelo infinitesimal no qual ele afirmou que, assumindo-se um número infinito de loci, com infinitos valores e, cada um deles com um efeito infinitesimalmente pequeno, o resultado é uma distribuição normal (FISHER, 1919). Neste mesmo paper ele apresentou ainda o conceito de análise de variância para decompor a variância genética em: variância genética aditiva e componentes de dominância e epistasia (FISHER, 1919; GIANOLA; ROSA, 2015).

O trabalho de Fisher, publicado em 1919 tem um impacto tão grande na área de genética quantitativa que muitos o consideram o paper de maior influência nesta área. Uma curiosidade e, de certo modo, um incentivo, é que quando Fisher enviou seu manuscrito para publicação, ele foi rejeitado, e os revisores disseram que seu trabalho era totalmente inútil, e eles (os

revisores) não viam alguma forma de que seus resultados servissem a algum propósito útil (*informação verbal*)⁶.

Utilizando path Analysis e correlação, à luz do trabalho de Fisher, Wright (1922) deduziu o coeficiente de endogamia e apresentou a métrica mais usada de parentesco aditivo entre dois indivíduos, mais tarde revisitada por Malécot (1948), que empregou, ao invés de correlação, probabilidades para encontrar tal métrica que, permanece até os dias atuais como “duas vezes a probabilidade de que um alelo em um locus aleatório seja idêntico por descendência em um par de indivíduos” (GIANOLA; ROSA, 2015, tradução nossa).

Em uma série de trabalhos, como, por exemplo, *Evolution in Mendelian Populations* (Evolução em Populações Mendelianas em tradução livre) (WRIGHT, 1931), Wright apresentou as propriedades de populações mendelianas, tomando como base os trabalhos de Hardy (1908) e Weinberg (1908). Ele ainda teve um impacto muito grande ao mostrar que o equilíbrio das frequências alélicas em uma população finita, diferente do proposto no equilíbrio de Hardy-Weinberg (população infinitamente grande), responderia a equação de Fokker-Planck da difusão na Física.

Juntamente com Fisher e Wright, John Burdon Sanderson Haldane é reconhecido como um dos “pais fundadores” da genética de populações (CHARLESWORTH, 2017). Em uma série de trabalhos publicados entre os anos 1920 e 1930, Haldane apresentou importantes resultados na formulação matemática da seleção, bem como no cálculo da taxa de mutação espontânea dos genes (CHARLESWORTH, 2017; COOK; TURNER, 2020). Sua série de trabalhos *A mathematical theory of natural and artificial Selection* (Uma teoria matemática para a seleção natural e artificial em tradução livre) foi sumarizada, por ele mesmo, em um livro *The Causes of Evolution* (As Causas da Evolução - HALDANE, 1932).

De acordo com Gianola e Rosa (2015), outra ideia de Fisher a impactar o melhoramento é o “efeito médio de substituição alélica” (FISHER, 1919, 1930), essa ideia foi introduzida formalmente aos melhoristas por Jay Laurence Lush, em diversos cursos, tarefa na qual foi assistido, mais tarde, por um excepcional livro escrito por Douglas Scott Falconer. Falconer teve várias outras contribuições para a genética de populações e para a genética de camundongos em particular, entretanto seu livro *Introduction to Quantitative Genetics* (Introdução a Genética Quantitativa - FALCONER, 1960) foi sua maior contribuição (HILL; MACKAY, 2004).

⁶ Informação fornecida por Gianola, D. na palestra “A brief and ongoing history of statistics in agricultural breeding”, durante o congresso How the Biotechnology can Vanish World Hunger, Pirassununga, 2019.

Lush foi o grande responsável por colocar as ideias de Fisher, Wright e Haldane juntas em um contexto agropecuário. Durante os anos em que esteve na Iowa State University, Lush contribuiu, excepcionalmente, para o desenvolvimento metodológico-científico do melhoramento genético enquanto ciência. Suas notas de aula foram publicadas como um livro: *Animal Breeding Plans* (Planos de Melhoramento Animal - LUSH, 1937) e, em uma série de trabalhos publicados entre 1921 e 1968 ele desenvolveu o conhecimento sobre a formação de linhagens e sobre o uso da endogamia. De acordo com ele mesmo, seu trabalho mais importante foi *Family merit and individual merit as bases for selection* (Méritos familiar e individual como base para seleção - LUSH, 1947).

Sob a orientação de Lush, Lanoy Nelson Hazel apresentou a primeira aplicação dos Índices de Seleção para predição do mérito genético em animais (HAZEL; LUSH, 1942; HAZEL, 1943), adaptando a metodologia proposta, inicialmente para seleção de plantas, por Smith (1936) e, introduziu a ideia, e o conceito de correlação genética, no melhoramento genético. Ainda sob a orientação de Lush, Charles Roy Henderson deduziu as equações de modelos mistos (HENDERSON, 1949, 1950) para predição do valor genético, que foram introduzidas em dois congressos pelo próprio Henderson (SCHAEFFER, 1991).

De acordo com Schaeffer (1991), Henderson foi inspirado por um paper de Lush de 1933: *The Bull Index Problem in the Light of Modern Genetics* (O problema do Índice de touro à luz da genética moderna - LUSH, 1933) no processo de derivação das equações de modelos mistos (MME). De fato, Henderson foi um marco divisor no melhoramento genético, suas MME hoje chamadas de BLUP (do inglês: *best linear unbiased prediction* – melhor predição linear não-viesada, em tradução livre) se tornaram o método mais utilizado na predição do valor genético. De acordo com o próprio Henderson, ele tratou de reunir o poder dos mínimos quadrados generalizados, para estimar os efeitos fixos, com os recursos atraentes dos índices de seleção para predizer os efeitos aleatórios (HENDERSON, 1990).

Embora as MME tenham sido apresentadas em 1949 (e seu método de derivação em 1950), a prova de que a solução para os efeitos fixos era a mesma que nos mínimos quadrados generalizados veio somente em 1959 (HENDERSON et al., 1959), com ajuda de um dos orientados de Henderson, Shayle Robert Searle (SCHAEFFER, 1991) e, em 1963 Henderson publicou a prova de que a solução para os efeitos aleatórios de suas MME era a mesma que a dos índices de seleção (HENDERSON, 1963). Mesmo com essas provas de que as MME geravam soluções do tipo BLUE (do inglês, *best linear unbiased estimator* – melhor estimador linear não viesado em tradução livre) para os efeitos fixos, e do tipo BLUP para os efeitos aleatórios, elas não foram utilizadas para avaliação genética até 1970, devido às limitações

computacionais e necessidade de um sólido conhecimento em álgebra matricial (HENDERSON, 1990):

Eu tentei sem muito sucesso aparente persuadir melhoristas de que o método [suas MME] não era nada mais do que o índice de seleção usando $y - X\beta^0$ no lugar de $y - X\beta$ e, que isso levava a menores variâncias dos erros de predição do que o índice de seleção usando qualquer outro $y - X\beta^*$, em que β^0 é uma solução de mínimos quadrados generalizados e, $X\beta^*$ é qualquer outro estimador não viesado de $X\beta$. É claro que deveria ser reconhecido que poucos melhoristas eram competentes em álgebra linear, particularmente em álgebra matricial em 1949 e, além disso, nenhum tinha instalações computacionais que poderiam lidar com as grandes matrizes necessárias. De fato, o método [suas MME] não foi usado na avaliação de touros de inseminação artificial em Cornell até 1970, embora estivéssemos avaliando touros desde o início da inseminação artificial em Nova York no início dos anos 1940. BLUP por métodos de modelos mistos [em referência a suas MME] é agora o método padrão para avaliação em gado de leite, gado de corte e suínos na maioria dos países (HENDERSON, 1990, p. 8, tradução nossa).

Nos anos 60, Henderson introduziu o modelo vaca que ficou mais tarde conhecido como modelo animal (GIANOLA; ROSA, 2015). Neste caso, ao invés de apenas se predizerem os valores genéticos dos touros, prediziam-se os valores genéticos de todos os animais envolvidos em análise, isto é, progênie, ancestrais e colaterais. Tal fato foi conseguido incorporando-se a matriz de relacionamento genético (A) entre todos os indivíduos nas MME (HENDERSON, 1963, 1973). Entretanto, era preciso encontrar a inversa da matriz do numerador do parentesco (A^{-1}) de Wright (WRIGHT, 1922), o que demanda muito poder computacional se o número de indivíduos é grande. Para contornar este problema, Henderson e Quaas apresentaram um algoritmo para obtenção direta de A^{-1} da partir de um arquivo de pedigree (HENDERSON, 1975, 1976; QUAAS, 1976).

Ainda nos anos 60, muito interesse se despertou nos efeitos genéticos maternos em diversas espécies de produção (GIANOLA; ROSA, 2015). À luz dos trabalhos de Dickerson; Kempthorne; Koch; e Clark, que introduziram os conceitos por trás das características com efeitos maternos (DICKERSON, 1947; KOCH; CLARK, 1955; KEMPTHORNE, 1957), Willham propõe, em 1963, uma decomposição para o fenótipo de um indivíduo (digamos X , filho de uma mãe W), em: $P_X = G_{dx} + E_{dx} + G_{mW} + E_{mW}$, desta forma o fenótipo do indivíduo X , passa a ser o resultado da contribuição direta de seu valor genético (G_{dx}) mais a contribuição direta do ambiente onde ele é criado (E_{dx}); mais as contribuições maternas (mãe W): genéticas (G_{mW}) e ambientais (E_{mW}) (WILLHAM, 1963).

Falconer propôs, ainda, incluir o fenótipo das mães como covariável nos modelos de avaliação genética, permitindo assim levar em consideração os efeitos maternos (FALCONER, 1965). Em 1970, Van Vleck apresentou as equações dos índices de seleção para características com efeito materno (VAN VLECK, 1970) e, em seu livro *Selection Index and Introduction to Mixed Model Methods* (Índice de Seleção e Introdução aos Métodos de Modelos Mistos) Van Vleck apresentou, de maneira extensiva e didática, como incluir os efeitos maternos nas MME (VAN VLECK, 1993). Em 1997, Koerhuis e Thompson revisaram uma série de modelos e algoritmos para a estimação dos efeitos maternos, bem como seus componentes de variância, conseguindo integrar os modelos direto de Falconer e o modelo de Willham (KOERHUIS; THOMPSON, 1997).

Em 1976, a metodologia BLUP pelas MME foi estendida para avaliação multicaracterística por Henderson e Quaas, considerando informações de diferentes características mensuradas em diferentes indivíduos geneticamente relacionados (HENDERSON; QUAAS, 1976). Devido ao grande número de equações na matriz de coeficientes considerando o modelo animal e o modelo animal multicaracterística, em 1980 Quaas e Pollak apresentaram o modelo animal reduzido, considerando na matriz **A**, apenas os animais com progênie tendo dados, mas não considerando as progênies dessas progênies (QUAAS; POLLAK, 1980). Na prática é construída a matriz **A**, mas apenas para os animais com informação fenotípica, não considerando-se seus ancestrais no processo de predição (*informação verbal*)⁷.

Outra importante contribuição de Henderson foi na estimação de componentes de variância, uma vez que tanto para os clássicos índices de seleção, quanto para suas MME era preciso que os componentes de variância (residual e de cada possível efeito aleatório) fossem conhecidos. O primeiro método para estimarem-se os componentes de variância foi o método dos momentos, igualando-se os quadrados médios da tabela de ANOVA à suas esperanças, mas como uma consequência de dados desbalanceados várias aproximações eram utilizadas (HENDERSON, 1990). Georg Waddel Snedecor descreveu, em 1946, como obter as esperanças dos quadrados médios em modelos aninhados (SNEDECOR, 1946) e Hazel, juntamente com Clair E. Terril, estenderam essa ideia para modelos mistos com dados desbalanceados e apenas um efeito aleatório (HAZEL; TERRILL, 1945).

⁷ Informação fornecida por Albuquerque, L.G. durante a disciplina “Teoria dos Índices de Seleção e Introdução aos Modelos Mistos” ministrada na Faculdade de Ciências Agrárias e Veterinárias da Universidade Estadual Paulista Júlio de Mesquita Filho, Jaboticabal, 2019.

Em 1953, Henderson introduziu os métodos I, II e III para estimação dos componentes de variância corrigindo as esperanças dos quadrados médios no método II (HENDERSON, 1953) e, de acordo com Henderson (1990), Searle advogou em favor destes métodos em seus artigos e alguns livros, como por exemplo seu livro *Variance Components* (Componentes de Variância - SEARLE; CASELLA; MCCULLOCH, 1992), publicado mais tarde. Tratando do problema de minimizar a variância de amostragem em dados desbalanceados, Calyampudi Radhakrishna Rao introduziu os estimadores quadráticos não viesados, de norma mínima (MINQUE) e de variância mínima (MIVQUE), em uma série de quatro publicações (RAO, 1970, 1971a, 1971b, 1972).

De acordo com Searle (1995), as primeiras aplicações do método de máxima verossimilhança (FISHER, 1925) para estimação de componentes de variância foram propostas por S. Lee Crump em dois trabalhos (CRUMP, 1947, 1951), mas devido a condições de saúde ele não pôde continuar seu trabalho (HENDERSON, 1990). Foi somente em 1967 que uma solução geral para as estimativas dos componentes de variância nos modelos mistos foi deduzida por Herman Otto Hartley e Jonnagadda Nalini Kanth Rao (HARTLEY; RAO, 1967). Henderson (1973), encontrou o mesmo resultado integrando o vetor de efeitos aleatórios e o vetor de termos residuais nas MME e, mais tarde, tanto os resultados de Hartley e Rao, quanto os de Henderson foram comprovados pelo algoritmo de *expectation maximization* (maximização de expectativa, em tradução livre) adaptado por Dempster, Laird e Rubin (1977).

Contornando o viés da máxima verossimilhança (ML), em 1971 a máxima verossimilhança restrita (REML do inglês, *Restricted Maximum Likelihood*) foi proposta por Robin Thompson que, motivado por um paper de Henderson (CUNNINGHAM; HENDERSON, 1968), se aproximou de Desmond Patterson para escrever uma nota de correção (THOMPSON, 1969), durante as discussões para a nota de correção, Thompson resolveu propor uma metodologia para estimar os componentes de variância com dados desbalanceados (THOMPSON, 2019). Inspirado em um trabalho anterior de Patterson (PATTERSON, 1964), Thompson e Patterson resolvem utilizar a função de verossimilhança, não dos dados completos, mas em contrastes dos resíduos, dando origem a REML (PATTERSON; THOMPSON, 1971).

Em uma série de trabalhos publicados em 1986, Gianola, Foulley e Fernando introduziram a Inferência Bayesiana no melhoramento animal para estimação de componentes de variância e predição dos valores genéticos (FERNANDO; GIANOLA, 1986; GIANOLA; FERNANDO, 1986; GIANOLA; FOULLEY; FERNANDO, 1986). Entretanto, estes trabalhos se suportam em estudos anteriores que mostraram que o Índice de Seleção, derivado pelas MME, é um estimador de Bayes (RØNNINGEN, 1971) e na relação entre o BLUP e os

estimadores Bayesianos, pela qual, se os efeitos genéticos aditivos e residuais são normais e independentes e, a informação apriori é vaga, o BLUP e a metodologia Bayesiana fornecem os mesmos estimadores (DEMPFLE, 1977).

Ainda em 1986, Fernando e Gianola, derivaram uma série de propriedades da média condicional como critério de seleção (FERNANDO; GIANOLA, 1986). De certa forma, um método ótimo para seleção deveria maximizar, não somente a média dos indivíduos selecionados, mas também a probabilidade de correta classificação dos indivíduos com base no valor genético predito, com respeito a seu valor genético verdadeiro (PORTNOY, 1982). No entanto, seja trabalhando com fenótipos ajustados para os efeitos fixos ($y - X\hat{\beta}$), seja utilizando a ML (ou REML), os estimadores BLUP apresentam apenas a primeira propriedade (FERNANDO; GIANOLA, 1986). Desta forma, o método BLUP seria não viesado para a média, ou valor esperado, mas não para o valor verdadeiro.

Para contornar o problema do aparente viés na classificação dos animais ao fazemos o ranqueamento com $E(u|y) = \text{BLUP}(u)$, Fernando e Gianola (1986) invocam a propriedade de translação invariante nas soluções BLUE e BLUP (HARVILLE, 1977). Essa propriedade, entre outras coisas, nos garante que independente da função/transformação empregada para encontrar $y - X\hat{\beta}$, a função de máxima verossimilhança resultante é invariante. Assim, as inferências sobre as variâncias na distribuição conjunta de u e $y - X\hat{\beta}$, seriam também invariantes (FERNANDO; GIANOLA, 1986).

Por consequência, sob pressuposição de normalidade é possível mostrar que mesmo quando a esperança do valor genético verdadeiro é desconhecida (ou varia de animal para animal), podemos utilizar a metodologia BLUP (DEMPFLE, 1977; GIANOLA; FERNANDO, 1986), mas devemos ter em mente que a classificação dos animais fica então dependente da variabilidade associada às predições (FERNANDO; GIANOLA, 1986). Assim, utilizamos o valor genético predito dos animais associado a uma medida de dispersão em torno desta predição, a qual por convenção chamamos acurácia (essa métrica será definida mais adiante).

Uma extensão do BLUP tradicional, para medidas repetidas no mesmo animal resultou no modelo de repetibilidade (QUAAS, 1984), entretanto como neste modelo medidas adjacentes são assumidas igualmente correlacionadas, uma abordagem longitudinal foi proposta em 1982 de forma independente por três pesquisadores (SCHAEFFER, 2004). Em setembro de 1982, Henderson Jr. argumentou que quando existe um coeficiente de regressão para cada indivíduo, devido às medidas repetidas, se os indivíduos são tratados como aleatórios,

tais coeficientes de regressão devem ser tratados, também, como aleatórios (HENDERSON JR, 1982). Laird e Ware (1982), apresentaram uma extensão do método de ML via algoritmo de maximização de esperanças para dados longitudinais e alternativas Bayesianas para a estimação dos parâmetros.

Em 1989, foi proposto o modelo de funções de covariância para “reduzir” a dimensão das equações na análise via modelos de dimensão infinita, por Kirkpatrick e Heckman, considerando medidas repetidas descritas por uma função n-dimensional (KIRKPATRICK; HECKMAN, 1989). Meyer e Hill mostraram que uma função de covariância nada mais é do que a “versão”, de dimensão infinita, de uma matriz de covariâncias, portanto, são um caso especial dos modelos de regressão aleatória, quando as informações fenotípicas são tratadas por meio de funções contínuas (MEYER; HILL, 1997). Posteriormente, Ignacy Misztal mostrou que modelos de regressão aleatória são mais acurados que modelos de repetibilidade, ou multicaracterística, mas são marginalmente mais complicados (MISZTAL, 2006).

Em 1977 foi introduzido o algoritmo de maximização de esperanças (EM) para resolução iterativa de sistemas de equações complexos por Dempster, Laird e Rubin, inicialmente via ML (DEMPSTER; LAIRD; RUBIN, 1977), mais tarde estendido para REML (EM-REML). Em essência, o algoritmo EM encontra um parâmetro de dispersão Φ , que maximiza a esperança condicional de uma função dos dados, $f(y|\Phi)$, condicionada às observações y . Dez anos após, foi apresentado por Graser, Smith e Tier um algoritmo livre de derivadas via REML (DF-REML), utilizando eliminação de Gauss para resolução do sistema (GRASER; SMITH; TIER, 1987). Embora o algoritmo DF-REML tenha facilitado a implementação de avaliações com maior número de animais ele somente foi estendido, em 1991 por Meyer, para avaliações multicaracterística (MEYER, 1991).

Para evitar o cálculo do traço de matrizes grandes como nos algoritmos EM-REML ou DF-REML, em 1995 foi proposto o algoritmo de informação média (AI, do inglês *Average Information*) via REML (AI-REML) por Johnson e Thompson, substituindo as segundas derivadas, já tradicionais, pela média da informação observada e esperada com respeito aos parâmetros de interesse (JOHNSON; THOMPSON, 1995). A partir desses algoritmos uma série de softwares foi desenvolvida para estimação de componentes (co)variância e predição de valores genéticos sob uma diversidade de modelos, como por exemplo: DFREML (MEYER, 1988); MTDFREML (BOLDMAN; VAN VLECK, 1991; VAN VLECK; BOLDMAN, 1993); DMU (JENSEN; MADSEN, 1994); VCE (GROENEVELD, 1994); ASREML (GILMOUR et

al., 1996); MTGSAM (VAN TASSELL; VAN VLECK, 1996); BLUPF90 (MISZTAL et al., 2002); WOMBAT (MEYER, 2007); entre outros.

Recordando os trabalhos de Neimann-Sorensen e Robertson (1961) e Smith (1967), em 1990, Lande e Thompson introduzem o conceito de seleção assistida por marcadores (MAS, do inglês *Marker Assisted Selection*). A ideia central deste método era incluir nas tradicionais MME, marcadores com efeito direto nas características quantitativas, o que deveria permitir um aumento da acurácia de predição e, além disso, marcadores sem efeito poderiam ser incluídos desde que estivessem em desequilíbrio de ligação com os marcadores com efeito direto (LANDE; THOMPSON, 1990). Mais tarde, este conceito permitiu o desenvolvimento da seleção genômica, mas antes era preciso que, a partir dos marcadores fosse possível inferir o relacionamento entre os animais.

Em 1997, um estudante de doutorado da *University of Guelph* introduziu em sua tese, *Alternative methods for defining relationship, assigning haplotypes and measuring linkage in animal breeding* (Métodos alternativos para definição de relacionamento, atribuir haplótipos e mensurar ligação em melhoramento animal, em tradução livre), uma maneira de se utilizar informações genóticas para se estimar o parentesco entre indivíduos, publicando no mesmo ano um paper (NEJATI-JAVAREMI; SMITH; GIBSON, 1997) que é tido como a primeira proposição de seleção genômica (*informação verbal*)⁸.

Em 2001, Meuwissen, Hayes e Goddard mostram que era possível estimar o valor genético dos indivíduos a partir apenas de informações fenotípicas e genóticas (MEUWISSEN; HAYES; GODDARD, 2001). O princípio se baseava no fato de que “definido em termos de efeitos médios, o valor genético de um indivíduo é igual à soma dos efeitos médios dos genes que ele carrega, sendo a soma feita para cada par de alelos em cada locus (FALCONER; MACKAY, 1996).

Zhang et al. (2007), mostraram que é possível incluir matrizes de relacionamento genômico nas MME via algoritmo DF-REML e, em 2008 VanRaden introduziu a matriz de relacionamento genômico **G** e o método GBLUP (do inglês *Genomic Best Linear Unbiased Predictor*), no qual a tradicional matriz de parentesco **A** é substituída pela matriz **G** (VANRADEN, 2008). A matriz **G**, proposta por VanRaden, é a mais utilizada nos estudos genômicos atualmente; sua criação se baseia em:

⁸ Informação fornecida por SARGOLZAEI, M. durante o curso “Applied Genomic Selection in Livestock: Methods and Tools”, Universidade Estadual Paulista Júlio de Mesquita Filho, Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal, 2019.

Seja $\mathbf{M}$ uma matriz que especifica quais alelos cada indivíduo herdou. As dimensões de $\mathbf{M}$ são o número de indivíduos (n) pelo número de *loci* (m) [isto é $\mathbf{M}$ é ($n \times m$)]. [...] Se os valores de $\mathbf{M}$ são -1, 0 e 1 para homozigoto, heterozigoto e o outro homozigoto, respectivamente [...] [e] $\mathbf{P}$ contém as frequências alélicas expressas como desviadas de 0.5 e multiplicadas por 2, assim a coluna i de $\mathbf{P}$ é $2(p_i - 0.5)$ [p_i é a frequência alélica no *locus* i]. Subtraindo $\mathbf{P}$ de $\mathbf{M}$, $\mathbf{Z}$ é obtida, a qual define os valores médios dos efeitos dos alelos como 0. [...] $\mathbf{G}$ pode ser obtida por ao menos 3 métodos [aqui apresentado apenas o mais usado]. O primeiro usa a fórmula $\mathbf{G} = \mathbf{Z}\mathbf{Z}^T / 2 \sum_{i=1}^m p_i(1-p_i)$. A divisão por $2 \sum_{i=1}^m p_i(1-p_i)$ escala $\mathbf{G}$ para ser análoga à matriz de numerador de parentesco $\mathbf{A}$ (VANRADEN, 2008, p. 4416, tradução nossa).

VanRaden introduziu ainda em seu trabalho de 2008 o modelo SNP-BLUP (do inglês *Single Nucleotide Polymorphism Best Linear Unbiased Prediction*), no qual ao invés da predição dos valores genéticos, são preditos os efeitos de substituição alélica para cada *locus* (LIU et al., 2016). Em 2009, Strandén e Garrick mostraram que os métodos GBLUP e SNP-BLUP são equivalentes para os animais de referência (STRANDÉN; GARRICK, 2009) e, mais tarde, em 2014 veio a prova, por Taylor, de que estes métodos também são equivalentes para animais jovens genotipados e sem fenótipo (TAYLOR, 2014).

A equivalência entre estes dois métodos pode ser mostrada provando que eles dois modelam a mesma razão de variâncias, pois a variância dos marcadores é igual a variância genética aditiva dividida pelo somatório do produto das frequências alélicas multiplicado por dois; isto é: $\sigma_a^2 = \sigma_u^2 / 2 \sum_{i=1}^m p_i(1-p_i) \therefore \sigma_e^2 / \sigma_u^2 = 2 \sum_{i=1}^m p_i(1-p_i) \sigma_e^2 / \sigma_a^2$, em que: σ_a^2 é a variância devida aos marcadores; σ_u^2 é a variância genética aditiva da característica; σ_e^2 é a variância residual da característica e p_i é a frequência alélica do i -ésimo marcador (VANRADEN, 2008; STRANDÉN; GARRICK, 2009; TAYLOR, 2014).

Possibilitando incluir, em uma única avaliação animais não genotipados e genotipados, em 2009 Misztal, Legarra e Aguilar apresentaram, em dois trabalhos, a melhor predição linear não viesada genômica de passo único (ssGBLUP, do inglês *single-step Genomic Best Linear Predictor*), propondo uma maneira de combinar de forma eficiente as matrizes $\mathbf{A}$ e $\mathbf{G}$ em uma única matriz, chamada de $\mathbf{H}$ (LEGARRA; AGUILAR; MISZTAL, 2009; MISZTAL; LEGARRA; AGUILAR, 2009). A ideia inicial era particionar a matriz $\mathbf{A}$ (considere 1 para animais não genotipados e 2 para animais genotipados) de tal sorte que a submatriz correspondente aos animais genotipados fosse substituída pela matriz $\mathbf{G}$, isto é:

$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{G} \end{bmatrix}$ (MISZTAL; LEGARRA; AGUILAR, 2009), mas, utilizando um argumento

Bayesiano, e maximizando a distribuição dos valores genéticos dos indivíduos não genotipados condicional aos indivíduos genotipados, isto é, $p(\mathbf{u}_1 | \mathbf{u}_2) = N(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2, \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})$, resultou na matriz (LEGARRA; AGUILAR; MISZTAL, 2009):

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{21}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{bmatrix}.$$

Em 2010, de forma independente, dois trabalhos mostram que a inversa de $\mathbf{H}$ é bastante simples. Aguilar, Misztal, Johnson, Legarra, Tsuruta e Lawlor utilizaram-se da matriz $\mathbf{H}$ proposta por Legarra, Aguilar e Misztal (2009) e calcularam sua inversa utilizando propriedades da inversa de matrizes particionadas (AGUILAR et al., 2010). Christensen e Lund tratando a criação de $\mathbf{H}$ como um problema de dados perdidos (*missing data*), a encontram como sendo

$$(\text{CHRISTENSEN; LUND, 2010}): \mathbf{H} = \begin{bmatrix} \mathbf{G}_w & \mathbf{G}_w\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \\ \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{G}_w & \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{G}_w\mathbf{A}_{11}^{-1}\mathbf{A}_{12} + \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \end{bmatrix},$$

em que $\mathbf{G}_w = (1-w)\mathbf{G} + w\mathbf{A}_{11}$ (w aqui pode ser interpretado como um ponderador para o efeito poligênico contido em $\mathbf{A}_{11}$). Em ambos os casos a inversa de $\mathbf{H}$ é dada por:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix} \quad (\text{AGUILAR et al., 2010; CHRISTENSEN; LUND, 2010}).$$

Uma vez de posse de matriz $\mathbf{H}^{-1}$, diversos autores trataram de estudar os efeitos da compatibilidade entre $\mathbf{G}$ e $\mathbf{A}$, na predição dos valores genéticos genômicos, bem como na acurácia de predição (POWELL; VISSCHER; GODDARD, 2010; CHEN et al., 2011; VITEZICA et al., 2011). Estes trabalhos mostraram que incluir uma pequena parte da matriz $\mathbf{A}_{22}$ na matriz $\mathbf{G}$ leva a inclusão de um ligeiro efeito poligênico residual na nova matriz $\mathbf{G}$ (CHRISTENSEN; LUND, 2010); Vitezica et al. (2011), mostraram que, se a parcela da população genotipada está sob seleção é plausível incluir a média do valor genético dos animais genotipados no modelo, o que resulta em: $\mathbf{G}^* = \mathbf{G} + \mathbf{1}\mathbf{1}^T (\overline{\mathbf{A}_{22}} - \overline{\mathbf{G}})$; Powell, Visscher e Goddard (2010) e Vitezica et al. (2011), implementaram ajustes através do parâmetro F_{ST} (WRIGHT, 1949) e; Chen et al. (2011), mostraram que diferentes formas de se construir a matriz $\mathbf{G}$ levam a diferentes compatibilidades com a matriz $\mathbf{A}_{22}$.

Com efeito, ao se considerar os resultados destes estudos, a matriz $\mathbf{H}^{-1}$ passou a ser construída da seguinte maneira (MISZTAL et al., 2010, 2013):

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{A}_{12}^{-1} \\ \mathbf{A}_{21}^{-1} & \mathbf{A}_{22}^{-1} + \left\{ \tau(\alpha \mathbf{G} + \beta \mathbf{A}_{22} + \gamma \mathbf{I} + \delta) - \omega \mathbf{A}_{22}^{-1} \right\} \end{bmatrix}, \text{ em que: } \alpha \text{ e } \beta \text{ controlam o quanto do}$$

efeito poligênico residual de $\mathbf{A}_{22}$ é incluído em $\mathbf{G}$ (CHRISTENSEN; LUND, 2010); γ e δ controlam o viés nas predições (CHEN et al., 2011; VITEZICA et al., 2011); τ e ω controlam o “peso” da informação genômica nas predições (MISZTAL et al., 2010, 2013). Para uma discussão mais aprofundada acerca da determinação dos parâmetros α , β , γ , δ , τ e ω , recomenda-se a leitura de Misztal et al. (2010, 2013); Vitezica et al. (2011) e Lourenço et al. (2020).

Recentemente, o método ssGBLUP foi estendido para uma abordagem Bayesiana, permitindo estimar diretamente os efeitos de substituição alélica. Liu, Goddard, Reinhardt e Reents propuseram anexar ao vetor $\mathbf{u}$ (vetor de solução para os valores genéticos nas MME), o vetor $\mathbf{g}$ (vetor dos efeitos de substituição alélica), desta forma é possível solucionar as MME para obter, não somente os valores genéticos, mas também, os efeitos de substituição alélica, de tal sorte que para o vetor $\mathbf{g}$ é possível utilizar algum método de regressão Bayesiana e $\hat{\mathbf{u}}$ leva em conta $\hat{\mathbf{g}}$ (LIU et al., 2014).

Assumindo que $\mathbf{h}^T = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \mathbf{g}]$, esses autores mostraram que: $\text{Var}(\mathbf{h}^T) = \mathbf{H}\sigma_u^2$,

$$\text{sendo que } \mathbf{H}^{-1} \text{ é dada por (LIU et al., 2014): } \mathbf{H}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{A}_{12}^{-1} & \mathbf{0} \\ \mathbf{A}_{21}^{-1} & \mathbf{A}_{22}^{-1} + \left(\frac{1}{k} - 1 \right) \mathbf{A}_{22}^{-1} & -\frac{1}{k} \mathbf{A}_{22}^{-1} \mathbf{Z} \\ \mathbf{0} & -\frac{1}{k} \mathbf{Z}^T \mathbf{A}_{22}^{-1} & \mathbf{B}^{-1} + \frac{1}{k} \mathbf{Z}^T \mathbf{A}_{22}^{-1} \mathbf{Z} \end{bmatrix},$$

em que $\mathbf{B}^{-1}$ permite que seja assumido qualquer modelo de regressão Bayesiano para $\mathbf{g}$.

Fernando, Dekkers e Garrick, propuseram imputar as covariáveis de cada marcador para os animais não genotipados através de: $\hat{\mathbf{M}}_1 = \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{M}_2$, em que $\hat{\mathbf{M}}_1$ é a matriz de covariáveis imputadas para os animais não genotipados e, $\mathbf{M}_2$ é a matriz de covariáveis observadas para os animais genotipados (FERNANDO; DEKKERS; GARRICK, 2014). Fernando et al. (2016), revisitaram esses métodos e suas particularidades, seja do proposto por Liu et al. (2014), seja do proposto por Fernando, Dekkers e Garrick (2014), concluindo que o segundo é mais versátil em modelar os efeitos dos marcadores diretamente, além de não necessitar de inversão de matrizes.

Por fim, esta revisão não tem como objetivo abranger todos os aspectos minuciosamente do melhoramento genético enquanto ciência, mas fornecer o contexto histórico em que os principais avanços metodológicos se deram, na opinião do autor. Para leituras mais aprofundadas o leitor é encorajado a ler a revisão escrita por Gianola e Rosa (2015), bem como a literatura citada.

2.2. Alguns conceitos estatísticos

2.2.1. Métodos multivariados

A maioria das técnicas estatísticas tratam somente da variação em uma única variável aleatória, sendo chamadas de univariadas. Por outro lado, existem técnicas capazes de considerar, simultaneamente, diversas variáveis interrelacionadas; técnicas as quais são chamadas de multivariadas (MANLY, 2008). De acordo com Jhonson e Wichern (2018), o conjunto de técnicas estatísticas conhecidas como métodos multivariados se presta a problemas relacionados a: 1) Estrutura de Covariância de um vetor aleatório (por meio da matriz de covariâncias e/ou de correlações); 2) Agrupamentos; e 3) Reconhecimento e Classificação de Padrões. Nesta sessão serão abordadas duas técnicas que incorrem no item 1, são elas: Análise de Componentes Principais e Análise de Fatores.

a) Análise de componentes principais

A análise de componentes principais (PCA – *Principal Components Analysis*) foi inicialmente descrita em 1901, quando Pearson, trabalhando com ajuste do plano de regressão em regressões múltiplas, propõe que “o plano de melhor ajuste a um sistema de pontos é perpendicular ao menor eixo do elipsoide de correlações [...] [e] a reta de melhor ajuste a um sistema de pontos coincide em direção com o eixo máximo do elipsoide de correlações” (PEARSON, 1901, p. 569, tradução nossa).

Entretanto, uma aplicação prática veio muito mais tarde em 1933, quando Hotelling publicou um artigo no *Journal of Educational Psychology* (HOTELLING, 1933), tratando de encontrar um conjunto de variáveis independentes que explicassem um outro conjunto de variáveis dependentes entre si, de modo que o conjunto independente não excedesse, em número, o conjunto dependente. A este conjunto de variáveis independentes deu-se o nome de componentes principais (PCs – *Principal Components*).

Pelo nome da revista, na qual foi publicado o primeiro trabalho sobre a PCA, tem-se uma ideia da área na qual as aplicações da PCA foram mais expressivas (BARTHOLOMEW, 2010). Como, em essência, trata-se de uma técnica para reexpressar o conjunto de dados original, com menos variáveis (LATTIN; CARROLL; GREEN, 2011a); desde o seu desenvolvimento a PCA tem sido usada em diversas áreas e com diversas finalidades, desde redução de dimensão, até tornar possível a visualização gráfica de problemas multivariados complexos.

Por conta das vantagens de visualização gráfica da PCA, ela já foi aplicada em estudos de diversidade genética (por exemplo, HONDA et al., 2002; COSTA et al., 2005), até mesmo com dados genômicos (ZHANG et al., 2018; GRILZ-SEGER et al., 2019; LI; RALPH, 2019). Aplicada tanto a matriz de parentesco aditivo via pedigree, quanto na matriz de parentesco genômico é possível explicar como os animais se inter-relacionam de forma fácil e isso se deve ao fato de que a PCA também é uma técnica para identificar padrões nos dados (IEZZONI; PRITTS, 1991; LATTIN; CARROLL; GREEN, 2011a).

Para entender como funciona a PCA é necessário, porém, recuperarmos alguns conceitos de álgebra matricial, todavia uma conceituação formal e rígida poderia deixar o assunto tedioso. Podemos definir uma matriz por um conjunto de dados arranjados em linhas e colunas (RENCHEER; SCHAAALJE, 2008) e, em uma definição menos formal, podemos dizer que “matrizes fazem coisas a vetores” (STAR, 2019, tradução nossa), por “coisas”, deve-se entender escalonamento e rotação, dentre outros.

Para uma exemplificação considere os vetores $\mathbf{a}^T = [1 \ 0]$, $\mathbf{b}^T = [2 \ 1]$, $\mathbf{c}^T = [1 \ 1]$ e $\mathbf{d}^T = [0 \ 1]$. Seja a matriz $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, sejam os vetores $\mathbf{a}^*$, $\mathbf{b}^*$, $\mathbf{c}^*$ e $\mathbf{d}^*$, resultantes da pré-multiplicação por $\mathbf{A}$ ($\mathbf{a}^* = [2 \ 1]$, $\mathbf{b}^* = [5 \ 4]$, $\mathbf{c}^* = [3 \ 3]$ e $\mathbf{d}^* = [1 \ 2]$). Representando os vetores e o resultado de todas as multiplicações em um plano cartesiano tem-se (Figura 8):

Figura 8 - Representação cartesiana de vetores aleatórios e sua respectiva pré-multiplicação por uma dada matriz

Fonte: Bussiman (2021).

É possível notar que o vetor c^* não foi rotacionado, permanecendo com a mesma orientação do vetor c , embora tenha sido “alongado” (escalonado) exatamente ao triplo de seu comprimento inicial. Em linhas gerais qualquer vetor que não seja rotacionado por uma matriz é chamado de autovetor desta matriz e, o quanto este vetor é escalonado (neste caso 3) é chamado de autovalor (RENCER; SCHAALJE, 2008; STAR, 2019).

Agora podemos conceituar formalmente a PCA como sendo uma técnica multivariada que objetiva reduzir a dimensão de um conjunto de dados com q variáveis, através de combinações lineares dessas variáveis, em m PCs obtidos através da multiplicação das variáveis originais pelo respectivo autovetor (BARTHOLOMEW, 2010; LATTIN; CARROLL; GREEN, 2011a; HONGYU; SANDANIELO; OLIVEIRA-JR, 2015; JOLLIFFE; CADIMA, 2016).

Assim, a PCA permite reduzir a dimensão inicial do conjunto de dados (q), de modo que a variância explicada pelas novas variáveis (PCs) seja máxima, o que só é possível, pois, os PCs são combinações lineares das variáveis originais (JOLLIFFE; CADIMA, 2016). Uma das principais vantagens da PCA, além da redução de dimensão, sem perda da variabilidade inicial é a construção de índices e agrupamentos de indivíduos (SAVEGNAGO et al., 2011, 2013; BUZANSKAS et al., 2013; HONGYU; SANDANIELO; OLIVEIRA-JR, 2015), tal

raciocínio pode ainda ser estendido às variáveis (COSTA et al., 2005; JOLLIFFE; CADIMA, 2016).

Desta forma, a PCA permite um olhar simplificado nas relações entre indivíduos (de acordo com os valores observados para todas as variáveis conjuntamente), bem como entre as variáveis (de acordo com os valores observados para todos os indivíduos conjuntamente). Mas, nem sempre a PCA funciona, isto é, nem sempre a redução de dimensionalidade dos dados é possível, principalmente, nos casos em que o número de variáveis é maior do que as observações (isto é, quando $q > n$), ou quando as variáveis originais não são suficientemente correlacionadas (MANLY, 2008; LATTIN; CARROLL; GREEN, 2011a).

De acordo com Bartholomew (2010), o ponto de partida para a PCA é a matriz de correlações/covariâncias entre as variáveis originais. Esta, por sua vez, pode ser obtida através de operações matriciais, como segue:

$$(\mathbf{X} - \bar{\mathbf{X}})^T (\mathbf{X} - \bar{\mathbf{X}}) / (n-1), \quad (2.1)$$

em que: $\mathbf{X} = [\mathbf{X}_1 \ \cdots \ \mathbf{X}_q]$ é a matriz de dados originais com dimensão $n \times q$; $\bar{\mathbf{X}} = [\bar{\mathbf{X}}_1 \ \cdots \ \bar{\mathbf{X}}_q]$ é uma matriz com as médias amostrais para cada variável com dimensão $n \times q$ e n é o número total de informações. O resultado da equação (2.1) é uma matriz $q \times q$, definida como:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1q} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1q} & \sigma_{2q} & \cdots & \sigma_q^2 \end{bmatrix}, \quad (2.2)$$

em que: σ_i^2 é a variância da i -ésima variável e σ_{ij} é a covariância entre a i -ésima e a j -ésima variáveis.

É digno de nota que, utilizando-se os dados padronizados e posteriormente calculando-se $\mathbf{\Sigma}$, ou calculando-se a matriz de correlações ($\mathbf{P}$) a partir dos dados originais, os resultados da PCA em ambas as matrizes é o mesmo (*informação verbal*)⁹. Uma vez calculada $\mathbf{\Sigma}$, seus autovalores (λ_i) são obtidos facilmente por meio da seguinte equação (GENTLE, 2007; RENCHER; SCHAALJE, 2008; SEARLE; KHURI, 2017):

$$\det[\mathbf{\Sigma} - \lambda \mathbf{I}] = 0, \quad (2.3)$$

⁹ Informação fornecida por Lima, C. G. durante o curso Matrizes e Suas Aplicações em Estatística, Pirassununga, 2020.

em que: $\mathbf{I}$ é uma matriz identidade de dimensão $q \times q$.

Desta forma, podemos representar os autovalores em uma matriz diagonal (Λ) como sendo:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_q \end{bmatrix}, \quad (2.4)$$

De posse de Λ , obtém-se os autovetores ($\mathbf{v}$) de Σ , solucionando-se: $(\lambda \mathbf{I} - \Sigma) \mathbf{v} = \mathbf{0}$, de tal sorte que, cada autovetor ($\mathbf{v}_i$) pode ser agrupado em uma matriz $\mathbf{E} = [\mathbf{v}_1 \quad \cdots \quad \mathbf{v}_q]$, assim podemos escrever a, assim chamada, decomposição espectral de Σ , como (GENTLE, 2007; RENCHER; SCHAALJE, 2008; SEARLE; KHURI, 2017):

$$\Sigma = \mathbf{E} \Lambda \mathbf{E}^T, \quad (2.5)$$

Uma vez que se obtém os autovetores de Σ é possível obter cada um dos componentes principais fazendo-se o produto: $\mathbf{E}^T \mathbf{X} = \mathbf{Z}$, o que equivale a dizer que cada componente é calculado como a soma ponderada dos valores observados para cada variável, isto é (MANLY, 2008; HONGYU; SANDANIELO; OLIVEIRA-JR, 2015; JOLLIFFE; CADIMA, 2016):

$$\mathbf{Z}_i = \mathbf{v}_i^T \mathbf{X} = \mathbf{v}_{i1} \mathbf{X}_1 + \mathbf{v}_{i2} \mathbf{X}_2 + \cdots + \mathbf{v}_{iq} \mathbf{X}_q, \quad (2.6)$$

Neste sentido $\mathbf{Z}_i$ pode ser interpretada como uma variável latente (i.e., não diretamente mensurável) (JOHNSON; WICHERN, 2007; HONGYU; SANDANIELO; OLIVEIRA-JR, 2015). Objetiva-se obter $\mathbf{Z}_i$ de modo que os pontos coordenados originais estejam projetados no plano de ajuste de $\mathbf{Z}_i$ o mais distante possível, o que equivale a maximizar a variância de $\mathbf{Z}_i$ (JOHNSON; WICHERN, 2007; HONGYU; SANDANIELO; OLIVEIRA-JR, 2015).

A variância de $\mathbf{Z}_i$ é dada por (HONGYU; SANDANIELO; OLIVEIRA-JR, 2015):

$$\text{Var}(\mathbf{Z}_i) = \text{Var}(\mathbf{v}_i^T \mathbf{X}) = \mathbf{v}_i^T \text{Var}(\mathbf{X}) \mathbf{v}_i = \mathbf{v}_i^T \Sigma \mathbf{v}_i, \quad (2.7)$$

Se utilizamos a decomposição espectral de Σ , podemos provar que a variância total contida nos dados equivale a soma dos autovalores de Σ (JOHNSON; WICHERN, 2007):

$$\begin{aligned} \text{tr}(\Sigma) &= \sum_{i=1}^q \sigma_{ii}^2 = \text{tr}(\mathbf{E} \Lambda \mathbf{E}^T) = \text{tr}(\Lambda \mathbf{E}^T \mathbf{E}) = \text{tr}(\Lambda) = \sum_{i=1}^q \lambda_i, \\ \therefore \sum_{i=1}^q \sigma_{ii}^2 &= \sum_{i=1}^q \lambda_i \end{aligned}, \quad (2.8)$$

em que: $\text{tr}(\cdot)$ denota o traço matricial (i.e., soma dos valores na diagonal principal).

Fica evidente agora que, dado o fato de que a cada autovetor está associado um autovalor, como consequência, a cada componente principal (PC) está associado seu respectivo autovalor. Desta forma, a contribuição de cada PC para a variância total pode ser obtida fazendo-se a razão entre a variância de $\mathbf{Z}_i$ e $\sum_{i=1}^q \lambda_i$:

$$C_i = \frac{\mathbf{v}_i^T \Sigma \mathbf{v}_i}{\sum_{i=1}^q \lambda_i} = \frac{\lambda_i}{\sum_{i=1}^q \lambda_i}, \quad (2.9)$$

em que: C_i representa a proporção de variância explicada do i -ésimo componente principal. É ainda arbitrária, a porcentagem de variância total mínima aceita para que sejam utilizados os componentes principais ao invés dos dados originais, mas parece existir um consenso na literatura pela busca de “mágicos” 70% de variância retidos no menor número possível de PCs (JOLLIFFE; CADIMA, 2016).

É possível ainda calcular quanto da variância de cada variável é explicado em cada PC, pelo quadrado das cargas vetoriais de cada variável, isto é, o quadrado da correlação entre a variável e o componente principal, dado como (ABDI; WILLIAMS, 2010): $C_{qi} = v_{qi}^2 \times 100$, em que: C_{qi} é a proporção de variância da q -ésima variável explicada pelo i -ésimo PC.

Uma forma de visualização gráfica da associação entre cada uma das q variáveis pode ser conseguida utilizando-se um gráfico bidimensional (*biplot*) (GABRIEL, 1971, 1981), no qual as coordenadas de cada uma das q variáveis podem ser obtidas por (dentre outros métodos - KASSAMBARA, 2017): $\text{coord}_{iq} = v_{iq} \sqrt{\lambda_i}$, em que: coord_{iq} é o valor da coordenada na i -ésima dimensão (PC) para a q -ésima variável. Da mesma maneira, é possível computar a contribuição de cada variável para cada uma das dimensões (KASSAMBARA, 2017):

$$c_{qi} = \frac{\text{coord}_{iq}^2}{\sum_{q=1}^q \text{coord}_{iq}^2}, \quad (2.10)$$

em que: c_{qi} é a contribuição a q -ésima variável para o i -ésimo componente principal.

O uso de *biplots* para interpretar os resultados de PCA é bastante difundido na literatura, uma vez que as conclusões fornecidas por esse tipo de gráfico são visualmente deduzidas: indivíduos ou variáveis muito próximos apresentam similaridade (ou correlação), tão maior a distância entre indivíduos ou variáveis da origem, maior a importância dos mesmos para a variância do PC em questão, bem como, são melhores explicados pelos PCs utilizados no *biplot*

(GABRIEL, 1971, 1981; COSTA et al., 2005; BARTHOLOMEW, 2010; JOLLIFFE; CADIMA, 2016).

Ao final, o que nos resta é verificar quais PCs devem ser mantidos e adequabilidade dos dados à PCA. Segundo Cattell (1966) e Meyer (2005), uma das maneiras de se decidir o número de PCs a serem mantidos para descrever um conjunto de dados é o *scree test*, o qual, basicamente, é feito graficamente. Para conduzir o *scree test* basta plotar os sucessivos autovalores e verificar a divisão entre os PCs importantes e não importantes no ponto em que a curva gerada começa a ficar paralela ao eixo x (ponto de inflexão).

Kaiser (1960), propôs outra regra prática para determinar o número de PCs a serem mantidos em uma PCA: considere que os dados estejam padronizados ou que se use a matriz de correlações como ponto inicial. Neste caso, tanto em Σ , quanto em $\mathbf{P}$, os valores na diagonal principal são 1. Desta forma os PCs a serem utilizados são aqueles que possuem autovalor maior, ou igual, a 1 (i.e., $\lambda_i \geq 1$). Segundo Lattin, Carroll e Green (2011a), a regra reflete bom-senso, pois, como os PCs são uma medida de variância comum, eles deveriam explicar tantas variações quanto qualquer uma das variáveis originais.

Testar a adequabilidade de um conjunto de dados à PCA envolve verificar se as variáveis apresentam correlações suficientes para favorecer a decomposição espectral de Σ (MANLY, 2008). Para tal finalidade, o Teste de Esfericidade de Bartlett (BARTLETT, 1950) é o mais amplamente difundido, entretanto Lattin, Carroll e Green (2011a) argumentam que existem outras formas, como o Limite de Horn (HORN, 1965) e as validações Jackknife e Bootstrap (EFRON; GONG, 1983).

Aqui trataremos apenas do Teste de Esfericidade de Bartlett; para os outros métodos recomendamos a leitura de Lattin, Carroll e Green (2011a); Horn (1965) e Efron e Gong (1983). Segundo Lattin, Carroll e Green (2011a, p. 88), o Teste de Bartlett propõe explicitamente a pergunta: “a matriz de correlações deve ser decomposta em fatores, em primeiro lugar?” Este teste se baseia na distribuição de qui-quadrado (χ^2) e é aplicado a matriz $\mathbf{P}$ (ou a matriz Σ obtida com os dados padronizados); é calculado como:

$$\chi^2_{\left[\frac{(q^2-q)}{2}\right]} = -\left[(n-1) - \frac{2q-5}{6}\right] \ln|\mathbf{P}|, \quad (2.11)$$

em que: $\chi^2_{\left[\frac{(q^2-q)}{2}\right]}$ é o valor de qui-quadrado com $\left[\frac{(q^2-q)}{2}\right]$ graus de liberdade; q é o número de variáveis; n é o número de observações e $\ln|\mathbf{P}|$ é o logaritmo natural do determinante da matriz de correlações.

É fácil provar que $|\mathbf{P}| = \prod_{i=1}^i \lambda_i$, desta forma, se as variáveis são independentes $\mathbf{P} \approx \mathbf{I}$, consequentemente $|\mathbf{P}| \approx 1$ e $\ln|\mathbf{P}| \approx 0$, quando isso acontece o diagrama de dispersão dos dados é esférico (dando origem ao nome do teste), então, o teste de Bartlett verifica a adequação da matriz de correlações a uma matriz identidade, isto é, a um conjunto esférico (BARTLETT, 1950; LATTIN; CARROLL; GREEN, 2011a). Valores significativos para este teste são desejados, pois, indicam que a PCA terá sucesso em reduzir a dimensionalidade dos dados.

b) *Análise de fatores*

A metodologia foi criada por Spearman em 1904, quando, para tratar de mensurar corretamente a inteligência, criou os procedimentos hoje conhecidos como análise de fatores. Spearman lançou a hipótese de que “todos os ramos da atividade intelectual têm em comum uma função fundamental (ou um grupo de funções), enquanto os elementos restantes ou específicos da atividade parecem, em todo caso, ser completamente diferentes uns dos outros” (SPEARMAN, 1904, p. 284, tradução nossa). Esta afirmação deu origem ao que entendemos serem fatores comuns e específicos.

Embora tenha sido bastante eficiente, a proposição de Spearman foi revista em 1919 por Garnett, quando surge o conceito de fatores múltiplos (GARNETT, 1919), pois, Spearman propunha que as m variáveis observadas eram causadas por um único fator, ou grupo de fatores, em comum. Entretanto, foi somente com Thurstone, em 1931, que surgiu o termo Análise de Fatores Múltiplos (*Multiple Factor Analysis*) (THURSTONE, 1931). A partir de então, numerosas foram as aplicações em áreas como psicologia, ciências sociais, estatística e, eventualmente, zootecnia e medicina veterinária, chegando ao melhoramento animal.

De acordo com Manly (2008), a Análise de Fatores (FA – *Factor Analysis*) tem objetivos similares a PCA. A ideia básica é descrever um conjunto de q variáveis por meio de um conjunto menor, de índices ou fatores e, neste processo, obter uma melhor compreensão do relacionamento entre as variáveis. Desta forma, é possível dizer que a FA se dedica a explicar

a estrutura de covariância entre as q variáveis ao invés de olhar apenas para as variâncias, como faz a PCA (THOMPSON, 2004; MANLY, 2008).

Segundo Loehlin (2004), a FA não se inicia com um modelo específico, mas com especificações gerais acerca do modelo que se busca, então é preciso encontrar o modelo e as estimativas de seus coeficientes. Bartholomew et al. (2008), adicionam, ainda, a esta definição que, por se tratar de uma técnica baseada em modelagem probabilística, a FA envolve pressuposições sobre a distribuição conjunta das variáveis envolvidas. Além disso, os parâmetros do modelo fatorial podem ser estimados por máxima verossimilhança ou algum método equivalente (BARTHOLOMEW, 2010), nos permitindo inferência sobre a população, baseado em qualidade de ajuste, significância estatística e precisão das predições (BARTHOLOMEW et al., 2008).

Vale ainda acrescentar que, de modo similar à PCA, a FA é um método dos assim chamados problemas de variáveis latentes (FORD; MACCALLUM; TAIT, 1986; LOEHLIN, 2004; THOMPSON, 2004), isto é, trata de mensurar variáveis que não são diretamente mensuráveis, mas possuem aspectos observáveis. Desta forma, as q variáveis são assumidas como sendo o reflexo de um fenômeno subliminar (latente) complexo (PEÑAGARICANO et al., 2015) e a variância dos dados é particionada em m fatores comuns e s fatores específicos (MEYER, 2009), de modo que $m < q$ e $s = q$.

O ponto de partida da FA é o modelo fatorial. Para defini-lo, porém, vamos pressupor um conjunto de q variáveis X , medidas em n indivíduos; ordenadas por indivíduos (por conveniência) e por variável, desta forma tem-se: $\mathbf{w}^T = [\mathbf{X}_1^T \quad \cdots \quad \mathbf{X}_q^T]$; para cada variável suponha uma média (μ), desta forma o vetor das médias pode ser representado por: $\boldsymbol{\mu}^T = [\boldsymbol{\mu}_1^T \quad \cdots \quad \boldsymbol{\mu}_q^T]$.

Assumindo que, cada uma das variáveis é causada por m Fatores Comuns (F_m) e q Fatores Específicos (δ_q), tem-se (MAITI, 2014a) (Figura 9):

Figura 9 - Representação visual esquemática da análise de fatores

Fonte: Maiti (2014a).

Assim, o modelo fatorial pode ser escrito como (MEYER, 2009; MAITI, 2014a):

$$\begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_q \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_q \end{bmatrix} + \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1m} \\ \vdots & \ddots & \vdots \\ \lambda_{q1} & \cdots & \lambda_{qm} \end{bmatrix} \begin{bmatrix} F_1 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} \boldsymbol{\delta}_1 \\ \vdots \\ \boldsymbol{\delta}_m \end{bmatrix}, \quad (2.12)$$

$$\therefore \mathbf{w} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{F} + \boldsymbol{\delta}$$

em que: $\mathbf{w}$ é o vetor de observações, com dimensão $q \times 1$; $\boldsymbol{\mu}$ é o vetor de médias, com dimensão $q \times 1$; $\boldsymbol{\Lambda}$ é a matriz de cargas fatoriais (λ_{qm}), com dimensão $q \times m$; $\mathbf{F}$ é o vetor de soluções para os Fatores Comuns, com dimensão $m \times 1$ e $\boldsymbol{\delta}$ é o vetor de soluções para os Fatores Específicos, com dimensão $q \times 1$.

É preciso agora pressupor que o valor esperado de $\mathbf{w}$ é $\boldsymbol{\mu}$, ou seja $E[\mathbf{w}] = \boldsymbol{\mu}$. Desta forma, podemos escrever (2.12) removendo o efeito de $\boldsymbol{\mu}$, $\mathbf{w} - \boldsymbol{\mu} = \boldsymbol{\Lambda}\mathbf{F} + \boldsymbol{\delta}$; deste modo: $E[\mathbf{F}] = E[\boldsymbol{\delta}] = 0$ porque $E[\mathbf{w} - \boldsymbol{\mu}] = 0$. Convenientemente, é assumido que: $\text{Cov}(\mathbf{F}\boldsymbol{\delta}^T) = 0$ e $\text{Var}(\boldsymbol{\delta}) = E[\boldsymbol{\delta}\boldsymbol{\delta}^T] = \boldsymbol{\Psi}$, em que $\boldsymbol{\Psi}$ é uma matriz diagonal com as variâncias específicas das q variáveis. Assim, é possível mostrar que $\text{Var}(\mathbf{w} - \boldsymbol{\mu}) = \text{Var}(\boldsymbol{\Lambda}\mathbf{F} + \boldsymbol{\delta})$, o que resulta na decomposição fatorial da matriz de (co)variâncias ($\boldsymbol{\Sigma}$) entre q variáveis (LOEHLIN, 2004; BARTHOLOMEW et al., 2008):

$$\begin{aligned}
\text{Var}(\mathbf{w} - \boldsymbol{\mu}) &= \text{Var}(\mathbf{\Lambda F} + \boldsymbol{\delta}) \\
E[(\mathbf{w} - \boldsymbol{\mu})(\mathbf{w} - \boldsymbol{\mu})^T] &= [(\mathbf{\Lambda F} + \boldsymbol{\delta})(\mathbf{\Lambda F} + \boldsymbol{\delta})^T], \\
\boldsymbol{\Sigma} &= \mathbf{\Lambda \Phi \Lambda}^T + \boldsymbol{\Psi}
\end{aligned} \tag{2.13}$$

Se assumirmos que os m Fatores Comuns são não-correlacionados, chamamos o modelo fatorial de ortogonal e, escrevemos $\boldsymbol{\Phi} = E[\mathbf{F F}^T] = \mathbf{I}$ de tal sorte que: $\boldsymbol{\Sigma} = \mathbf{\Lambda \Lambda}^T + \boldsymbol{\Psi}$. Por outro lado, se assumirmos que exista covariância entre os Fatores, chamamos o modelo fatorial de oblíquo e, escrevemos a decomposição fatorial como: $\boldsymbol{\Sigma} = \mathbf{\Lambda \Phi \Lambda}^T + \boldsymbol{\Psi}$ (MAITI, 2014a). Thompson (2004) argumenta que modelos fatoriais ortogonais são mais parcimoniosos do que os oblíquos quando $m \geq 2$, permanecendo em questão, quando sacrificar a parcimônia para um melhor ajuste do modelo.

Valendo-se do caso ortogonal, podemos mostrar que, do ponto de vista das variâncias das q variáveis, estamos decompondo a variância da q -ésima variável em duas porções, sendo uma devido aos m Fatores Comuns e outra devido a seu Fator Específico. Vamos olhar para o conteúdo da matriz resultante da decomposição fatorial $\boldsymbol{\Sigma} = \mathbf{\Lambda \Lambda}^T + \boldsymbol{\Psi}$:

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1q} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1q} & \sigma_{2q} & \cdots & \sigma_q^2 \end{bmatrix} = \begin{bmatrix} \sum_{m=1}^m \lambda_{1m}^2 + \psi_{11} & \sum_{m=1}^m \lambda_{1m} \lambda_{2m} & \cdots & \sum_{m=1}^m \lambda_{1m} \lambda_{qm} \\ \sum_{m=1}^m \lambda_{1m} \lambda_{2m} & \sum_{m=1}^m \lambda_{2m}^2 + \psi_{22} & \cdots & \sum_{m=1}^m \lambda_{2m} \lambda_{qm} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{m=1}^m \lambda_{1m} \lambda_{qm} & \sum_{m=1}^m \lambda_{2m} \lambda_{qm} & \cdots & \sum_{m=1}^m \lambda_{qm}^2 + \psi_{qq} \end{bmatrix}, \tag{2.14}$$

Consequentemente:

$$\begin{cases} \sigma_q^2 = \sum_{m=1}^m \lambda_{qm}^2 + \psi_{qq} \\ \sigma_{(q-1)q} = \sum_{m=1}^m \lambda_{(q-1)m} \lambda_{qm} \end{cases}, \tag{2.15}$$

Assim, a porção $\sum_{m=1}^m \lambda_{qm}^2$ equivale ao que é explicado de σ_q^2 pelos m Fatores Comuns, sendo chamada de Comunalidade (usualmente representada por: h_q^2), já a porção ψ_{qq} é o quanto de σ_q^2 não é explicado pelos fatores, sendo chamada de Especificidade. Por sua vez, as covariâncias são assumidas como perfeitamente explicadas pelos m Fatores Comuns. Fica evidente que os valores de h_q^2 aceitáveis são os que satisfazem: $h_q^2 \leq 1$ (assumindo padronização das variáveis); Loehlin (2004), argumentou que comunalidades iguais a um não são plausíveis, pois, tal fato implica em dois fenômenos: 1) a variável é perfeitamente explicada pelos m fatores e 2) a variável não possui especificidade.

Comunalidades maiores que 1 também podem ser encontradas nos chamados casos de Heywood (HEYWOOD, 1931); definidos como inadmissíveis por Thompson (2004), o qual argumenta que estes casos indicam problemas fundamentais semelhantes a estimativas de variâncias negativas. Para evitar esses tipos de problemas ($h_q^2 = 1$ ou $h_q^2 > 1$), o número de fatores a serem extraídos deve ser corretamente determinado (LOEHLIN, 2004; BARTHOLOMEW et al., 2008), bem como o número de iterações para solucionar-se o modelo fatorial pode ser reduzido (LOEHLIN, 2004), ou ainda, utilizar-se de estimação iterativa diretamente das comunalidades (THOMPSON, 2004).

Da maneira como é formulado, o modelo fatorial admite infinitas soluções, equivalentes em termos de reprodutibilidade de Σ , pois, a orientação da solução fatorial é arbitrária, o que é conhecido como Indeterminação Rotacional (LATTIN; CARROLL; GREEN, 2011b). Darton (1980) e Anderson e Rubin (1956), explicaram ainda que tal indeterminação decorre do fato de se expressar q variáveis em m fatores e q erros. Este “problema” foi solucionado por Thurstone em 1947, quando propôs o critério de estrutura simples (*Simple Structure Criteria*) (THURSTONE, 1947): um algoritmo eficiente para encontrar a solução mais parcimoniosa, ou seja, de mais fácil explicação (simples).

Tomando o caso ortogonal como exemplo: seja $\hat{\Sigma} - \hat{\Psi} = \hat{\Lambda}\hat{\Lambda}^T$ uma solução do modelo fatorial; seja ainda a matriz T ($m \times m$), ortogonal e não-singular (i.e. $|T| \neq 0$), é possível provar que, para qualquer T satisfazendo $\hat{\Lambda}_i = \hat{\Lambda}T$; $\hat{\Sigma} - \hat{\Psi} = \hat{\Lambda}_i\hat{\Lambda}_i^T$ é também uma solução, equivalente, do mesmo modelo fatorial (DARTON, 1980; LATTIN; CARROLL; GREEN, 2011b). Para o caso oblíquo, é possível escrever $\hat{\Lambda}_i = \hat{\Lambda}T^{-1}$ e $\hat{\Phi}_i = T\Phi T^{-1}$, consequentemente: $\hat{\Sigma} - \hat{\Psi} = \hat{\Lambda}_i\hat{\Phi}_i\hat{\Lambda}_i^T$ (DARTON, 1980). Desta forma, é plausível escolher uma matriz T , de tal forma que a estrutura dos fatores (Λ) seja simplificada (LATTIN; CARROLL; GREEN, 2011b), a este processo dá-se o nome de Rotação de Fatores.

Graficamente, é fácil entender o princípio da rotação de fatores. O que se busca é um posicionamento dos vetores base (eixos ou fatores) de tal sorte que uma determinada variável x tenha suas cargas fatoriais maximizadas em um determinado fator F e minimizadas em outro (DARTON, 1980; LATTIN; CARROLL; GREEN, 2011b). Desta forma, suponha um conjunto de 6 variáveis ($X = \{X_1, X_2, X_3, X_4, X_5, X_6\}$), do qual, após FA, foram extraídos dois fatores; $F = \{F_1, F_2\}$. Após a rotação fatorial, é possível encontrar a solução $F^* = \{F_1^*, F_2^*\}$ que facilita a interpretação, pois, as cargas fatoriais das variáveis x_1, x_2 e x_5 foram maximizadas no primeiro

fator, enquanto as cargas fatoriais das variáveis x_3 , x_4 e x_6 foram maximizadas no segundo fator (Figura 10):

Figura 10 - Representação cartesiana da rotação ortogonal de fatores

Fonte: Bussiman (2021).

Existem várias técnicas para a rotação de fatores, mas elas podem ser agrupadas em dois grandes grupos: rotações ortogonais e rotações oblíquas (DARTON, 1980; LOEHLIN, 2004). Quando o ângulo entre os fatores rotacionados é mantido em 90° a rotação é dita ortogonal, caso contrário, trata-se de uma rotação oblíqua (MAITI, 2014b). Para o caso da rotação ortogonal, uma possibilidade é valer-se do ângulo desejado entre o fator original e o fator rotacionado (Θ), assim, $\mathbf{T}$ tem forma conhecida (LATTIN; CARROLL; GREEN, 2011b):

$$\mathbf{T} = \begin{bmatrix} \cos \Theta & -\sin \Theta \\ \sin \Theta & \cos \Theta \end{bmatrix}, \text{ para outros tipos de rotação, bem como uma revisão bastante}$$

aprofundada do assunto, recomenda-se a leitura de Darton (1980).

Para se predizerem os Escores (*Factor Scores*), ou seja, os valores para as m variáveis latentes (Fatores Comuns), o método mais comum é a regressão de mínimos quadrados (LOEHLIN, 2004; THOMPSON, 2004), mas segundo Thompson (2004) existem outros métodos, como o proposto por Bartlett (1937) e o proposto por Anderson e Rubin (1956), ou ainda, o método de Thompson (1993), entre outros. Assumindo que os fatores são variáveis

aleatórias, a melhor predição linear não viesada (BLUP – *Best linear unbiased prediction*) de $\mathbf{F}$ é dada por (BARTHOLOMEW et al., 2008; LOEHLIN, 2004):

$$f_k = b_{k1}w_1 + b_{k2}w_2 + \dots + b_{kq}w_q, \quad (2.16)$$

em que: f_k representa o escore predito para o k -ésimo fator; b_{kq} representa o ponderador do k -ésimo fator na q -ésima variável e, w_q representa o valor observado para a q -ésima variável desviado da média.

Sob notação matricial, a equação (2.16) pode ser escrita como $\mathbf{F} = \mathbf{b}^T (\mathbf{w} - \boldsymbol{\mu})$ e, $\mathbf{b}^T$ pode ser obtido via mínimos quadrados (LOEHLIN, 2004), por meio de:

$$\mathbf{b}^T = \left[\boldsymbol{\Sigma}^{-1} \text{Cov}(\mathbf{w}, \mathbf{F}^T) \right]^T, \quad (2.17)$$

Desenvolvendo-se (2.17) para o caso ortogonal obtém-se: $\mathbf{b}^T = \left[(\boldsymbol{\Lambda} \boldsymbol{\Lambda}^T + \boldsymbol{\Psi})^{-1} \boldsymbol{\Lambda} \right]^T = \boldsymbol{\Lambda}^T (\boldsymbol{\Lambda} \boldsymbol{\Lambda}^T + \boldsymbol{\Psi})^{-1}$; para o caso oblíquo, basta incluir $\boldsymbol{\Phi}$, de modo que: $\mathbf{b}^T = \left[(\boldsymbol{\Lambda} \boldsymbol{\Phi} \boldsymbol{\Lambda}^T + \boldsymbol{\Psi})^{-1} \boldsymbol{\Lambda} \boldsymbol{\Phi} \right]^T = \boldsymbol{\Phi} \boldsymbol{\Lambda}^T (\boldsymbol{\Lambda} \boldsymbol{\Phi} \boldsymbol{\Lambda}^T + \boldsymbol{\Psi})^{-1}$, em ambos os casos, utilizando-se esses métodos para se obter $\mathbf{b}^T$, os escores são definidos como $E[\mathbf{F} | \mathbf{x}]$ (PEÑAGARICANO et al., 2015), desta forma é possível obter-se mensurações para os m $\mathbf{F}$ que, essencialmente, são imensuráveis.

Para se ter uma ideia de o quão bom esses ponderadores foram preditos, Rozeboom (1978), propôs um coeficiente de validade cruzada (*cross-validity coefficient*) que pode ser obtido por:

$$R_c = \sqrt{1 - (1 - R^2)(N + q) / (N - q)}, \quad (2.18)$$

em que: R_c é uma medida de qualidade da predição dos escores para cada fator; R^2 é o quadrado do coeficiente de correlação múltipla entre os fatores e as variáveis originais, obtidos como valores na diagonal principal da matriz: $\mathbf{b}^T \text{Cov}(\mathbf{w}, \mathbf{F}^T)$; N é o tamanho amostral (número de informações) e q é o número de variáveis.

Ao final, nos resta então, determinar a adequabilidade do modelo fatorial, bem como testar sua significância. Para tal finalidade são válidos os testes de esfericidade de Bartlett (BARTLETT, 1950), já definido; o critério de Kaiser que, na análise de fatores, foi proposto por Guttman (1954): somente devem ser extraídos fatores que possuam autovalor superior a 1.0; o scree test (CATTELL, 1966), já definido; e a inspeção da matriz de correlações residuais, buscando valores, em módulo, menores que 0.05 (THOMPSON, 2004). Há ainda outra forma

de se determinar o número de fatores e, por consequência, melhorar o ajuste do modelo fatorial: a Análise Paralela (HORN, 1965), figurando entre os melhores métodos para decisão do número de fatores (ZWICK; VELICER, 1986). Tal técnica envolve considerar o erro de amostragem que influencia um dado conjunto de variáveis (THOMPSON, 2004).

Existem ainda muitas outras formas de se checar a adequabilidade do modelo fatorial, como por exemplo: Raiz do Erro Quadrático Médio por Aproximação (*root-mean-square-error of approximation* – RMSEA), proposto por Steiger (1990); a Raiz da média quadrática residual padronizada (*standardized root-mean-square residual* – SRMR), proposto por Bentler (1995); o Índice de Tucker-Lewis (*Tucker-Lewis index* – TLI), proposto por Tucker e Lewis (1973); o Índice Comparativo de Ajuste (*comparative fit index* – CFI), proposto por Bentler (1990); o Índice Normalizado de Ajuste (*normed fit index* – NFI), proposto por Bentler e Bonett (1980); Índice de Qualidade de Ajuste (*goodness-of-fit index* – GFI), proposto por Hu e Bentler (1999); entre outros.

Todos esses índices alternativos para avaliar a qualidade do ajuste do modelo fatorial são mais usuais nas análises confirmatórias de fatores, as quais podem ser consideradas um caso especial da análise de fatores com modelo fatorial oblíquo, pois a única diferença para a análise exploratória de fatores é que uma mesma variável não é influenciada por mais de um fator (conhecimento *a priori*) (MAITI, 2014b). Desta forma, recomenda-se a leitura do Capítulo 3, página 67, de Brown (2015), para uma informação detalhada acerca das diversas medidas de ajuste, bem como Hu e Bentler (1999) para valores de corte dos índices comparativos de ajuste.

2.2.2. Modelos mistos

Provavelmente a maior disputa da estatística tenha sido a descoberta dos mínimos quadrados (LS, do inglês *Least Squares*): enquanto Legendre, em 1805, publica a primeira introdução formal do método dos LS (LEGENDRE, 1805), Gauss apresenta em 1809 o mesmo método e argumenta que vinha utilizando-o desde 1795 (GAUSS, 1809). O método dos LS não fazia nenhuma suposição da distribuição da variável resposta, ou das variáveis explanatórias, Pearson apresentou, em 1903, as médias e variâncias condicionais da distribuição normal multivariada (PEARSON, 1903); o que, segundo Henderson (1990), foi o pontapé inicial para o desenvolvimento dos modelos mistos.

Os modelos mistos são assim chamados por considerar de forma conjunta variáveis/fatores de efeito fixo e variáveis/fatores de efeito aleatório. Existe uma grande discussão sobre a determinação de efeitos fixos e aleatórios. Em geral, as definições mais

clássicas são baseadas em número de níveis (SCHAEFFER, 2019). De maneira geral, é possível designar os fatores como fixo ou aleatório de acordo com Schaeffer (2019) e (*informação verbal*)¹⁰:

- **Efeito Aleatório:** quando for considerado apropriado usar uma função de distribuição de probabilidades (f.d.p.) para descrever os efeitos associados com a população de níveis. Desta forma, efeitos aleatórios influenciam somente a variância da variável resposta; existe uma possibilidade infinita de níveis possíveis para o fator, ou seja, os níveis observados são resultado de uma amostra de uma população de níveis;
- **Efeito Fixo:** quando for considerado apropriado que as diferenças entre os níveis do fator ocorrem em uma maneira arbitrária, ou seja, ao invés do uso de uma f.d.p. para descrever os efeitos, o interesse é em um conjunto específico de níveis do fator. Desta forma, fatores de efeito fixo influenciam apenas a média da variável resposta, além disso, os níveis de interesse são finitos e não vem de uma população de possíveis níveis (na prática o objetivo é estimação do efeito de níveis em particular).

Schaeffer (2011) apresenta alguns exemplos de fatores de efeito fixo e/ou aleatório. Este autor considerou que fatores, tais como: Dietas; Raças; Sexo; Classes de Idade, Tanques e Estações, possuiriam efeito fixo e, fatores, tais como: Animais; Grupos de Contemporâneos; Interação Rebanho-Ano-Estação, Ambiente Permanente; Efeitos Maternos e Leitegada, possuiriam efeito aleatório. Cabe, ainda, ressaltar que a determinação dos fatores fixos e/ou aleatórios dependerá da experiência do pesquisador, bem como, das condições de cada banco dados disponíveis para análise.

Fisher foi, provavelmente, o primeiro a utilizar um modelo misto quando ele propõe a análise de variância (FISHER, 1919) – mais tarde ele também foi o responsável pelo descobrimento da máxima verossimilhança (FISHER, 1925), usada para o desenvolvimento das MME – contudo, foi Eisenhart em 1947 que fez a primeira distinção formal entre efeitos fixos e aleatórios na análise de variância (EISENHART, 1947). Entretanto, mesmo com esses avanços os modelos mistos eram pouco utilizados devido a necessidade de se inverter a matriz de covariâncias para o vetor de observações (definida a frente), o que nem sempre era tarefa fácil (*informação verbal*)⁶.

¹⁰ Informação fornecida por Demétrio, C. G. B. durante o curso “Modelos Mistos e Componentes de Variância”, Universidade de São Paulo, Escola Superior de Agricultura “Luiz de Queiroz”, Piracicaba, 2020.

Para contornar o problema, Henderson (1949, 1950) apresentou as MME, reunindo o poder dos LS para estimação dos efeitos fixos, com a versatilidade dos índices de seleção para predição dos efeitos aleatórios (HENDERSON, 1990). Em 1959 foi publicada a prova de que as soluções para os efeitos fixos, por meio das MME, eram as mesmas que nos mínimos quadrados generalizados (HENDERSON et al., 1959), e em 1963 C. Henderson apresenta a completa equivalência entre o índice de seleção com a solução para o efeito aleatório de animal nas MME (HENDERSON, 1963).

Para entender como se dá a estimação dos efeitos fixos, e a predição dos efeitos aleatórios, vamos primeiro analisar os resultados de Pearson (1903), adotando-se a notação matricial tradicional do modelo linear misto clássico no melhoramento animal (HENDERSON, 1990):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad (2.19)$$

em que: $\mathbf{y}$ representa o vetor de observações da variável dependente; $\boldsymbol{\beta}$ representa o vetor de efeitos fixos; $\mathbf{u}$ representa o vetor de efeitos aleatórios (usualmente, genético aditivo de animal); $\mathbf{X}$ e $\mathbf{Z}$ são matrizes conhecidas que relacionam os elementos de $\boldsymbol{\beta}$ e $\mathbf{u}$, respectivamente; $\mathbf{e}$ representa o vetor de termos aleatórios residuais.

Assumindo que $\mathbf{u}$ e $\mathbf{e}$ são variáveis aleatórias, com médias 0 e variâncias: $\text{Var}(\mathbf{u}) = \mathbf{G}$ e $\text{Var}(\mathbf{e}) = \mathbf{R}$, se $\text{Cov}(\mathbf{u}, \mathbf{e}^T) = 0$, os resultados de Pearson são:

$$\begin{aligned} E(\boldsymbol{\beta}|\mathbf{y}) &= (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y} \\ E(\mathbf{u}|\mathbf{y}) &= \mathbf{GZ}^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned} \quad (2.20)$$

em que: $\mathbf{V} = \text{Var}(\mathbf{y}) = \mathbf{ZGZ}^T + \mathbf{R}$. Já as médias e variâncias incondicionais de $[\mathbf{y} \quad \mathbf{u} \quad \mathbf{e}]^T$ são:

$$E \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad (2.21)$$

$$\text{Var} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{ZGZ}^T + \mathbf{R} & \mathbf{ZG} & \mathbf{R} \\ \mathbf{GZ}^T & \mathbf{G} & \mathbf{0} \\ \mathbf{R} & \mathbf{0} & \mathbf{R} \end{bmatrix}, \quad (2.22)$$

em que: $\mathbf{G}$ é a matriz de covariâncias do efeito aleatório genético aditivo de animal, isto é: $\mathbf{G} = \mathbf{A}\sigma_u^2$ (sendo σ_u^2 a variância genética aditiva da característica em análise); $\mathbf{R}$ é a matriz de covariâncias residuais, isto é: $\mathbf{R} = \mathbf{I}\sigma_e^2$ (sendo σ_e^2 a variância residual da característica em análise).

Por outro lado, os índices seleção pressupunham que $\mathbf{X}\boldsymbol{\beta}$ era conhecido, o que na prática não era realidade, portanto o que se usava era encontrar uma estimativa para $\mathbf{X}\boldsymbol{\beta}$, em geral por mínimos quadrados, através do modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$; como resultado, $\mathbf{X}\boldsymbol{\beta}$ era substituído por $\mathbf{X}\boldsymbol{\beta}^*$ (que é alguma estimativa não viesada de $\mathbf{X}\boldsymbol{\beta}$). Henderson (1963), mostrou que o índice podia ser calculado aplicando-se ponderadores econômicos a $\hat{\mathbf{u}}$ e Henderson (1990), mostrou que, obter $\hat{\mathbf{u}}$ utilizando-se $\mathbf{X}\boldsymbol{\beta}^*$ não era equivalente a obter $\hat{\mathbf{u}}$ utilizando-se $\mathbf{X}\boldsymbol{\beta}$, isto é: $f(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}, \mathbf{u}) \neq f(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*, \mathbf{u})$ (HENDERSON, 1990).

A solução seria, então, aplicar os resultados da distribuição normal multivariada de Pearson, para estimar de maneira conjunta $\boldsymbol{\beta}$ e $\mathbf{u}$. Os resultados de Pearson, apesar de simples, dificultam sua implementação, uma vez que $\mathbf{V}$ em geral é esparsa e bastante grande, e sua inversão é muito custosa. Henderson (1949, 1950), aplicou a metodologia de máxima verossimilhança para maximizar a distribuição conjunta de $\mathbf{y}$ e $\mathbf{u}$, com respeito a $\boldsymbol{\beta}$ e $\mathbf{u}$.

Seja a distribuição de $\mathbf{y}|\boldsymbol{\beta}, \mathbf{V}$ dada por $l(\mathbf{y}|\boldsymbol{\beta}, \mathbf{V}) = 2\pi^{-\frac{n}{2}} |\mathbf{G}|^{-\frac{1}{2}} e^{\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}}$; a distribuição de $\mathbf{u}|\mathbf{G}$ dada por $l(\mathbf{u}|\mathbf{G}) = 2\pi^{-\frac{n}{2}} |\mathbf{G}|^{-\frac{1}{2}} e^{\left\{-\frac{1}{2}\mathbf{u}^T \mathbf{G}^{-1}\mathbf{u}\right\}}$ e, a distribuição de $\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \mathbf{R}$ dada por $l(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \mathbf{R}) = 2\pi^{-\frac{n}{2}} |\mathbf{R}|^{-\frac{1}{2}} e^{\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})^T \mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})\right\}}$, então a distribuição conjunta de $\mathbf{y}$ e $\mathbf{u}$ pode ser escrita como: $l(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \mathbf{R}, \mathbf{G}) = l(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \mathbf{R})l(\mathbf{u}|\mathbf{G})$. Tomando-se o ln de $l(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \mathbf{R}, \mathbf{G})$ tem-se: $\ln[l(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \mathbf{R}, \mathbf{G})] = \ln[l(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \mathbf{R})] + \ln[l(\mathbf{u}|\mathbf{G})]$, o que é igual a:

$$L = -n \ln(2\pi) - \frac{1}{2} \ln|\mathbf{R}| - \frac{1}{2} \ln|\mathbf{G}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) - \frac{1}{2} \mathbf{u}^T \mathbf{G}^{-1} \mathbf{u}, \quad (23)$$

Removendo-se a parte constante $-n \ln(2\pi)$ e multiplicando-se por -2, (23) pode ser escrita como:

$$-2L \propto \ln|\mathbf{R}| + \ln|\mathbf{G}| + \mathbf{y}^T \mathbf{R}^{-1} \mathbf{y} - 2\mathbf{y}^T \mathbf{R}^{-1} \mathbf{X}\boldsymbol{\beta} - 2\mathbf{y}^T \mathbf{R}^{-1} \mathbf{Z}\mathbf{u} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{R}^{-1} \mathbf{X}\boldsymbol{\beta} + 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z}\mathbf{u} + \mathbf{u}^T \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z}\mathbf{u} + \mathbf{u}^T \mathbf{G}^{-1} \mathbf{u}, \quad (24)$$

Derivando-se (24) com respeito a $\boldsymbol{\beta}$ e $\mathbf{u}$ tem-se:

$$\begin{aligned}
\begin{bmatrix} \frac{\partial - 2L}{\partial \boldsymbol{\beta}} \\ \frac{\partial - 2L}{\partial \mathbf{u}} \end{bmatrix} &= \begin{bmatrix} \frac{\partial - 2\mathbf{y}^T \mathbf{R}^{-1} \mathbf{X} \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} + \frac{\partial \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} + \frac{\partial 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z} \mathbf{u}}{\partial \boldsymbol{\beta}} \\ \frac{\partial - 2\mathbf{y}^T \mathbf{R}^{-1} \mathbf{Z} \mathbf{u}}{\partial \mathbf{u}} + \frac{\partial 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z} \mathbf{u}}{\partial \mathbf{u}} + \frac{\partial \mathbf{u}^T \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} \mathbf{u}}{\partial \mathbf{u}} + \frac{\partial \mathbf{u}^T \mathbf{G}^{-1} \mathbf{u}}{\partial \mathbf{u}} \end{bmatrix}, \\
&= \begin{bmatrix} -2\mathbf{X}^T \mathbf{R}^{-1} \mathbf{y} + 2\mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} \boldsymbol{\beta} + 2\mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z} \mathbf{u} \\ -2\mathbf{Z}^T \mathbf{R}^{-1} \mathbf{y} + 2\mathbf{Z}^T \mathbf{R}^{-1} \mathbf{X} \boldsymbol{\beta} + 2\mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} \mathbf{u} + 2\mathbf{G}^{-1} \mathbf{u} \end{bmatrix}
\end{aligned} \quad (2.25)$$

Igualando-se (2.25) a zero e, rearranjando-se os termos “surtem” as MME (HENDERSON, 1949, 1950):

$$\begin{bmatrix} \mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \mathbf{R}^{-1} \mathbf{y} \\ \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{y} \end{bmatrix}, \quad (2.26)$$

em que: $\mathbf{y}$, $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{R}$ e $\mathbf{G}$ são definidas como em (2.19); $\hat{\boldsymbol{\beta}}$ representa o vetor de soluções para $\boldsymbol{\beta}$ e $\hat{\mathbf{u}}$ representa o vetor de soluções para $\mathbf{u}$.

Isolando-se $\hat{\mathbf{u}}$, na segunda equação em (2.26), e substituindo-o na primeira, obtém-se $\hat{\boldsymbol{\beta}}$ e $\hat{\mathbf{u}}$:

$$\begin{cases} \hat{\boldsymbol{\beta}} = \left\{ \mathbf{X}^T \left[\mathbf{R}^{-1} - \mathbf{R}^{-1} \mathbf{Z} (\mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}^T \mathbf{R}^{-1} \right] \mathbf{X} \right\}^{-1} \mathbf{X}^T \left[\mathbf{R}^{-1} - \mathbf{R}^{-1} \mathbf{Z} (\mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}^T \mathbf{R}^{-1} \right] \mathbf{y} \\ \hat{\mathbf{u}} = (\mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}) \end{cases}, \quad (2.27)$$

De acordo com os resultados de Pearson, ou no método dos mínimos quadrados generalizados (GLM, do inglês *Generalized Least Squares*), $\hat{\boldsymbol{\beta}}$ é dado por: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}$, portanto, para provar que $\hat{\boldsymbol{\beta}}$ nas MME é uma solução GLM, basta mostrar que $\mathbf{V}^{-1} = \left[\mathbf{R}^{-1} - \mathbf{R}^{-1} \mathbf{Z} (\mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}^T \mathbf{R}^{-1} \right]$; o que pode ser feito como em

Henderson et al. (1959), provando que $\mathbf{V} \left[\mathbf{R}^{-1} - \mathbf{R}^{-1} \mathbf{Z} (\mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}^T \mathbf{R}^{-1} \right] = \mathbf{I}$:

$$\begin{aligned}
& (\mathbf{Z} \mathbf{G} \mathbf{Z}^T + \mathbf{R}) \left[\mathbf{R}^{-1} - \mathbf{R}^{-1} \mathbf{Z} (\mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}^T \mathbf{R}^{-1} \right] \\
&= \mathbf{Z} \mathbf{G} \mathbf{Z}^T \mathbf{R}^{-1} - \mathbf{Z} \mathbf{G} \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} (\mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}^T \mathbf{R}^{-1} + \mathbf{I} - \mathbf{Z} (\mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}^T \mathbf{R}^{-1} \\
&= \mathbf{Z} \mathbf{G} \mathbf{Z}^T \mathbf{R}^{-1} - \mathbf{Z} (\mathbf{I} + \mathbf{G} \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z}) (\mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}^T \mathbf{R}^{-1} + \mathbf{I} \\
&= \mathbf{Z} \mathbf{G} \mathbf{Z}^T \mathbf{R}^{-1} - \mathbf{Z} \mathbf{G} (\mathbf{G}^{-1} + \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z}) (\mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}^T \mathbf{R}^{-1} + \mathbf{I} \\
&= \mathbf{Z} \mathbf{G} \mathbf{Z}^T \mathbf{R}^{-1} - \mathbf{Z} \mathbf{G} \mathbf{Z}^T \mathbf{R}^{-1} + \mathbf{I} \\
&= \mathbf{I}
\end{aligned} \quad (2.28)$$

Uma vez que $\hat{\boldsymbol{\beta}}$ nas MME é uma solução GLM, é possível mostrar que $\hat{\boldsymbol{\beta}}$ é o BLUE (do inglês, *Best Linear Unbiased Estimator*) dos efeitos fixos, pois, $E[\hat{\boldsymbol{\beta}}] = E\left[(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}\right] = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \boldsymbol{\beta} = \boldsymbol{\beta}$.

Para mostrar que $\hat{\mathbf{u}}$, nas MME equivale aos resultados de Pearson é possível seguir a prova de Henderson (1963), mas aqui, por conveniência, usaremos a propriedade de que sejam as matrizes $\mathbf{P}$ e $\mathbf{R}$, positivas definidas, então (PETERSEN; PEDERSEN, 2006):

$$(\mathbf{P}^{-1} + \mathbf{B}^T \mathbf{R}^{-1} \mathbf{B})^{-1} \mathbf{B}^T \mathbf{R}^{-1} = \mathbf{P} \mathbf{B} (\mathbf{B} \mathbf{P} \mathbf{B}^T + \mathbf{R})^{-1}, \quad (2.29)$$

Aplicando-se (2.29), na segunda equação em (2.27), tem-se:

$$\begin{aligned} (\mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}^T \mathbf{R}^{-1} &= (\mathbf{G}^{-1} + \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{R}^{-1} \\ &= \mathbf{G} \mathbf{Z}^T (\mathbf{Z} \mathbf{G} \mathbf{Z}^T + \mathbf{R})^{-1}, \\ &= \mathbf{G} \mathbf{Z}^T \mathbf{V}^{-1} \end{aligned} \quad (2.30)$$

Desta forma, substituindo-se (2.30) em (2.27), obtém-se exatamente (2.20), isto é, $\hat{\mathbf{u}} = \mathbf{G} \mathbf{Z}^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})$, assim, é possível mostrar que $\hat{\mathbf{u}}$ nas MME é o BLUP (do inglês, *Best Linear Unbiased Predictor*) de $\mathbf{u}$, pois, $E[\hat{\mathbf{u}}] = E[\mathbf{G} \mathbf{Z}^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})] = 0 = E[\mathbf{u}]$.

As MME propostas por C. R. Henderson são uma importante ferramenta para predição dos valores genéticos, bem como valem-se para estimação de componentes variância. Apesar da teoria por traz desse conjunto de equações ser bastante complexa, como visto acima, seu uso é bastante simplificado pela implementação do conjunto de equações (2.26), que pode ser reduzido para o caso unicaracterística, com variâncias residuais homogêneas (2.31), ou estendido para o caso multicaracterística (2.32):

$$\begin{bmatrix} \mathbf{X}^T \mathbf{X} & \mathbf{X}^T \mathbf{Z} \\ \mathbf{Z}^T \mathbf{X} & \mathbf{Z}^T \mathbf{Z} + \frac{\sigma_e^2}{\sigma_u^2} \mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \mathbf{y} \\ \mathbf{Z}^T \mathbf{y} \end{bmatrix}, \quad (2.31)$$

$$\begin{bmatrix} \mathbf{X}^T (\mathbf{R}_0^{-1} \otimes \mathbf{I}) \mathbf{X} & \mathbf{X}^T (\mathbf{R}_0^{-1} \otimes \mathbf{I}) \mathbf{Z} \\ \mathbf{Z}^T (\mathbf{R}_0^{-1} \otimes \mathbf{I}) \mathbf{X} & \mathbf{Z}^T (\mathbf{R}_0^{-1} \otimes \mathbf{I}) \mathbf{Z} + \mathbf{G}_0^{-1} \otimes \mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T (\mathbf{R}_0^{-1} \otimes \mathbf{I}) \mathbf{y} \\ \mathbf{Z}^T (\mathbf{R}_0^{-1} \otimes \mathbf{I}) \mathbf{y} \end{bmatrix}, \quad (2.32)$$

em que: σ_e^2 é a variância residual da característica em análise; σ_u^2 é a variância genética aditiva da característica em análise; $\mathbf{R}_0$ é a matriz de covariâncias residuais entre as características,

isto é: $\mathbf{R}_0 = \begin{bmatrix} \sigma_{e_1}^2 & \sigma_{e_{12}} & \cdots & \sigma_{e_{1n}} \\ \sigma_{e_{12}} & \sigma_{e_2}^2 & \cdots & \sigma_{e_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{e_{1n}} & \sigma_{e_{2n}} & \cdots & \sigma_{e_n}^2 \end{bmatrix}$; $\mathbf{G}_0$ é a matriz de covariâncias genéticas aditivas entre as

características, isto é: $\mathbf{G}_0 = \begin{bmatrix} \sigma_{u_1}^2 & \sigma_{u_{12}} & \cdots & \sigma_{u_{1n}} \\ \sigma_{u_{12}} & \sigma_{u_2}^2 & \cdots & \sigma_{u_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{u_{1n}} & \sigma_{u_{2n}} & \cdots & \sigma_{u_n}^2 \end{bmatrix}$; $\mathbf{A}$ é a matriz de parentesco aditivo.

Em (2.32), as matrizes de incidência $\mathbf{X}$ e $\mathbf{Z}$, bem como os vetores $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{u}}$ e $\mathbf{y}$ passam a ser escritos como composições, de acordo com cada característica em análise (MRODE, 2014),

$$\text{isto é: } \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_n \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_n \end{bmatrix}, \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_n \end{bmatrix}, \hat{\mathbf{u}} = \begin{bmatrix} \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \\ \vdots \\ \hat{\mathbf{u}}_n \end{bmatrix} \text{ e } \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix}. \text{ Por}$$

fim, é possível mostrar que a variância de amostragem, ou variância dos erros de predição, para $\hat{\boldsymbol{\beta}}$ e $\hat{\mathbf{u}}$ é dada pela inversa da matriz de coeficientes das MME (RESENDE; SILVA; AZEVEDO, 2014):

$$\text{Var} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} - \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-1} = \mathbf{C}^{-1}, \quad (2.33)$$

Provar que (2.33) é verdadeira, envolve um pequeno “artifício matemático”. Fazendo $\mathbf{W} = [\mathbf{X} \quad \mathbf{Z}]$ e $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix}$, (2.19) pode ser escrito como $\mathbf{y} = \mathbf{W}\boldsymbol{\theta} + \mathbf{e}$. Assumindo que

$\text{Var}(\boldsymbol{\theta}) = \boldsymbol{\Lambda} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{bmatrix}$, as MME podem então ser reescritas como:

$$(\mathbf{W}^T \mathbf{R}^{-1} \mathbf{W} + \boldsymbol{\Lambda}^{-1}) \hat{\boldsymbol{\theta}} = \mathbf{W}^T \mathbf{R}^{-1} \mathbf{y}, \quad (2.34)$$

Fazendo-se $\mathbf{C} = (\mathbf{W}^T \mathbf{R}^{-1} \mathbf{W} + \boldsymbol{\Lambda}^{-1})$ facilmente se obtém (2.33), e a variância de $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ é dada por:

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) &= \text{Var} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} - \mathbf{u} \end{bmatrix} \\
&= \text{Var}(\mathbf{C}^{-1} \mathbf{W}^T \mathbf{R}^{-1} \mathbf{y} - \boldsymbol{\theta}) \\
&= \text{Var}[\mathbf{C}^{-1} \mathbf{W}^T \mathbf{R}^{-1} (\mathbf{W} \boldsymbol{\theta} + \mathbf{e}) - \boldsymbol{\theta}] \\
&= \text{Var}[\mathbf{C}^{-1} \mathbf{W}^T \mathbf{R}^{-1} \mathbf{W} \boldsymbol{\theta} - \boldsymbol{\theta} + \mathbf{C}^{-1} \mathbf{W}^T \mathbf{R}^{-1} \mathbf{e}] \\
&= \text{Var}[\mathbf{C}^{-1} (\mathbf{W}^T \mathbf{R}^{-1} \mathbf{W} - \mathbf{I}) \boldsymbol{\theta} + \mathbf{C}^{-1} \mathbf{W}^T \mathbf{R}^{-1} \mathbf{e}] \\
&= \text{Var}[\mathbf{C}^{-1} (\mathbf{W}^T \mathbf{R}^{-1} \mathbf{W} - \mathbf{C}) \boldsymbol{\theta} + \mathbf{C}^{-1} \mathbf{W}^T \mathbf{R}^{-1} \mathbf{e}] \quad , \tag{2.35} \\
&= \text{Var}[\mathbf{C}^{-1} \boldsymbol{\Lambda}^{-1} \boldsymbol{\theta} + \mathbf{C}^{-1} \mathbf{W}^T \mathbf{R}^{-1} \mathbf{e}] \\
&= \text{Var}(\mathbf{C}^{-1} \boldsymbol{\Lambda}^{-1} \boldsymbol{\theta}) + \text{Var}(\mathbf{C}^{-1} \mathbf{W}^T \mathbf{R}^{-1} \mathbf{e}) \\
&= \mathbf{C}^{-1} \boldsymbol{\Lambda}^{-1} \boldsymbol{\Lambda} \boldsymbol{\Lambda}^{-1} \mathbf{C}^{-1} + \mathbf{C}^{-1} \mathbf{W}^T \mathbf{R}^{-1} \mathbf{R} \mathbf{R}^{-1} \mathbf{W} \mathbf{C}^{-1} \\
&= \mathbf{C}^{-1} \boldsymbol{\Lambda}^{-1} \mathbf{C}^{-1} + \mathbf{C}^{-1} \mathbf{W}^T \mathbf{R}^{-1} \mathbf{W} \mathbf{C}^{-1} \\
&= \mathbf{C}^{-1} (\boldsymbol{\Lambda}^{-1} + \mathbf{W}^T \mathbf{R}^{-1} \mathbf{W}) \mathbf{C}^{-1} \\
&= \mathbf{C}^{-1} \mathbf{C} \mathbf{C}^{-1} \\
&= \mathbf{C}^{-1}
\end{aligned}$$

Uma vez que conhecemos a variância dos estimadores de $\boldsymbol{\beta}$ e $\mathbf{u}$, podemos conceitualizar a acurácia de predição dos efeitos aleatórios. De acordo com Resende, Silva e Azevedo (2014), um método ótimo de predição deve minimizar o erro quadrático médio (EQM), o qual é dado por: $\text{EQM} = \delta^2 + \text{PEV}$, em que δ é o viés e PEV representa a variância dos erros de predição. Conforme visto anteriormente, a própria derivação das MME garante a ausência de viés, de tal sorte que $\text{EQM} = \text{PEV}$.

Se $\mathbf{R}_0 \neq \mathbf{I} \sigma_e^2$ então $\text{PEV} = \mathbf{C}^{-1}$, caso contrário $\text{PEV} = \mathbf{C}^{-1} \sigma_e^2$ (RESENDE; SILVA; AZEVEDO, 2014). Podemos entender então que a PEV seja ainda equivalente a: $\text{PEV} = (1 - r_{\hat{\mathbf{u}}\mathbf{u}}^2) \sigma_u^2$, em que $r_{\hat{\mathbf{u}}\mathbf{u}}^2$ representa o quadrado do coeficiente de correlação entre o valor de $\mathbf{u}$ predito e o verdadeiro (em inglês *reliability*) (MRODE, 2014; RESENDE; SILVA; AZEVEDO, 2014). Igualando-se as duas formas de mensurarmos a PEV e desenvolvendo para $r_{\hat{\mathbf{u}}\mathbf{u}}^2$, encontramos a acurácia de predição teórica:

$$\begin{aligned}
c_{22}^{-1}\sigma_e^2 &= (1 - r_{\hat{u}u}^2)\sigma_u^2 \\
\frac{c_{22}^{-1}\sigma_e^2}{\sigma_u^2} &= (1 - r_{\hat{u}u}^2) \\
\frac{c_{22}^{-1}\sigma_e^2}{\sigma_u^2} - 1 &= -r_{\hat{u}u}^2, \\
r_{\hat{u}u}^2 &= 1 - \frac{c_{22}^{-1}\sigma_e^2}{\sigma_u^2} \\
r_{\hat{u}u} &= \sqrt{1 - \frac{c_{22}^{-1}\sigma_e^2}{\sigma_u^2}}
\end{aligned} \tag{2.36}$$

Em que: c_{22}^{-1} é o elemento diagonal, correspondente ao efeito aleatório, de C^{-1} . Vale ressaltar que, no caso do melhoramento genético, geralmente a variável aleatória de interesse para predição é o valor genético aditivo. Quando o indivíduo é endogâmico a acurácia pode ser ajustada para a endogamia (F), fazendo com que suas predições não sejam tão penalizadas (MRODE, 2014), assim tem-se:

$$r_{\hat{u}u} = \sqrt{1 - \frac{c_{22}^{-1}\sigma_e^2}{(1-F)\sigma_u^2}}, \tag{2.37}$$

2.3. “Um pouquinho de inspiração”

No momento da defesa, a banca sugeriu que eu encerrassem esta revisão de literatura com uma reflexão sobre os *papers* que eu considerava os mais relevantes para minha tese neste momento. Resolvi preparar uma pequena lista com os trabalhos mais inspiradores que eu li ao longo da condução deste doutorado, juntamente com alguma reflexão, quando oportuno, sobre eles:

- MISZTAL, I. et al. Emerging issues in genomic selection. **Journal of Animal Science**, v. 99, n. 6, 2021.
- MEYER, K.; KIRKPATRICK, M. Perils of Parsimony: Properties of Reduced-Rank Estimates of Genetic Covariance Matrices. **Genetics**, v. 180, n. 2, p. 1153–1166, 2008.
- GIANOLA, D.; VAN KAAM, J. B. C. H. M. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. **Genetics**, v. 178, n. 4, p. 2289–2303, 2008.
- VALENTE, B. D. et al. Is Structural Equation Modeling Advantageous for the Genetic

Improvement of Multiple Traits? **Genetics**, v. 194, n. 3, p. 561–572, 2013.

- LOURENÇO, D. et al. Single-Step Genomic Evaluations from Theory to Practice: Using SNP Chips and Sequence Data in BLUPF90. **Genes**, v. 11, n. 7, p. 790, 2020.

Nesses trabalhos pude encontrar a fundamentação teórica que faltava quando algum resultado não fazia muito sentido, ou mesmo quando alguma análise não convergia. O paper do Prof. Ignacy Misztal é assombrosamente inspirador, à medida em que mostra que, por vezes, precisamos estar cientes de que “causamos os problemas, nós mesmos”. O trabalho da Profa. Karin Meyer é de longe o mais teórico, todos os seus *papers* costumam ser bastante densos e informativos, neste que separei e faço destaque, ela apresenta uma série de propriedades muito importantes dos métodos de redução de *rank*. Para quem me conhece, não podia faltar um trabalho do Prof. Daniel Gianola. Este trabalho me tirou o sono, e me tomou várias semanas de estudo e dedicação para entender todas as temidas fórmulas e deduções. Sem sombra de dúvidas, um esforço recompensado ao final do trabalho.

O trabalho seguinte, do Prof. Bruno Valente nem é o que fundamentou o uso de modelos de equações estruturais, sob meu ponto de vista. Mas, enquanto em todos os trabalhos, os autores se preocupavam apenas em encontrar os efeitos causais, neste são apresentadas as vantagens e desvantagens dessa classe de modelos do ponto de vista de seleção. Algo que simplesmente “abriu minha mente”. Por fim, mas não menos importante, o *paper* mais didático que eu já li sobre *single-step*. De autoria da Profa. Daniela Lourenço (sou sua fã), neste trabalho ela trata com maestria os detalhes práticos da utilização de dados genômicos e dos programas da família BLUPF90 para predições genômicas via *single-step*. Vale muito a leitura e o estudo, principalmente a parte em que são apresentadas as matrizes **G** e **H**.

Como não seria eu sem um pouco de referência histórica, acrescento a esta lista alguns trabalhos que dispensam apresentação, mas que precisariam ser lidos por absolutamente todos os estudantes de mestrado e doutorado em melhoramento genético:

- FALCONER, D. S.; MACKAY, T. F. C. **Introduction to Quantitative Genetics**. 4. ed. Essex: Pearson Education Ltda, 1996.
- VAN VLECK, L. D. **Selection Index and Introduction to Mixed Models Methods**. 1. ed. Boca Raton: CRC Press, 1993.
- SEARLE, S. R.; CASELLA, G.; MCCULLOCH, C. E. **Variance Components**. 1. ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 1992.

Por fim, toda essa divagação sobre o que deveria servir de inspiração, me levou a rever dois “achados” da minha coleção de trabalhos impressos, os quais eu diria, sem medo nenhum, servirão ainda durante muitos anos para o melhoramento genético animal e para os futuros pesquisadores:

- SEARLE, S. R. The matrix Handling of BLUE and BLUP in the mixed linear model. **Linear Algebra and its Applications**, v. 264, p. 291–311, 1997.
- VAN VLECK, L. D. Contemporary Groups for Genetic Evaluations. **Journal of Dairy Science**, v.70, p. 2456–1464, 1987.

2.4. Referências

ABCCCAMPOLINA, Associação Brasileira dos Criadores do Cavalo Campolina. **História da Raça**. 2017a. Disponível em: <<http://www.campolina.org.br/sobre.php>>. Acesso em: 22 ago. 2020.

ABCCCAMPOLINA, Associação Brasileira dos Criadores do Cavalo Campolina. **O legado de Cassiano Campolina**. 2017b. Disponível em: <<http://www.campolina.org.br/hospital.php>>. Acesso em: 23 ago. 2020.

ABCCCAMPOLINA, Associação Brasileira dos Criadores do Cavalo Campolina. **Regulamento do Serviço de Registro Genealógico do cavalo Campolina – SRGCC**. Brasil: Ministério da Agricultura, Pesca e Abastecimento, 2018.

ABDI, H.; WILLIAMS, L. J. Principal component analysis. **Wiley Interdisciplinary Reviews: Computational Statistics**, v. 2, n. 4, p. 433–459, 2010.

AGUILAR, I. et al. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Dairy Science**, v. 93, n. 2, p. 743–752, 2010.

ANDERSON, T. W.; RUBIN, H. Statistical Inference in Factor Analysis. In: BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 3. **Proceedings...** Berkeley, CA: University of California Press, 1956. Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry.

ARAÚJO, N. **Cavalos marchadores brasileiros**. Rio de Janeiro, Brazil: Globo Rural, 2013. Disponível em: <<https://globoplay.globo.com/v/2328452/>>. Acesso em: 10 de jun. de 2021.

BARTHOLOMEW, D. J. et al. **Analysis of multivariate social science data**. 2. ed. London: Chapman & Hall/CRC, 2008.

BARTHOLOMEW, D. J. Principal components analysis. In: **International encyclopedia of education**. Oxford: Elsevier, 2010. p. 374–377.

- BARTLETT, M. S. The statistical conception of mental factors. **British Journal of Psychology. General Section**, v. 28, n. 1, p. 97–104, 1937.
- BARTLETT, M. S. Tests of significance in factor analysis. **British Journal of Statistical Psychology**, v. 3, n. 2, p. 77–85, 1950.
- BENTLER, P. M. Comparative fit indexes in structural models. **Psychological Bulletin**, v. 107, n. 2, p. 238–246, 1990.
- BENTLER, P. M. **EQS structural equations program manual**. 1. ed. Encino, CA, 1995.
- BENTLER, P. M.; BONETT, D. G. Significance tests and goodness of fit in the analysis of covariance structures. **Psychological Bulletin**, v. 88, n. 3, p. 588–606, 1980.
- BOLDMAN, K. G.; VAN VLECK, L. D. Derivative-Free Restricted Maximum Likelihood Estimation in Animal Models with a Sparse Matrix Solver. **Journal of Dairy Science**, v. 74, n. 12, p. 4337–4343, 1991.
- BRASIL. Ministério da Agricultura Pecuária e Abastecimento. **Revisão do estudo do complexo do agronegócio do cavalo**. Brasília-DF: MAPA, 2016.
- BROWN, T. A. **Confirmatory Factor Analysis for Applied Research**. 2 ed. New York: The Guilford Press, 2015.
- BUZANSKAS, M. E. et al. Genetic parameter estimates and principal component analysis of breeding values of reproduction and growth traits in female Canchim cattle. **Reproduction, Fertility and Development**, v. 25, n. 5, p. 775, 2013.
- CATTELL, R. B. The scree test for the number of factors. **Multivariate Behavioral Research**, v. 1, n. 2, p. 245–276, 1966.
- CHARLESWORTH, B. Haldane and modern evolutionary genetics. **Journal of Genetics**, v. 96, n. 5, p. 773–782, 24 2017.
- CHEN, C. Y. et al. Effect of different genomic relationship matrices on accuracy and scale1. **Journal of Animal Science**, v. 89, n. 9, p. 2673–2679, 2011.
- CHRISTENSEN, O. F.; LUND, M. S. Genomic prediction when some animals are not genotyped. **Genetics Selection Evolution**, v. 42, n. 1, p. 2, 2010.
- COOK, L. M.; TURNER, J. R. G. Fifty per cent and all that: what Haldane actually said. **Biological Journal of the Linnean Society**, v. 129, n. 3, p. 765–771, 2020.
- CORRENS, C. G. Mendel's Regel über das Verhalten der Nachkommenschaft der Rassenbastarde. **Berichte der Deutschen Botanischen Gesellschaft**, v. 18, n. 156, p. 158–168, 1900.
- COSTA, M. D. et al. Estudo da subdivisão genética da raça Mangalarga Marchador. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 57, n. 2, p. 272–280, 2005.
- CROW, J. F. Hardy, Weinberg and language impediments. **Genetics**, v. 152, n. 3, p. 821–825, 1999.

CRUMP, S. L. **The estimation of components of variance in multiple classifications**. 1947. 78 f. Tese (Doutorado) – Iowa State University, Ames: Iowa, 1947.

CRUMP, S. L. The Present Status of Variance Component Analysis. **Biometrics**, v. 7, n. 1, p. 1–16, 1951.

CUNNINGHAM, E. P.; HENDERSON, C. R. An Iterative Procedure for Estimating Fixed Effects and Variance Components in Mixed Model Situations. **Biometrics**, v. 24, n. 1, p. 13–25, 1968.

DARTON, R. A. Rotation in Factor Analysis. **The Statistician**, v. 29, n. 3, p. 167, 1980.

DARWIN, C. **On the origin of species by means of natural selection**. 1. ed. London: John Murray, 1859.

DE OLIVEIRA BUSSIMAN, F. **Estudo de associação genômica ampla para as diferenças genéticas entre as marchas batida e picada em equinos Mangalarga Marchador**. 2018. 140 f. Dissertação (Mestrado) – Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo, Pirassununga, 2018.

DE OLIVEIRA BUSSIMAN, F. et al. Análise genética multicaracterística para escores vizuais de andamento em cavalos da raça Campolina. In: SIMMELHOR - I SIMPÓSIO DE GENÉTICA E MELHORAMENTO ANIMAL, 1. **Anais...** Viçosa: Universidade Federal de Viçosa, 2018a.

DE OLIVEIRA BUSSIMAN, F. et al. Pedigree analysis and inbreeding effects over morphological traits in Campolina horse population. **Animal**, v. 12, n. 11, p. 2246–2255, 2018b.

DE OLIVEIRA BUSSIMAN, F. et al. Genetic analysis of morphological and functional traits in Campolina horses using Bayesian multi-trait model. **Livestock Science**, v. 216, n. April, p. 119–129, 2018c.

DE OLIVEIRA BUSSIMAN, F. et al. Allelic and genotypic frequencies of the DMRT3 gene in the Brazilian horse breed Mangalarga Marchador and their association with types of gait. **Genetics and Molecular Research**, v. 18, n. 1, 2019.

DE VRIES, H. Sur les unités des caractères spécifiques et leur application à l'étude des Hybrides. **Revue Générale de Botanique**, v. 12, p. 257–271, 1900a.

DE VRIES, H. Sur la loi de disjonction des Hybrides. **Comptes Rendus de Académie des Sciences**, v. 130, p. 845–847, 1900b.

DE VRIES, H. Das Spaltungsgesetz der Bastarde. **Berichte der Deutschen Botanischen Gesellschaft**, v. 18, n. 3, p. 83–90, 1900c.

DEL PRIORE, M. L. M. **Festas e utopias no Brasil colonial**. 1. ed. São Paulo: Brasiliense, 1994.

DEMPFLE, L. Relation entre BLUP (Best Linear Unbiased Prediction) et estimateurs bayésiens. **Genetics Selection Evolution**, v. 9, n. 1, p. 27, 1977.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 39, n. 1, p. 1–38, 1977.

DICKERSON, G. E. Composition of hog carcasses as influenced by heritable differences in rate and economy of gain. **Iowa Agricultural Experiment Station Research Bulletin**, v. 354, n. 354, p. 492–524, 1947.

DUARTE, Foto. **Ferro Elétrico do Haras Ventania dun. H., Campolina, 2014**. 2014. Disponível em: <https://www.allbreedpedigree.com/ferro+eletrico+do+haras+ventania>. Acesso em: 26 de ago. 2020.

EDWARDS, E. H. **Cavalos: Um guia ilustrado com mais de 100 raças de cavalos de todo o mundo**. 1. ed. Rio de Janeiro: Ediouro S.A., 1994.

EFRON, B.; GONG, G. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. **The American Statistician**, v. 37, n. 1, p. 36, 1983.

EISENHART, C. The Assumptions Underlying the Analysis of Variance. **Biometrics**, v. 3, n. 1, p. 1–21, 1947.

FALCONER, D. S. **Introduction to Quantitative Genetics**. 1. ed. London: The Ronald Press Company, 1960.

FALCONER, D. S. Maternal effects and selection response. In: (S. J. GEERT, Ed.) **GENETICS TODAY: PROCEEDINGS OF THE XI INTERNATIONAL CONGRESS OF GENETICS**, 11. **Anais...** Hagues: Oxford, Pergamon Press, 1965.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to Quantitative Genetics**. 4. ed. Essex: Pearson Education Ltda, 1996.

FERNANDO, R. L. et al. Computational strategies for alternative single-step Bayesian regression models with large numbers of genotyped and non-genotyped animals. **Genetics Selection Evolution**, v. 48, n. 1, p. 96, 2016.

FERNANDO, R. L.; DEKKERS, J. C.; GARRICK, D. J. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. **Genetics Selection Evolution**, v. 46, n. 1, p. 50, 2014.

FERNANDO, R. L.; GIANOLA, D. Optimal properties of the conditional mean as a selection criterion. **Theoretical and Applied Genetics**, v. 72, n. 6, p. 822–825, 1986.

FISHER, R. A. The Correlation between Relatives on the Supposition of Mendelian Inheritance. **Transactions of the Royal Society of Edinburgh**, v. 52, n. 2, p. 399–433, 1919.

FISHER, R. A. Theory of Statistical Estimation. **Mathematical Proceedings of the Cambridge Philosophical Society**, v. 22, n. 5, p. 700–725, 1925.

FISHER, R. A. **The genetical theory of natural selection**. 1. ed. Oxford: Clarendon Press, 1930.

FISHER, R. A. Natural Selection from the genetical standpoint. **Australian Journal of**

Science, v. 22, p. 16–17, 1959.

FORD, J. K.; MACCALLUM, R. C.; TAIT, M. The Application of Exploratory Factor Analysis in Applied Psychology: a critical review and analysis. **Personnel Psychology**, v. 39, n. 2, p. 291–314, 1986.

GABRIEL, K. R. The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. **Biometrika**, v. 58, n. 3, p. 453, 1971.

GABRIEL, K. R. Biplot display of multivariate matrices for inspection of data and diagnosis. In: BARNETT, V. (Ed.) **Interpreting Multivariate Data**. 1. ed. Chichester: Wiley, 1981. p. 147–173.

GALTON, F. Regression Towards Mediocrity in Hereditary Stature. **The Journal of the Anthropological Institute of Great Britain and Ireland**, v. 15, p. 246–263, 1886.

GARNETT, J. C. M. On certain independent factors in mental measurements. **Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character**, v. 96, n. 675, p. 91–111, 1919.

GAUSS, C. F. **Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium**. 1. ed. Hamburg: Hamburgi Sumtibus Frid. Perthes et I. H. Besser, 1809.

GENTLE, J. E. Basic Properties of Matrices. In: CASELLA, G.; FIENBERG, S.; OLKIN, I. (Eds.). **Matrix Algebra: theory, computations, and applications in statistics**. 1. ed. New York: Springer-Verlag, 2007. p. 41–144.

GIANOLA, D.; FERNANDO, R. L. Bayesian Methods in Animal Breeding Theory. **Journal of Animal Science**, v. 63, n. 1, p. 217–244, 1986.

GIANOLA, D.; FOULLEY, J.; FERNANDO, R. Prediction of breeding values when variances are not known. **Genetics Selection Evolution**, v. 18, n. 4, p. 485, 1986.

GIANOLA, D.; ROSA, G. J. M. One Hundred Years of Statistical Developments in Animal Breeding. **Annual Review of Animal Biosciences**, v. 3, n. 1, p. 19–56, 2015.

GILMOUR, A. R. et al. ASREML, an average information REML program. **New South Wales Agriculture Biometric Bulletin**, n. 3, 1996.

GRASER, H.-U.; SMITH, S. P.; TIER, B. A Derivative-Free Approach for Estimating Variance Components in Animal Models by Restricted Maximum Likelihood. **Journal of Animal Science**, v. 64, n. 5, p. 1362–1370, 1987.

GRILZ-SEGER, G. et al. High-resolution population structure and runs of homozygosity reveal the genetic architecture of complex traits in the Lipizzan horse. **BMC Genomics**, v. 20, n. 1, p. 174, 2019.

GROENEVELD, E. REML VCE - a multivariate multimodel restricted maximum likelihood (co)variance component estimation package. In: EC Symposium on Application of Mixed Linear Models in the Prediction of Genetic Merit in Pigs, 1. **Proceedings...** Mariensee: Groeneveld, 1994.

GROSU, H. et al. **History of Genetic Evaluation Methods in Dairy Cattle**. 1. ed. Bucharest: Sc Editura Ceres SA, 2014.

GUTTMAN, L. Some necessary conditions for common-factor analysis. **Psychometrika**, v. 19, n. 2, p. 149–161, 1954.

HAAVELMO, T. The Statistical Implications of a System of Simultaneous Equations. **Econometrica**, v. 11, n. 1, p. 1, 1943.

HALDANE, J. B. S. **The Causes of Evolution**. 1. ed. London: Longmans, Green and Co. Ltd., 1932.

HARAS São Judas Tadeu. **Lançamento! Cotas de Coberturas do Tri Grande Campeão da Raça e Bi Grande Marchador Nacional**. 2020. 1 fotografia. Disponível em: <https://www.facebook.com/harassaojudas/photos/a.202373300149242/1101689630217600>. Acesso em: 26 de ago. 2020.

HARDY, G. H. Mendelian proportions in a mixed population. **Science**, v. 28, n. 706, p. 49 LP – 50, 1908.

HARTLEY, H. O.; RAO, J. N. K. Maximum-Likelihood Estimation for the Mixed Analysis of Variance Model. **Biometrika**, v. 54, n. 1/2, p. 93, 1967.

HARVILLE, D. A. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. **Journal of the American Statistical Association**, v. 72, n. 358, p. 320–338, 1977.

HAZEL, L. N. The Genetic Basis for Constructing Selection Indexes. **Genetics**, v. 28, n. 6, p. 476–490, 1943.

HAZEL, L. N.; LUSH, J. L. The efficiency of three methods of selection. **Journal of Heredity**, v. 33, n. 11, p. 393–399, 1942.

HAZEL, L. N.; TERRILL, C. E. Effects of Some Environmental Factors on Weanling Traits of Range Rambouillet Lambs. **Journal of Animal Science**, v. 4, n. 4, p. 331–341, 1945.

HENDERSON, A. C. R. A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. **Biometrics**, v. 32, n. 1, p. 69–83, 1976.

HENDERSON, C. R. Estimation of Changes in Herd Environment. In: FORTY-FOURTH ANNUAL MEETING OF THE AMERICAN DAIRY SCIENCE ASSOCIATION, 44. **Book of Abstracts...** Minneapolis, Minnesota: Elsevier, 1949.

HENDERSON, C. R. Estimation of Genetic Parameters. In: THE ANNALS OF MATHEMATICAL STATISTICS. **Anais...** Chapel Hill, North Carolina: The Annals of Mathematical Statistics, 1950.

HENDERSON, C. R. Estimation of Variance and Covariance Components. **Biometrics**, v. 9, n. 2, p. 226–252, jun. 1953.

HENDERSON, C. R. et al. The Estimation of Environmental and Genetic Trends from

Records Subject to Culling. **Biometrics**, v. 15, n. 2, p. 192, 1959.

HENDERSON, C. R. Selection index and expected genetic advance. In: HANSON, W. D.; ROBINSON, H. F. (Eds.) **Statistical genetics and plant breeding**. Washington, DC: The Nacional Academies Press, 1963. p. 141–163.

HENDERSON, C. R. Sire evaluation and genetic trends. **Journal of Animal Science**, v. 1973, n. Symposium, p. 10–41, 1973.

HENDERSON, C. R. Rapid method for computing the inverse of a relationship matrix. **Journal of Dairy Science**, v. 58, n. 11, p. 1727–1730, 1975.

HENDERSON, C. R. Statistical Methods In Animal Improvement: Historical Overview. In: GIANOLA, D.; HAMMOND, K. (Eds.) **Advances in Statistical Methods for Genetic Improvement of Livestock**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1990. p. 2–14.

HENDERSON, C. R.; QUAAS, R. L. Multiple Trait Evaluation Using Relatives' Records. **Journal of Animal Science**, v. 43, n. 6, p. 1188–1197, 1976.

HENDERSON JR, C. R. Analysis of Covariance in the Mixed Model: Higher-Level, Nonhomogeneous, and Random Regressions. **Biometrics**, v. 38, n. 3, p. 623, 1982.

HEYWOOD, H. B. On Finite Sequences of Real Numbers. **Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character**, v. 134, n. 824, p. 486–501, 1931.

HILL, W. G.; MACKAY, T. F. C. D. S. Falconer and Introduction to Quantitative Genetics. **Genetics**, v. 167, n. 4, p. 1529–1536, 2004.

HONDA, T. et al. Pedigree analysis of genetic subdivision in a population of Japanese Black cattle. **Animal Science Journal**, v. 73, n. 6, p. 445–452, 2002.

HONGYU, K.; SANDANIELO, V. L. M.; OLIVEIRA-JR, G. J. Análise de Componentes Principais : resumo teórico, aplicação e interpretação. **E&S - Engineering and Science**, v. 1, n. 5, p. 83–90, 2015.

HORN, J. L. A rationale and test for the number of factors in factor analysis. **Psychometrika**, v. 30, n. 2, p. 179–185, 1965.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, v. 24, n. 6, p. 417–441, 1933.

HU, L.; BENTLER, P. M. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. **Structural Equation Modeling: A Multidisciplinary Journal**, v. 6, n. 1, p. 1–55, 1999.

IEZZONI, A. F.; PRITTS, M. P. Applications of Principal Component Analysis to Horticultural Research. **HortScience**, v. 26, n. 4, p. 334–338, 1991.

JENSEN, J.; MADSEN, P. DMU: A package for the analysis of multivariate mixed models. In: 5th CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION. **Proceedings...** Guelph: Wageningen University, 1994.

JHONSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 6. ed. Upper Saddle River, New Jersey: Pearson Education Limited, 2018.

JOHNSON, D. L.; THOMPSON, R. Restricted Maximum Likelihood Estimation of Variance Components for Univariate Animal Models Using Sparse Matrix Techniques and Average Information. **Journal of Dairy Science**, v. 78, n. 2, p. 449–456, 1995.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 6. ed. Upper Saddle River: Pearson Education Ltda, 2007.

JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 374, n. 2065, p. 20150202, 2016.

KAISER, H. F. The Application of Electronic Computers to Factor Analysis. **Educational and Psychological Measurement**, v. 20, n. 1, p. 141–151, 1960.

KAMETARO, T. On the Hybridology of the Silkworm. **Bulletin of the Tokyo Imperial University College of Agriculture**, v. 168, n. 1, p. 259–393, 1906.

KASSAMBARA, A. **Practical guide to Principal Component Methods in R**. 1. ed. Marseille: STHDA, 2017.

KEMPTHORNE, O. **An Introduction to Genetic Statistics**. 1. ed. New York: John Wiley & Sons, Inc., 1957.

KIRKPATRICK, M.; HECKMAN, N. A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. **Journal of Mathematical Biology**, v. 27, n. 4, p. 429–450, 1989.

KOCH, R. M.; CLARK, R. T. Genetic and Environmental Relationships among Economic Characters in Beef Cattle III. Evaluating Maternal Environment. **Journal of Animal Science**, v. 14, n. 4, p. 979–996, 1955.

KOERHUIS, A.; THOMPSON, R. Models to estimate maternal effects for juvenile body weight in broiler chickens. **Genetics Selection Evolution**, v. 29, n. 2, p. 225, 1997.

LAIRD, N. M.; WARE, J. H. Random-Effects Models for Longitudinal Data. **Biometrics**, v. 38, n. 4, p. 963, 1982.

LANDE, R.; THOMPSON, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. **Genetics**, v. 124, n. 3, p. 743–756, 1990.

LATTIN, J. M.; CARROLL, J. D.; GREEN, P. E. Análise de Componentes Principais. In: LA ROSA, P.; QUIRINO DE SOUZA, G. G.; ALENCAR, F. A. (Eds.) **Análise de Dados Multivariados**. 1. ed. São Paulo: Cengage Learning, 2011a. p. 67–101.

LATTIN, J. M.; CARROLL, J. D.; GREEN, P. E. Análise fatorial exploratória. In: LA ROSA, P.; QUIRINO DE SOUZA, G.; ALENCAR, F. (Eds.) **Análise de Dados Multivariados**. 1. ed. São Paulo: Cengage Learning, 2011b. p. 103–137.

LEGARRA, A.; AGUILAR, I.; MISZTAL, I. A relationship matrix including full pedigree

and genomic information. **Journal of Dairy Science**, v. 92, p. 4656–4663, 2009.

LEGENDRE, A. M. **Nouvelles Méthodes pour la détermination des orbites des comètes**. 1. ed. Paris: Firmin Didot, Libraire pour les Mathématiques, 1805.

LI, H.; RALPH, P. Local PCA Shows How the Effect of Population Structure Differs Along the Genome. **Genetics**, v. 211, n. 1, p. 289–304, 2019.

LIU, Z. et al. A single-step genomic model with direct estimation of marker effects. **Journal of Dairy Science**, v. 97, n. 9, p. 5833–5850, 2014.

LIU, Z. et al. Technical note: Equivalent genomic models with a residual polygenic effect. **Journal of Dairy Science**, v. 99, n. 3, p. 2016–2025, 2016.

LOEHLIN, J. C. **Latent Variable Models: an introduction to factor, path, and structural equation analysis**. 4. ed. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc., 2004.

LOURENÇO, D. et al. Single-Step Genomic Evaluations from Theory to Practice: Using SNP Chips and Sequence Data in BLUPF90. **Genes**, v. 11, n. 7, p. 790, 2020.

LUSH, J. L. The Bull Index Problem in the Light of Modern Genetics. **Journal of Dairy Science**, v. 16, n. 6, p. 501–522, 1933.

LUSH, J. L. **Animal Breeding Plans**. 1. ed. Ames: The Collegiate Press, Inc. Iowa State College, 1937.

LUSH, J. L. Family Merit and Individual Merit as Bases for Selection. **The American Naturalist**, v. 81, n. 799, p. 241–261, 1947.

MAITI, J. **Mod-01 Lec-33 Factor Analysis**, 9 maio 2014a. Disponível em: <<https://www.youtube.com/watch?v=n3y3xLNoPk4&list=WL&index=5&t=1301s>>. Acesso em: 8 ago. 2020

MAITI, J. **Mod-01 Lec-35 Factor Analysis -- Model Adequacy, rotation, factor scores & case study**, 9 maio 2014b. Disponível em: <<https://www.youtube.com/watch?v=ZkJ5tqG6uY0&list=WL&index=3&t=3050s>>. Acesso em: 8 ago. 2020

MALÉCOT, G. **Les mathématiques de l'hérédité**. 1. ed. Paris: Barnéoud frères, 1948.

MANLY, B. J. F. **Métodos Estatísticos Multivariados uma introdução**. 3. ed. Porto Alegre: Bookman, 2008.

MARIOTTI, Paulo. **Gata do Anaití**. 2019. 10 fotografias. Disponível em: <https://pt.calameo.com/read/005431395403d43fa84e2>. Acesso em: 25 de ago. 2020.

MARTINS DA SILVA, M. As cavalladas de pirenópolis: um estudo sobre sociedade, festas e espaço urbano (1940-1988). **História Revista**, v. 6, n. 1, p. 135–162, 2010.

MENDEL, G. **Versuche über Pflanzen-Hybriden**. Verhandlungen des naturforschenden Vereines in Brünn. Brünn: 1866

- MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819–1829, 2001.
- MEYER, K. DFREML-A Set of Programs to Estimate Variance Components Under an Individual Animal Model. **Journal of Dairy Science**, v. 71, n. Suppl., p. 33–34, 1988.
- MEYER, K. Estimating variances and covariances for multivariate animal models by restricted maximum likelihood. **Genetics Selection Evolution**, v. 23, n. 1, p. 67, 1991.
- MEYER, K. Genetic principal components for live ultrasound scan traits of Angus cattle. **Animal Science**, v. 81, n. 3, p. 337–345, 2005.
- MEYER, K. WOMBAT-A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). **Journal of Zhejiang University SCIENCE B**, v. 8, n. 11, p. 815–821, 2007.
- MEYER, K. Factor-analytic models for genotype x environment type problems and structured covariance matrices. **Genetics Selection Evolution**, v. 41, n. 1, p. 21, 2009.
- MEYER, K.; HILL, W. G. Estimation of genetic and phenotypic covariance functions for longitudinal or ‘repeated’ records by restricted maximum likelihood. **Livestock Production Science**, v. 47, n. 3, p. 185–200, 1997.
- MISZTAL, I. et al. BLUPF90 and related programs (BGF90). In: 7th WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 7. **Proceeding...** Montpellier: Wageningen University, 2002.
- MISZTAL, I. Properties of random regression models using linear splines. **Journal of Animal Breeding and Genetics**, v. 123, n. 2, p. 74–80, 2006.
- MISZTAL, I. et al. Choice of parameters for a single-step genomic evaluation. In: 61th ANNUAL MEETING OF THE EUROPEAN FEDERATION OF ANIMAL SCIENCE, 61. **Book of Abstracts...** Heraklion, Greece: Wageningen Academic Publishers, 2010.
- MISZTAL, I. et al. Unknown-parent groups in single-step genomic evaluation. **Journal of Animal Breeding and Genetics**, v. 130, n. 4, p. 252–258, 2013.
- MISZTAL, I.; LEGARRA, A.; AGUILAR, I. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. **Journal of Dairy Science**, v. 92, n. 9, p. 4648–4655, 2009.
- MRODE, R. A. **Linear models for the prediction of animal breeding values**. 3. ed. Oxford: CABI Publishing, 2014.
- NEIMANN-SORENSEN, A.; ROBERTSON, A. The Association between Blood Groups and Several Production Characteristics in Three Danish Cattle Breeds. **Acta Agriculturae Scandinavica**, v. 11, n. 2, p. 163–196, 1961.
- NEJATI-JAVAREMI, A.; SMITH, C.; GIBSON, J. P. Effect of total allelic relationship on accuracy of evaluation and response to selection. **Journal of Animal Science**, v. 75, n. 7, p. 1738, 1997.

NICODEMUS, M. C.; CLAYTON, H. M. Temporal variables of four-beat, stepping gaits of gaited horses. **Applied Animal Behaviour Science**, v. 80, n. 1, p. 133–142, 2003.

OLDENBROEK, K.; VAN DER WAAIJ, L. **Textbook Animal breeding and Genetics for BSc students**. 1. ed. Wageningen: Centre for Genetic Resources The Netherlands and Animal Breeding and Genomics Centre, 2015.

PATTERSON, H. D. Theory of Cyclic Rotation Experiments. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 26, n. 1, p. 1–36, 1964.

PATTERSON, H. D.; THOMPSON, R. Recovery of Inter-Block Information when Block Sizes are Unequal. **Biometrika**, v. 58, n. 3, p. 545, 1971.

PEARSON, K. Contributions to the mathematical theory of evolution. **Philosophical Transactions of the Royal Society of London. (A.)**, v. 185, p. 71–110, 1894.

PEARSON, K. On lines and planes of closest fit to systems of points in space. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, v. 2, n. 11, p. 559–572, 1901.

PEARSON, K. Mathematical contributions to the theory of evolution. —XI. On the influence of natural selection on the variability and correlation of organs. **Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character**, v. 200, p. 1–66, 1903.

PEÑAGARICANO, F. et al. Searching for causal networks involving latent variables in complex traits: Application to growth, carcass, and meat quality traits in pigs¹. **Journal of Animal Science**, v. 93, n. 10, p. 4617–4623, 2015.

PETERSEN, K. B.; PEDERSEN, M. S. **The Matrix Cookbook**. Version: F ed. Kongens Lyngby: Technical University of Denmark, 2006.

PORTNOY, S. Maximizing the probability of correctly ordering random variables using linear predictors. **Journal of Multivariate Analysis**, v. 12, n. 2, p. 256–269, 1982.

POWELL, J. E.; VISSCHER, P. M.; GODDARD, M. E. Reconciling the analysis of IBD and IBS in complex trait studies. **Nature Reviews Genetics**, v. 11, n. 11, p. 800–805, 2010.

PROCÓPIO, A. M. M.; BERGMANN, J. A. G. A. G.; COSTA, M. D. D. Formação e demografia da raça Campolina. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 55, n. 3, p. 361–365, 2003.

QUAAS, R. L. Computing the Diagonal Elements and Inverse of a Large Numerator Relationship Matrix. **Biometrics**, v. 32, n. 4, p. 949, 1976.

QUAAS, R. L. Linear Prediction. In: **BLUP school handbook: Use of mixed models for prediction and for estimation of (co)variance components**. 1. ed. Armidale: Animal Genetics and Breeding Unit, University of New England, 1984. p. 1–76.

QUAAS, R. L.; POLLAK, E. J. Mixed Model Methodology for Farm and Ranch Beef Cattle Testing Programs. **Journal of Animal Science**, v. 51, n. 6, p. 1277–1287, 1980.

RAO, C. R. Estimation of Heteroscedastic Variances in Linear Models. **Journal of the American Statistical Association**, v. 65, n. 329, p. 161–172, 1970.

RAO, C. R. Estimation of variance and covariance components—MINQUE theory. **Journal of Multivariate Analysis**, v. 1, n. 3, p. 257–275, 1971a.

RAO, C. R. Minimum variance quadratic unbiased estimation of variance components. **Journal of Multivariate Analysis**, v. 1, n. 4, p. 445–456, 1971b.

RAO, C. R. Estimation of Variance and Covariance Components in Linear Models. **Journal of the American Statistical Association**, v. 67, n. 337, p. 112–115, 1972.

RENCHE, A. C.; SCHAALJE, G. B. Matrix Algebra. In: **Linear Models in Statistics**. 2. ed. Hoboken, New Jersey: John Wiley & Sons, 2008. p. 5–68.

RESENDE, M. D. V.; SILVA, F. F.; AZEVEDO, C. F. Modelos Lineares Mistos. In: **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**. 1. ed. Viçosa: Suprema Gráfica e Editora Ltda, 2014. p. 159–190.

REZENDE, A. S. C.; MOURA, R. S. **Raças de equídeos no Brasil**. Belo Horizonte: Federal University of Minas Gerais, 2004.

RØNNINGEN, K. Some Properties of the Selection Index derived by “Henderson’s Mixed Model Method”. **Zeitschrift für Tierzüchtung und Züchtungsbiologie**, v. 88, n. 1–4, p. 186–193, 1971.

ROZEBOOM, W. W. Estimation of cross-validated multiple correlation: A clarification. **Psychological Bulletin**, v. 85, n. 6, p. 1348–1351, 1978.

SANTIAGO, J. M. M. et al. Comparação entre as medidas morfométricas de equinos mangalarga marchador de marcha batida e marcha picada. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 66, n. 2, p. 635–639, 2014.

SAVEGNAGO, R. P. et al. Estimates of genetic parameters, and cluster and principal components analyses of breeding values related to egg production traits in a White Leghorn population. **Poultry Science**, v. 90, n. 10, p. 2174–2188, 2011.

SAVEGNAGO, R. P. et al. Estimates of genetic parameters and eigenvector indices for milk production of Holstein cows. **Journal of Dairy Science**, v. 96, n. 11, p. 7284–7293, 2013.

SCHAEFFER, L. R. C. R. Henderson: Contributions to Predicting Genetic Merit. **Journal of Dairy Science**, v. 74, n. 11, p. 4052–4066, 1991.

SCHAEFFER, L. R. Application of random regression models in animal breeding. **Livestock Production Science**, v. 86, n. 1–3, p. 35–45, 2004.

SCHAEFFER, L. R. **Animal Models - Course Notes (ANSC 6370)**. 2. ed. Guelph: University of Guelph, 2011.

SCHAEFFER, L. R. **Animal Models**. 1. ed. Guelph: L. Schaeffer, 2019.

- SEARLE, S. R. An overview of variance component estimation. **Metrika**, v. 42, n. 1, p. 215–230, 1995.
- SEARLE, S. R.; CASELLA, G.; MCCULLOCH, C. E. **Variance Components**. 1. ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 1992.
- SEARLE, S. R.; KHURI, A. I. Eigenvalues and Eigenvectors. In: BALDING, D. J. et al. (Eds.). **Matrix Algebra useful for statistics**. 1. ed. Hoboken, New Jersey: John Wiley & Sons, 2017. p. 227–267.
- SMITH, C. Improvement of metric traits through specific genetic loci. **Animal Science**, v. 9, n. 3, p. 349–358, 1967.
- SMITH, H. F. A discriminant function for plant selection. **Annals of Eugenics**, v. 7, n. 3, p. 240–250, 1936.
- SNEDECOR, G. W. **Statistical Methods**. 1. ed. Ames: Iowa State University Press, 1946.
- SOLAR DIAZ, I. D. P. et al. Principal components for morphometric traits in Campolina horses. **Journal of Animal Breeding and Genetics**, v. 00, p. 1–9, 2 nov. 2020.
- SPEARMAN, C. “General Intelligence,” Objectively Determined and Measured. **The American Journal of Psychology**, v. 15, n. 2, p. 201–293, 1904.
- STAR, Z. **The Applications of Matrices | What I wish my teachers told me way earlier**, 11 out. 2019. Disponível em: <<https://www.youtube.com/watch?v=rowWM-MijXU&list=WL&index=25&t=206s>>. Acesso em: 27 jul. 2020. Zach Star: The Applications of Matrices | What I wish my teachers told me way earlier.
- STEIGER, J. H. Structural Model Evaluation and Modification: An Interval Estimation Approach. **Multivariate Behavioral Research**, v. 25, n. 2, p. 173–180, abr. 1990.
- STRANDÉN, I.; GARRICK, D. J. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. **Journal of Dairy Science**, v. 92, n. 6, p. 2971–2975, 2009.
- TAYLOR, J. F. Implementation and accuracy of genomic selection. **Aquaculture**, v. 420–421, p. S8–S14, 2014.
- TEEGEN, R.; EDEL, C.; THALLER, G. Population structure of the Trakehner Horse breed. **animal**, v. 3, n. 1, p. 6–15, 2009.
- THOMPSON, B. Calculation of Standardized, Noncentered Factor Scores: An Alternative to Conventional Factor Scores. **Perceptual and Motor Skills**, v. 77, n. 3_suppl, p. 1128–1130, 1993.
- THOMPSON, B. **Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications**. 1. ed. Washington, DC: American Psychological Association, 2004.
- THOMPSON, R. Iterative Estimation of Variance Components for Non-Orthogonal Data. **Biometrics**, v. 25, n. 4, p. 767–773, 1969.

THOMPSON, R. Desert island papers—A life in variance parameter and quantitative genetic parameter estimation reviewed using 16 papers. **Journal of Animal Breeding and Genetics**, v. 136, n. 4, p. 230–242, 2019.

THURSTONE, L. L. Multiple factor analysis. **Psychological Review**, v. 38, n. 5, p. 406–427, 1931.

THURSTONE, L. L. **Multiple Factor Analysis: A development and expansion of the Vectors of Mind**. 1. ed. Chicago: The University of Chicago Press, 1947.

TSCHERMAK-SEYSENEGG, E. VON. Über künstliche Kreuzung bei *Pisum sativum*. **Zeitschrift für das landwirthschaftliche Versuchswesen in Österreich**, v. 3, n. 6, p. 465–555, 1900.

TUCKER, L. R.; LEWIS, C. A reliability coefficient for maximum likelihood factor analysis. **Psychometrika**, v. 38, n. 1, p. 1–10, 1973.

TÚLIO, Marcos. **Rei das Marias**. 2019. 1 fotografia. Disponível em: <https://www.facebook.com/harassantaanna/photos/a.105554580931976/105554977598603/?type=3&theater>. Acesso em: 24 ago. 2020.

VAN TASSELL, C. P.; VAN VLECK, L. D. Multiple-trait Gibbs sampler for animal models: flexible programs for Bayesian and likelihood-based (co)variance component inference. **Journal of Animal Science**, v. 74, n. 11, p. 2586, 1996.

VALENTE, B. D. et al. Searching for Recursive Causal Structures in Multivariate Quantitative Genetics Mixed Models. **Genetics**, v. 185, n. 2, p. 633 LP – 644, 2010.

VALENTE, B. D. et al. Is Structural Equation Modeling Advantageous for the Genetic Improvement of Multiple Traits? **Genetics**, v. 194, n. 3, p. 561–572, 2013.

VALENTE, B. D.; ROSA, G. J. DE M. Mixed Effects Structural Equation Models and Phenotypic Causal Networks. In: GONDRO, C.; VAN DER WERF, J.; HAYES, B. (Eds.) **Genome-Wide Association Studies and Genomic Prediction**. Totowa, NJ: Humana Press, 2013. p. 449–464.

VAN VLECK, L. D. Index Selection for Direct and Maternal Genetic Components of Economic Traits. **Biometrics**, v. 26, n. 3, p. 477, 1970.

VAN VLECK, L. D. **Selection Index and Introduction to Mixed Models Methods**. 1. ed. Boca Raton: CRC Press, 1993.

VAN VLECK, L. D.; BOLDMAN, K. G. Sequential transformation for multiple traits for estimation of (co)variance components with a derivative-free algorithm for restricted maximum likelihood. **Journal of Animal Science**, v. 71, n. 4, p. 836–844, 1993.

VANRADEN, P. M. Efficient Methods to Compute Genomic Predictions. **Journal of Dairy Science**, v. 91, n. 11, p. 4414–4423, 2008.

VITEZICA, Z. G. et al. Bias in genomic predictions for populations under selection. **Genetics Research**, v. 93, n. 5, p. 357–366, 2011.

WEINBERG, W. Über Vererbungsgesetze beim Menschen. **Zeitschrift für induktive Abstammungs- und Vererbungslehre**, v. 1, n. 1, p. 440–460, 1908.

WILLHAM, A. R. L. The Covariance between Relatives for Characters Composed of Components Contributed by Related Individuals. **Biometrics**, v. 19, n. 1, p. 18–27, 1963.

WRIGHT, S. On the Nature of Size Factors. **Genetics**, v. 3, n. 4, p. 367–374, 1918.

WRIGHT, S. Correlation and Causation. **Journal of Agriculture Research**, v. 201, p. 557–585, 1921.

WRIGHT, S. Coefficients of Inbreeding and Relationship. **The American Naturalist**, v. 56, n. 645, p. 330–338, 1922.

WRIGHT, S. Evolution in Mendelian Populations. **Genetics**, v. 16, n. 2, p. 97 LP – 159, 1931.

WRIGHT, S. The Genetical Structure of Populations. **Annals of Eugenics**, v. 15, n. 1, p. 323–354, 1949.

YULE, G. U. Mendel's laws and their probable relations to intra-racial heredity. **New Phytologist**, v. 1, n. 9, p. 193–207, nov. 1902.

YULE, G. U. **On the theory of inheritance of quantitative compound characters on the basis of Mendel's law - a preliminary note.** (W. Wilks, Ed.) Report on the Third International Conference on Genetics. London: Royal Horticultural Society, 1906.

ZHANG, C. et al. Detecting the Population Structure and Scanning for Signatures of Selection in Horses (*Equus caballus*) From Whole-Genome Sequencing Data. **Evolutionary Bioinformatics**, v. 14, p. 1–9, 2018.

ZHANG, Z. et al. Technical note: Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood. **Journal of Animal Science**, v. 85, n. 4, p. 881–885, 2007.

ZWICK, W. R.; VELICER, W. F. Comparison of five rules for determining the number of components to retain. **Psychological Bulletin**, v. 99, n. 3, p. 432–442, 1986.

3. FOUNDER ANALYSIS AND FAMILY IDENTIFICATION IN THE CURRENT CAMPOLINA HORSE POPULATION

*“From a few basic rules you can generate a cosmos. Some call these basic rules axioms. Some call them algorithms. But don't let the fancy names fool you. They're just simple rules”
(BLOOM, 2012, p. 24)¹¹.*

F. de Oliveira Bussiman¹, R.S.B. Carvalho², R.V. Ventura¹, E.C. Mattos³, J.B.S. Ferraz³, J.P. Eler³, F.F. Silva (*in memoriam*)⁴, J.C.C. Balieiro¹

¹*Bioinformatic and Animal Breeding Lab., Department of Animal Nutrition and Production (VNP/FMVZ-USP), College of Veterinary Medicine and Animal Science, University of São Paulo, Av. Duque de Caxias Norte, 225, 13635-900, Pirassununga-SP, Brazil;*

²*Department of Basic Sciences (ZAB/FZEA-USP), College of Animal Science and Food Engineering, University of São Paulo, Av. Duque de Caxias Norte, 225, 13635-900, Pirassununga-SP, Brazil;*

³*Group of Animal Breeding and Biotechnology, Department of Veterinary Medicine (ZMV/FZEA-USP), College of Animal Science and Food Engineering, University of São Paulo, Av. Duque de Caxias Norte, 225, 13635-900, Pirassununga-SP, Brazil;*

⁴*Department of Animal Science, Federal University of Viçosa (DZO-UFV), Av. P H Rolfs, s/n, 36.570-900, Viçosa-MG, Brazil.*

RESUMO: Estudar a contribuição dos fundadores em uma população não é um conceito novo, mas ainda é importante, pois, qualquer estudo sobre perda na diversidade genética, ou aumento na endogamia, precisa começar com uma população de base, no caso de um pedigree completo, a população fundadora. Neste estudo, depois de analisar um pedigree contendo 107.951 cavalos, juntamente com um banco de dados incluindo 19 características mensuradas em 18.695 cavalos (da população de referência – definida a frente), nós demonstramos a formação de diferentes linhagens e avaliamos seu impacto na predição de valores genéticos para cada animal em estudo. Os cavalos nascidos após 1990 (N = 65.634) foram tratados como população de referência e o pedigree para esses animais foi extraído considerando o ano referido como critério de corte para criar a população se base (no total, 81.651 cavalos foram incluídos no pedigree final). Os fundadores mais influentes foram considerados como variáveis, e seu coeficiente parentesco com cada cavalo da população de referência foi tratado como observações em uma análise de componentes principais. Três linhagens foram encontradas na população recente, formadas por 6 garanhões e 8 éguas; esses fundadores explicaram 42,53% do *pool* gênico da população presente. Quase todos os fundadores tiveram valores genéticos negativos para escores de andamento, o que levou a médias negativas para a população recente, nesta característica. Adicionalmente, após conduzir análises dentro dos grupos/linhagens, as linhagens afetaram as correlações genéticas entre os fenótipos ajustados para todas as características em análise. Nossos resultados sugerem que a constituição genética dos fundadores permaneceu importante na população, mesmo após muitas gerações.

Palavras-chave: ACP; Linhagens Fundadoras; Contribuição Genética; *Pool* Gênico; Cavalo.

¹¹ BLOOM, H. **The God problem:** how a godless cosmos creates. 1. ed. Amherst: Prometheus Books, 2012.

ABSTRACT: Studying the founders' contribution in a population is not a new concept, but still being important because any study involving genetic diversity loss, or inbreeding increase (heterozygosity loss), must begin with a baseline population, in case of a full pedigree, the founder population. In this study, after studying a pedigree comprised of 107,951 horses, along with a dataset including 19 traits measured in 18,695 horses (from the reference population – defined latter), we demonstrated the formation of different lineages and evaluated its impact on the breeding values for each considered animal. The horses born after 1990 ($N = 65,634$) were treated as the reference population and the pedigree for those animals was extracted considering the referred year as the cutoff to compute the base population (in total, 81,651 horses were included in the final pedigree set). The most influential founders were considered as variables, and their kinship coefficient among all other horses from the reference population was treated as observations in a principal component analysis. Three lineages were found in the current population, formed by 6 stallions and 8 mares; those founders explained about 42.53% of the gene pool in the present population. Almost all founders had negative breeding values for gait scores, which lead to negative genetic averages in the current population. Additionally, after performing analyses within lineages/groups, the lineages affected genetic correlations among adjusted phenotypes for all studied traits. Our results suggest that the founders' genetic constitution remained important in the population, even after many generations.

Keywords: PCA, Founder lineages, Genetic contribution, Gene pool, Horse.

3.1. Introduction

The Brazilian Campolina gaited horse has its origins in a crossbred stud of mares (Andalusian, Lusitano, and Berber) by using European stallions (Anglo-Norman, Clydesdale, Holsteiner, Percheron, and Thoroughbred) at the Minas Gerais State, around the year 1870 (PROCÓPIO; BERGMANN; COSTA, 2003). The breeders' association was established in 1951 (Campolina Horse Breeders Association – ABCCCampolina, <http://www.campolina.org.br/portal/>), and nowadays this breed is the sixth-largest horse stud in Brazil (third of the Brazilian gaited horses). In the Minas Gerais State, this breed was historically related to both economic and social development, and recently, this breed was exported to other countries achieving more global presence.

According to Lacy (1989), captive populations lost genetic diversity because of drift. By that, alleles present in each generation, become by chance, more (or less) abundant (even “extinct”) in subsequent generations. The same author argued that the only possible biological forces that can restore genetic diversity (GD) are mutation and immigration. Analysis of GD must begin with the founder population because all determination of losses is relative to founder GD. Such studies are dependent on parameters derived from probabilities of gene origin, in which the effective number of founders is used to determine if founder contributions are balanced (BOICHARD; MAIGNEL; VERRIER, 1997; LACY, 1989). According to TANG et

al. (2013), when the founder contributions are balanced, the total number of founders is equal to the effective number of founders, and the presence of unequal founder contributions led to an inbreeding increase.

The occurrence of family/lineages in the pedigree is related to the unbalanced use of parents. That increases coancestry among relatives but decreases in the total population. As a result, GD can be lost within the family (because of susceptibility to drift), which will promote the loss of GD on the entire population, particularly if the variation on family size is inflated. Given this relationship between family formation and loss of GD, investigating the population stratification allows us to better design breeding programs to minimize coancestry/inbreeding. Therefore, the aims of this study were to assess the founder's genetic contributions and to verify the subsequent formation of founder lineages/families in the current population of Campolina horses.

3.2. Material and Methods

3.2.1. Pedigree file

Phenotypic and genealogical data were provided by the ABCCCampolina. The pedigree file was comprised of 107,951 horses born between 1920 and 2014 (across 13 generations). In total, this file contained about 4,253 stallions and 26,755 mares. From the total number of stallions, 105 were also founders (parents of 2,346 progeny), and the number of mare founders was 9,636 (originating 28,934 offspring). The average number of complete generations was 2.96 (with a maximum of 7.50) and the average inbreeding level was 2.44% (considering the entire pedigree). A deeper description of the used pedigree can be found in de Oliveira Bussiman et al. (2018a). Concerning the reference population (defined later), the average number of complete generations was 3.46 (maximum of 7.50), and the average of inbreeding was 3.54%.

3.2.2. Founders and contributions

Animals born after 1990 ($N = 65,634$) were treated as the reference population (RP). Those animals were chosen based on the retrospective inspection of the pedigree structure (27 years in the past defined as threshold), a constant that was defined after multiplying the Campolina generation interval (9.00 ± 0.09 years), previously reported in de Oliveira Bussiman

et al. (2018a), by 3. The next step consisted of a pedigree reconstruction, formed by a subset obtained from the original genealogical structure, along with its ancestors and founders. The final pedigree (used subsequently for all remaining analyses) contained 81,651 horses (65,634 from RP, plus its ancestors/founders). We computed genetic contribution from each ancestor/founder related to the RP as proposed by Lacy (1989). To accomplish that, the function “*probability of gene origins*” from *CFC* software (SARGOLZAEI; IWASAKI; COLLEAU, 2006) was used. This gives us the founders’ genetic contribution, along with the contribution of founder genes to the inbreeding coefficient. After the computation of founders’ contribution, values were transformed in percentage by dividing each contribution by the sum of all contributions, and founders showing a contribution value above 1% were selected to investigate the lineage formation across the RP.

The genetic diversity (GD) on RP was calculated as: $GD = 1 - (1/2f_g)$ (LACY, 1989), where f_g is the founder genome equivalent. The loss in GD of RP was calculated as: $1 - GD$ (SANTANA; BIGNARDI, 2015), and the loss due to the unequal founder contributions was calculated as: $1 - GD^*$ (CABALLERO; TORO, 2000; SANTANA; BIGNARDI, 2015), where $GD^* = 1 - (1/2f_e)$ (CABALLERO; TORO, 2000). The effective number of founders was calculated by $f_e = 1 / \sum_{k=1}^f q_k^2$ (LACY, 1989), where q_k is the proportion of the genomic content carried by the descendant population, contributed by the founder k . The proportion of total GD lost due to unequal founder contributions was calculated by $(1 - GD^*) / (1 - GD)$.

3.2.3. Lineage discovery

After the identification of founders that were responsible for at least 1% of genetic variation in the RP, a principal component analysis (PCA) was conducted. Each founder was treated as variable, and the observations were the numerator relationship coefficients (WRIGHT, 1922) for the respective founders with every other horse in the RP. Because PCA is a technique to identify patterns among variables, the results can be interpreted as correlations among variables due to similarity between observations.

In this study, the pattern was defined as the relationship between reference individuals and founders of the RP. Two founders may remain in the biplot, due to the similarity in terms of shared genetic material among the descendants of those two founders. When the variables

(founders in the case of the present study) are far from the origin in the biplot, more importance those variables have in the total variance of the dataset (DESTEFANIS et al., 2000).

After extracting the principal components (PC), by plotting the first two PCs, correlated founders were assumed to form a lineage across the RP, and to assign each reference animal to a specific lineage, the maximum value of kinship between this animal and each founder was used. The function “*principal*” from the *R* (R CORE TEAM, 2016) package “*psych*” (REVELLE, 2017) was used to produce the PCA results, considering a varimax rotation. The reference animals that had dubious lineage assignments were excluded from the dataset, remaining after that, 18,695 horses in the final dataset.

3.2.4. Phenotypes and data editing

From the RP (N = 65,634), 18,695 horses, born between 1990 and 2014, had information available for the 19 following traits: Height at Withers (HW); Height at Back (HB); Height at Croup (HC); Ratio between HW and HC (HWHC); Height at Chest (HCh); Leg Length (LL); Head Width (HWi); Chest Width (CW); Buttock Width (BW); Head Length (HL); Neck Length (NL); Back-Loins Length (BLL); Croup Length (CL); Shoulder Length (SL); Body Length (BL); Ratio between BLL and BL (BLLBL); Heart Girth (HG); Cannon Girth (CG) and Gait total Score (GtS), for a deeper description of those traits, please read de Oliveira Bussiman et al. (2018b). Except for HWHC, BLLBL, and GtS (which doesn't have a unit of measurement), all traits were measured in centimeters (cm).

To investigate lineage effects, all phenotypes were corrected for fixed and random effects (except the animal random additive genetic effect), through the following models:

$$y_{ij} = \mu + \beta_1 \text{Age} + \beta_2 \text{Age}^2 + \text{GC}_i + \text{S}_j + \varepsilon_{ij}, \quad (3.1)$$

$$y_{ijk} = \mu + \beta_1 \text{Age} + \beta_2 \text{Age}^2 + \text{GC}_i + \text{S}_j + t_k + \varepsilon_{ijk}, \quad (3.2)$$

Where: y_{ij} represents the phenotypic information for each morphometric trait (HW, HB, HC, HWHC, HCh, LL, HWi, CW, BW, HL, NL, BLL, CL, SL, BL, BLLBL, HG, and CG) and y_{ijk} represents the gait information (GtS); β_1 and β_2 are the regression coefficients of the covariate age at measurement (Age), linear and quadratic, respectively; GC_i is the fixed effect of i^{th} contemporary group, defined by the concatenation of year and season of birth, sex, and year of registry; S_j is the fixed effect of the j^{th} stud of birth; t_k is the random effect of k^{th} technician (gait evaluator, only for GtS); ε_{ij} and ε_{ijk} are the residual terms.

In matrix notation, models (3.1) and (3.2) can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{t} + \boldsymbol{\varepsilon}, \quad (3.3)$$

Where: $\mathbf{y}$ is the vector of phenotypic observations; $\boldsymbol{\beta}$ is the vector of solutions for fixed effects; $\mathbf{t}$ is the vector of solutions for the random technician effect (only for GtS); $\mathbf{X}$ and $\mathbf{Z}_1$ are the incidence matrices of fixed and random effects, respectively; $\boldsymbol{\varepsilon}$ is the vector of residual terms.

It was expected that the vector $\boldsymbol{\varepsilon}$ contained the estimated breeding value (EBV) and the residual ($\mathbf{e}$) once the animal effect was not included in the model. Then adjusted phenotypes could be obtained by:

$$\mathbf{y}^* = \boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}_1\mathbf{t} = \mathbf{Z}_2\mathbf{u} + \mathbf{e}, \quad (3.4)$$

Where: $\mathbf{y}^*$ is the vector of adjusted phenotypes; $\mathbf{u}$ is the vector of random additive animal genetic effect; $\mathbf{Z}_2$ is the incidence matrix of $\mathbf{u}$; $\mathbf{e}$ is the vector of random residual terms.

Contemporary groups and studs with less than 5 records and without at least 2 stallions, also considering a minimum value of 2 progeny, were excluded from the dataset. Thus, the final dataset consisted of 18,695 records for all morphometric traits (across 786 GCs and 1,813 studs), whereas there were 5,462 records for GtS, across 244 GCs and 301 studs. The description of the final dataset (before the adjustment for fixed and random effects) is presented in Table 1. The statistical models were implemented by the functions “*lm*”, “*lme*” (BATES et al., 2015), and “*residuals*”, available on *R* software (R CORE TEAM, 2016).

Table 1 - Descriptive statistics (before adjustment for fixed/random effects) of the traits in Campolina horse breed used in present study

Trait	Mean	SD	Min	Max
HW	156.60	5.17	145.00	173.00
HB	149.20	5.01	134.00	163.00
HC	156.31	4.97	137.00	172.00
HWHC ¹	100.19	1.01	93.67	110.00
HCh	65.13	3.17	50.00	79.00
LL	91.47	4.71	72.00	110.00
Hwi	20.88	1.09	17.00	25.00
CW	39.54	2.95	30.00	48.00
BW	52.68	2.82	44.00	60.00
HL	61.71	2.75	52.00	69.00
NL	63.64	3.38	52.00	74.00
BLL	58.05	5.26	40.00	80.00
CL	50.86	4.42	37.00	64.00
SL	55.94	3.54	44.00	68.00
BL	159.43	5.86	138.00	176.00
BLLBL ¹	36.43	3.28	25.16	50.32
HG	183.09	7.58	159.00	205.00
CG	19.25	1.08	15.00	23.00
GtS	139.17	11.67	85.00	173.00

Source: Bussiman (2021). ¹This trait was multiplied by 100 to achieve the same level of variation of the other morphological traits. Abbreviations: HW = Height at Withers; HB = Height at Back; HC = Height at Croup; HWHC = Ratio between HW and HC; HCh = Height at Chest; LL = Leg Length; Hwi = Head Width; CW = Chest Width; BW = Buttock Width; HL = Head Length; NL = Neck Length; BLL = Back-Loins Length; CL = Croup Length; SL = Shoulder Length; BL = Body Length; BLLBL = Ratio between BLL and BL; HG = Heart Girth; CG = Cannon Girth; and GtS = Gait total Score. Except for HWHC, BLLBL, and GtS, all traits were measured in cm.

3.2.5. Lineage effect

To investigate if the detected lineages affected the EBV for the studied traits, two approaches were proposed. Firstly, GtS was studied alone because it is the principal goal part of the indirect selection currently implemented by the Campolina breeders. A simple linear regression of the EBVs computed for GtS (EBV_{GtS}) over the lineages (coded as numeric) was applied and the regression coefficient was tested through analysis of variance. The line effect is a cross-classified effect, thus the coding to estimate the regression coefficient was made by an analysis of variance ($P < 0.001$, for the lineage effect), that is, EBV_{GtS} were regressed on least squares means for lineages. To verify the model adjustment, we considered the same components, and lineage was treated as a random effect. The described models are presented as follows:

$$EBV_{GtS_i} = \beta_0 + \beta_1 \text{Lineage}_i + e_i, \quad (3.5)$$

$$EBV_{GtS_i} = \mu + \text{Lineage}_i + e_i, \quad (3.6)$$

Where: EBV_{GtS_i} is the EBV for GtS from i^{th} animal, without considering lineage effects in the model (coming from DE OLIVEIRA BUSSIMAN et al., 2018b); β_0 is the intercept; β_1 is the regression coefficient of EBV_{GtS} over lineage lsmean for the i^{th} animal; μ is the overall mean fixed effect; $Lineage_i$ in the model (3.6), is the random effect of lineage and e_i is the residual associated with each animal EBV.

Another approach to investigate the lineage effect was performing PCA, it was conducted considering the set of adjusted phenotypes (as reported in the 4th model, evaluating the 19 traits across each lineage). To accomplish that, the original dataset was splitted into one subgroup for each lineage and the approach proposed by Josse and Husson (2016) was used to handle the missing values in GtS. According to this procedure, missing observations were imputed based on the data's multivariate normal distribution, considering the obtained similarity among variables and observations. After that, the dataset was centralized based on a mean value equal to 0 and scaled to a variance of 1. According to this principle, the magnitude of each variable should not affect the results. The *R* (R CORE TEAM, 2016) function “*principal*” from the package “*psych*” (REVELLE, 2017) was used at this point, adopting a varimax rotation. For all PCA analyses, the Kaiser-Meyer-Olkins (KMO) criterion (KAISER, 1960) and Bartlett's test (BARTLETT, 1950) were checked, to verify if the PCA was applicable to the data.

3.3. Results

3.3.1. Founders and contributions

In total, we identified 16,318 animals as founders of the RP. From this group, only 47.19% showed a genetic contribution higher than zero. Summing the values from founders that presented a contribution greater than 1%, we observed that 14 founders were responsible for 42.53% of the genetic diversity in the RP (Table 2). The founder with the smallest number of direct connections was related with 20.93% of the RP, and a value almost 4-fold higher was noted for the animal with the highest connectivity (82.15%). Those average relationship coefficients ranged from 0.03 to 0.12, specifically for the founders 04/000138 and 03/000308, respectively.

Table 2 - Founders with genetic contribution to the recent Campolina population higher than 1%, their sex, year of birth, genetic contribution, progeny number, number of recent animals related (kinship > 0) to them, and average relationship between them and the recent population

Founder	Sex	YOB ¹	Genetic contribution (%)	Progeny	N ²	$\bar{f}^3$
03/000308	M	1963	10.11	63	53,918 (82.15%)	0.12 ± 0.06
04/000926	F	1963	6.60	8	42,535 (64.81%)	0.10 ± 0.05
04/000163	F	1948	4.11	3	51,771 (78.88%)	0.05 ± 0.03
03/000130	M	1956	3.42	107	46,239 (70.45%)	0.05 ± 0.03
03/000069	M	1951	3.15	4	51,572 (78.58%)	0.04 ± 0.03
03/000166	M	1960	2.37	163	30,992 (47.22%)	0.05 ± 0.04
04/000099	F	1945	1.94	7	28,378 (43.24%)	0.04 ± 0.03
04/000549	F	1960	1.77	10	28,031 (42.71%)	0.04 ± 0.03
03/000081	M	1954	1.68	2	35,792 (54.53%)	0.03 ± 0.02
04/000172	F	1947	1.68	3	35,347 (53.85%)	0.03 ± 0.02
04/000891	F	1960	1.62	13	24,727 (37.67%)	0.04 ± 0.02
04/000138	F	1948	1.46	5	32,796 (49.97%)	0.03 ± 0.02
04/001127	F	1971	1.32	7	13,740 (20.93%)	0.06 ± 0.03
03/000315	M	1964	1.30	89	16,503 (25.14%)	0.05 ± 0.05

Source: Bussiman (2021). ¹Year of birth; ²Number and percentage of animals with kinship between founder and recent animals higher than 0; ³Average kinship between founder and recent population.

The average number of progeny generated from each founders' offspring ranged from 1.62 to 89.33. In addition, 8 mares and 6 stallions formed a group of 14 founders that contributed more than 1% to the RP gene pool. The contribution of founder genes to the inbreeding coefficient ranged from 0.72% to 28.43%, and together, the founders that were responsible for more than 1% of the RP gene pool, contributed with 79.87% of genes that impacted the inbreeding coefficient. The f_g found in the RP was 23.90; thus, the total loss in GD was 2.09%. The f_e was 46.33, then GD lost due to unequal founder contributions was 1.08%, which represents 51.67% of the total lost in GD.

3.3.2. Lineage discovery and its effect

The Bartlett's test was successfully applied on the dataset ($P < 0.001$) and the KMO value was 0.65 (varying from 0.29 to 0.84, for founder 04/000099 and 04/000891, respectively). Based on the scree-plot (Figures 1A and 1B), five PCs should be enough to explain the original data, and these first five PCs explained 73.83% of the total variance. Using the first two PCs (explaining 46.50% of the total variance), three founder lineages were found and the result is shown graphically (Figure 1C). We did not identify a single group that was formed only by

mares (or only by stallions), and the lineage with the highest number of founders was composed of two mares and four stallions (the other lineages had two mares and two stallions).

Figure 1 - Principal Component Analysis of kinship patterns among the most influential founders applied to the current Campolina horse population. A: Scree plot of Principal Components; B: Percentage of variance explained by each principal component and cumulated explained variance; and C: Lineage identification

Source: Bussiman (2021).

The regression coefficient of EBV_{GtS} over lineage was -0.15 ($P < 0.001$), and the solutions for each lineage were -0.72 , -0.74 , and -1.00 (from model 3.6), for groups A, B, and C, respectively. In addition, lineages A and B, had the same variance of EBV_{GtS} (2.30), whereas line C had the highest variance (4.02). The pattern of correlations among the adjusted phenotypes was changed according to lineages (Figure 2). Some studied traits were not affected by the lineage effect, i.e., HW; HB; HC; LL; and HCh, whereas BLLBL, HWHB and GtS varied

on biplots according to lineages (Figure 2). The main changes in terms of correlation patterns across lineages were related to the importance of each variable for the total variance, as well as the correlations among variables. It seems that lineage B had more correlated adjusted phenotypes, whereas lineage C was more similar to lineage A (Figure 2).

Figure 2 - Correlation plot (Biplot) for studied traits according to each found lineage. Height at Withers (HW); Height at Back (HB); Height at Croup (HC); Ratio between HW and HC (HWHC); Height at Chest (HCh); Leg Length (LL); Head Width (HWi); Chest Width (CW); Buttock Width (BW); Head Length (HL); Neck Length (NL); Back-Loins Length (BLL); Croup Length (CL); Shoulder Length (SL); Body Length (BL); Ratio between BLL and BL (BLLBL); Heart Girth (HG); Cannon Girth (CG) and Gait total Score (GtS)

Source: Bussiman (2021).

The only lineage that had a positive association between GtS and other traits was lineage B, but correlations were close to zero. It was also in lineage B that GtS had the highest importance to the total variance. Additionally, the average relationship coefficient among animals within each lineage ranged from 0.03 to 0.08, for lineages B and C, respectively (Table 3). The inbreeding coefficient of animals within each lineage varied from 2.40% to 4.80%, for lines B and C, respectively (Table 3). As the maximum relationship between a given reference animal and each founder was used to assign each reference animal to each line, animals from distinct lines were related (Figure 3). The animals which were not founders or RP (coded as 0 on Figure 3A) were less related to each other than animals in the RP.

The number of animals was too large ($N = 81,651$, considering all RP, and $N = 18,695$, considering only phenotyped) for complete visualization of relationships. Thus, a heatmap of average relationships among animals within (and between) lineages is being presented (Figure

3A), along with a “three-dimensional” scatter plot of the first three principal components from the additive relationship matrix for phenotyped RP (N = 18,695; Figure 3B). The kinship was, on average, higher for horses within lineages than between horses from different lineages (Figure 3A). The occurrence of matting of horses coming from different lineages causes a distribution of relationships among all phenotyped animals, but those animals from different lineages remained grouped in the population (Figure 3B). Additionally, the first three principal components are being showed on Figure 4.

Table 3 - Average relationship and inbreeding coefficients of the animals within each found line

Line	Number of Animals	$\bar{f}^1$	Inbreds	$\bar{F}^2$	$\bar{F}_I^3$
A	3,485	0.05	2,407 (96.86%)	3.30	4.80
B	2,548	0.03	1,475 (57.18%)	2.40	4.20
C	12,662	0.08	10,525 (83.12%)	4.80	5.73

Source: Bussiman (2021). ¹Average relationship coefficient; ²Average inbreeding coefficient (in percentage); ³Average inbreeding coefficients on inbreds (in percentage).

Figure 3 - Relatedness among horses in the reference population (born after 1990). A: Heatmap of average kinship within (and between) found lines. 0 = non-founders and not reference population animals; RP_A = animals descending from line A; RP_B = animals descending from line B; RP_C = animals descending from line C; F_A = founders of line A; F_B = founders of line B; and F_C = founders of line C; and B: Three-dimensional scatter plot of first three principal components of the relationship matrix of the phenotyped reference population (N = 18,695)

Source: Bussiman (2021).

Figure 4 - First three principal components from the additive relationship matrix among all phenotyped animals in the reference population.

Source: Bussiman (2021).

In addition, EBV_{GTS} (and accuracy) were estimated for the most influential founders (Table 4). These ranged from -2.87 to 2.62 , for founders 04/000926 and 03/000308, respectively. Accuracy varied between 0.21 and 0.62 , for founders 04/000138 and 03/000308, respectively. For the groups of founders forming each lineage, averages of EBV_{GTS} were -0.43 , -0.44 , and -0.33 , for lineages A, B, and C, respectively; and the accuracy average for each line was around 0.35 .

Table 4 - Estimated breeding value for gait total score and accuracy of prediction for the most influential founders of the recent population in the Campolina breed

Founder	EBV _{GtS} ¹	ACC ²
03/000308	2.62	0.62
04/000926	-2.87	0.49
04/000163	-0.62	0.36
03/000130	-1.42	0.37
03/000069	-0.01	0.32
03/000166	-0.88	0.43
04/000099	0.57	0.39
04/000549	-1.29	0.35
03/000081	-0.27	0.24
04/000172	-0.02	0.24
04/000891	-1.41	0.29
04/000138	-0.22	0.21
04/001127	-0.85	0.28
03/000315	-0.71	0.43

Source: Bussiman (2021). ¹Estimated breeding value for gait total score; ²Individual accuracy, calculated as:

$ACC_i = \sqrt{1 - \left(\frac{PEV_i}{\sigma_u^2(1 - F_i)} \right)}$, where PEV_i is the individual prediction error variance; σ_u^2 is the additive genetic variance, and F_i is the individual inbreeding coefficient.

3.4. Discussion

The observed reduction from the total number of founders to the effective number of founders could be explained by unequal founders' contribution (LACY, 1989; MELKA; SCHENKEL, 2010; TANG et al., 2013). If the founders contributed equally to RP, the effective number of founders and the total number of founders would be equal (BOICHARD; Maignel; VERRIER, 1997). According to Tang et al. (2013), the unequal contribution of founders directly causes an increase of inbreeding in the reference population, which could explain inbreeding averages found in this study. Another consequence of unequal genetic contribution of founders is a reduction of genetic variability, which can happen joint with bottleneck and drift events (DURU, 2017).

The loss in GD due to unequal founders' contribution found in this study (51.67%) was lower than found by (SANTANA; BIGNARDI, 2015) in Pêga donkey (70%), but our estimate of the total loss of GD (2.09%) was higher than found by the same authors (1.29%). Given the geographical proximity of the origin of both breed/species (Pêga donkey and Campolina horse), de Oliveira Bussiman et al. (2018a) speculated about the genetic diversity of the Campolina population in comparison with Pêga donkey. Here, we used the geographical origin (Minas Gerais State) to hypothesize that breeders designed mating of the animals in the same way, allowing this comparison.

Vostrá-Vydrová et al. (2016), working with an endangered horse population reported losses in GD above 10%, which shows that RP of Campolina horse is not at risk due to unequal founder contributions and drift. Studying GD for the entire population, de Oliveira Bussiman et al. (2018a) reported losses of 1.08% for the total loss of GD, and 55.56% of the losses were due to unequal founder contributions, these values suggest that in the RP (animals born after 1990) inbreeding is increased, much because of gene drift but also by unbalanced use of descendants from specific founders, causing the increase in total loss of GD.

The reduction in the proportion of GD lost due to unequal founder contributions (from 55.56% to 51.67%) may be related to the mating of animals from different founder lines in the current population. Fernández et al. (2005), stated that genetic drift is the main factor causing loss of GD within breeds. The results found in this study may suggest an increase of drift in RP, which increased the total loss of GD, since the proportion of GD lost due to unequal founder contributions was decreased.

The number of founders with genetic contribution higher than 1% to the RP (14 founders, explaining 42.53% of RP gene pool), for our analysis, was lower than found by Faria et al. (2018a), who found that 500 founders explained 62.92% of the genes of RP in Quarter Horses. In the Brazilian Lusitano horse, the numbers of most influential founders were 10 and 50, explaining 43.5% and 78.2% of the RP gene pool (FARIA et al., 2018b). The reason for that may be the inbreeding, whereas the Quarter Horse presents an inbreeding coefficient of 1.56% (FARIA et al., 2018a), the Lusitano horse presents an inbreeding of 4.46% (FARIA et al., 2018b), a value close to the inbreeding coefficient in Campolina RP (3.54%). Duru (2017), reported that 14 founders were responsible for 50% of GD in Turkish Arab horses, and the average inbreeding coefficient (in this population) was 4.60% these values are close to those found in this study support the found number of founders because if the founders contributed equally to the RP, the inbreeding would be decreased (LACY, 1989).

Costa et al. (2005), stated when a founder has a large number of progeny and has relationship coefficients different from zero with a large part of the RP, but its genetic contribution is low, it must be because his progeny are not extensively used. This fact was also found in this study, for example, founder 03/000166 had the largest number of progeny (163) but his genetic contribution was 2.37% (6th most influential founder), whereas founder 03/000308 had only 63 progeny but its genetic contribution was 10.11%, being the most influential founder of the Campolina horse. This is related to non-equal founders' contribution because the descendants of founder 03/000308 were more intensively used than founder's 03/000166 descendants.

The PCA has been applied to study genetic diversity and population structure (COSTA et al., 2005; HONDA et al., 2002) even with genomic data (GRILZ-SEGER et al., 2019; LI; RALPH, 2019; ZHANG et al., 2018), and it is because this technique allows researchers to reduce high dimensionality problems into low dimension scale. Costa et al. (2005), studied by PCA the formation of lines across the population of Mangalarga Marchador horses, in a way very close to the used in the present study. These authors found the existence of five lines in the studied population, whereas in our study, it was found three founder lines.

In general, the PCA is applied in a set of animals, either in the numerator relationship matrix (HENDERSON, 1975) or in the genomic relationship matrix (VANRADEN, 2008), to reduce dimensionality and to visualize how animals are close to each other. Essentially, PCA is a technique to identify patterns in the data (IEZZONI; PRITTS, 1991; LATTIN; CARROLL; GREEN, 2011), and in the case of this study, identified patterns were about the relationship between RP animals and influential founders, but also how the descendants of each founder were related to each other.

Founders that are correlated at biplot may form a family/lineage on RP. In general, those animals descending from one lineage were more related to those inside the same lineage, which supports the PCA results. The kinship between RP and influential founders was high within the lineages, but groups A and B had some kinship with lineage C, indicating that despite preferred mating within lines, descendants from founders forming line C were widely used on RP. De Oliveira Bussiman et al. (2018a) found results that support this reasoning because those authors reported that the kinship between different states was increased in the RP (constructed in the same way as in this study), probably those animals that are responsible for this increase in kinship would be the descendants of founders forming line C.

The importance of family/lineage formation in horse pedigrees was studied (COSTA et al., 2005; FLETCHER, 1946; RHOAD; KLEBERG, 1946) because horse breeders tend to prefer to use inbreeding as a tool to achieve phenotype uniformity. Thus, knowledge of the founder lineages/families should help breeders make decisions to preserve, as much as possible, the founders' genetic diversity. The founders usually are assumed as having 100% of genetic diversity and every loss in GD, since the population foundation, is expressed in terms of loss from the founders' GD (LACY, 1989).

In the past, the founders used in domestic populations should have different phenotypes, as well as different genetic merit for those traits considered at that moment. If the founder more used has negative genetic merit for an important trait, his descendants will have, on average, negative genetic merit also. Thus, it was found an effect of lineage over adjusted

phenotypes, once each founder group was expected to have different genetic merit. The variance in EBV_{GtS} was different according to the lineages, which could be related to uniformity within the lineages. The correlations between traits were also affected, showing that the decision at the beginning of the population remained important and affected even present animals.

The negative effect of lineages over GtS could be explained by the EBV_{GtS} of the founders forming each lineage. If the average of EBV_{GtS} from founders forming each lineage is considered, it can be said that the lines started at the same value of EBV_{GtS} , because the lines were very close. Those negative values led to a decrease in the averages for GtS within each lineage, and the regression coefficient found (-0.15) should be interpreted as a negative association between the founder lineage and the RP. Additionally, solutions for line effect and averages of EBV_{GtS} for RP animals within the lineages were the same, justifying the negative effect of line over GtS.

The founders' influence over correlations among adjusted phenotypes could be due to differences in the genetic constitution of each founders group. These founders had good values of EBV for morphological traits, which could explain the increase in genetic merit for most morphological traits in the Campolina breed (DE OLIVEIRA BUSSIMAN et al., 2018b). Under that reasoning, traits that were not influenced by founder lines must be due to similarity among most influential founders. Remarkably, the traits that were more influenced by founder lines (HWHB and BLLBL), were those traits more uncorrelated with others, and those that showed no increase in EBV (DE OLIVEIRA BUSSIMAN et al., 2018b). In addition, the most positively correlated traits (HB, HW, and HC) were the most non-changed ones.

3.5. Conclusion

The Campolina horse breeders have used inbreeding mainly as a tool to achieve uniformity, and therefore, different families were formed in the population. Those families were related to the founders used in the base population and our results suggested that there are three founder lineages in the current Campolina population. Those lineages accounted for almost 50% of the present gene pool and affected the estimated breeding values and genetic correlations among all phenotypes in this breed. Our results suggested that the founders' genetic constitution remained important in the population, even after many generations.

3.6. Implications

The study of founder contribution and family formation allows both researchers and breeders to make better decisions in terms of breeding programs and genetic conservation. This study offered an application of principal component analysis to identify lineages/families in a pedigree, based on the most influential founders. The presented results will serve as a reference for future researchers to better develop new populations, considering the genetic merit of each founder, as it may impact the subsequent offsprings and will remain important even after many generations. In addition, this study presented a simple framework to identify founder lineages: i) compute founder contributions; ii) obtain the kinship coefficient among every influential founder and the remaining set of individuals in the reference population; iii) perform a principal component analysis to check for possible founder lineages; iv) investigate, if possible, the lineage effect on the reference individual phenotypes and breeding values.

3.7. References

- BARTLETT, M. S. Tests of significance in factor analysis. **British Journal of Statistical Psychology**, v. 3, n. 2, p. 77–85, 1950.
- BATES, D. et al. Fitting Linear Mixed-Effects Models using lme4. **Journal of Statistical Software**, v. 67, n. 1, 2015.
- BOICHARD, D.; MAIGNEL, L.; VERRIER, É. The value of using probabilities of gene origin to measure genetic variability in a population. **Genetics Selection Evolution**, v. 29, n. 1, p. 5, 1997.
- CABALLERO, A.; TORO, M. A. Interrelations between effective population size and other pedigree tools for the management of conserved populations. **Genetical Research**, v. 75, n. 3, p. 331–343, 2000.
- COSTA, M. D. et al. Estudo da subdivisão genética da raça Mangalarga Marchador. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 57, n. 2, p. 272–280, 2005.
- DE OLIVEIRA BUSSIMAN, F. et al. Pedigree analysis and inbreeding effects over morphological traits in Campolina horse population. **Animal**, v. 12, n. 11, p. 2246–2255, fev. 2018a.
- DE OLIVEIRA BUSSIMAN, F. et al. Genetic analysis of morphological and functional traits in Campolina horses using Bayesian multi-trait model. **Livestock Science**, v. 216, n. April, p. 119–129, 2018b.
- DESTEFANIS, G. et al. The use of principal component analysis (PCA) to characterize beef. **Meat Science**, v. 56, n. 3, p. 255–259, 2000.

- DURU, S. Pedigree analysis of the Turkish Arab horse population: structure, inbreeding and genetic variability. **Animal**, v. 11, n. 9, p. 1449–1456, 2017.
- FARIA, R. A. da S. et al. Assessment of pedigree information in the Quarter Horse: Population, breeding and genetic diversity. **Livestock Science**, v. 214, p. 135–141, ago. 2018a.
- FARIA, R. A. da S. et al. Genetic Diversity of Lusitano Horse in Brazil Using Pedigree Information. **Journal of Equine Veterinary Science**, v. 69, p. 149–158, 2018b.
- FERNÁNDEZ, J. et al. Efficiency of the Use of Pedigree and Molecular Marker Information in Conservation Programs. **Genetics**, v. 170, n. 3, p. 1313–1321, 2005.
- FLETCHER, J. L. A study of the first fifty years of Tennessee Walking Horse breeding. **Journal of Heredity**, v. 37, n. 12, p. 369–373, 1946.
- GRILZ-SEGER, G. et al. High-resolution population structure and runs of homozygosity reveal the genetic architecture of complex traits in the Lipizzan horse. **BMC Genomics**, v. 20, n. 1, p. 174, 2019.
- HENDERSON, C. R. Rapid method for computing the inverse of a relationship matrix. **Journal of Dairy Science**, v. 58, n. 11, p. 1727–1730, 1975.
- HONDA, T. et al. Pedigree analysis of genetic subdivision in a population of Japanese Black cattle. **Animal Science Journal**, v. 73, n. 6, p. 445–452, 2002.
- IEZZONI, A. F.; PRITTS, M. P. Applications of Principal Component Analysis to Horticultural Research. **HortScience**, v. 26, n. 4, p. 334–338, 1991.
- JOSSE, J.; HUSSON, F. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. **Journal of Statistical Software**, v. 70, n. 1, 2016.
- KAISER, H. F. The Application of Electronic Computers to Factor Analysis. **Educational and Psychological Measurement**, v. 20, n. 1, p. 141–151, 1960.
- LACY, R. C. Analysis of founder representation in pedigrees: Founder equivalents and founder genome equivalents. **Zoo Biology**, v. 8, n. 2, p. 111–123, 1989.
- LATTIN, J. M.; CARROLL, J. D.; GREEN, P. E. **Analyzing multivariate data**. 1. ed. São Paulo: Cengage Learning, 2011.
- LI, H.; RALPH, P. Local PCA Shows How the Effect of Population Structure Differs Along the Genome. **Genetics**, v. 211, n. 1, p. 289–304, 2019.
- MELKA, M. G.; SCHENKEL, F. Analysis of genetic diversity in four Canadian swine breeds using pedigree data. **Canadian Journal of Animal Science**, v. 90, n. 3, p. 331–340, 2010.
- PROCÓPIO, A. M. M.; BERGMANN, J. A. G. A. G.; COSTA, M. D. D. Formação e demografia da raça Campolina. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 55, n. 3, p. 361–365, 2003.
- R CORE TEAM. **R: A Language and Environment for Statistical Computing** (R. D. C. Team,

Ed.) R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. 2016.

REVELLE, W. R. **psych**: Procedures for Personality and Psychological Research. Evanston, Illinois, USAR Foundation for Statistical Computing, 2017.

RHOAD, A. O.; KLEBERG, R. J. The development of a superior family in the modern Quarter Horse. **Journal of Heredity**, v. 37, n. 8, p. 227–238, 1946.

SANTANA, M. L.; BIGNARDI, A. B. Status of the genetic diversity and population structure of the Pêga donkey. **Tropical Animal Health and Production**, v. 47, n. 8, p. 1573–1580, 2015.

SARGOLZAEI, M.; IWASAKI, H.; COLLEAU, J. J. CFC: A tool for Monitoring genetic diversity. In: 8th WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 8. **Proceedings...** Belo Horizonte: 2006.

TANG, G. Q. et al. Inbreeding and Genetic Diversity in Three Imported Swine Breeds in China Using Pedigree Data. **Asian-Australasian Journal of Animal Sciences**, v. 26, n. 6, p. 755–765, 2013.

VANRADEN, P. M. Efficient Methods to Compute Genomic Predictions. **Journal of Dairy Science**, v. 91, n. 11, p. 4414–4423, 2008.

VOSTRÁ-VYDROVÁ, H. et al. Pedigree analysis of the endangered Old Kladruber horse population. **Livestock Science**, v. 185, p. 17–23, 2016.

WRIGHT, S. Coefficients of Inbreeding and Relationship. **The American Naturalist**, v. 56, n. 645, p. 330–338, 1922.

ZHANG, C. et al. Detecting the Population Structure and Scanning for Signatures of Selection in Horses (*Equus caballus*) From Whole-Genome Sequencing Data. **Evolutionary Bioinformatics**, v. 14, p. 1–9, 2018.

4. REDUCED RANK ANALYSIS OF MORPHOMETRIC AND FUNCTIONAL TRAITS IN CAMPOLINA HORSES: PRINCIPAL COMPONENTS AND FACTOR ANALYTICAL MODELS

"I am a firm believer, that without speculation there is no good and original observation"
(Charles Darwin)¹².

F. de Oliveira Bussiman¹, R.S.B. Carvalho², F.F. Silva (*in memoriam*)³, R.V. Ventura¹, J.B.S. Ferraz⁴, E.C. Mattos⁴, J.P. Eler⁴, J.C.C. Balieiro¹

¹*Bioinformatic and Animal Breeding Lab., Department of Animal Nutrition and Production, College of Veterinary Medicine and Animal Science University of São Paulo (BIOMA-VNP/FMVZ-USP), Av. Duque de Caxias Norte, 225, 13.635-900, Pirassununga-SP, Brazil;*

²*Department of Basic Sciences, College of Animal Science and Food Engineering, University of São Paulo (ZAB/FZEA-USP), Av. Duque de Caxias Norte, 225, 13.635-900, Pirassununga-SP, Brazil;*

³*Department of Animal Science, Federal University of Viçosa (DZO/UFV), Av. P H Rolfs, s/n, 36.750-900, Viçosa-MG, Brazil;*

⁴*Group of Animal Breeding and Biotechnology, Department of Veterinary Medicine, College of Animal Science and Food Engineering, University of São Paulo (GMAB-ZMV/FZEA-USP), Av. Duque de Caxias Norte, 225, 13.635-900, Pirassununga-SP, Brazil.*

RESUMO: Modelos multicaracterística podem aumentar a acurácia de predição dos valores genéticos e reduzir o viés devido à seleção ao usar características medidas antes e após ela ter ocorrido. Entretanto, a medida que o número de características cresce, uma tendência similar é esperada para o número de parâmetros a serem estimados, o que afeta diretamente o poder computacional e a quantidade de dados requeridos. O objetivo do presente estudo foi aplicar modelos de *rank* reduzido (modelos de componentes principais – PCM) e modelos fatores analíticos (FAM) para estimar os componentes de (co)variância para dezenove características, avaliadas de maneira conjunta em uma única análise em cavalos Campolina. Um total de dezoito características morfométricas (MT) e um escore visual de andamento (GtS), junto com dados genealógicos de 48.806 cavalos foi analisado sob máxima verossimilhança restrita. Foram ajustados aos dados nove PCM, nove FAM e um modelo multicaracterística padrão (MTM), estes foram comparados para encontrar o melhor modelo. Baseado no critério Bayesiano de Schwarz, o melhor modelo foi a opção FAM, considerando cinco fatores comuns (FAM5). Após conduzir uma análise intraclasse, nenhuma das MT foi genética-negativamente correlacionada, enquanto GtS foi negativa-geneticamente correlacionado com todas as MT, exceto para a correlação genética entre GtS e BLL, e entre GtS e BLLBL (0.01 e 0.10, respectivamente). De todas as MT, duas foram derivadas calculando-se razões envolvendo outras características, essas tiveram correlações genéticas negativas com as outras MT, mas todas favoráveis à seleção. Padrões similares foram observados entre os parâmetros genéticos obtidos pelo MTM ou pelo FAM5, respectivamente. As estimativas de herdabilidade variaram de 0,09 (largura de cabeça) a 0,47 (altura de cernelha). Nossos resultados indicaram que os FAM foram efetivos em reduzir a dimensionalidade de modelos multicaracterística, e por isso,

¹² DARWIN, C. [Carta enviada a Alfred R. Wallace]. Destinatário: Alfred Russel Wallace. Downe, 22 dez. 1857. 1 carta. Disponível em: <<https://www.darwinproject.ac.uk/letter/DCP-LETT-2192.xml>>. Acesso em: 02 set. 2021.

as características podem ser combinadas baseando-se nos primeiros três autovetores da matriz de (co)variância genética aditiva. Além disso, para fins de seleção, existe variabilidade suficiente na população estudada, beneficiando a potencial instalação de um programa de melhoramento.

Palavras-chave: Genética equina; parâmetros genéticos; componentes principais; modelos fatores analíticos; correlação de ranque.

ABSTRACT: Multitrait models can increase the accuracy of breeding value prediction and reduce bias due to selection by using traits measured before and after it has occurred. However, as the number of traits grows, a similar trend is expected for the number of parameters to be estimated, which directly affects the computing power and the amount of data required. The aim of the present study was to apply reduced rank (principal components model - PCM) and factor analytical models (FAM), to estimate (co)variance components for nineteen traits, jointly evaluated in a single analysis in Campolina horses. A total of eighteen morphometric traits (MT) and one gait visual score (GtS), along with genealogical records of 48,806 horses, were analyzed under a restricted maximum likelihood framework. Nine PCM, nine FAM, and one standard multitrait model (MTM) were fitted to the data and compared to find the best suitable model. Based on Bayesian information criterion, the best model was the FAM option, considering five common factors (FAM5). After performing an intraclass analysis, none of MT were genetically negatively correlated, whereas GtS was negatively related to all MT, except for the genetic correlations among GtS and BLL, and between GtS and BLLBL (0.01 and 0.10, respectively). From all MT, two traits were derived computing ratios involving other traits, those had negative correlations with others MT, but all favorable for selection. Similar patterns were observed between the genetic parameters obtained from MTM and FAM5, respectively. The heritability estimates ranged from 0.09 (head width) to 0.47 (height at withers). Our results indicated that FAM was efficient to reduce the multitrait analysis dimensionality, and therefore, traits can be combined based on the first three eigenvectors from the additive genetic (co)variance matrix. In addition, there was sufficient genetic variation for selection, benefiting its potential implementation in a breeding program.

Keywords: Horse genetics, genetic parameters, principal components, factor analytical models, rank correlation.

4.1. Introduction

The Campolina horse is one of the Brazilian gaited horse breeds. Horses perform (naturally) four types of gaits: walk, intermediate gait, canter, and gallop. When the horse exhibit trot as intermediate gait, the horse is called “non-gaited”, if the intermediate gait is a four-beat gait (called “marcha” in Brazil), the horse is named “gaited”. The Campolina origins are traced back to the horses brought from the Iberian Peninsula (Andalusian, Lusitano, Berber and Sorraia). According to Procópio, Bergmann, and Costa (2003), these horses formed a crossbred stud of mares, and by using European stallions (such as: Anglo-Norman, Clydesdale, Holsteiner, Percheron, and Thoroughbred), in 1870 at Minas Gerais State (MG), the Campolina

breed emerged. In 1951 the breeders' association (Campolina Horse Breeders Association – ABCCCampolina, <http://www.campolina.org.br/portal/>) was founded.

This breed is historically related to the social-economic development of MG (REZENDE; MOURA, 2004), and currently, this breed is the sixth larger purebred Brazilian horse breed (VIEIRA et al., 2015). The president of the ABCCCampolina stated that animals from this breed were exported to Mexico, Venezuela, USA, and Germany, thus this breed is receiving attention from the international horse market. The Campolina population has been selected through phenotypic breeders' preferences and by gait competitions results, and inbreeding is widely used by the breeders.

In horses, it is common to use multitrait mixed models (MTM) for genetic evaluations, mainly because most morphometric traits are genetically positively correlated (MOLINA et al., 1999; RUSTIN et al., 2009; VICENTE et al., 2014a, 2014b; VOSTRY et al., 2017). An advantage of using a single analysis including several traits is the selection bias reduction (BOLIGON et al., 2013, 2015; POLLAK; VAN DER WERF; QUAAS, 1984; SCHAEFFER, 1984), by using traits measured before and after selection (BOLIGON et al., 2009; MEYER, 1995).

Concerning the selection bias, suppose there are three traits genetically correlated to some degree (T1, T2, and T3). Let's say T2 is under selection to higher values, then some animals are often discarded when they do not pass the selection criterion. This causes a bias on prediction for T2. But suppose T1 was measured before discarding animals for T2. Thus, some animals that does not have record for T2, may have for T1. If we do a bitrait model, the selection bias on T2 is then reduced, by the indirect prediction for T2 through the genetic correlation with T1. Now, suppose T3 is measured after we apply selection for T2. Then, the animals that have phenotypes for T1 may also have for T3, but those discarded for T2 will not have observations for T3.

Through genetic correlations among those three traits, the predictions for T2, therefore, can be assumed corrected for selection bias, because we will have more accurate predictions for those animals discarded for T2. In fact, (HENDERSON, 1975a) proposed a way to deal with the selection bias. Posteriorly, other investigations proposed the concept of ignorability (IM; FERNANDO; GIANOLA, 1989; RUBIN, 1976), in which the selection can be ignored when the genetic relationship matrix is included in the MME and there are correlated traits measured before and after selection has been applied in a trait.

According to (THOMPSON; MEYER, 1986; VAN DER WERF; VAN ARENDONK; DE VRIES, 1992) another benefit of adopting MTM is an increase on the prediction accuracy,

by taking into account genetic correlations among all considered traits, combining direct and indirect information for breeding goal purposes. The advantages of MTM are dependent on the stability of genetic correlation coefficients over time, which will be affected in the presence of strong selection. If that is the case, the MTM must account for potential changes in the genetic correlation, otherwise, a single trait model should be preferred.

The adoption of MTM may impact the required computational power, as well as the appropriate data set volume to support an efficient estimation process (MEYER; KIRKPATRICK, 2008). Another limitation of MTM is that while the number of effects in the model tend to be linear with the number of traits, the number of non-zero elements in the mixed model equations (MME) increases more rapidly (MEYER, 2005), thus more calculations are required for genetic evaluations.

To overcome the dimensionality issue associated with MTM, (KIRKPATRICK; MEYER, 2004) proposed a reduced rank model that used principal component analysis (PCA) during the (co)variance components estimation (which is called principal components model – PCM), allowing a reduction on number of parameters to estimated. (SMITH; CULLIS; THOMPSON, 2001), further used factor analysis (FA) instead of PCA proposing the factor analytical model (FAM), and (MEYER, 2007b) extended this idea to a general model, achieving more parsimony along with reduction in the number of parameters. However, the dimensionality reduction must be carefully tested because if important principal components (or factors) are removed, the reduced rank models can result in biased (co)variance components.

Despite the promising advantages of PCM and FAM, few studies exist comparing both with MTM in horses. Sánchez-Guerrero et al. (2017) used partial least squares regression, with factorial extraction, to investigate the relationship among morphometric and dressage traits; along with to decide which traits should be included in MTM analysis. This procedure involves, in a first moment, the prediction of breeding values, and subsequently, its application as an input for PCA (or FA) method. However, the PCM and FAM allow to do the rank reduction together with the breeding value prediction, or (co)variance components estimation, in a single analysis. Given that, the aim of this study was to estimate genetic parameters for MT and gait visual scores, considering 19 traits (defined latter) simultaneously, measured in Campolina horses, fitting PCM and FAM to evaluate the possible use of those methodologies in genetic evaluations of horses.

4.2. Material and Methods

4.2.1. Pedigree and phenotypic data

For the current investigation, pedigree and phenotypic information were obtained from the ABCCCampolina. It was comprised of 48,717 animals, born between 1950 and 2013, with average age at measurement of 39.57 ± 4.44 months (ranging between 21.72 and 51.98 months); and records for 19 traits, as follows: Height at Withers (HW); Height at Back (HB); Height at Croup (HC); Height at Chest (HCh); Leg Length (LL); Head Width (HWi); Chest Width (CW); Buttock Width (BW); Head Length (HL); Neck Length (NL); Back-Loins Length (BLL); Croup Length (CL); Shoulder Length (SL); Body Length (BL); Heart Girth (HG); and Cannon Girth (CG); Ratio between HW and HC (HWHC, calculated as $HWHC = 100[HW/HC]$); Ratio between BLL and BL (BLLBL, i.e., $BLLBL = 100[BLL/BL]$); and Gait Total Scores (GtS).

For all traits, contemporary groups (GC) were formed by concatenating the following information: birth year (64 levels – from 1950 to 2013); season of birth (4 levels - summer, autumn, winter, and spring); sex (2 levels – male and female); and year of registry (64 levels – from 1953 to 2016). It must be noticed that stud (herd) was not included in the concatenation because its inclusion would lead to several GC with one single animal. The GC were used to allow the breeding value prediction as the individuals' performance deviated from the mean obtained from all animals that were directly comparable. That is, animals in the same environment opportunities (born in the same year and season, with the same sex, and registered in the same year of registry).

In the Brazilian horse industry, at least for gaited horses, animals from different sex are raised together, thus they form a stud (herd) with both males and females. When the males are in the reproductive age the stud is then splitted according to sex, but the management conditions in the whole farm remain the same for males and females, allowing us to differ between sex and stud effects. Any phenotypic record above or below three standard deviations from the phenotypic mean was excluded from the dataset, as well as GCs and studs with less than 5 observations or without phenotypic variance. The descriptive statistic of the final dataset is presented in Table 1. In addition, the number of records in common for every two pair of traits is shown in Table 2.

Table 1 - Descriptive statistics of the used dataset

Trait	N	$\bar{X}$	SD	Min.	Max.	N _{GC}	N _{STUD}	N _F	N _M
HW	42,880	153.84	5.80	137.00	170.00	1,369	1,675	35,279	7,601
HB	44,045	146.56	5.56	130.00	162.00	1,426	1,731	36,300	7,745
HC	43,089	153.67	5.60	137.00	170.00	1,381	1,702	35,429	7,660
HCh	45,839	64.61	2.99	56.00	73.00	1,588	1,769	37,920	7,919
HWHC	46,956	100.04	0.88	97.30	102.84	1,678	1,835	38,776	8,180
LL	41,017	89.08	5.20	74.00	104.00	1,313	1,608	34,355	6,662
HWi	47,432	20.91	1.04	18.00	24.00	1,673	1,893	39,096	8,336
CW	44,533	38.83	2.96	30.00	47.00	1,508	1,719	36,578	7,955
BW	41,633	51.99	2.84	44.00	60.00	1,340	1,624	34,564	7,069
HL	42,093	60.61	2.84	53.00	68.00	1,298	1,684	34,578	7,515
NL	40,719	63.02	3.81	52.00	74.00	1,260	1,685	34,210	6,509
BLL	43,746	60.29	6.19	42.00	78.00	1,507	1,655	36,048	7,698
CL	39,846	51.17	3.88	40.00	62.00	1,224	1,630	32,868	6,978
SL	41,454	56.30	3.88	45.00	67.00	1,380	1,652	33,958	7,496
BL	45,670	156.26	6.75	136.00	176.00	1,510	1,823	37,775	7,895
BLLBL	41,139	38.78	4.34	26.09	51.47	1,445	1,512	34,017	7,122
HG	45,940	181.80	7.43	160.00	204.00	1,565	1,824	38,032	7,908
CG	43,210	18.96	1.09	16.00	22.00	1,413	1,728	35,310	7,900
GtS	3,243	139.51	12.66	88.00	175.00	133	201	2,142	1,101

Source: Bussiman (2021). N = number of observations; $\bar{X}$ = phenotypic average; SD = standard deviation; Min. = minimum record value; Max. = maximum record value; N_{GC} = number of contemporary groups; N_{STUD} = number of studs; N_F = number of females; and N_M = number of males. Abbreviations: HW, Height at Withers; HB, Height at Back; HC, Height at Croup; HCh, Height at Chest; HWHC, Ratio between HW and HC; LL, Leg Length; HWi, Head Width; CW, Chest Width; BW, Buttock Width; HL, Head Length; NL, Neck Length; BLL, Back-Loins Length; CL, Croup Length; SL, Shoulder Length; BL, Body Length; BLLBL, Ratio between BLL and BL; HG, Heart Girth; CG, Cannon Girth; and GtS, Gait Total Scores.

Table 2 - Number of records for each trait (diagonal) and common for every two traits (above diagonal)

Traits	Traits													
	HW	HB	HC	HCh	HWHC	LL	HWi	CW	BW	HL	NL	BLL	CL	SL
HW	42,880	41,588	41,409	41,054	41,604	38,588	41,994	40,217	38,102	38,753	37,075	39,272	36,557	37,554
HB		44,045	41,608	42,136	42,747	38,825	43,156	41,204	38,973	39,507	37,960	40,263	37,240	38,542
HC			43,089	41,262	41,858	38,634	42,247	40,508	38,385	38,824	37,138	39,451	36,859	37,864
HCh				45,839	44,433	39,422	44,864	42,507	39,871	40,238	38,765	41,769	38,088	39,589
HWHC					46,956	39,853	45,851	43,168	40,390	40,744	39,410	42,366	38,596	40,151
LL						41,017	40,244	38,454	36,452	36,650	35,445	37,629	34,982	36,079
HWi							47,432	43,626	40,834	41,311	39,898	42,886	39,140	40,623
CW								44,533	39,347	39,341	37,956	40,736	37,450	38,912
BW									41,633	37,524	36,223	38,141	35,899	36,950
HL										42,093	36,670	38,553	35,947	36,888
NL											40,719	37,165	34,733	35,569
BLL												43,746	36,425	37,977
CL													39,846	35,448
SL														41,454
BL														45,670
BLLBL														41,139
HG														45,940
CG														43,210
GtS														3,243

Source: Bussiman (2021). Abbreviations: HW, Height at Withers; HB, Height at Back; HC, Height at Croup; HCh, Height at Chest; HWHC, Ratio between HW and HC; LL, Leg Length; HWi, Head Width; CW, Chest Width; BW, Buttock Width; HL, Head Length; NL, Neck Length; BLL, Back-Loins Length; CL, Croup Length; SL, Shoulder Length; BL, Body Length; BLLBL, Ratio between BLL and BL; HG, Heart Girth; CG, Cannon Girth; and GtS, Gait Total Scores.

The pedigree was comprised of 3 generations, thus the additive genetic relationship matrix **A** (HENDERSON, 1975; WRIGHT, 1922) had 48,806 horses, born between 1900 and 2013. There were 2,784 stallions with an average progeny of 11.69 offspring/stallion; and 14,525 mares with an average progeny of 2.23 offspring/mare. The average inbreeding coefficient was 2.05% (complete pedigree), and the average additive relationship coefficient (excluding self-relationships) was 0.02. In addition, 16,263 animals had both parents unknown, and from this set of individuals, 4,421 became parents as part of the following generations; no animal had only sire unknown, and 164 individuals had only dam unknown.

Concerning the horses without parentage information, considered as population founders, the Campolina breed has a closed stud book; that is, no animal from another breed (or without both parents known) can be registered. Thus, the ancestors with unknown parents were not assumed connected (which could be done by using UPG according to sex, but this was not the purpose of this study). Finally, this breed is not a composite breed, it may arose from crosses, as well as the most Brazilian horse breeds, but as soon the breeders achieved the desired standard (around the year 1950) no horse of other breed was used in this population. From the total of measured horses for GtS (N = 3,243), around 93% also had phenotypes for the other traits (Table 2) and, in general, approximately 90% of all measured horses had phenotypes for all MT.

4.2.2. Statistical models

To analyze all traits, the following two models were implemented for full-rank and alternative analysis:

$$y_{ijk} = \mu + \beta_1 (Age_k - \overline{Age}) + \beta_2 (Age_k - \overline{Age})^2 + CG_i + Stud_j + u_k + e_{ijk}, \quad (4.1)$$

$$y_{ijkl} = \mu + \beta_1 (Age_k - \overline{Age}) + \beta_2 (Age_k - \overline{Age})^2 + CG_i + Stud_j + u_k + t_l + e_{ijkl}, \quad (4.2)$$

Where: y_{ijk} is the observed value of MT (HW; HB; HC; HCh; HWHC; LL; HWi; CW; BW; HL; NL; BLL; CL; SL; BL; BLLBL; HG; and CG); y_{ijkl} is the observed value of gait visual scores (GtS); μ is a constant; β_1 and β_2 are the regression coefficients of covariate age at measurement (Age) of the k^{th} animal (included in the model to correct for the fact that young horses will be smaller, in conformation, then completely developed animals); $\overline{Age}$ is the average of age at measurement; GC_i is the fixed effect of the i^{th} contemporary group; $Stud_j$ is

the fixed effect of the j^{th} Stud (herd) of birth; u_k is the random additive genetic effect of the k^{th} animal; t_l is the uncorrelated random effect of the l^{th} technician (person who gave the scores, only for GtS), included in the model to account for the non-genetic differences among animals caused by technician subjective preferences; e_{ijk} and e_{ijkl} are the random residual terms. Assumptions about these models will be given further. In addition, for all studied traits the normality assumptions were checked, and it holds for all traits.

4.2.3. Multitrait mixed model (MTM)

Under a vector-matrix notation models (4.1) and (4.2) can be generally written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u} + \mathbf{Z}_2\mathbf{t} + \mathbf{e}, \quad (4.3)$$

Where: $\mathbf{y}$ is the vector of phenotypic observations, ordered by trait within animal; $\boldsymbol{\beta}$ is the vector of fixed effects, ordered by trait; $\mathbf{u}$ is the vector of additive genetic random effect, ordered by trait within animal; $\mathbf{t}$ is the vector of technician uncorrelated random effect (only for GtS); $\mathbf{e}$ is the vector of random residual terms, ordered by trait within animal; $\mathbf{X}$, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are the incidence matrices for fixed effects, additive genetic random effect, and technician uncorrelated random effect, respectively.

It was assumed that $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, $\text{Var}(\mathbf{u}) = \mathbf{G}_u \otimes \mathbf{A}$, $\text{Var}(\mathbf{t}) = \mathbf{G}_t \otimes \mathbf{I}$, $\text{Var}(\mathbf{e}) = \mathbf{R} \otimes \mathbf{I}$

and therefore $\text{Var}(\mathbf{y}) = \text{Var} \begin{pmatrix} \mathbf{u} \\ \mathbf{t} \\ \mathbf{e} \end{pmatrix} = \begin{bmatrix} \mathbf{G}_u \otimes \mathbf{A} & 0 & 0 \\ 0 & \mathbf{G}_t \otimes \mathbf{I} & 0 \\ 0 & 0 & \mathbf{R} \otimes \mathbf{I} \end{bmatrix}$, where $\mathbf{G}_u$ is the additive

genetic (co)variance matrix (19 x 19) among all traits; $\mathbf{G}_t$ is the technician (co)variance matrix (1 x 1); $\mathbf{R}$ is the residual (co)variance matrix (19 x 19) among all traits; $\mathbf{I}$ is an identity matrix with proper order; and $\otimes$ is the Kronecker product. Under those assumptions, the MME are given by:

$$\begin{bmatrix} \mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z}_1 & \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z}_2 \\ \mathbf{Z}_1^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{Z}_1^T \mathbf{R}^{-1} \mathbf{Z}_1 + \mathbf{G}_0^{-1} \otimes \mathbf{A}^{-1} & \mathbf{Z}_1^T \mathbf{R}^{-1} \mathbf{Z}_2 \\ \mathbf{Z}_2^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{Z}_2^T \mathbf{R}^{-1} \mathbf{Z}_1 & \mathbf{Z}_2^T \mathbf{R}^{-1} \mathbf{Z}_2 + \mathbf{G}_1^{-1} \otimes \mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{t}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \mathbf{R}^{-1} \mathbf{y} \\ \mathbf{Z}_1^T \mathbf{R}^{-1} \mathbf{y} \\ \mathbf{Z}_2^T \mathbf{R}^{-1} \mathbf{y} \end{bmatrix}, \quad (4.4)$$

As there were two indexes' traits (HWHC and BLLBL), before proceeding to the dimensionality reduction, the additive genetic (co)variance matrix from the MTM was submitted to the Bartlett's (BARTLETT, 1950) and Kaiser-Meyer-Olkin's (KAISER, 1960) tests to verify the PCA or FA adequacy under this condition. The matrix passed Bartlett's test

($P < 0.001$) and the KMO value was 0.895 (varying from 0.752 – for HWHC – to 0.995 – for HW).

4.2.4. Rationale for rank reduction

Suppose an analysis considering q traits, and only two random effects (additive genetic and residual effects, respectively), then there are two (co)variance matrices (with dimension $q \times q$) to be estimated (one for each effect). Those matrices are assumed unstructured; i.e., the number of distinct (co)variance components is $q(q+1)/2$ (KIRKPATRICK; MEYER, 2004; MEYER, 2005; MEYER; KIRKPATRICK, 2008), thus the final number of parameters to be estimated is $q(q+1)$: for instance the number of (co)variance components increases quadratically with the number of traits.

Kirkpatrick and Meyer (2004), proposed to reduce dimensionality in MTM fitting only the m leading first principal components (PC) for a set of traits. The PCM is based on the fact that unless the traits were uncorrelated, the covariance matrices have some eigenvalues close to zero (MEYER, 2005). According to Meyer, (2005) and Meyer and Kirkpatrick (2008), fitting only first m PC results in: 1) a more parsimonious model, by reducing the number of parameters from $q(q+1)/2$ to $m(2q-m+1)/2$; 2) if omitted PC have eigenvalues close to zero, bias due this omission is negligible and is outweighed by reducing sample variances (implying in smaller mean square errors); and 3) computational power required is reduced.

The PCA is in essence a technique to reduce the dimensionality of the data by replacing the original variables by orthogonal linear combinations of them. Similar in technique (but differing in concept) the FA seeks to explain the covariance structure among variables, by finding common factors (CF) and specific factors among traits (MANLY; NAVARRO ALBERTO, 2016; THOMPSON, 2004). Smith, Cullis, and Thompson (2001) suggested a FAM to model genotype by environment interaction, and later on this model was extended to achieve more parsimony from MTM (MEYER, 2007b).

According to Meyer and Kirkpatrick (2008), the rationale under a FAM is that the original variables are modeled by common independent factors that cause covariance among traits, and specific factors (in general, assumed as uncorrelated). Predictions of common factors are, like in PCM, independent and linear combinations of the original traits. The main difference between PCM and FAM is that while under a FAM, original traits are modeled by a latent variable model (i.e., original traits are expressed as the sum of common and specific factors),

under a PCM they are not (BOLIGON et al., 2013; MEYER, 2009). Suppose a FAM with m common factors, then the number of parameters to be estimated become $q + m(2q - m + 1)/2$ which cannot exceed $q(q + 1)/2$, thus if specific factors are assumed as zero, FAM becomes PCM, and the covariance structure is the same as there were fitted only first m leading PCs (MEYER; KIRKPATRICK, 2008).

4.2.5. Principal components model (PCM)

Let $\mathbf{G}_u$ be written in terms of its eigenvalues decomposition, thus $\mathbf{G}_u = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T$ (MEYER; KIRKPATRICK, 2005a), with $\mathbf{\Lambda}$ being a diagonal matrix of eigenvalues λ_i , and $\mathbf{E}$ the matrix of eigenvectors $\mathbf{e}_i$. If elements λ_i are ordered, fitting the first m elements of $\mathbf{\Lambda}$ can be done by replacing $\mathbf{E}$ by $\mathbf{E}_m$ and $\mathbf{\Lambda}$ by $\mathbf{\Lambda}_m$, then the incidence matrix $\mathbf{Z}_1$ and the random effect $\mathbf{u}$ can be written in terms of eigenvectors of $\mathbf{G}_u$: $\mathbf{Z}^* = \mathbf{Z}_1(\mathbf{E}_m \otimes \mathbf{I})$ and $\mathbf{u}^* = (\mathbf{E}_m^T \otimes \mathbf{I})$, and the breeding values can be predicted by $\hat{\mathbf{u}} = (\mathbf{E}_m \otimes \mathbf{I})\hat{\mathbf{u}}^*$ (MEYER, 2007b; MEYER; KIRKPATRICK, 2005a). Under those assumptions, $\text{Var}(\mathbf{u}^*) = \mathbf{I} \otimes \mathbf{\Lambda}$ and the linear mixed model for this analysis was given by:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1^*\mathbf{u}^* + \mathbf{Z}_2\mathbf{t} + \mathbf{e}, \quad (4.5)$$

Where: $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, $\mathbf{Z}_2$, and $\mathbf{e}$ are the same as defined in (4.3); $\mathbf{Z}_1^*$ and $\mathbf{u}^*$ were defined above.

For this model, the MME were as follows:

$$\begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}_1^* & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}_2 \\ (\mathbf{Z}_1^*)^T\mathbf{R}^{-1}\mathbf{X} & (\mathbf{Z}_1^*)^T\mathbf{R}^{-1}\mathbf{Z}_1^* + \mathbf{\Lambda}_m^{-1} \otimes \mathbf{A}^{-1} & (\mathbf{Z}_1^*)^T\mathbf{R}^{-1}\mathbf{Z}_2 \\ \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{Z}_1^* & \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{Z}_2 + \mathbf{G}_1^{-1} \otimes \mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}^* \\ \hat{\mathbf{t}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ (\mathbf{Z}_1^*)^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}, \quad (4.6)$$

4.2.6. Factor analytical model (FAM)

For the following model steps, we assumed that the additive genetic effects could be partitioned into m common (c) and q specific (s) factors (SMITH; CULLIS; THOMPSON, 2001; THOMPSON et al., 2003); and $\mathbf{G}_u$ written as $\mathbf{G}_u = \mathbf{\Gamma}\mathbf{\Gamma}^T + \mathbf{\Psi}$ (MEYER, 2009). Replacing $\mathbf{\Gamma}$ by $\mathbf{\Gamma}_m$ allow estimation of first m factors, maintaining $\mathbf{\Psi}$ as a diagonal matrix of

specific factors. Under this reasoning, the incidence matrix $\mathbf{Z}_1$ and the random effect $\mathbf{u}$ can be written in terms of $\mathbf{\Gamma}_m$, as follows: $\mathbf{Z}^+ = \mathbf{Z}_1(\mathbf{\Gamma}_m \otimes \mathbf{I})$; $\mathbf{u} = (\mathbf{\Gamma}_m \otimes \mathbf{I})\mathbf{c} + \mathbf{s}$ (MEYER, 2009), with $\text{Var}(\mathbf{c}) = \mathbf{I}$; and $\text{Var}(\mathbf{s}) = \mathbf{\Psi} \otimes \mathbf{I}$. Given those assumptions, the linear mixed model was as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1^+\mathbf{c} + \mathbf{Z}_2\mathbf{s} + \mathbf{Z}_3\mathbf{t} + \mathbf{e}, \quad (4.7)$$

Where: $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, $\mathbf{Z}_1$, $\mathbf{Z}_2$, and $\mathbf{e}$ are the same as defined in (4.3); $\mathbf{Z}^+$, $\mathbf{c}$ and $\mathbf{s}$ are defined above.

This model was implemented through the following MME:

$$\begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}_1^+ & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}_2 & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}_3 \\ (\mathbf{Z}_1^+)^T\mathbf{R}^{-1}\mathbf{X} & (\mathbf{Z}_1^+)^T\mathbf{R}^{-1}\mathbf{Z}_1^+ + \mathbf{I}_m \otimes \mathbf{A}^{-1} & (\mathbf{Z}_1^+)^T\mathbf{R}^{-1}\mathbf{Z}_2 & (\mathbf{Z}_1^+)^T\mathbf{R}^{-1}\mathbf{Z}_3 \\ \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{Z}_1^+ & \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{Z}_2 + \mathbf{\Psi}^{-1} \otimes \mathbf{A}^{-1} & \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{Z}_3 \\ \mathbf{Z}_3^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_3^T\mathbf{R}^{-1}\mathbf{Z}_1^+ & \mathbf{Z}_3^T\mathbf{R}^{-1}\mathbf{Z}_2 & \mathbf{Z}_3^T\mathbf{R}^{-1}\mathbf{Z}_3 + \mathbf{G}_1^{-1} \otimes \mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{c}} \\ \hat{\mathbf{s}} \\ \hat{\mathbf{t}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ (\mathbf{Z}_1^+)^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_3^T\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}, \quad (4.8)$$

4.2.7. Recall of breeding value prediction

Meyer (2009) showed that when $\mathbf{G}_u$ is adequately modelled by a factor analytical structure, the MTM and FAM are directly equivalent. If the PCM consider all q principal components, then the MTM, FAM, and PCM are all equivalents (MEYER, 2009; MEYER; KIRKPATRICK, 2008). From (4.3), the breeding values are given by:

$$\hat{\mathbf{u}} = \hat{\mathbf{G}}_u \mathbf{Z}_1^T \hat{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \quad (4.9)$$

Where: $\hat{\mathbf{u}}$ is the predicted breeding value under a full rank MTM; $\hat{\mathbf{G}}_u$ is the estimated additive (co)variance matrix; $\mathbf{Z}_1^T$ is defined in (3); $\hat{\mathbf{V}}^{-1}$ is the inverse of the estimated (co)variance matrix of the vector $\mathbf{y}$ (i.e., $\mathbf{V} = \text{Var}[\mathbf{y}]$), which is the vector of phenotypic observations; $\mathbf{X}$ is the incidence matrix of fixed effects; and $\hat{\boldsymbol{\beta}}$ is the vector of estimated solutions for fixed effects (i.e., BLUE of $\boldsymbol{\beta}$).

From (4.5) and (4.7) the breeding values are predicted as (MEYER, 2009):

$$\hat{\mathbf{u}} = \begin{cases} (\mathbf{E}_m \otimes \mathbf{I}_n) \hat{\mathbf{u}}^* & \text{for the PCM} \\ (\mathbf{I}_n \otimes \mathbf{\Gamma}_m) \hat{\mathbf{c}} + \hat{\mathbf{s}} & \text{for the FAM} \end{cases}, \quad (4.10)$$

Where: $\hat{\mathbf{u}}$ is the predicted breeding value under a reduced rank by PCM or FAM, respectively; $\mathbf{I}_n$ is an identity matrix of proper order; $\mathbf{E}_m$ is the eigenvector matrix fitting only first m

principal components; $\hat{\mathbf{u}}^*$ is the vector of solutions for the random effect under a PCM; Γ_m is the factor loading matrix fitting only m factors; $\hat{\mathbf{c}}$ and $\hat{\mathbf{s}}$ are the vector of solutions for the m common and specific factors under a FAM.

Finally, provide the equivalence between those models, from a full rank MTM solution, the vectors $\hat{\mathbf{c}}$ and $\hat{\mathbf{s}}$ can be predicted as (MEYER, 2009; SMITH; CULLIS; THOMPSON, 2001):

$$\begin{cases} \hat{\mathbf{c}} = \left[\mathbf{I}_n \otimes \Gamma_m^T (\Gamma_m \Gamma_m^T + \Psi)^{-1} \right] \hat{\mathbf{u}} \\ \hat{\mathbf{s}} = \left[\mathbf{I}_n \otimes \Psi (\Gamma_m \Gamma_m^T + \Psi)^{-1} \right] \hat{\mathbf{u}} \end{cases}, \quad (4.11)$$

Where: $\hat{\mathbf{u}}$ is the predicted breeding value under a full rank MTM; $\mathbf{I}_n$ are Γ_m defined above; and Ψ is a diagonal matrix with the specific factors variance.

4.2.8. Model comparison

All models (MTM, PCM and FAM) were implemented through Average Information (JOHNSON; THOMPSON, 1995; MEYER, 1997) using Restricted Maximum Likelihood Estimation (AI-REML) at *WOMBAT* software (MEYER, 2007a), considering for convergence the default software parameters. For PCM and FAM m was from 1 to 9. Models were compared by the number of parameters, the AI-REML maximum log of likelihood ($\log L$; MEYER; KIRKPATRICK, 2005a), and two information criteria: the Akaike's information criterion (AKAIKE, 1973) corrected for the sample size (AIC_C - BURNHAM; ANDERSON, 2004) and the Schwarz's Bayesian information criterion (BIC - SCHWARZ, 1978; WOLFINGER, 1993). In addition, the dissimilarity (Dsim) between each PCM/FAM $\mathbf{G}_u$ and MTM $\mathbf{G}_u$ was evaluated by the Schaeffer method (SCHAEFFER, 1986); as well as by computing the square root of the average squared deviation of correlations ($\sqrt{\Delta r}$), as proposed by (MEYER, 2007b). Those were calculated as follows:

$$AIC_C = -2 \log L + 2p \left(1 + \frac{p+1}{N-p-1} \right), \quad (4.12)$$

$$BIC = -2 \log L + p \log(N - r(\mathbf{X})), \quad (4.13)$$

$$Dsim = \frac{\|\hat{\mathbf{G}}_{0,m} - \hat{\mathbf{G}}_{0,MTM}\|}{\|\hat{\mathbf{G}}_{0,MTM}\|}, \quad (4.14)$$

$$\sqrt{\Delta r} = \sqrt{\sum_{i=1}^{19} \sum_{j=i+1}^{19} \frac{(r_{gij,m} - r_{gij,MTM})^2}{171}}, \quad (4.15)$$

Where: $-2 \log L$ is the AI-REML maximum log of likelihood; p is the number of parameters; N is the total number of records; $r(\mathbf{X})$ is the rank of fixed effects coefficient matrix; $\hat{\mathbf{G}}_{u,m}$ is the estimated additive genetic (co)variance matrix fitting m PC or CF; $\hat{\mathbf{G}}_{u,MTM}$ is the estimated additive genetic (co)variance from MTM; and $\|\bullet\|$ denotes the Euclidean norm; $r_{ij,m}$ is the estimated genetic correlation coefficient between traits i and j from analysis fitting m PC or CF; $r_{ij,MTM}$ is the same coefficient but from the full rank MTM analysis; and 171 is the number of correlations being compared.

Furthermore, the angle (in degrees) between the first three eigenvectors ($\mathbf{e}_i$) of the additive genetic (co)variance matrix, from all alternative analysis, and MTM were estimated according to (KIRKPATRICK; MEYER, 2004; MEYER, 2007b). Concerning the angles, it was calculated the angle between $\mathbf{e}_{1MTM}$ (first eigenvector from MTM) and $\mathbf{e}_{1PCM/FAM}$ (first eigenvector from PCM or FAM), between $\mathbf{e}_{2MTM}$ (second eigenvector from MTM) and $\mathbf{e}_{2PCM/FAM}$ (second eigenvector from PCM or FAM), and between $\mathbf{e}_{3MTM}$ (third eigenvector from MTM) and $\mathbf{e}_{3PCM/FAM}$ (third eigenvector from PCM or FAM). Graphical visualization of results was performed by the “*ggplot*” function from the “*ggplot2*” *R* package (WICKHAM, 2016) from *R language* (R CORE TEAM, 2020). All analyses were performed in an Intel Xeon server with 3.60 GHz clock speed; 15 MB of cache memory; and 188 GB RAM memory. The final model was chosen based on BIC value: models with smaller BIC are better supported by the data.

4.3. Results

4.3.1. Model comparisons

Maximum $\log L$ increased according to the number of PC or CF fitted, being highest for the MTM (Figure 1A). The AIC_C decreased as the number of parameters increased, stabilizing at the FAM7 (Factor Analytical Model with 7 common factors – Figure 1C), whereas the BIC (with a more stringent penalty in the number of parameters) indicated model FAM5 as the best one (Figure 1B). Moreover, for the same number of CF and PC, the FAM showed

higher values of $\log L$, and lower values of AIC_C and BIC. As stated before, the final model was chosen based on BIC value (balancing the model's adequacy and complexity), which turned out to be the model FAM5.

Figure 1 - Fitting statistics for all tested models. Maximum logarithm of likelihood (A), Bayesian information criteria (B), and Akaike's criteria, corrected for the sample size (C). Order of fit 1 to 9 is the number of principal components (or common factors) fitted, and 19 is the standard multitrait mixed model

Source: Bussiman (2021).

The first three eigenvalues (λ_i) of additive genetic (co)variance matrix from MTM accounted for 54.33%, 23.03%, and 6.91% of the total genetic variation, respectively, whereas for the FAM5, the proportion of the additive genetic variance explained by λ_i was 77.19%, 10.86%, and 6.93%, respectively (Table 3). The dissimilarity values ranged from 0.08 (FAM9) to 0.49 (PCM1); in other hand, $\sqrt{\Delta r}$ values were substantially smaller, varying between 0.03 and 0.65, FAM9 and PCM1, respectively (Table 3). For residual (co)variance matrix, after 4 CF, Dsim values and $\sqrt{\Delta r}$ decreased substantially and tended to be close to zero (Table 3).

The total additive genetic variation, among all animals, was 72.42 (MTM), and FAM5 captured 67.75% of it, and by comparing the residual variation it was 256.21 (MTM) while FAM5 estimated 99.49% of it (Table 3). All PCM overestimated residual variance, whereas all FAM, starting from 3 CF, underestimated it. The angles between eigenvectors $\mathbf{e}_{1MTM}$ (first eigenvector from MTM) and $\mathbf{e}_{1PCM/FAM}$ (first eigenvector from PCM or FAM) ranged from 2.49 (PCM9) to 9.76 (PCM1), whereas for the second and third eigenvectors it varied more (Table 3).

Table 3 - Dissimilarity (Dsim) between a given genetic additive (co)variance matrix and the one from a multitrait mixed model analysis; square root of average squared deviation of correlations ($\sqrt{\Delta r}$); proportion of additive genetic variance (λ_i in %) explained by first three eigenvalues; estimates of total variation ($\sum_{i=1}^n \lambda_i$); angles between estimates of i^{th} eigenvectors (α_i in degrees); and

Model ^a	Additive genetic (co)variance matrix										Residual (co)variance matrix		
	Dsim	$\sqrt{\Delta r}$	$\lambda_1\%$	$\lambda_2\%$	$\lambda_3\%$	$\sum_{i=1}^n \lambda_i$	α_1	α_2	α_3	P	Dsim	$\sqrt{\Delta r}$	$\sum_{i=1}^n \lambda_i$
PCM1	0.49	0.65	100.00	--	--	31.91	9.76	--	--	210	0.46	0.03	295.51
PCM2	0.44	0.46	92.93	7.07	--	37.98	6.88	76.58	--	228	0.42	0.02	289.89
PCM3	0.33	0.35	78.97	16.94	4.09	44.93	5.29	32.26	131.80	245	0.32	0.02	283.23
PCM4	0.31	0.30	77.62	18.07	3.88	47.55	4.81	30.48	133.00	261	0.30	0.02	280.80
PCM5	0.32	0.28	77.43	17.89	3.77	47.80	4.95	33.53	51.05	276	0.31	0.02	280.45
PCM6	0.29	0.24	73.61	17.73	5.12	49.87	4.75	30.51	33.11	290	0.29	0.02	278.29
PCM7	0.29	0.23	72.22	17.09	5.26	49.76	4.78	30.53	33.35	303	0.28	0.01	277.82
PCM8	0.27	0.20	71.72	16.03	4.90	53.42	2.87	31.01	38.18	315	0.26	0.01	274.80
PCM9	0.24	0.18	68.37	17.54	5.92	56.24	2.49	25.78	37.95	326	0.23	0.01	272.19
MTM	0.00	0.00	54.33	23.03	6.91	72.42	0.00	0.00	0.00	381	0.00	0.00	256.21
FAM1	0.26	0.28	100.00	--	--	34.54	9.26	--	--	229	0.22	0.02	267.08
FAM2	0.15	0.17	87.86	12.14	--	42.24	6.63	70.43	--	247	0.12	0.02	260.28
FAM3	0.15	0.13	84.34	11.85	3.81	44.00	6.58	73.79	89.40	264	0.12	0.02	258.67
FAM4	0.11	0.07	79.26	10.93	6.44	48.07	6.36	71.37	66.85	280	0.08	0.01	255.28
FAM5	0.10	0.05	77.19	10.86	6.93	49.07	6.64	70.17	107.25	295	0.07	0.01	254.94
FAM6	0.10	0.05	75.40	11.04	6.57	50.26	6.26	65.10	104.80	309	0.08	0.01	254.44
FAM7	0.10	0.04	73.97	11.43	6.74	51.48	5.17	58.47	110.35	322	0.08	0.00	253.85
FAM8	0.09	0.04	71.26	11.33	7.60	54.26	5.22	51.58	123.64	334	0.07	0.00	253.58
FAM9	0.08	0.03	56.79	25.03	6.43	69.92	3.38	8.71	15.67	345	0.05	0.00	256.24

Source: Bussiman (2021). ^aPCM n = principal components model fitting n principal components; FAM n = factor analytical model fitting n common factors. Bold font means that the model was the best based on BIC values.

The sum of specific variances was 24.95 (FAM5) which represents 34.45% of the total additive genetic variance among all animals. This value decreased (until 3.39 – FAM9) as the number of CF increased. Specific variances given as proportion of phenotypic variance (uniqueness) for each trait ranged from 0.01 to 15.60%, for BLL and GtS, respectively. In addition, proportion of uniqueness is presented in Table 4.

Table 4 - Values of specific variances (S) and uniqueness (U – S as proportion of phenotypic variance)

Trait	S	U
HW	0.02	0.14
HB	0.22	1.32
HC	0.03	0.18
HCh	0.02	0.27
HWHC	0.00	0.13
LL	0.05	0.34
HWi	0.04	4.57
CW	0.44	7.45
BW	0.41	6.89
HL	0.48	9.63
NL	0.70	7.16
BLL	0.00	0.01
CL	0.55	5.38
SL	0.84	8.82
BL	0.26	0.95
BLLBL	0.05	0.55
HG	1.52	3.59
CG	0.05	5.55
GtS	19.26	15.60

Source: Bussiman (2021). Abbreviations: HW, Height at Withers; HB, Height at Back; HC, Height at Croup; HCh, Height at Chest; HWHC, Ratio between HW and HC; LL, Leg Length; HWi, Head Width; CW, Chest Width; BW, Buttock Width; HL, Head Length; NL, Neck Length; BLL, Back-Loins Length; CL, Croup Length; SL, Shoulder Length; BL, Body Length; BLLBL, Ratio between BLL and BL; HG, Heart Girth; CG, Cannon Girth; and GtS, Gait Total Scores.

4.3.2. Genetic parameters

Moderate to high heritability estimates were obtained for the height-related traits (HW, HB, HC, and HCh), ranging from 0.18 (HCh) to 0.43 (HW). Ratio traits were in the same moderate level of heritability: 0.12 (HWHC) and 0.19 (BLLBL); the same occurred with perimeter traits, which were 0.10 (CG) and 0.14 (HG). Length traits had heritability estimates varying from 0.11 (CL) to 0.31 (BL), whereas the width traits varied the most: from 0.09 (HWi) to 0.18 (BW) (Figure 2).

Figure 2 - Genetic parameter estimates from factor analytical model fitting five common factors. Genetic correlations (above diagonal – A) and residual correlations (below diagonal – A); standardized variance components (B) together with the heritability (numbers over each bar). Size of circles is proportional to magnitude of the correlation coefficient

Source: Bussiman (2021). Abbreviations: HW, Height at Withers; HB, Height at Back; HC, Height at Croup; HCh, Height at Chest; HWHC, Ratio between HW and HC; LL, Leg Length; HWi, Head Width; CW, Chest Width; BW, Buttock Width; HL, Head Length; NL, Neck Length; BLL, Back-Loins Length; CL, Croup Length; SL, Shoulder Length; BL, Body Length; BLLBL, Ratio between BLL and BL; HG, Heart Girth; CG, Cannon Girth; and GtS, Gait Total Scores.

The GtS heritability was 0.16 and the proportion of phenotypic variance explained by the technician was 12.90%. Genetic correlations were between -0.32 and 0.98, when considering all studied traits. Negative genetic correlations were found only between ratio traits and MT, and between GtS and all MT (Figure 2, and Table 5). The strongest genetic correlations were obtained between the pairs HW and HB (0.98) and HW and HC (0.98) (Table 5).

Around 80% of the total genetic correlation estimates were positive (Figure 2, and Table 5). GtS and all other traits were negatively genetic correlated (except for correlations between GtS and BLLBL – 0.10 – and between GtS and BLL – 0.01), which coefficients ranged from -0.19 (GtS and LL) to -0.07 (GtS and HWi) (Figure 2, and Table 5). Excluding ratio traits, all MT traits were positively correlated (Figure 2).

Table 5A - Estimated genetic parameters from a full rank multitrait mixed model. Heritability (diagonal, bold); genetic correlations (above diagonal); and residual correlations (below diagonal)

Traits	Traits														
	HW	HB	HC	HCh	HWHC	LL	HWi	CW	BW	HL	NL	BLL	CL	SL	BL
HW	0.43	0.97	0.98	0.68	0.20	0.89	0.52	0.56	0.67	0.69	0.68	0.34	0.62	0.51	0.80
HB	0.86	0.40	0.97	0.69	0.12	0.86	0.51	0.50	0.64	0.69	0.69	0.35	0.62	0.52	0.76
HC	0.91	0.84	0.42	0.67	0.01	0.87	0.53	0.56	0.67	0.69	0.68	0.33	0.64	0.51	0.80
HCh	0.38	0.36	0.38	0.18	0.10	0.29	0.50	0.44	0.57	0.62	0.46	0.24	0.40	0.47	0.61
HWHC	0.24	0.07	-0.17	0.03	0.12	0.20	0.03	0.01	0.05	0.11	0.12	0.07	-0.04	0.04	0.07
LL	0.70	0.58	0.62	-0.39	0.22	0.28	0.37	0.44	0.52	0.53	0.61	0.30	0.57	0.40	0.68
HWi	0.27	0.25	0.26	0.23	0.04	0.09	0.10	0.40	0.50	0.67	0.38	0.22	0.54	0.36	0.56
CW	0.41	0.35	0.41	0.30	0.02	0.18	0.28	0.16	0.63	0.50	0.37	0.07	0.50	0.36	0.63
BW	0.47	0.41	0.46	0.37	0.03	0.18	0.31	0.55	0.18	0.65	0.54	0.23	0.53	0.45	0.69
HL	0.47	0.42	0.44	0.41	0.07	0.15	0.35	0.37	0.42	0.29	0.69	0.31	0.53	0.47	0.71
NL	0.44	0.40	0.42	0.29	0.06	0.22	0.19	0.33	0.33	0.49	0.18	0.39	0.53	0.49	0.65
BLL	0.24	0.23	0.24	0.10	0.01	0.16	0.08	0.06	0.13	0.11	0.26	0.19	0.28	0.29	0.34
CL	0.37	0.33	0.36	0.20	0.02	0.21	0.27	0.39	0.38	0.32	0.34	0.15	0.12	0.47	0.63
SL	0.40	0.36	0.42	0.21	-0.03	0.24	0.14	0.39	0.39	0.25	0.33	0.21	0.37	0.14	0.49
BL	0.63	0.53	0.60	0.37	0.07	0.34	0.30	0.42	0.51	0.45	0.39	0.20	0.39	0.35	0.31
BLLBL	0.00	0.02	0.01	-0.05	-0.02	0.03	-0.04	-0.10	-0.06	-0.06	0.12	0.92	0.01	0.08	-0.20
HG	0.40	0.36	0.40	0.44	0.00	0.06	0.18	0.40	0.41	0.31	0.26	0.12	0.28	0.30	0.40
CG	0.24	0.23	0.24	0.18	0.00	0.10	0.24	0.21	0.20	0.21	0.14	-0.08	0.21	0.15	0.27
GtS	0.12	0.08	0.08	0.09	0.09	0.05	-0.03	0.21	0.12	0.14	0.04	-0.06	0.09	0.05	0.10

Source: Bussiman (2021). Abbreviations: HW, Height at Withers; HB, Height at Back; HC, Height at Croup; HCh, Height at Chest; HWHC, Ratio between HW and HC; LL, Leg Length; HWi, Head Width; CW, Chest Width; BW, Buttock Width; HL, Head Length; NL, Neck Length; BLL, Back-Loins Length; CL, Croup Length; SL, Shoulder Length; BL, Body Length; BLLBL, Ratio between BLL and BL; HG, Heart Girth; CG, Cannon Girth; and GtS, Gait Total Scores.

Table 5B - Estimated genetic parameters from a factor analytical model, fitting only 5 common factors. Heritability (diagonal, bold); genetic correlations (above diagonal); and residual correlations (below diagonal)

Traits	Traits														
	HW	HB	HC	HCh	HWHC	LL	HWi	CW	BW	HL	NL	BLL	CL	SL	BL
HW	0.43	0.98	0.98	0.68	0.19	0.89	0.55	0.57	0.69	0.72	0.71	0.34	0.64	0.51	0.80
HB	0.86	0.40	0.97	0.69	0.12	0.85	0.54	0.54	0.67	0.70	0.69	0.36	0.62	0.51	0.76
HC	0.91	0.85	0.42	0.67	-0.01	0.87	0.57	0.59	0.70	0.71	0.71	0.33	0.66	0.52	0.81
HCh	0.38	0.36	0.38	0.18	0.10	0.29	0.59	0.47	0.61	0.66	0.52	0.25	0.43	0.48	0.61
HWHC	0.24	0.07	-0.17	0.03	0.12	0.19	-0.02	0.00	0.03	0.08	0.10	0.06	-0.08	0.01	0.07
LL	0.70	0.58	0.62	-0.39	0.22	0.28	0.37	0.46	0.53	0.53	0.62	0.30	0.57	0.38	0.68
HWi	0.27	0.25	0.25	0.22	0.05	0.10	0.09	0.46	0.52	0.54	0.47	0.23	0.44	0.38	0.62
CW	0.41	0.34	0.41	0.30	0.02	0.18	0.27	0.15	0.54	0.53	0.46	0.07	0.46	0.34	0.67
BW	0.46	0.40	0.46	0.36	0.03	0.18	0.31	0.56	0.18	0.62	0.56	0.23	0.52	0.43	0.74
HL	0.47	0.42	0.44	0.40	0.08	0.16	0.37	0.36	0.42	0.27	0.58	0.31	0.53	0.45	0.75
NL	0.44	0.40	0.42	0.28	0.07	0.22	0.18	0.31	0.33	0.52	0.17	0.41	0.52	0.42	0.69
BLL	0.24	0.22	0.24	0.10	0.01	0.16	0.08	0.06	0.13	0.11	0.26	0.19	0.29	0.31	0.35
CL	0.37	0.33	0.36	0.20	0.03	0.21	0.29	0.39	0.39	0.32	0.35	0.15	0.11	0.38	0.66
SL	0.41	0.37	0.42	0.21	-0.02	0.24	0.14	0.39	0.39	0.26	0.34	0.21	0.38	0.13	0.50
BL	0.63	0.53	0.60	0.37	0.07	0.34	0.29	0.41	0.50	0.44	0.38	0.19	0.38	0.34	0.31
BLLBL	0.00	0.01	0.00	-0.05	-0.02	0.03	-0.04	-0.10	-0.06	-0.06	0.12	0.92	0.01	0.07	-0.20
HG	0.40	0.36	0.40	0.44	0.01	0.06	0.18	0.41	0.42	0.31	0.26	0.12	0.29	0.32	0.39
CG	0.24	0.24	0.25	0.18	0.00	0.10	0.24	0.20	0.20	0.22	0.14	-0.08	0.20	0.16	0.28
GtS	0.11	0.07	0.08	0.07	0.10	0.05	-0.02	0.20	0.13	0.14	0.04	-0.05	0.10	0.04	0.09

Source: Bussiman (2021). Abbreviations: HW, Height at Withers; HB, Height at Back; HC, Height at Croup; HCh, Height at Chest; HWHC, Ratio between HW and HC; LL, Leg Length; HWi, Head Width; CW, Chest Width; BW, Buttock Width; HL, Head Length; NL, Neck Length; BLL, Back-Loins Length; CL, Croup Length; SL, Shoulder Length; BL, Body Length; BLLBL, Ratio between BLL and BL; HG, Heart Girth; CG, Cannon Girth; and GtS, Gait Total Scores.

Estimates of eigenvectors, scaled by the corresponding λ_i (BUZANSKAS et al., 2013); i.e., the weights of each trait in order to form an aggregate genotype from MTM and FAM5 are presented in Figure 3. A very similar pattern was found on the first eigenvector (Figure 3A), while the main differences on the second were related to BLL, SL, BLLBL, HG, and GtS (Figure 3B). On the other hand, the third eigenvector presented notorious differences for HW, HB, HC, HCh, LL, HG, and GtS (Figure 3C).

Figure 3 - Weights for each trait based on first three eigenvectors from additive genetic (co)variance matrix from a factor analytical model fitting five common factors and a standard multitrait mixed model, scaled by the square root of respective eigenvalue. First eigenvector (A), second eigenvector (B), and third eigenvectors (C)

Source: Bussiman (2021).

There was an overall agreement between the additive genetic (co)variance components obtained from FAM5 (best BIC) and MTM (Table 3). After comparing elements from the matrix generated by both methods (Figure 4), we identified that in general, FAM overestimated the reported values. In general, heritability estimates did not differ, except for GtS which was increased in 0.02 from MTM to FAM5 (Table 5). The most affected correlation estimates were between GtS and CG (increased 0.39); GtS and BLLBL (decreased 0.13); GtS and HWi (decreased 0.18); HG and SL (decreased 0.12); NL and HL (decreased 0.11); and HL and HWi (decreased 0.13).

Figure 4 - Scatter Plot of estimated genetic parameters from a factor analytical model fitting five common factors and from a standard multitrait mixed model (A); and of estimated sampling errors in the same analyzes (B)

Source: Bussiman (2021).

4.3.3. Sampling errors and rank correlations

The difference (squared) between the prediction errors obtained from both models, MTM and FAM5, was 0.01. The FAM5 overestimated the sampling errors, especially for the values ranging between 0.07 and 0.13. Due to the large number of covariance components (342), sampling errors are presented in Table 6.

Table 6A - Estimated sampling errors (S.E.) from a full rank multitrait mixed model. S.E. of heritability (diagonal, bold); S.E. of genetic correlations (above diagonal); and S.E. of residual correlations (below diagonal)

Traits	Traits														
	HW	HB	HC	HCh	HWHC	LL	HWi	CW	BW	HL	NL	BLL	CL	SL	BL
HW	0.01	0.00	0.00	0.02	0.03	0.01	0.03	0.02	0.02	0.01	0.02	0.02	0.02	0.02	0.01
HB	0.00	0.01	0.00	0.02	0.03	0.01	0.03	0.02	0.02	0.01	0.02	0.02	0.03	0.02	0.01
HC	0.00	0.00	0.01	0.02	0.03	0.01	0.03	0.02	0.02	0.01	0.02	0.02	0.02	0.02	0.01
HCh	0.01	0.01	0.01	0.01	0.04	0.03	0.04	0.03	0.03	0.02	0.03	0.03	0.04	0.03	0.02
HWHC	0.01	0.01	0.01	0.01	0.01	0.03	0.05	0.04	0.04	0.04	0.04	0.04	0.05	0.05	0.03
LL	0.00	0.01	0.01	0.01	0.01	0.01	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.03	0.02
HWi	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.04	0.04	0.03	0.04	0.04	0.04	0.05	0.03
CW	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.03	0.03	0.04	0.03	0.04	0.02
BW	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.03	0.03	0.03	0.03	0.02
HL	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.03	0.03	0.03	0.02
NL	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.03	0.03	0.03	0.02
BLL	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.04	0.04	0.03
CL	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.04	0.03
SL	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.03
BL	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
BLLBL	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.01
HG	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
CG	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
GtS	0.03	0.03	0.03	0.02	0.02	0.03	0.02	0.02	0.02	0.03	0.03	0.03	0.02	0.02	0.03

Source: Bussiman (2021). Abbreviations: HW, Height at Withers; HB, Height at Back; HC, Height at Croup; HCh, Height at Chest; HWHC, Ratio between HW and HC; LL, Leg Length; HWi, Head Width; BW, Chest Width; CW, Chest Width; BW, Buttock Width; HL, Head Length; NL, Neck Length; BLL, Back-Loins Length; CL, Croup Length; SL, Shoulder Length; BL, Body Length; BLLBL, Ratio between BLL and BL; HG, Heart Girth; CG, Cannon Girth; and GtS, Gait Total Scores.

Table 6B - Estimated sampling errors (S.E.) from a factor analytical model, fitting only 5 common factors. S.E. of heritability (diagonal, bold); S.E. of genetic correlations (above diagonal); and S.E. of residual correlations (below diagonal)

Traits	Traits														
	HW	HB	HC	HCh	HWHC	LL	HWi	CW	BW	HL	NL	BLL	CL	SL	BL
HW	0.01	0.00	0.00	0.02	0.03	0.01	0.03	0.03	0.02	0.01	0.01	0.02	0.02	0.03	0.01
HB	0.00	0.01	0.00	0.02	0.03	0.01	0.04	0.03	0.02	0.01	0.02	0.02	0.02	0.03	0.01
HC	0.00	0.00	0.01	0.02	0.03	0.01	0.03	0.02	0.02	0.01	0.02	0.02	0.02	0.03	0.01
HCh	0.01	0.01	0.01	0.01	0.04	0.03	0.03	0.04	0.03	0.02	0.04	0.03	0.05	0.04	0.02
HWHC	0.01	0.01	0.01	0.01	0.01	0.03	0.07	0.06	0.05	0.04	0.05	0.04	0.06	0.08	0.03
LL	0.00	0.01	0.01	0.01	0.01	0.01	0.05	0.04	0.03	0.03	0.03	0.03	0.03	0.05	0.02
HWi	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.03	0.01	0.01	0.03	0.06	0.04	0.03	0.03
CW	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.03	0.05	0.03	0.04	0.01
BW	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.04	0.02	0.02	0.01
HL	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.03	0.02	0.02	0.01
NL	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.04	0.02	0.02	0.02
BLL	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.05	0.06	0.03
CL	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.03	0.02
SL	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.03
BL	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
BLLBL	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.01
HG	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
CG	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
GtS	0.03	0.03	0.03	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.03	0.02	0.02	0.03

Source: Bussiman (2021). Abbreviations: HW, Height at Withers; HB, Height at Back; HC, Height at Croup; HCh, Height at Chest; HWHC, Ratio between HW and HC; LL, Leg Length; HWi, Head Width; BW, Buttock Width; HL, Head Length; NL, Neck Length; BLL, Back-Loins Length; CL, Croup Length; SL, Shoulder Length; BL, Body Length; BLLBL, Ratio between BLL and BL; HG, Heart Girth; CG, Cannon Girth; and GtS, Gait Total Scores.

In addition, the overall mean squared difference between the predicted breeding values (EBV) from MTM and FAM5 varied from 0.00 (HWHC) to 1.38 (GtS). The rank correlations between EBVs from MTM and FAM5 ranged between 0.50 (GtS) and 1.00 (BLL), when considering solely the 2% superior animals. Considering the 10% superior horses, rank correlations ranged from 0.31 (GtS) to 1.00 (HW, HC, and BLL). Except for HWi, exclusively for the 2% best animals, which showed a correlation of 0.71, all other rank correlations for MT were higher than 0.80 (Table 7).

Table 7 - Rank correlations between predicted breeding value (EBV) from a full rank multitrait mixed (MTM) model and from a factor analytical model, fitting only five common factors (FAM5), i.e., $r_s(EBV_{MTM}, EBV_{FAM5})$

Trait	Percentage and number of selected animals				
	All horses	2% (976 horses)	10% (4,880 horses)	20% (9,761 horses)	50% (24,403 horses)
HW	1.00	0.99	1.00	1.00	1.00
HB	1.00	0.98	0.99	1.00	1.00
HC	1.00	0.99	1.00	1.00	1.00
HCh	1.00	0.97	0.99	0.99	1.00
HWHC	1.00	0.98	0.99	0.99	0.99
LL	1.00	0.98	0.99	1.00	1.00
HWi	0.97	0.71	0.81	0.85	0.93
CW	0.99	0.83	0.93	0.96	0.99
BW	0.99	0.88	0.94	0.96	0.99
HL	1.00	0.91	0.96	0.98	0.99
NL	0.99	0.80	0.90	0.94	0.98
BLL	1.00	1.00	1.00	1.00	1.00
CL	0.99	0.85	0.93	0.95	0.98
SL	0.98	0.82	0.89	0.93	0.96
BL	1.00	0.98	0.99	1.00	1.00
BLLBL	1.00	0.97	0.99	0.99	1.00
HG	0.99	0.87	0.93	0.95	0.98
CG	0.99	0.85	0.91	0.95	0.98
GtS	0.82	0.50	0.31	0.32	0.49

Source: Bussiman (2021). Abbreviations: HW, Height at Withers; HB, Height at Back; HC, Height at Croup; HCh, Height at Chest; HWHC, Ratio between HW and HC; LL, Leg Length; HWi, Head Width; CW, Chest Width; BW, Buttock Width; HL, Head Length; NL, Neck Length; BLL, Back-Loins Length; CL, Croup Length; SL, Shoulder Length; BL, Body Length; BLLBL, Ratio between BLL and BL; HG, Heart Girth; CG, Cannon Girth; and GtS, Gait Total Scores.

4.4. Discussion

De Oliveira Bussiman et al. (2018), proposed the use of HWHC and BLLBL in the Campolina breed. Those authors conducted a MTM (co)variance components estimation under a Bayesian approach for the same traits as in the present study. These two traits are computed

from other traits (i.e., HWHC is the ratio between HW and HC; and BLLBL is the ratio between BLL and BL), which inserts redundancy among all (co)variance components. Considering the selection for higher values of HW, and its negative genetic correlation with GtS; after generations, a reduction on economic value for GtS, in the Campolina horse, is expected. However, if we select based on BLLBL (for lower values), the result is an increase on the proportion of BLL over BL (which is good from mechanical aspects of horse locomotion), causing an increase in GtS (by correlated response). Furthermore, the selection for lower values HWHC will cause minor effects on GtS (due to the low magnitude of the genetic correlation coefficient between HWHC and GtS), but its inclusion on analysis led to an increase of breeding value accuracy for GtS.

The decision to keep HWHC and BLLBL in the analysis, was based on the above mentioned (supported by the results from Bartlett's and Kaiser-Meyer-Olkin's tests of PCA – or FA – adequacy), as well as the possibility to directly evaluate the time reduction for (co)variance components estimation when fitting a reduced rank structure comparing our results with those from (DE OLIVEIRA BUSSIMAN et al., 2018). In our previous published results, the time to estimate all (co)variance components was around a month, which was due to limited computing power. For the present study, the slowest analysis was the MTM, which took about 10 days to converge. Concerning the FAM5, as the MTM (co)variance components were used as starting values, the time for convergence was 5 days.

The reduced number of records evaluated for GtS reflects late data collection involving the gait score traits that only happened after 1998, carried out by the ABCCCampolina. It is well known that a large (and representative) dataset is needed for the correct genetic covariance estimation (HILL, 2013); in particular, if we consider the number of animals phenotyped simultaneously for at least two traits. In this study, the overall number of horses phenotyped for all MT was high (80%), thus the differences on sampling errors of the genetic correlations among models were not expected to be derived from the analyzed dataset. Boligon et al. (2015), after conducting a series of reduced rank analysis, investigated a dataset presenting a substantially lower percentage of animal phenotyped for a pair of traits; and despite of it, reported very low values of sampling errors.

The observed increase in the logarithm of maximum likelihood ($\log L$) in our study corroborates with previous studies using PCM or FAM (BIGNARDI et al., 2012, 2014; BOLIGON et al., 2013, 2015; MEYER, 2005, 2007b). According to (MEYER; KIRKPATRICK, 2008), if the rank's model is correct, AIC_C is at minimum value, and the sum of eigenvalues stabilizes, as well as the magnitude of leading principal components. This may

indicate that the re-partition process (from additive genetic variance into residual variance) was not applied in this situation, and is often the case when too few principal components are fitted (MEYER, 2007b). In addition, beyond the optimal model, the improvement on $\log L$ is expected to be negligible (TYRISEVÄ et al., 2011).

A negligible increase in the $\log L$ was noted fitting more than seven common factors; and as expected, the same behavior was presented by the AIC_C . On the other hand, BIC imposes a stronger penalty on the number of parameters, thus the model with better balance (between model fitting and complexity) was the one using only five common factors (based on BIC value). Bignardi et al. (2012) and Meyer (2007b) also found discrepancy between AIC_C and BIC on best model determination, but Boligon et al. (2015) found concordance between those criteria. This may indicate, in our models, that the optimal rank is between five and seven common factors.

The PCM showed no fit to the data at all, similar to the results reported by (Meyer, 2005), fitting reduced rank models to study live ultrasound scan traits in Angus cattle. This author also reported that the sum of eigenvectors increased with the number of fitted genetic principal components, which was evidence that, at a reduced rank, the genetic variance was underestimated. In this study, the total genetic variation increased from 31.91 (PCM1) to 56.24 (PCM9), and the proportion of additive genetic variance explained by the third eigenvalue of the additive (co)variance matrix, tended to increase across all PCM analysis.

In the opposite direction, Bignardi et al. (2012, 2014); Boligon et al. (2013, 2015) and Meyer (2007b), found that reduced rank models were able to reduce the dimensionality of MTM analysis. In the present study, it is possible that the number of negative and low genetic correlations between some traits (HWHC and all MT; BLLBL and all MT; and GtS and all MT – except with BLLBL and BLL) affected the PCM method. According to (KIRKPATRICK; MEYER, 2004), the PCM is a special case of a standard PCA, and it is well known that PCA is more effective (also advantageous) when the variables are strongly correlated (LATTIN; CARROLL; GREEN, 2011; MANLY; NAVARRO ALBERTO, 2016). To overcome this situation, (MEYER, 2005) suggested the use of scree test (CATTELL, 1966) from factor analysis to investigate MTM results and further determine the appropriate submodel, but the author also states that there are no appropriate criterion for it.

Despite in this study no PCM was able to reduce the additive genetic (co)variance matrix dimensionality, in a recent study; Solar Diaz et al. (2020) successfully applied PCA to the additive genetic (co)variance matrix of MT on Campolina horses. Those authors found that the

first eigenvalue explained 97.75% of the additive genetic variance. Applying the same procedure to the estimated (co)variance matrix from FAM5 (only for common factors), the first eigenvalue explained only 52.23% of the additive genetic variance, which reflects the more complex structure when HWHC, BLLBL, and GtS were included on the same analysis.

Nascimento et al. (2014), reported the equivalence on performing PCA with EBV and PCM. However, Solar Diaz et al. (2020) and Vargas et al. (2018), suggested rather than use of EBV to conduct PCA, first to estimate (co)variance components and then to perform PCA on the additive genetic (co)variance matrix. Thus, despite the substantial difference associated with the proportion of the variance explained by each eigenvector, all the eigenvalues should be in the same direction among all these methods, once the covariance among EBV approximates the additive genetic (co)variance matrix.

Differently from PCA, the FA assumes the covariance structure among the traits is a result of latent uncorrelated common variables (called factors) and the partition of each trait variance that is not explained by the factors is accounted by specific components in each trait (MANLY; NAVARRO ALBERTO, 2016; MEYER, 2007b; SMITH; CULLIS; THOMPSON, 2001). Thus, in general, FAM tends to better fit the data, when there are traits which are weakly correlated or having specific “causes”. The reason for it, is that FAM does not repartition the genetic variance (not accounted for a limited number of PC) into residual variance, instead it goes to specific factors variances (MEYER, 2007b).

For our analysis, the proportion of phenotypic variance explained by the specific factors was much higher than for Bignardi et al. (2012) and Boligon et al. (2013), but lower than reported by Meyer (2007b). This may be related with data structure and also associated with the traits itself, once all MT were expected to share common factors. In fact, the proportion of additive genetic variance explained by the first three eigenvalues of the fitted (co)variance matrix of common factors (from FAM5) was 94.98%. Moreover, when considered together, the common and specific genetic factors, the total estimated genetic variance was 74.02, which is overestimated compared with MTM (72.42).

To fit a reduced covariance structure for the additive genetic (co)variance matrix, is important to find a balance between how much bias (when important PCs or factors are omitted) and sampling variance (when fitting additional PCs or factors explaining negligible variance) are allowed (MEYER, 2007b). Boligon et al. (2013), also argued about the dependence on the amount of information that is included in the analysis. By using simulated data; Meyer and Kirkpatrick (2005a, 2005b), showed models fitting a reduced rank (and selected based on minimum BIC) yield genetic covariance estimation with smallest mean square error. Thus, the

dissimilarity values found in this study may be in accordance with the little disturbing permitted when fitting a reduced rank genetic (co)variance structure, with the gain of small sampling errors.

The angles between each first three eigenvectors from a PCM (or FAM) and the corresponding eigenvector from MTM were smaller than reported by Meyer (2007b) and similar to those presented by Meyer (2005). The higher divergences were for second and third eigenvector, showing that leading eigenvector was estimated in the same direction comparing MTM and alternative models. As a result of it, the correlation coefficients between these eigenvectors were 0.99, 0.34, and -0.30, for first, second and third eigenvector respectively (FAM5 vs MTM), which shows clearly that direction was changed in the third eigenvector.

When considered together, common, and specific factors, the eigenvectors become highly correlated (0.99, 0.99, and 0.98, respectively), showing that the directions of second and third eigenvectors were changed due to specific factors that are needed to explain the most variation. The referred scenario results in similar weights for each trait comparing FAM5 and MTM, which can benefit selection strategies for an aggregate genotype (BUZANSKAS et al., 2013), once the original nineteen traits can be resumed in three genetic linear combinations, retaining around 94% of all variation among all horses.

De Oliveira Bussiman et al. (2018), found similar genetic parameters (in terms of heritability estimates and genetic correlations) as in our analysis. However, in the present study, GtS presented a substantially higher heritability (0.16 vs 0.07). This can be related to model differences. Specifically, the GC used by these authors considered only year and season of birth, whereas in our analysis, GC were formed by year and season of birth, sex, and year of registry. Despite these differences in modelling, the signal of genetic correlation between GtS and other traits did not change, while the magnitude was changed.

According to Cabral et al. (2004), HW is one of the most important traits for Brazilian horse selection. In fact, this trait is the only MT appearing in the registry rule of ABCCCampolina with a minimum value, and all other MT are just controlled. In our analysis, HW had heritability of 0.43, which is similar to the results found by Solar Diaz et al. (2020) (0.47) in the same population. Solar Diaz et al. (2020), recently presented a multitrait analysis for fifteen MT; those authors found heritability estimates ranging from 0.20 (HG) to 0.47 (HW and HC), whereas in the present study heritability for same traits were 0.14, 0.43, and 0.40, respectively. These differences are explained by the number of records and contemporary groups differing from the present study.

Ablondi et al. (2020), studying MT in Bardigiano horses found higher heritability estimates for HW (0.63), HG (0.41), and CG (0.31) than found in this study, but for gait traits their estimates were from 0.09 to 0.12, similar to found in the present study. Vostry et al. (2017), found heritability of 0.14 for CL in Belgian warmblood, and Rustin et al. (2009), found heritability of 0.27 for NL in Czech draught horse; those values were similar to which was found in our analysis. For gait traits, Vostry et al. (2017) reported heritabilities from 0.05 to 0.25, and Rustin et al. (2009) reported estimates ranging between 0.33 and 0.52. Those results may reinforce the idea that the Campolina horse shows the same pattern of genetic parameters as some draught horses, which can be explained by its process of formation, and the use of breeds such as Percheron and Clydesdale to achieve the desired morphological type (PROCÓPIO; BERGMANN; COSTA, 2003).

The genetic correlations among all MT estimates were positive, which is in accordance with various studies (DE OLIVEIRA BUSSIMAN et al., 2018; GIONTELLA et al., 2020; SOLAR DIAZ et al., 2020; VICENTE et al., 2014b; VOSTRY et al., 2017). The magnitude of estimates presented in this study differ from those presented by Solar Diaz et al. (2020), which is attributed not only to the amount of data (and number of records), but also to the number of analyzed traits. Our analysis was fitted considering 19 traits while Solar Diaz et al. (2020) considered 15 (only MT). Positive genetic correlations between gait traits and MT were found by Molina et al. (1999) in Andalusian horses and by Vicente et al. (2014b) in Lusitano horses, however those authors used an objective way to measure gait, and in this study, it was used visual scores which are known to suffer from subjectivity (RUSTIN et al., 2009; VICENTE et al., 2014b, 2014a).

After analyzing gait scores, Rustin et al. (2009) and Vostry et al. (2017) also found negative genetic correlations between gait traits and MT. This may indicate limitations on the use of visual subjective scores on phenotyping horses for gait. Another disadvantage of visual scores is the influence of the appraiser/technician, which was 12.90% of phenotypic variation in this study. Velie, Hamilton, and Wade (2015) found also high values of human interference (rider) on phenotype in Australian Thoroughbred horses (32.41%). The estimated negative genetic correlation between GtS and all MT in the present study may be supported by Sousa et al. (2018). These authors found that the most frequent gait champions from the Campolina breed had shorter HW, which appears to disagree with the selection practiced by the Campolina breeders for higher HW (SOLAR DIAZ et al., 2020).

Rank correlations between EBVs from PCM (or FAM) with MTM were investigated in bovine (BIGNARDI et al., 2012, 2014; BOLIGON et al., 2013, 2015), and the authors showed

that fitting a PCM or a FAM did not substantially affect the ranking of the animals. For our analysis, rank correlations were also high, except for GtS, once this trait had the fewer number of records and higher proportion of specific variance; in addition, was negatively correlated to the other traits, which certainly influenced the EBV, and consequently its rank, causing more re-ranking across all studied selection intensities. Another possibility is that sampling errors are increased in the (co)variance components related to GtS, which would lead to uncertainty on EBV estimation for GtS. FAM5 caused an increase in some sampling errors, but it must be kept in mind that those values were estimated using linear approximation of nonlinear functions, through large sample theory, which perfectly accommodates some errors (MEYER, 2007b).

4.5. Conclusion

Results of this study shows that considering five common factors in a Factor Analytical model is sufficient to model the additive genetic (co)variance of the nineteen studied traits in Campolina horses. Due to the presence of some weak correlations, principal components models were not capable of reducing the data dimensionality. Therefore, considering the appeal of multitrait models, we can recommend FAM for genetic evaluations of morphometric and gait traits on horses. Estimates of genetic parameters indicated enough additive genetic variation that could be explored for selection purposes on the studied traits, but the high proportion of phenotypic variance explained by technician on GtS, and its negative association with morphometry, suggests that the development of new phenotyping strategies for this trait could benefit selection. In addition, to reduce complications due to the high number of traits, we can suggest the use of the first three eigenvectors of additive genetic (co)variance matrix from FAM to compute aggregate genotypes for all traits and those be used as pseudo-phenotypes.

4.6. References

ABLONDI, M. et al. Genetic parameters estimation in an Italian horse native breed to support the conversion from agricultural uses to riding purposes. **Journal of Animal Breeding and Genetics**, v. 137, n. 2, p. 200–210, 2020.

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: INTERNATIONAL SYMPOSIUM ON INFORMATION THEORY, 2. **Book of Abstracts...** Tsahkadsor, Armenia: Akadémiai Kiadó, 1973, p. 267–281.

BARTLETT, M. S. Tests of significance in factor analysis. **British Journal of Statistical**

Psychology, v. 3, n. 2, p. 77–85, 1950.

BIGNARDI, A. B. et al. Short communication: Principal components and factor analytic models for test-day milk yield in Brazilian Holstein cattle. **Journal of Dairy Science**, v. 95, n. 4, p. 2157–2164, 2012.

BIGNARDI, A. B. et al. Models for genetic evaluation of growth of Brazilian Bonsmara cattle. **Livestock Science**, v. 162, p. 50–58, 2014.

BOLIGON, A. A. et al. Herdabilidades e correlações entre pesos do nascimento à idade adulta em rebanhos da raça Nelore. **Revista Brasileira de Zootecnia**, v. 38, n. 12, p. 2320–2326, 2009.

BOLIGON, A. A. et al. Principal components and factor analytic models for birth to mature weights in Nelore cattle. **Livestock Science**, v. 152, n. 2–3, p. 135–142, 2013.

BOLIGON, A. A. et al. Reduced-rank models of growth and reproductive traits in Nelore cattle. **Theriogenology**, v. 83, n. 8, p. 1338–1343, 2015.

BURNHAM, K. P.; ANDERSON, D. R. Multimodel Inference: Understanding AIC and BIC in model selection. **Sociological Methods & Research**, v. 33, n. 2, p. 261–304, 2004.

BUZANSKAS, M. E. et al. Genetic parameter estimates and principal component analysis of breeding values of reproduction and growth traits in female Canchim cattle. **Reproduction, Fertility and Development**, v. 25, n. 5, p. 775, 2013.

CABRAL, G. C. et al. Avaliação morfométrica de equinos da raça Mangalarga Marchador: índices de conformação e proporções corporais. **Revista Brasileira de Zootecnia**, v. 33, n. 6 suppl 1, p. 1798–1805, 2004.

CATTELL, R. B. The Scree Test For The Number Of Factors. **Multivariate Behavioral Research**, v. 1, n. 2, p. 245–276, 1966.

DE OLIVEIRA BUSSIMAN, F. et al. Genetic analysis of morphological and functional traits in Campolina horses using Bayesian multi-trait model. **Livestock Science**, v. 216, n. April, p. 119–129, 2018.

GIONTELLA, A. et al. Genetic Parameters and Inbreeding Effect of Morphological Traits in Sardinian Anglo Arab Horse. **Animals**, v. 10, n. 5, p. 791, 2020.

HENDERSON, C. R. Rapid method for computing the inverse of a relationship matrix. **Journal of Dairy Science**, v. 58, n. 11, p. 1727–1730, 1975.

HILL, W. G. Genetic Correlation. In: MALOY, S.; HUGHES, K. (Eds.). **Brenner's Encyclopedia of Genetics**. Second ed. San Diego, US: Elsevier, 2013. p. 237–239.

JOHNSON, D. L.; THOMPSON, R. Restricted Maximum Likelihood Estimation of Variance Components for Univariate Animal Models Using Sparse Matrix Techniques and Average Information. **Journal of Dairy Science**, v. 78, n. 2, p. 449–456, 1995.

KAISER, H. F. The Application of Electronic Computers to Factor Analysis. **Educational and Psychological Measurement**, v. 20, n. 1, p. 141–151, 1960.

KIRKPATRICK, M.; MEYER, K. Direct estimation of genetic principal components: Simplified analysis of complex phenotypes. **Genetics**, v. 168, n. 4, p. 2295–2306, 2004.

LATTIN, J. M.; CARROLL, J. D.; GREEN, P. E. **Analyzing multivariate data**. 1. ed. São Paulo: Cengage Learning, 2011.

MANLY, B. J. F.; NAVARRO ALBERTO, J. A. **Multivariate Statistical Methods: A primer**. 4. ed. Boca Raton: Chapman & Hall/CRC, 2016.

MEYER, K. Estimates of genetic parameters for mature weight of Australian beef cows and its relationship to early growth and skeletal measures. **Livestock Production Science**, v. 44, n. 2, p. 125–137, 1995.

MEYER, K. An “average information” restricted maximum likelihood algorithm for estimating reduced rank genetic covariance matrices or covariance functions for animal models with equal design matrices. **Genetics Selection Evolution**, v. 29, n. 2, p. 97, 1997.

MEYER, K. Genetic principal components for live ultrasound scan traits of Angus cattle. **Animal Science**, v. 81, n. 3, p. 337–345, 2005.

MEYER, K. WOMBAT—A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). **Journal of Zhejiang University SCIENCE B**, v. 8, n. 11, p. 815–821, 2007a.

MEYER, K. Multivariate analyses of carcass traits for Angus cattle fitting reduced rank and factor analytic models. **Journal of Animal Breeding and Genetics**, v. 124, n. 2, p. 50–64, 2007b.

MEYER, K. Factor-analytic models for genotype $\times$ environment type problems and structured covariance matrices. **Genetics Selection Evolution**, v. 41, n. 1, p. 21, 2009.

MEYER, K.; KIRKPATRICK, M. Restricted maximum likelihood estimation of genetic principal components and smoothed covariance matrices. **Genetics Selection Evolution**, v. 37, n. 1, p. 1–30, 2005a.

MEYER, K.; KIRKPATRICK, M. Up hill, down dale: quantitative genetics of curvaceous traits. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 360, n. 1459, p. 1443–1455, 2005b.

MEYER, K.; KIRKPATRICK, M. Perils of Parsimony: Properties of Reduced-Rank Estimates of Genetic Covariance Matrices. **Genetics**, v. 180, n. 2, p. 1153–1166, 2008.

MOLINA, A. et al. Genetic parameters of morphofunctional traits in Andalusian horse. **Livestock Production Science**, v. 60, n. 2–3, p. 295–303, 1999.

NASCIMENTO, G. B. et al. Genetic parameter estimates and principal component analysis on performance and carcass traits of a terminal pig sire line. **Acta Agriculturae Scandinavica, Section A — Animal Science**, v. 64, n. 3, p. 137–144, 2014.

POLLAK, E. J.; VAN DER WERF, J.; QUAAS, R. L. Selection Bias and Multiple Trait Evaluation. **Journal of Dairy Science**, v. 67, n. 7, p. 1590–1595, 1984.

PROCÓPIO, A. M. M.; BERGMANN, J. A. G. A. G.; COSTA, M. D. D. Formação e demografia da raça Campolina. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 55, n. 3, p. 361–365, 2003.

R CORE TEAM. **R: A Language and Environment for Statistical Computing** (R. D. C. Team, Ed.) R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. 2020.

REZENDE, A. S. C.; MOURA, R. S. **Raças de equídeos no Brasil**. Belo Horizonte: Federal University of Minas Gerais, 2004.

RUSTIN, M. et al. Multi-trait animal model estimation of genetic parameters for linear type and gait traits in the Belgian warmblood horse. **Journal of Animal Breeding and Genetics**, v. 126, n. 5, p. 378–386, 2009.

SÁNCHEZ-GUERRERO, M. J. et al. Designing an early selection morphological linear traits index for dressage in the Pura Raza Español horse. **Animal**, v. 11, n. 6, p. 948–957, 2017.

SCHAEFFER, L. R. Sire and Cow Evaluation Under Multiple Trait Models. **Journal of Dairy Science**, v. 67, n. 7, p. 1567–1580, 1984.

SCHAEFFER, L. R. Estimation of Variances and Covariances Within the Allowable Parameter Space. **Journal of Dairy Science**, v. 69, n. 1, p. 187–194, 1986.

SCHWARZ, G. Estimating the Dimension of a Model. **The Annals of Statistics**, v. 6, n. 2, p. 461–464, 1978.

SMITH, A.; CULLIS, B.; THOMPSON, R. Analyzing Variety by Environment Data Using Multiplicative Mixed Models and Adjustments for Spatial Field Trend. **Biometrics**, v. 57, n. 4, p. 1138–1147, 2001.

SOLAR DIAZ, I. D. P. et al. Principal components for morphometric traits in Campolina horses. **Journal of Animal Breeding and Genetics**, v. 00, p. 1–9, 2 nov. 2020.

SOUSA, A. S. DE et al. How is the Morphometry of Stallions and Mares Show-Winning and Nonwinning Campolina Brazilian Breed with Batida and Picada Gaits? **Journal of Equine Veterinary Science**, v. 64, p. 34–40, 2018.

THOMPSON, B. **Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications**. 1. ed. Washington, DC: American Psychological Association, 2004.

THOMPSON, R. et al. A Sparse Implementation of the Average Information Algorithm for Factor Analytic and Reduced Rank Variance Models. **Australian New Zealand Journal of Statistics**, v. 45, n. 4, p. 445–459, 2003.

THOMPSON, R.; MEYER, K. A review of theoretical aspects in the estimation of breeding values for multi-trait selection. **Livestock Production Science**, v. 15, n. 4, p. 299–313, 1986.

TYRISEVÄ, A.-M. et al. Principal component approach in variance component estimation for international sire evaluation. **Genetics Selection Evolution**, v. 43, n. 1, p. 21, 2011.

VAN DER WERF, J. H. J.; VAN ARENDONK, J. A. M.; DE VRIES, A. G. Improving selection of pigs using correlated characters. In: 43th ANNUAL MEETING OF EUROPEAN FEDERATION OF ANIMAL SCIENCE, 43. **Proceedings...** Madrid: Wageningen Academic Publishers, 1992.

VARGAS, G. et al. Unravelling biological biotypes for growth, visual score and reproductive traits in Nelore cattle via principal component analysis. **Livestock Science**, v. 217, p. 37–43, 2018.

VELIE, B. D.; HAMILTON, N. A.; WADE, C. M. Heritability of racing performance in the Australian Thoroughbred racing population. **Animal Genetics**, v. 46, n. 1, p. 23–29, 2015.

VICENTE, A. A. et al. Selection for morphology, gaits and functional traits in Lusitano horses: I. Genetic parameter estimates. **Livestock Science**, v. 164, n. 1, p. 1–12, 2014a.

VICENTE, A. A. et al. Selection for morphology, gaits and functional traits in Lusitano horses: II. Fixed effects, genetic trends and selection in retrospect. **Livestock Science**, v. 164, n. 1, p. 13–25, 2014b.

VIEIRA, E. R. et al. Caracterização da equideocultura no estado de Minas Gerais. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 67, n. 1, p. 319–323, 2015.

VOSTRY, L. et al. Genetic Parameters for Linear Type Traits in Three Czech Draught Horse Breeds. **Agriculturae Conspectus Scientificus**, v. 82, n. 2, p. 111–115, 2017.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. 2. ed. New York: Springer International Publishing, 2016.

WOLFINGER, R. Covariance structure selection in general mixed models. **Communications in Statistics - Simulation and Computation**, v. 22, n. 4, p. 1079–1106, 1993.

WRIGHT, S. Coefficients of Inbreeding and Relationship. **The American Naturalist**, v. 56, n. 645, p. 330–338, 1922.

5. CONFIRMATORY FACTOR ANALYSIS AND STRUCTURAL EQUATION MODELS TO DISSECT THE RELATIONSHIP BETWEEN GAIT AND MORPHOLOGY IN HORSES

*“Science may set limits to knowledge, but should not set limits to imagination”
(Bertrand Russell)¹³.*

F. de Oliveira Bussiman¹; F.F. Silva (*in memoriam*)²; R.S.B. Carvalho³; R.V. Ventura¹; E.C. Mattos⁴; J.B.S. Ferraz⁴; J.P. Eler⁴; J.C.C. Balieiro¹

¹*Bioinformatic and Animal Breeding Lab., Department of Animal Nutrition and Production, College of Veterinary Medicine and Animal Science University of São Paulo (BIOMA-VNP/FMVZ-USP), Av. Duque de Caxias Norte, 225, 13.635-900, Pirassununga-SP, Brazil;*

²*Department of Animal Science, Federal University of Viçosa (DZO/UFV), Av. P H Rolfs, s/n, 36.750-900, Viçosa-MG, Brazil;*

³*Department of Basic Sciences, College of Animal Science and Food Engineering, University of São Paulo (ZAB/FZEA-USP), Av. Duque de Caxias Norte, 225, 13.635-900, Pirassununga-SP, Brazil;*

⁴*Group of Animal Breeding and Biotechnology, Department of Veterinary Medicine, College of Animal Science and Food Engineering, University of São Paulo (GMAB-ZMV/FZEA-USP), Av. Duque de Caxias Norte, 225, 13.635-900, Pirassununga-SP, Brazil.*

RESUMO: É bem sabido que andamento e morfologia são relacionados. Diversos estudos foram feitos ao longo dos anos mostrando a correlação genética entre eles em cavalos. Mas, nenhum pôde revelar as possíveis ligações causais entre essas características nesta espécie. O objetivo deste estudo foi aplicar inferência causal para dissecar as possíveis ligações causais entre morfologia e andamento em cavalos Campolina. Foram utilizados 49.994 cavalos com dados para 28 características. Todos os dados fenotípicos foram ajustados para os efeitos fixos (grupo contemporâneo, rebanho e idade), bem como para os efeitos aleatórios (técnico) exceto o efeito genético aditivo de animal. Todas as informações perdidas foram removidas e o arquivo de dados final ficou com 640 animais com observações ajustadas para todas as características. Uma análise confirmatória de fatores (CFA) foi conduzida para construir um modelo de mensuração para três variáveis latentes: Morfologia (M), Qualidade Morfológica (Q) e Locomoção (L). Aqui foi assumido que morfologia e andamento não podem ser diretamente mensurados, mas apenas aspectos morfológicos ou de andamento. Um modelo animal multicaracterística foi utilizado para se obterem os valores genéticos (EBV) para essas três variáveis latentes, e então os escores dos fatores da CFA foram ajustados para os EBVs. Os escores ajustados foram usados como *input* em modelos de equações estruturais para investigar todos as possíveis ligações causais entre essas características latentes. O modelo de mensuração mostrou ajuste aceitável aos dados e as diferenças entre as matrizes de (co)variância amostral e implícita pelo modelo foram baixas (0,01). L teve baixa herdabilidade (0,03) enquanto M e Q tiveram estimativas moderada (0,21 e 0,36, respectivamente). Todas as correlações genéticas foram negativas, mas as entre Q e as demais foram favoráveis à seleção. Existe um efeito causal direto de Q em M (-0,30) e um indireto sobre L (-0,10) e foi encontrado um efeito causal de M sobre L (0,35). Conhecer a rede causal entre as características ajudará criadores a tomar

¹³ RUSSELL, B. **History of western Philosophy**. 1. ed. London: George Allen and Unwin Ltda, 1946.

decisões de seleção e talvez seja importante para o desenvolvimento de estratégias de melhoramento em cavalos.

Palavras-chave: Inferência causal, análise confirmatória de fatores, modelos de equações estruturais, variáveis latentes, genética equina.

ABSTRACT: It is well known that locomotion and morphology are related. Several studies have been performed over the years showing the genetic correlation between them in horses. However, none could reveal the possible causative link between those traits in this species. The aim of this study was to apply statistical causal inference to dissect the possible causative links between morphology and locomotion traits in Campolina horses. It was used 49,994 horses with records for 28 traits. All phenotypic records were adjusted for fixed (contemporary group, stud, and age) and random (technician) effects except the animal genetic effect. All missing observations were removed, and the final dataset was left with 640 adjusted records for all traits. A confirmatory factor analysis (CFA) was conducted to build a measurement model for three latent traits: Morphology (M), Morphological quality (Q) and Locomotion traits (L). Here it was assumed that morphology and locomotion cannot be directly measured, but only morphological and locomotion aspects. A multitrait animal model was performed to obtain breeding values (EBV) for the three latent traits, and factor scores from CFA were adjusted for the EBV. The adjusted factor scores were used as input in structural equation models to investigate all possible causative links among latent traits. The measurement model showed acceptable fit to the data, and the differences between sample and model-implied (co)variance matrices were low (0.01). L had low heritability estimate (0.03) whereas M and Q had moderate estimates (0.21 and 0.36, respectively). All genetic correlations were negative, but those between Q and the others were favorable for selection. There is a causative direct effect of Q over M (-0.30) and an indirect over L (-0.10). It was found a causative effect of M on L (0.35). Knowing causal paths among traits will help breeders to take selection decisions and may have important implications for strategies in horse improvement.

Keywords: Causal inference, confirmatory factor analysis, structural equation models, latent traits, horse genetics.

5.1. Introduction

Several studies have been carried out in order to identify the genetic relationship between locomotion traits and morphology in Horses (for example: MEDEIROS et al., 2019; RUSTIN et al., 2009; SCHÖPKE; WENSCH-DORENDORF; SWALVE, 2013; VICENTE et al., 2014; VOSTRY et al., 2017). Because morphology has been used for selection, breed identification, and as indicator for future performance ability; it is generally acknowledged that body conformation can influence locomotion traits of the horse (STAIGER et al., 2016).

Comparing gaited and non-gaited horses, Staiger et al. (2016) found that gaited animals expose specific morphological features. However, no morphological differences were found by comparing horses with two different gait patterns in two Brazilian gaited horse breeds

(SANTOS et al., 2018). This kind of work was based on correlation analysis, but “correlation does not imply causation” (PEARL, 2009), and thus, other types of inference are needed to assess possible causative links between locomotion traits and morphology.

It can be assumed that locomotion traits and morphological measurements are the observable result of complex unmeasurable phenomena, called locomotion and morphology, respectively. To deal with traits that are not directly measurable, Confirmatory Factor Analysis (CFA) could be a powerful tool (BROWN, 2015). Under this framework, traits are used to build a measurement model (MM) for latent variables (defined as factors in common factor analysis) (BROWN, 2015; HU; BENTLER, 1999; SCHUMACKER; LOMAX, 2015).

Inferring causal structures, from observational data, is a difficult task because of the presence of potential confounders among variables, however it may be the ultimate goal in studies involving complex traits (PEÑAGARICANO et al., 2015). For this kind of investigation, Structural Equation Models (SEM), which are traced back to Wright (1921, 1918) and Haavelmo (1943), have been successfully applied in many areas, especially social sciences, and psychology, among others, such as economy and epidemiology.

SEMs were first adapted to quantitative genetics and mixed models by Gianola and Sorensen (2004), and after they were applied in several cases in plant and animal breeding (for example: DE LOS CAMPOS et al., 2006; DE MATURANA et al., 2009; INOUE et al., 2016; MOMEN et al., 2018, 2019; PEÑAGARICANO et al., 2015; VALENTE et al., 2010, 2011, 2013; WU et al., 2007). Rosa et al. (2011), presented a deep review about SEMs, and Valente and Rosa (2013), wrote an extensive tutorial on fitting SEMs.

Peñagaricano et al. (2015), proposed a general framework to assess causative links among complex latent traits. This methodology is based on three main steps: first, a CFA is carried out to build a MM for latent variables; second, predicted factor scores are used as pseudo-phenotypes to estimate (co)variance components and to predict breeding values (EBV); and third, scores are adjusted for the EBVs, and the adjusted factor scores are used as input in a SEM to investigate possible causative links among them.

Given that, the aim of the present study was to apply confirmatory factor analysis for morphological and locomotion traits in Campolina horses, to reduce the number of variables and to prospect the possible causative links among locomotion and morphology on this horse breed.

5.2. Material and Methods

5.2.1. Dataset and phenotype adjustment

For this study, the complete dataset of the Campolina Horse Breeders Association (ABCCCampolina, <http://www.campolina.org.br/portal/>) was used. It was comprised of 49,994 horses born between 1950 and 2013, with records for 28 traits, which were as follows:

- **Morphological measurements:** Height at Withers (HW); Height at Back (HB); Height at Croup (HC); Height at Chest (HCh); Leg Length (LL); Head Width (HWi); Chest Width (CW); Buttock Width (BW); Head Length (HL); Neck Length (NL); Back-Loins Length (BLL); Croup Length (CL); Shoulder Length (SL); Body Length (BL); Heart Girth (HG); and Cannon Girth (CG).
- **Locomotion visual scores:** Dissociation (Di); Style (S); Regularity (R); Development (De); Comfort (C); and Gait Total Scores (GtS; i.e., $GtS = C + S + R + De + Di$). For a better description on how these traits are measured please see de Oliveira Bussiman et al., 2018).
- **Conformation harmony traits:** Ratio between HW and HC (HWHC, calculated as $HWHC = HW/HC$); Ratio between BLL and BL (BLLBL; i.e., $BLLBL = BLL/BL$); Difference between HW and HC (dHWHC, obtained by $dHWHC = |HW - HC|$); Difference between HL and NL (HLNL, calculated as $HLNL = |HL - NL|$); Difference between HL and SL (HLSL; i.e., $HLSL = |HL - SL|$); and the overall Harmony (Harm), which was treated as binary trait, that is: 1 if $HW = HC = BL$, and 0 otherwise.

For all traits, Contemporary Groups (GC) were made by the concatenation of the following information: Year of Birth (64 levels – between 1950 and 2013), Season of Birth (4 levels – summer, autumn, winter, and spring), Sex (2 levels – male and female), and Year of Registry (64 levels – between 1953 and 2016). All GC with less than five observations (or with no phenotypic variation) were removed from the dataset, as well as any phenotypic record deviating 3 times above or below phenotypic mean. Additionally, all Studs with less than five observations (or without phenotypic variation) were removed from the dataset. Descriptive statistics of the final dataset is presented in Table 1, and the correlation circle plot (raw coefficients) is presented in Figure 1 (the correlation table is on Figure 2).

Table 1 - Descriptive statistics of the traits in Campolina breed used in this study

Trait	N	$\bar{X}$	SD	Min.	Max.	GC	Stud	F	M
HW (cm)	42,880	153.84	5.80	137.00	170.00	1,369	1,675	35,279	7,601
HB (cm)	44,045	146.56	5.56	130.00	162.00	1,426	1,731	36,300	7,745
HC (cm)	43,089	153.67	5.60	137.00	170.00	1,381	1,702	35,429	7,660
HCh (cm)	45,839	64.61	2.99	56.00	73.00	1,588	1,769	37,920	7,919
LL (cm)	41,017	89.08	5.20	74.00	104.00	1,313	1,608	34,355	6,662
HWi (cm)	47,432	20.91	1.04	18.00	24.00	1,673	1,893	39,096	8,336
CW (cm)	44,533	38.83	2.96	30.00	47.00	1,508	1,719	36,578	7,955
BW (cm)	41,633	51.99	2.84	44.00	60.00	1,340	1,624	34,564	7,069
HL (cm)	42,093	60.61	2.84	53.00	68.00	1,298	1,684	34,578	7,515
NL (cm)	40,719	63.02	3.81	52.00	74.00	1,260	1,685	34,210	6,509
BLL (cm)	43,746	60.29	6.19	42.00	78.00	1,507	1,655	36,048	7,698
CL (cm)	39,846	51.17	3.88	40.00	62.00	1,224	1,630	32,868	6,978
SL (cm)	41,454	56.30	3.88	45.00	67.00	1,380	1,652	33,958	7,496
BL (cm)	45,670	156.26	6.75	136.00	176.00	1,510	1,823	37,775	7,895
HG (cm)	45,940	181.80	7.43	160.00	204.00	1,565	1,824	38,032	7,908
CG (cm)	43,210	18.96	1.09	16.00	22.00	1,413	1,728	35,310	7,900
GtS (vs ¹)	3,243	139.51	12.66	88.00	175.00	133	201	2,142	1,101
C (vs)	4,958	46.42	5.21	25.00	60.00	216	263	3,502	1,456
S (vs)	4,047	29.09	5.47	0.00	51.00	180	218	2,833	1,214
R (vs)	3,569	22.12	2.82	14.00	36.00	161	208	2,517	1,052
De (vs)	2,295	22.36	3.08	14.00	35.00	123	127	1,590	705
Di (vs)	3,799	29.43	4.00	17.00	50.00	165	203	2,627	1,172
HWHC (cm)	46,956	100.04	0.88	97.30	102.84	1,678	1,835	38,776	8,180
BLLBL (cm)	41,139	38.78	4.34	26.09	51.47	1,445	1,512	34,017	7,122
dHWHC (cm)	45,681	0.93	0.87	0.00	3.00	1,603	1,812	37,888	7,793
HLNL (cm)	40,335	2.96	2.08	0.00	9.00	1,341	1,618	33,635	6,700
HLSL (cm)	40,655	5.19	3.32	0.00	14.00	1,443	1,445	33,122	7,533
Harm [*]	37,375	0.10 ²	2.72 ³	0.00	1.00	1,036	961	32,013	5,362

Source: Bussiman (2021). N is the number of observations; $\bar{X}$ is the phenotypic average; SD is the phenotypic standard deviation; Min. is the minimum record value; Max. is the maximum record value; GC is the number of contemporary groups; Stud is the number of studs (herds); F is the number of females; M is the number of males. Abbreviations: HW, Height at Withers; HB, Height at Back; HC, Height at Croup; HCh, Height at Chest; LL, Leg Length; HWi, Head Width; CW, Chest Width; BW, Buttock Width; HL, Head Length; NL, Neck Length; BLL, Back-Loins Length; CL, Croup Length; SL, Shoulder Length; BL, Body Length; HG, Heart Girth; CG, Cannon Girth; GtS, Gait Total Scores; C, Comfort; S, Style; R, Regularity; De, Development; Di, Dissociation; HWHC, Ratio between HW and HC; BLLBL, Ratio between BLL and BL; dHWHC, difference between HW and HC; HLNL, difference between HL and NL; HLSL, difference between HL and SL; and Harm, overall harmony. ¹vs = Visual Score; ²This value is the proportion of 1 (p) in the Harm records; ³The SD was calculated as if Harm were a continuous gaussian variable; ^{*}Harm does not have a unity of measurement as it is presence (1) or absence (0).

Figure 1 - Correlation circle plot of the raw data used for further analysis

Source: Bussiman (2021). Abbreviations: Q = Morphological quality; M = Morphology; L = Locomotion traits; HW = Height at Withers; HB = Height at Back; HC = Height at Croup; HCh = Height at Chest; LL = Leg Length; HWi = Head Width; CW = Chest Width; BW = Buttock Width; HL = Head Length; NL = Neck Length; BLL = Back-Loins Length; CL = Croup Length; SL = Shoulder Length; BL = Body Length; HG = Heart Girth; CG = Cannon Girth; Di = Dissociation; S = Style; R = Regularity; De = Development; C = Comfort; GtS = Gait Total Scores; HWHC = Ratio between HW and HC; BLLBL = Ratio between BLL and BL; dHWHC = Difference between HW and HC; HLNL = Difference between HL and NL; HLST = Difference between HL and SL; and Harm = Overall Harmony.

Figure 2 - Raw correlation coefficients among all studied traits in the used dataset

Source: Bussiman (2021). Abbreviations: HW = Height at Withers; HB = Height at Back; HC = Height at Croup; HCh = Height at Chest; LL = Leg Length; HWi = Head Width; CW = Chest Width; BW = Buttock Width; HL = Head Length; NL = Neck Length; BLL = Back-Loins Length; CL = Croup Length; SL = Shoulder Length; BL = Body Length; HG = Heart Girth; CG = Cannon Girth; Di = Dissociation; S = Style; R = Regularity; De = Development; C = Comfort; GtS = Gait Total Scores; HWHC = Ratio between HW and HC; BLLBL = Ratio between BLL and BL; dHWHC = Difference between HW and HC; HLNL = Difference between HL and NL; HLNL = Difference between HL and SL; and Harm = Overall Harmony.

All phenotypes were adjusted for fixed and random effects, except the additive genetic effect of animal, by ordinary least squares through the “*lm*” (or “*glm*” for Harm – with a binomial distribution and probit link function) function of the *R* software (R CORE TEAM, 2016), according to the following models:

$$y_{ijk} = \mu + \beta_1 \text{Age}_k + \beta_2 \text{Age}_k^2 + \text{GC}_i + \text{Stud}_j + \varepsilon_{ijk}, \quad (5.1)$$

$$y_{ijkl} = \mu + \beta_1 \text{Age}_k + \beta_2 \text{Age}_k^2 + \text{GC}_i + \text{Stud}_j + t_l + \varepsilon_{ijkl}, \quad (5.2)$$

Where: y_{ijk} is the vector of phenotypic records (Morphological measurements and Conformation harmony traits), and y_{ijkl} is the vector of phenotypic records for locomotion visual scores; μ is just a constant; β_1 and β_2 are the regressions coefficients of the covariate age at measurement (linear and quadratic, respectively) of the k^{th} animal; GC_i is the fixed effect of the i^{th} GC; Stud_j is the fixed effect of the j^{th} Stud of birth (same as herd of birth); t_l is the uncorrelated random effect of the l^{th} technician (only for locomotion visual scores); ε_{ijk} and ε_{ijkl} are the random residual terms (not true errors), which were expected to be: $\varepsilon_{ijk} = u_k + e_{ijk}$ and $\varepsilon_{ijkl} = u_k + e_{ijkl}$; where: u_k is the animal EBV; and e_{ijk} and e_{ijkl} are the true residual terms (SEARLE, 1997). After adjustment, all missing observations were removed, thus the final dataset was left with 640 animals with adjusted phenotypes for all studied traits. The significance level (P-values) of the included effects (fixed and random) is summarized in Table 2.

Table 2 - P-value of effects included in the statistical models for all studied traits.

Effect	P-value distribution				
	Max.	Min.	Mean	Median	SD
Contemporary Group	0.027	0.000	0.002	0.000	0.006
Stud of birth	0.000	0.000	0.000	0.000	0.000
Age at Measurement (linear effect)	0.045	0.000	0.013	0.008	0.014
Age at Measurement (quadratic effect)	0.048	0.000	0.011	0.008	0.012
Technician random effect ¹	0.000	0.000	0.000	0.000	0.000

¹This effect was only included for locomotion visual scores, and the P-value reflects the probability of σ_t^2 being equal to zero.

5.2.2. Confirmatory factor analysis

Three latent variables were used in this study, and they were defined based on biological meaning and economic importance. Morphology (M) was used for modeling morphological measurements; Morphological quality (Q) was related to conformation harmony traits. Finally, Locomotion traits (L) was used to account for locomotion visual scores. Those latent traits were simultaneously evaluated through the following measurement model (MM):

$$\varepsilon = \Lambda \xi + \delta, \quad (5.3)$$

Where: $\boldsymbol{\varepsilon}$ represent the vector (28 x 1) of adjusted phenotypes; $\boldsymbol{\Lambda}$ is the matrix (28 x 3) of factor loadings relating observed variables to the latent variables; $\boldsymbol{\xi}$ represent the vector (3 x 1) of latent variables; and $\boldsymbol{\delta}$ represent the vector (28 x 1) of residual terms. It was assumed that: $E[\boldsymbol{\delta}] = \mathbf{0}$; $V[\boldsymbol{\delta}] = \boldsymbol{\theta}$; $E[\boldsymbol{\xi}] = \mathbf{0}$; and $V[\boldsymbol{\xi}] = \boldsymbol{\Phi}$. It must be noticed that here $\boldsymbol{\theta}$ was not a diagonal matrix; the covariances between error terms were freely estimated when considered appropriate.

The MM was fitted using maximum likelihood estimation with robust standard error and a mean-variance-adjusted test statistic (SATORRA; BENTLER, 1994). The overall fit was assessed using the standardized root-mean-square residual (SRMR - BENTLER, 1995); root-mean-square of approximation (RMSEA - STEIGER, 1990); Tucker-Lewis index (TLI - TUCKER; LEWIS, 1973); comparative fit index (CFI - BENTLER, 1990); normed fit index (NFI - BENTLER; BONETT, 1980); goodness-of-fit index (GFI - HU; BENTLER, 1999) and the adjusted goodness-of-fit index (AGFI - BENTLER, 1983). In addition, it was used the ratio χ^2/df (WHEATON et al., 1977); where: χ^2 is the adjusted observed chi-square value; df are the degrees of freedom. For all goodness-of-fit indicators, it was assumed cutoff values suggested by Hu and Bentler (1999).

Factor scores (FS) for latent variables were predicted as follows (PEÑAGARICANO et al., 2015):

$$E[\boldsymbol{\xi} | \boldsymbol{\varepsilon}] = \hat{\boldsymbol{\xi}} = \boldsymbol{\Phi} \boldsymbol{\Lambda}^T (\boldsymbol{\Lambda} \boldsymbol{\Phi} \boldsymbol{\Lambda}^T + \boldsymbol{\theta})^{-1} (\boldsymbol{\varepsilon} - \boldsymbol{\mu}), \quad (5.4)$$

Where: $E[\boldsymbol{\xi} | \boldsymbol{\varepsilon}]$ is the expectation of the factor scores for latent variables ($\boldsymbol{\xi}$) given the adjusted phenotypes ($\boldsymbol{\varepsilon}$); $\hat{\boldsymbol{\xi}}$ are the estimated factor scores; $\boldsymbol{\Phi}$ is the variance-covariance matrix for latent variables; $\boldsymbol{\Lambda}$ is the same as defined in (5.3); $\boldsymbol{\theta}$ is the model-implied (co)variance matrix for adjusted phenotypes; and $\boldsymbol{\mu}$ is the vector of means for adjusted phenotypes. Those procedures (CFA and FS estimation) were conducted through the “*lavaan*” R package (ROSSEEL, 2012); and for graphical representation of the MM, the “*semPaths*” function from “*semPlot*” R package (EPSKAMP, 2015) was used. Additionally, the overall difference between sample ($\mathbf{S}$) and model-implied ($\boldsymbol{\Sigma}$) (co)variance matrix was estimated by applying the Schaeffer method ($\|\boldsymbol{\Sigma} - \mathbf{S}\| / \|\mathbf{S}\|$ - SCHAEFFER, 1986).

5.2.3. Multitrait mixed model

FS were used as pseudo-phenotypes on the following multitrait animal model (MTM):

$$\begin{bmatrix} \hat{\xi}_1 \\ \hat{\xi}_2 \\ \hat{\xi}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{1}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_3 \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_3 \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_3 \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{bmatrix}, \quad (5.5)$$

Where: $\hat{\xi}_i$ are the same as defined in (5.4); $\mathbf{1}_i$ are vectors of ones to assign the location parameters $\boldsymbol{\mu}_i$ (overall mean); $\mathbf{Z}_i$ are the incidence matrices for the random additive animal genetic effect $\mathbf{u}_i$; and $\mathbf{e}_i$ is the vector of random residual terms.

It was assumed that $E[\hat{\xi}] = \boldsymbol{\mu}$; $E[\mathbf{u}] = E[\mathbf{e}] = 0$; and the (co)variance matrix for the random effects was as follows:

$$V \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \\ \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{bmatrix} = V \begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{G}_0 \otimes \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_0 \otimes \mathbf{I} \end{bmatrix}, \quad (5.6)$$

Where: $\mathbf{A}$ is the numerator relationship matrix (3,665 x 3,665); $\mathbf{G}_0$ is the additive genetic (co)variance matrix (3 x 3) among latent variables; $\mathbf{R}_0$ is the residual (co)variance matrix (3 x 3) among latent variables; and $\mathbf{I}$ is an identity matrix (with proper order).

This model was implemented using Average Information Restricted Maximum Likelihood (AIREML – JOHNSON; THOMPSON, 1995) estimation at AIREMLF90 from BLUPF90 software family (MISZTAL et al., 2002); and as starting values was used $\Phi/2$, for both $\mathbf{G}_0$ and $\mathbf{R}_0$. After variance components estimation, the EBVs were used to correct FS for possible genetic confounders (PEÑAGARICANO et al., 2015; VALENTE et al., 2010) as: $\mathbf{w} = \hat{\xi} - \hat{\mathbf{u}}$; where: $\mathbf{w}$ is the vector of adjusted FS; $\hat{\xi}$ is the same as defined in (5.4); and $\hat{\mathbf{u}}$ is the vector of predicted breeding values.

5.2.4. Structural equation model

After adjusting the FS, all possible causative links among latent variables were assessed by the following SEM:

$$\mathbf{w} = \mathbf{B}\mathbf{w} + \mathbf{\Gamma}\mathbf{v} + \boldsymbol{\zeta}, \quad (5.7)$$

Where: $\mathbf{w}$ and $\mathbf{v}$ are vectors of latent variable adjusted scores; $\mathbf{B}$ and $\mathbf{\Gamma}$ are path (causal) coefficients between latent variables; and $\boldsymbol{\zeta}$ is a vector of disturbance terms. It was assumed that $\text{Cov}(\mathbf{v}, \boldsymbol{\zeta}) = 0$. The SEM was fitted using maximum likelihood estimation with Satterthwaite correction (mean-variance-adjusted test). The overall fit was evaluated using the same goodness-of-fit criteria as for CFA. For model comparison, it was applied the Akaike information criterion (AIC - AKAIKE, 1973). These analyses were also conducted using the “*lavaan*” R package (ROSSEEL, 2012), and for graphical visualization of the final model the “*semPaths*” function of the “*semPlot*” R package (EPSKAMP, 2015) was used. Given the number of variable (q) included in the SEM, and the number of estimated parameters (p), the SEM was also evaluated based on df , as follows:

$$df = [q(q+1)/2] - p, \quad (5.8)$$

When: $df \geq 1$ the model was supposed to be overidentified (SCHUMACKER; LOMAX, 2015). To do some kind of validation, the adjusted factor scores ($\mathbf{w}$) were submitted to the Grow-Shrink constraint-based structure learning algorithm (MARGARITIS, 2003) to verify if the learned model was similar to the final SEM.

5.3. Results

5.3.1. Measurement model

The factor loadings were, in general, significant except for HLNL, Harm, dHWHC, and HWHC (Table 3), however the MM showed acceptable fit to the data according to the adopted cutoff criteria for goodness of fit (HU; BENTLER, 1999). Overall fit indicators were: $\chi^2_{(331, N=640)} = 885,05$; $p < 0.01$; GFI = 0.95; AGFI = 0.94; CFI = 0.95; TLI = 0.95; NFI = 0.93; SRMR = 0.06; RMSEA = 0.05; and additionally, the ratio χ^2/df was 2.67. Based on the formula to obtain the chi-square (i.e., $(N-1)F_{ML}$; where: N is the number of observations; and F_{ML} is the objective function of the maximum likelihood estimator - BROWNE, 1984), the

maximum value N to achieve a P -value ≥ 0.05 was estimated (260, Figure 3). In addition, the overall difference between Σ and S was 0.01. The MM is presented in Figure 4.

Table 3 - Standardized parameter estimates from measurement model

Trait	Estimate ¹	SE ²	Z-value	P-value
Locomotion (L)				
C (vs)	0.69	0.04	16.69	0.000
S (vs)	0.71	0.05	13.61	0.000
R (vs)	0.72	0.04	19.88	0.000
De (vs)	0.64	0.04	17.34	0.000
Di (vs)	0.81	0.04	18.74	0.000
GtS (vs)	1.08	0.03	32.04	0.000
Morphology (M)				
HW (cm)	0.77	0.04	20.79	0.000
HB (cm)	0.72	0.04	18.70	0.000
HC (cm)	0.78	0.04	21.33	0.000
HCh (cm)	0.52	0.04	12.90	0.000
LL (cm)	0.43	0.04	9.99	0.000
HWi (cm)	0.36	0.04	8.64	0.000
CW (cm)	0.60	0.04	16.10	0.000
BW (cm)	0.62	0.04	16.25	0.000
HL (cm)	0.70	0.03	20.48	0.000
NL (cm)	0.66	0.04	18.25	0.000
CL (cm)	0.44	0.04	11.84	0.000
SL (cm)	0.54	0.04	13.98	0.000
BL (cm)	0.85	0.04	24.68	0.000
HG (cm)	0.59	0.04	14.02	0.000
CG (cm)	0.45	0.04	10.70	0.000
BLL (cm)	0.24	0.04	5.61	0.000
Morphological Quality (Q)				
HLNL (cm)	0.04	0.03	1.63	0.103
HLSL (cm)	0.15	0.05	2.77	0.006
Harm *	-0.02	0.03	-0.61	0.543
BLLBL (cm)	-0.35	0.10	-3.63	0.000
dHWHC (cm)	0.06	0.03	1.76	0.078
HWHC (cm)	-0.03	0.04	-0.65	0.519

Source: Bussiman (2021). Abbreviations: HW, Height at Withers; HB, Height at Back; HC, Height at Croup; HCh, Height at Chest; LL, Leg Length; HWi, Head Width; CW, Chest Width; BW, Buttock Width; HL, Head Length; NL, Neck Length; BLL, Back-Loins Length; CL, Croup Length; SL, Shoulder Length; BL, Body Length; HG, Heart Girth; CG, Cannon Girth; GtS, Gait Total Scores; C, Comfort; S, Style; R, Regularity; De, Development; Di, Dissociation; HWHC, Ratio between HW and HC; BLLBL, Ratio between BLL and BL; dHWHC, difference between HW and HC; HLNL, difference between HL and NL; HLSL, difference between HL and SL; and Harm, overall harmony. ¹Factor Loadings; ²Standard Error; *Harm does not have a unity of measurement as it is presence (1) or absence (0).

Figure 3 - P-value simulation to obtain maximum N to not reject null hypothesis

Source: Bussiman (2021).

Figure 4 - Measurement model from confirmatory factor analysis. Curved arrows between error terms represent possible confounders

Source: Bussiman (2021). Abbreviations: Q = Morphological quality; M = Morphology; L = Locomotion traits; HW = Height at Withers; HB = Height at Back; HC = Height at Croup; HCh = Height at Chest; LL = Leg Length; HWi = Head Width; CW = Chest Width; BW = Buttock Width; HL = Head Length; NL = Neck Length; BLL = Back-Loins Length; CL = Croup Length; SL = Shoulder Length; BL = Body Length; HG = Heart Girth; CG = Cannon Girth; Di = Dissociation; S = Style; R = Regularity; De = Development; C = Comfort; GtS = Gait Total Scores; HWHC = Ratio between HW and HC; BLLBL = Ratio between BLL and BL; dHWHC = Difference between HW and HC; HLNL = Difference between HL and NL; HLSL = Difference between HL and SL; and Harm = Overall Harmony.

5.3.2. Multitrait mixed model

Heritability estimates were from low to moderate for all latent variables, ranging from 0.03 (L) to 0.36 (Q) (Table 4). L presented a very low estimate for the additive genetic variance, whereas it was “acceptable” for M and Q. Genetic correlations among all latent variables ranged

from -0.67 to -0.23. The strongest additive genetic correlation was observed between Q and M (-0.67 ± 0.21). Residual correlations varied from -0.60 to 0.23 (Table 4) and were, in general, with low magnitude (except for the residual correlation between L and M; -0.60 ± 0.07). Additionally, to evaluate how the selection for each latent trait could affect the others, EBVs for L were plotted against EBVs for M (Figure 5), and they were colored by the “signal” of EBVs for Q (i.e., if EBVs for Q were negative or positive). The selection for positive EBVs for L and M implies in selecting only animals with negative EBVs for Q (Figure 5).

Table 4 - Genetic parameter estimates and standard errors (SE) from multitrait animal model. Heritability (h^2) estimates are presented on the diagonal; genetic correlations (r_g) above and residual correlations (r_e) below diagonal

Trait	Trait		
	Locomotion	Morphology	Morphological Quality
Locomotion	0.03 (0.01)	-0.58 (0.61)	-0.23 (0.41)
Morphology	0.23 (0.05)	0.21 (0.07)	-0.67 (0.21)
Morphological Quality	-0.11 (0.05)	-0.60 (0.07)	0.36 (0.07)

Source: Bussiman (2021). All parameters are presented as: parameter (SE).

Figure 5 - Selection criteria for latent traits in the Campolina population

Source: Bussiman (2021).

5.3.3. Structural equation model

L and M were positively correlated, whereas the correlations between L and Q, and Q and M, were negative (Table 5). The additive genetic covariances (Table 5) explained up to 30% of the phenotypic covariances between latent variables. The residual (co)variances among L, M and Q (Table 5) followed the same pattern of the phenotypic covariances. The final SEM (Figure 6) showed acceptable fit, that is: $\chi^2_{(1,N=640)} = 3,41$; $p = 0.07$; GFI = 0.99; AGFI = 0.97; CFI = 0.99; TLI = 0.95; NFI = 0.99; SRMR = 0.02; and RMSEA = 0.06. All other possible causal specifications (Figure 7) were also evaluated, but none presented AIC index smaller than the final model and acceptable goodness-of-fit indexes.

Table 5 - Model-implied, Additive genetic and residual variance-covariance matrices among latent traits and standard errors (SE)

Trait	Trait		
	Locomotion	Morphology	Morphological Quality
Phenotypic (Model-implied)			
Locomotion	1.12 (0.06)		
Morphology	0.18 (0.01)	1.13 (0.06)	
Morphological Quality	-0.30 (0.02)	-1.71 (0.10)	6.91 (0.39)
Additive genetic			
Locomotion	0.03 (0.01)		
Morphology	-0.05 (0.03)	0.24 (0.09)	
Morphological Quality	-0.06 (0.08)	-0.52 (0.21)	2.58 (0.76)
Residual			
Locomotion	1.09 (0.06)		
Morphology	0.23 (0.05)	0.91 (0.07)	
Morphological Quality	-0.25 (0.11)	-1.23 (0.15)	4.57 (0.59)

Source: Bussiman (2021). All parameters are presented as: parameter (SE).

Figure 6 - Structure equation model with the final causal structure among Morphology (M); Locomotion traits (L); and Morphological quality (Q). Arrows are weighted and colored according to the parameter estimates

Source: Bussiman (2021).

Figure 7 - Structural equation models tested for all possible causal structures

Source: Bussiman (2021).

The covariance between Q and L was not statistically different from zero, and thus it was removed from the final model. Q had a negative causal effect on M (Table 6), whereas M had a positive causal effect on L. The total causal effect of Q on L was also negative. All standardized parameters estimates were statistically different ($P < 0.001$) from zero (Table 6). The learned model (from the Grow-Shrink constraint-based Bayesian network algorithm) had the same directed structure as the final SEM (data not showed).

Table 6 - Standardized parameter estimates from final structural equation model

Path	Estimate	SE	Z-value	P-value
Q → M	-0.30	0.02	-19.06	0.000
M → L	0.35	0.05	7.75	0.000
Total effect of Q on L	-0.11	0.01	-7.18	0.000

Source: Bussiman (2021). Q = Morphological Quality; M = Morphology; and L = Locomotion.

5.4. Discussion

Selecting based on EBVs, and estimation of correlated response to selection (on MTM) does not require knowledge about the causal model, even when there are complex causal relationships among phenotypes (VALENTE et al., 2013). However, fitting MTM to predict EBVs can be disadvantageous when several traits are used, because of redundancy; computation (time and efficiency); and selection efficiency (which is difficult in presence of unfavorable genetic correlations).

Exploring the relationship among observed phenotypes, CFA can reduce the number of traits by defining a MM for latent traits (PEÑAGARICANO et al., 2015), assuming that observed variables are result of underlying traits. Here, the reduction of estimable (co)variance components (CC) was dramatic. By comparing the number of CCs using CFA, with the possible CCs for a full MTM (for all observed variables); the reduction was from 883 (406 additive genetic co-variances + 21 technician co-variances + 406 residual co-variances) to 12 (6 additive genetic co-variances + 6 residual co-variances).

The common statistic test for CFA is the chi-square, however it is inflated by the sample size (SCHUMACKER; LOMAX, 2015; WHEATON et al., 1977). Thus, even when the differences between sample and model-implied (co)variance matrices are negligible, the solutions are routinely rejected based on χ^2 (BROWN, 2015). This supports the results of our P-value simulation, which showed that the MM would not reject the null-hypothesis when $N \leq 260$. Here it was used 640 observations; thus, it was expected that the chi-square test would reject the null hypothesis. However, all the alternative fit indexes pointed to a good fit of the MM to the data. In addition, the overall differences between observed and model-implied (co)variance matrices were low, reflected by the low values of RMSEA (BROWN, 2015; SCHUMACKER; LOMAX, 2015) and Schaeffer criterium (SCHAEFFER, 1986).

Despite the reduced number of observations for CCs estimation ($N = 640$), the standard errors (SE) of heritability estimates were all close to zero (i.e., $SE \leq 0.07$), showing that this amount of data was enough for estimation. However, the SE for covariances/correlations were

high. Which was expected, once for covariance estimation, substantial amount of data is needed (HILL, 2013; ROBERTSON, 1959). The SE for genetic correlations could be also a result of the family structure in the used pedigree, and a very asymmetrical use of some stallions (increasing progeny number of specific individuals) (VISSCHER, 1998).

M and Q showed sufficient additive genetic variance for selection purposes, whereas the selection for L will be slow due to the amount of additive genetic variance reflected by the low heritability estimate (0.03). These results are in accordance with a previous study with the same population, in which was found that GtS has heritability of 0.07 and all morphological measurements had heritability estimates ranging from 0.11 to 0.43 (DE OLIVEIRA BUSSIMAN et al., 2018). It must be noticed that GtS had the highest factor loading on L, which could explain the similarity between heritability for these two traits (L and GtS).

Despite genetic correlations among latent traits with Q were all negative, these correlations were, in fact, favorable for selection, once the selection for smaller ratios and differences (dHWHC; HWHC; BLLBL; HLSL; and HLNL) will led to high EBVs for Harm (i.e., animals with more similarity are more harmonic), this reflects on the correlations between Q and other latent traits. Thus, the selection for low (or negative) EBVs for Q will led to high EBVs for L and M.

The genetic correlation between L and M was unfavorable for selection, once animals with high EBVs for locomotion traits will have low EBVs for morphology, but this can be solved by the actual ABCCCampolina “empirical selection process”, in which a morphological standard is fixed, and any animal with several deviation from it, is not used as Stallion or Mare. Thus, the selection for L can be made within the animals selected for M (which is in fact shown on Figure 5), like what as proposed by Ducro et al., (2007) and Holmström et al. (1990).

The negative genetic correlation between GtS and morphological measurements was attributed to the high effects of technician on GtS in Campolina horses (DE OLIVEIRA BUSSIMAN et al., 2018). These effects were also high over all other locomotion visual scores (data not being shown), thus the direction of genetic correlations between L and M can be a reflect of the technician effect over locomotion traits. Rustin et al. (2009), found evidence of heterogeneity of technician effect, which can reinforce the impact of this effect over EBV and genetic correlations. In addition, Vostry et al. (2017), also found negative genetic correlations between locomotion and morphological traits, even not including technician in the model of analysis.

Translation from causal hypothesis into statistical models is not perfect (PEÑAGARICANO et al., 2015), and this imperfection is due to the replacement of

asymmetrical relationships by symmetrical ones, involving probability distributions (SHIPLEY, 2003). As a consequence of that, different models can be statistically equivalent, by implying the same (co)variance matrices (PEÑAGARICANO et al., 2015). In practical terms, those models (that are equivalent) will present the same values of goodness-of-fit indexes (BROWN, 2015; SCHUMACKER; LOMAX, 2015).

Testing all possible causal structures led to find only one model fitting the data, thus no equivalent model can be discussed. The covariance between Q and L can be replaced by the indirect total effect of Q on L (-0.11 – Table 6). Despite genetic correlation between M and L was negative, it was found a positive direct effect of M on L (0.35) which is in accordance with previous studies that argue locomotion is associated with morphology (MEDEIROS et al., 2019; RUSTIN et al., 2009; SCHÖPKE; WENSCH-DORENDORF; SWALVE, 2013; STAIGER et al., 2016; VICENTE et al., 2014; VOSTRY et al., 2017).

The fact that the causal effect of M on L was positive and the genetic correlation between them was negative, can be explained by confounding genetic effects (VALENTE et al., 2010, 2013); that is, the genes controlling both traits may act in opposite direction, but when this effect was removed (by adjusting the FS for the EBVs), the causal effect of M on L was positive. This may indicate that considering both, genetic confounders and M causal effect, the resulting total effect of M on L is positive, which is a reflect of positive residual covariance between M and L.

Other possible explanation is that the SEM was fitted conditional do the genetic effects; thus, even the genetic covariance between M and L being negative, the causal effect of M on L, conditional on this covariance, is positive (PEARL, 1988, 2009; VALENTE et al., 2010). The negative covariance between Q and M can be substituted by the negative direct causal effect of Q on M (-0.30), which is in accordance with the selection goal for Q; i.e., selecting for low values of Q will led to high values of M.

Exploring genetic correlations, selection index theory (HAZEL, 1943; SMITH, 1936) was developed to combine different sources of information within a single trait (i.e., individual records and from its relatives) in an optimal weighting process, and also for phenotypic selection on multiple traits (HENDERSON, 1963; HENDERSON; QUAAS, 1976). The rationale under a selection index is quite simple: adjusted phenotypes (for correlated traits) are weighted by their relative economic importance to produce an aggregate genotype (HAZEL; DICKERSON; FREEMAN, 1994). However, the limitation is that often economical weights are unknow and difficult to be obtained.

The genetic gain would be slowed in the presence of unfavorable genetic correlations among the traits of interest, under the selection index method. To overcome this, multivariate techniques have been applied to combine different traits (or EBVs for those traits) without knowing its relative economic importance and solving the unfavorable genetic correlations problem. See, for example: factor analytical models (MEYER, 2007; SMITH; CULLIS; THOMPSON, 2001); principal components model (KIRKPATRICK; MEYER, 2004); and principal components aggregate genotype (BUZANSKAS et al., 2013; NASCIMENTO et al., 2014; SOLAR DIAZ et al., 2020; VARGAS et al., 2018).

The CFA is a special case of factor analysis, with the benefit of using previous biological knowledge to assign the observed variables to the factors (BROWN, 2015; HU; BENTLER, 1999; PEÑAGARICANO et al., 2015; SCHUMACKER; LOMAX, 2015). This is similar to the above-mentioned methods but allow more flexibility to determination of the factors. Moreover, the EBVs for the factors from a CFA could also be combined using selection indexes methods by weighting them by their relative importance. In that case, the correlations among observed traits would be optimal used, first to build the MM (reducing redundancy and the number of traits), and then to combine latent traits with biological meaning in an aggregate genotype.

Here, it was used the methodology proposed by Peñagaricano et al. (2015), and as discussed by the same authors, the main disadvantage of this framework is that the uncertainty of any statistical inference produced in one step, is not carried over the subsequent steps. However, the simple interpretation of the results provided by this technique, still being an advantage to investigate causal links among traits that are not directly measurable. To avoid the multi-step procedure, fitting CFA, MTM, and SEM at once could be performed, however it needs to be investigated the feasibility of such technique.

As stated before, knowing the causal structure is not necessary for selection purposes (VALENTE et al., 2013). The presented results showed it; the conventional MTM is sufficient for EBVs prediction, but the revealed causal structure have two main benefits: 1) it is easy to be interpreted and explained to the breeders; 2) if interventions to the selection process are made, the causal model is then more appropriated to take decisions, because the EBVs could be predicted knowing the causal paths (VALENTE et al., 2010, 2011, 2013). Finally, knowing the causal paths among traits allow researchers to better understand the existence of covariance among traits, and to improve selection methods by accounting for causal effects. It will also allow breeders to take decisions about selection, because knowing the causal paths, the selection intensity can be increased on the traits with more impact in the causal network.

5.5. Conclusion

Here it was demonstrated that confirmatory factor analysis can efficiently reduce data dimensionality while provides insights on complex phenomena. Allowed with structural equation models, these techniques could be a powerful tool to dissect the genetic covariances and explain, from the causal perspective, the relationships between latent traits. For selection purposes, predicted factor scores could be used, reducing selection complexity, and allowing breeders to obtain animals with “good locomotion pattern and morphology”. It seems that morphology has a positive causal effect on locomotion traits in Campolina horses, whereas the genes controlling both traits act in an opposite way. Understanding the causal paths between morphology and locomotion traits can contribute to the improvement of horse’s quality, while support researchers better design experiments to analyze locomotion and morphological differences.

5.6. References

- AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: INTERNATIONAL SYMPOSIUM ON INFORMATION THEORY, 2. **Book of Abstracts...** Tsahkadsor, Armenia: Akadémiai Kiadó, 1973, p. 267–281.
- BENTLER, P. M. Some contributions to efficient statistics in structural models: Specification and estimation of moment structures. **Psychometrika**, v. 48, n. 4, p. 493–517, 1983.
- BENTLER, P. M. Comparative fit indexes in structural models. **Psychological Bulletin**, v. 107, n. 2, p. 238–246, 1990.
- BENTLER, P. M. **EQS Structural Equations Program Manual**. 1. ed. Encino, CA: 1995.
- BENTLER, P. M.; BONETT, D. G. Significance tests and goodness of fit in the analysis of covariance structures. **Psychological Bulletin**, v. 88, n. 3, p. 588–606, 1980.
- BROWN, T. A. **Confirmatory Factor Analysis for Applied Research**. 2. ed. New York: The Guilford Press, 2015.
- BROWNE, M. W. Asymptotically distribution-free methods for the analysis of covariance structures. **British Journal of Mathematical and Statistical Psychology**, v. 37, n. 1, p. 62–83, 1984.
- BUZANSKAS, M. E. et al. Genetic parameter estimates and principal component analysis of breeding values of reproduction and growth traits in female Canchim cattle. **Reproduction, Fertility and Development**, v. 25, n. 5, p. 775, 2013.
- DE LOS CAMPOS, G. et al. A structural equation model for describing relationships between

somatic cell score and milk yield in dairy goats¹. **Journal of Animal Science**, v. 84, n. 11, p. 2934–2941, 2006.

DE MATURANA, E. L. et al. Exploring Biological Relationships Between Calving Traits in Primiparous Cattle with a Bayesian Recursive Model. **Genetics**, v. 181, n. 1, p. 277–287, 2009.

DE OLIVEIRA BUSSIMAN, F. et al. Genetic analysis of morphological and functional traits in Campolina horses using Bayesian multi-trait model. **Livestock Science**, v. 216, n. April, p. 119–129, 2018.

DUCRO, B. J. J. et al. Genetic relations of movement and free-jumping traits with dressage and show-jumping performance in competition of Dutch Warmblood horses. **Livestock Science**, v. 107, n. 2–3, p. 227–234, 2007.

EPSKAMP, S. semPlot: Unified Visualizations of Structural Equation Models. **Structural Equation Modeling: A Multidisciplinary Journal**, v. 22, n. 3, p. 474–483, 2015.

GIANOLA, D.; SORENSEN, D. Quantitative Genetic Models for Describing Simultaneous and Recursive Relationships Between Phenotypes. **Genetics**, v. 167, n. 3, p. 1407–1424, 2004.

HAAVELMO, T. The Statistical Implications of a System of Simultaneous Equations. **Econometrica**, v. 11, n. 1, p. 1, 1943.

HAZEL, L. N. The Genetic Basis for Constructing Selection Indexes. **Genetics**, v. 28, n. 6, p. 476–490, 1943.

HAZEL, L. N.; DICKERSON, G. E.; FREEMAN, A. E. The Selection Index-Then, Now, and for the Future. **Journal of Dairy Science**, v. 77, n. 10, p. 3236–3251, 1994.

HENDERSON, C. R. Selection index and expected genetic advance. In: HANSON, W. D.; ROBINSON, H. F. (Eds.) **Statistical genetics and plant breeding**. Washington, DC: The Nacional Academies Press, 1963. p. 141–163.

HENDERSON, C. R.; QUAAS, R. L. Multiple Trait Evaluation Using Relatives' Records. **Journal of Animal Science**, v. 43, n. 6, p. 1188–1197, 1976.

HILL, W. G. Genetic Correlation. In: MALOY, S.; HUGHES, K. (Ed.). **Brenner's Encyclopedia of Genetics**. 2. ed. San Diego, US: Elsevier, 2013. p. 237–239.

HOLMSTRÖM, M.; MAGNUSSON, L. E.; PHILIPSSON, J. Variation in conformation of Swedish Warmblood horses and conformational characteristics of elite sport horses. **Equine Veterinary Journal**, v. 22, n. 3, p. 186–193, 1990.

HU, L.; BENTLER, P. M. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. **Structural Equation Modeling: A Multidisciplinary Journal**, v. 6, n. 1, p. 1–55, 1999.

INOUE, K. et al. Inferring phenotypic causal structures among meat quality traits and the application of a structural equation model in Japanese Black cattle¹. **Journal of Animal Science**, v. 94, n. 10, p. 4133–4142, 2016.

JOHNSON, D. L.; THOMPSON, R. Restricted Maximum Likelihood Estimation of Variance Components for Univariate Animal Models Using Sparse Matrix Techniques and Average Information. **Journal of Dairy Science**, v. 78, n. 2, p. 449–456, 1995.

KIRKPATRICK, M.; MEYER, K. Direct estimation of genetic principal components: Simplified analysis of complex phenotypes. **Genetics**, v. 168, n. 4, p. 2295–2306, 2004.

MARGARITIS, D. **Learning Bayesian Network Model Structure**. Pittsburgh, PA: School of Computer Science, Carnegie Mellon University, 2003.

MEDEIROS, B. R. et al. Brazilian Sport Horse: genetic parameters for approval of Brasileiro de Hipismo stallions. **Tropical Animal Health and Production**, 19 dez. 2019.

MEYER, K. Multivariate analyses of carcass traits for Angus cattle fitting reduced rank and factor analytic models. **Journal of Animal Breeding and Genetics**, v. 124, n. 2, p. 50–64, 2007.

MISZTAL, I. et al. BLUPF90 and related programs (BGF90). In: 7th WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 7. **Proceeding...** Montpellier: Wageningen University, 2002.

MOMEN, M. et al. Including Phenotypic Causal Networks in Genome-Wide Association Studies Using Mixed Effects Structural Equation Models. **Frontiers in Genetics**, v. 9, 2018.

MOMEN, M. et al. Utilizing trait networks and structural equation models as tools to interpret multi-trait genome-wide association studies. **Plant Methods**, v. 15, n. 1, p. 107, 2019.

NASCIMENTO, G. B. et al. Genetic parameter estimates and principal component analysis on performance and carcass traits of a terminal pig sire line. **Acta Agriculturae Scandinavica, Section A — Animal Science**, v. 64, n. 3, p. 137–144, 2014.

PEARL, J. **Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference**. 2. ed. San Francisco: Morgan Kaufmann Publishers, INC., 1988.

PEARL, J. **Causality: Models, Reasoning and Inference**. 2. ed. Cambridge: Cambridge University Press, 2009.

PEÑAGARICANO, F. et al. Searching for causal networks involving latent variables in complex traits: Application to growth, carcass, and meat quality traits in pigs. **Journal of Animal Science**, v. 93, n. 10, p. 4617–4623, 2015.

R CORE TEAM. **R: A Language and Environment for Statistical Computing** (R. D. C. Team, Ed.) R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2016.

ROBERTSON, A. The Sampling Variance of the Genetic Correlation Coefficient. **Biometrics**, v. 15, n. 3, p. 469, 1959.

ROSA, G. J. M. et al. Inferring causal phenotype networks using structural equation models. **Genetics Selection Evolution**, v. 43, n. 1, p. 6, 2011.

ROSSEEL, Y. lavaan : An R Package for Structural Equation Modeling. **Journal of**

Statistical Software, v. 48, n. 2, 2012.

RUSTIN, M. et al. Multi-trait animal model estimation of genetic parameters for linear type and gait traits in the Belgian warmblood horse. **Journal of Animal Breeding and Genetics**, v. 126, n. 5, p. 378–386, 2009.

SANTOS, J. E. S. DOS et al. Effectiveness of the morphofunctional evaluation method of Campolina and Mangalarga Marchador breeds. **Revista Brasileira de Zootecnia**, v. 47, 2018.

SATORRA, A.; BENTLER, P. M. Corrections to test statistics and standard errors in covariance structure analysis. In: VON EYE, A.; CLOGG, C. C. (Eds.). **Latent Variable Analysis: Applications for developmental research**. 1. ed. Thousand Oaks, CA: SAGE Publications, Inc, 1994. p. 399–419.

SCHAEFFER, L. R. Estimation of Variances and Covariances Within the Allowable Parameter Space. **Journal of Dairy Science**, v. 69, n. 1, p. 187–194, 1986.

SCHÖPKE, K.; WENSCH-DORENDORF, M.; SWALVE, H. H. Genetic evaluations of the German Sport Horse: Population structure and use of data from foal and mare inspections and performance tests of mares. **Archives Animal Breeding**, v. 56, n. 1, p. 658–674, 2013.

SCHUMACKER, R. E.; LOMAX, R. G. **A Beginner's guide to Structural Equation Modeling**. 4. ed. New York: Routledge, 2015.

SEARLE, S. R. The matrix Handling of BLUE and BLUP in the mixed linear model. **Linear Algebra and its Applications**, v. 264, p. 291–311, 1997.

SHIPLEY, B. From biological hypotheses to structural equation models: the imperfection of causal translation. In: PUGESEK, B.; TOMER, A.; VON EYE, A. (Eds.) **Structural Equation Modeling**. Cambridge: Cambridge University Press, 2003. p. 194–211.

SMITH, A.; CULLIS, B.; THOMPSON, R. Analyzing Variety by Environment Data Using Multiplicative Mixed Models and Adjustments for Spatial Field Trend. **Biometrics**, v. 57, n. 4, p. 1138–1147, 2001.

SMITH, H. F. A discriminant function for plant selection. **Annals of Eugenics**, v. 7, n. 3, p. 240–250, 1936.

SOLAR DIAZ, I. D. P. et al. Principal components for morphometric traits in Campolina horses. **Journal of Animal Breeding and Genetics**, v. 00, p. 1–9, 2020.

STAIGER, E. A. et al. Morphological Variation in Gaited Horse Breeds. **Journal of Equine Veterinary Science**, v. 43, p. 55–65, 2016.

STEIGER, J. H. Structural Model Evaluation and Modification: An Interval Estimation Approach. **Multivariate Behavioral Research**, v. 25, n. 2, p. 173–180, 1990.

TUCKER, L. R.; LEWIS, C. A reliability coefficient for maximum likelihood factor analysis. **Psychometrika**, v. 38, n. 1, p. 1–10, 1973.

VARGAS, G. et al. Unravelling biological biotypes for growth, visual score and reproductive traits in Nellore cattle via principal component analysis. **Livestock Science**, v. 217, p. 37–43,

2018.

VALENTE, B. D. et al. Searching for Recursive Causal Structures in Multivariate Quantitative Genetics Mixed Models. **Genetics**, v. 185, n. 2, p. 633 LP – 644, 2010.

VALENTE, B. D. et al. Searching for phenotypic causal networks involving complex traits: an application to European quail. **Genetics Selection Evolution**, v. 43, n. 1, p. 37, 2011.

VALENTE, B. D. et al. Is Structural Equation Modeling Advantageous for the Genetic Improvement of Multiple Traits? **Genetics**, v. 194, n. 3, p. 561–572, 2013.

VALENTE, B. D.; ROSA, G. J. DE M. Mixed Effects Structural Equation Models and Phenotypic Causal Networks. In: GONDRO, C.; VAN DER WERF, J.; HAYES, B. (Eds.) **Genome-Wide Association Studies and Genomic Prediction**. Totowa, NJ: Humana Press, 2013. p. 449–464.

VICENTE, A. A. et al. Selection for morphology, gaits and functional traits in Lusitano horses: I. Genetic parameter estimates. **Livestock Science**, v. 164, n. 1, p. 1–12, 2014.

VISSCHER, P. M. On the sampling variance of intraclass correlations and genetic correlations. **Genetics**, v. 149, n. 3, p. 1605–1614, 1998.

VOSTRY, L. et al. Genetic Parameters for Linear Type Traits in Three Czech Draught Horse Breeds. **Agriculturae Conspectus Scientificus**, v. 82, n. 2, p. 111–115, 2017.

WHEATON, B. et al. Assessing Reliability and Stability in Panel Models. **Sociological Methodology**, v. 8, p. 84, 1977.

WRIGHT, S. On the Nature of Size Factors. **Genetics**, v. 3, n. 4, p. 367–374, 1918.

WRIGHT, S. Correlation and Causation. **Journal of Agriculture Research**, v. 201, p. 557–585, 1921.

WU, X.-L. et al. Inferring Relationships Between Somatic Cell Score and Milk Yield Using Simultaneous and Recursive Models. **Journal of Dairy Science**, v. 90, n. 7, p. 3508–3521, 2007.

6. BAYESIAN RECURSIVE AND STRUCTURAL EQUATION MODELS TO INFER CAUSAL LINKS AMONG GAIT VISUAL SCORES IN CAMPOLINA HORSES

*“É isso, sempre, que a gente precisa procurar na modelagem: tentar entender como foi o processo biológico”
(Clarice Demétrio)¹⁴.*

F. de Oliveira Bussiman¹; F.F. Silva (*in memoriam*)²; R.S.B. Carvalho³; R.V. Ventura¹; E.C. Mattos⁴; J.B.S. Ferraz⁴; J.P. Eler⁴; J.C.C. Balieiro¹

¹*Bioinformatic and Animal Breeding Lab., Department of Animal Nutrition and Production, College of Veterinary Medicine and Animal Science University of São Paulo (BIOMA-VNP/FMVZ-USP), Av. Duque de Caxias Norte, 225, 13.635-900, Pirassununga-SP, Brazil;*

²*Department of Animal Science, Federal University of Viçosa (DZO/UFV), Av. P H Rolfs, s/n, 36.750-900, Viçosa-MG, Brazil;*

³*Department of Basic Sciences, College of Animal Science and Food Engineering, University of São Paulo (ZAB/FZEA-USP), Av. Duque de Caxias Norte, 225, 13.635-900, Pirassununga-SP, Brazil;*

⁴*Group of Animal Breeding and Biotechnology, Department of Veterinary Medicine, College of Animal Science and Food Engineering, University of São Paulo (GMAB-ZMV/FZEA-USP), Av. Duque de Caxias Norte, 225, 13.635-900, Pirassununga-SP, Brazil;*

RESUMO: Escores visuais de andamento são amplamente usados no melhoramento de cavalos porque eles são uma estratégia rápida e fácil de fenotipagem, permitindo uma interpretação numérica de um fenômeno biológico complexo como a qualidade do andamento, mas eles sofrem de muita subjetividade ou elevada influência ambiental. O objetivo do presente trabalho foi investigar como cinco escores visuais de andamento se interagem para criar o escore de andamento total em cavalos Campolina. Foram usados dados de 5.475 cavalos, nascidos entre 1990 e 2013, com idade média a mensuração de $39,65 \pm 3,39$ meses, e ao menos um registro fenotípico para uma das seguintes características: Dissociação (Di); Cômulo (C); Estilo (S); Regularidade (R); Desenvolvimento (De); e Escore total de Andamento (GtS, calculado como a soma das anteriores). O pedigree utilizado conteve 3 gerações, assim, 14.079 cavalos foram incluídos na matriz de parentescos aditivos. Sob uma abordagem Bayesiana, os componentes de (co)variância foram estimados por meio de um modelo animal multicaracterística (MTM) e então foi aplicado o algoritmo de indução causal nas amostras da matriz de (co)variância residual. O gráfico não-direcionado obtido foi, então, orientado em 7 diferentes estruturas causais, e cada uma foi ajustada aos dados por meio de modelos de equações estruturais (SEM). Finalmente, a estrutura causal final foi escolhida baseada no critério de informação de *Deviance* (DIC). Foi encontrado que S tem um impacto maior na estrutura causal de andamento, afetando diretamente e indiretamente C. O efeito causal indireto de S em C foi através do efeito direto de S em De, então do efeito de De em R e, por fim do efeito direto de R em C. Di não causou nenhuma das outras características estudadas e foi causada por S, o que pode explicar a correlação genética entre Di e GtS, uma vez que quando os efeitos causais foram incluídos no modelo, tal correlação foi absorvida por eles. Esses *paths* têm sentido biológico nos

¹⁴ Informação fornecida por Demétrio, C. G. B. durante o curso “Modelos Lineares Generalizados”, Universidade de São Paulo, Escola Superior de Agricultura “Luiz de Queiroz”, Piracicaba, 2020.

movimentos dos cavalos e podem ajudar criadores e pesquisadores a entender melhor a complexa rede causal do andamento em cavalos.

Palavras-chave: Inferência causal, network causal de andamento, efeitos causais, indução causal, andamento de cavalos.

ABSTRACT: Gait visual scores are widely applied on horse breeding because they are fast and ease phenotyping strategy, allowing a numeric interpretation of a complex biological process such as gait quality; however, they may suffer from subjectivity or high environmental influence. This study aimed to investigate how five visual gait scores interact with each other, in terms of causal links, to create the gait total scores in Campolina horses. It was used data from 5,475 horses, born between 1990 and 2013, with an average age at the measurement of 39.65 ± 3.39 months, and records for at least one of the following traits: Dissociation (Di); Comfort (C); Style (S); Regularity (R); Development (De); and Gait total Scores (GtS, computed as the sum of all previous one). The used pedigree was comprised of 3 generations, i.e., 14,079 horses were included in the additive relationship matrix. Under a Bayesian framework, (co)variance components were estimated through a multitrait animal model (MTM) and then the inductive causation (IC) algorithm was applied to the samples of residual (co)variance matrix. The resulting undirected graph from the IC algorithm was directed in 7 possible causal structures and each one was fitted by a structural equation model (SEM). Finally, the final causal structure was chosen based on deviance information criteria (DIC). It was found that S has a major impact on the causal network of gait, affecting directly and indirectly C. The indirect causal effect of S on C was through the direct effect of S on De, and then the direct effect of De on R; and finally, the direct effect of R on C. Di did not cause any other studied trait, and was caused by S, which is the reason of genetic correlation between Di and GtS, since when causal effects are added on the model, the genetic correlation between Di and GtS was absorbed by them. Those paths have biological meaning on horse movements and can help breeders and researchers to better understand the complex causal network of gait in horses.

Keywords: Causal inference, gait causal network, causal effects; inductive causation, horse gait.

6.1. Introduction

The Campolina horse is a Brazilian gaited horse breed that arose from the brought horses from the Iberian Peninsula. In 1870 at Minas Gerais State, the Campolina horse emerged, and in 1951 the Breeders Association (Campolina Horse Breeders Association – ABCCCampolina, <http://www.campolina.org.br/portal/>) was formed (PROCÓPIO; BERGMANN; COSTA, 2003). Since 1996 all registered horses in the Campolina study book are phenotyped (around 36 months of age) for gait visual scores, and the overall score is simply the direct sum of them, however, those scores are used only to qualify the animals to be (or not) registered (ABCCCAMPOLINA, 2018). The selection for gait in the Campolina breed is based on

phenotypic competitions records and this population was never under a formal genetic improvement program.

It is common the use visual scores for gait evaluation on horses, mainly because of its low cost, speed of phenotyping, and logistics, however visual scores are well known to suffer from subjectivity (RUSTIN et al., 2009; VICENTE et al., 2014a, 2014b), that naturally occur when different people need to evaluate specific features. In addition to the subjectivity, the high environmental influence over gait phenotypes makes the selection for those traits difficult. Usually, those score traits are analyzed through multitrait models (MTM) allowing estimates of genetic (co)variance among them (MOLINA et al., 1999; RUSTIN et al., 2009; VICENTE et al., 2014a; VOSTRY et al., 2017) and with morphological traits.

MTM is often used on horses' genetic evaluations because in general morphological traits are genetically positively correlated which benefits prediction for gait traits. MTM is well known to increase the accuracy of prediction (THOMPSON; MEYER, 1986; VAN DER WERF; VAN ARENDONK; DE VRIES, 1992) and to reduce selection bias (MEYER, 1995; BOLIGON et al., 2009). Despite those advantages, MTM is limited when the number of traits is high, because of the increase in the mixed model equations (MME) (MEYER, 2005) and because of the presence of unfavorable genetic correlations among the traits.

Structural equation models (SEM) (WRIGHT, 1921; HAAVELMO, 1943) are a kind of MTM accounting for the causal relationships among traits (VALENTE et al., 2013). Those models were first adapted to the animal breeding context by Gianola and Sorensen (2004). To overcome the problem of causal relationship selection, Valente et al. (2010) presented an adaptation of the inductive causation (IC) algorithm (VERMA; PEARL, 1990; PEARL, 2009), allowing one to infer the causal structure among different traits in the context of MME.

Knowing the causal relationships among different traits is not necessary for selection purposes (VALENTE et al., 2013), but in the situations of interventions on the selection system, they can achieve more efficiency and accuracy (ROSA et al., 2011; VALENTE et al., 2013; VALENTE; ROSA, 2013). SEM can accommodate selection decisions (DE LOS CAMPOS; GIANOLA; HERINGSTAD, 2006; ROSA et al., 2011) and allow one to increase selection intensity on those traits that have a high impact on all other traits in the phenotypic network. Thus, this study aimed to investigate the causal relationship among six gait visual scores, being one the sum of all other five, in Campolina horses through Bayesian SEM.

6.2. Material and Methods

6.2.1. Phenotypic and pedigree data

For this study, the complete database of ABCCCampolina was provided. It contained 5,475 animals, born between 1990 and 2013 with an average age at the measurement of 39.65 ± 3.39 months, and records for five visual scores, as follows: Dissociation (Di); Comfort (C); Style (S); Regularity (R); Development (De); and Gait total Scores (GtS). According to ABCCCampolina, those scores are defined as (ABCCCAMPOLINA, 2018):

- I. Di: ranges from 0 (animal without dissociation – trot) to 40 (animal with clear visualization of triple-limb-support); it is related with the coordinated movements of thoracic and pelvic limb-pairs, with support and suspension of each pair causing triple-limb-support and guaranteeing the horse don't lose contact with the ground;
- II. C: it varies between 0 (animal with high impacts on saddle making the rider uncomfortable) and 60 (animal with no impacts on the saddle); it is related to the quality of the horse's movements, with no vertical, lateral, or frontal oscillations and impacts to the rider;
- III. S: ranges between 0 (animal with no beauty of movements) and 40 (animal with a good balance of limbs elevation and elegant movements); it represents the combination of posture, balance, and harmony of movements, which need to be elegant and with energy;
- IV. R: ranging from 0 (an animal that changes gait or loses rhythm) to 30 (an animal that is capable to perform the same gait for long periods), this score is associated with the maintenance of the same gait type, conserving itself defined, stable, rhythmic and with a good cadence;
- V. De: it varies from 0 (animal with high step frequency) to 30 (animal with low steps frequency); this score is related to the capability of the horse to cover long distances with few steps; and
- VI. GtS: is the overall gait evaluation computed as: $GtS = Di + C + S + R + De$; it ranges from 100 to 200 because any animal with less than 100 points is not registered in the Campolina study book.

For all analyses, the contemporary groups (CG) were defined as animals born in the same year (24 levels – from 1990 to 2013) and season (4 levels – summer, autumn, winter, and

spring), with the same sex (2 levels – male and female) and registered in the same year of registry (24 levels – from 1993 to 2016). The study (same as herd) was not included on CGs because its inclusion would lead to several CG with one single observation. Observations above or below three standard deviations from the phenotypic mean were excluded, as well as CG with less than five observations (or without phenotypic variance). Descriptive statistic of the final dataset is presented in Table 1, and the number of records in common with every pair of traits is shown in Table 2.

Table 1 - Descriptive statistic of the used dataset (after cleaning for outliers and for contemporary groups)

Traits	$\bar{X}$ ¹	SD ²	Min. ³	Max. ⁴	N _F ⁵	N _M ⁶	N _{GC} ⁷	N _{Stud} ⁸
Di	30.43	4.00	18	51	2,686	1,201	166	203
C	47.43	5.21	26	61	3,562	1,477	217	262
S	30.10	5.45	1	52	2,904	1,238	180	218
R	23.11	2.81	15	37	2,590	1,086	162	210
De	23.32	3.06	15	36	1,678	758	126	129
GtS	140.44	12.74	89	176	2,137	1,112	243	245

Source: Bussiman (2021). ¹Phenotypic mean; ²Phenotypic standard deviation; ³Minimum observable value; ⁴Maximum observable value; ⁵Number of females; ⁶Number of males; ⁷Number of contemporary groups; ⁸Number of Studs (herds). Abbreviations: Di = Dissociation; C = Comfort; S = Style; R = Regularity; De = Development; GtS = Gait total Scores.

Table 2 - Number of phenotypic records (diagonal bold) and number of animals with records for every pair of traits (above diagonal) for the studied traits

Traits	Traits					
	Di	C	S	R	De	GtS
Di	3,887	3,683	3,359	2,772	1,944	2,607
C		5,039	3,847	3,371	2,260	3,073
S			4,142	2,951	2,159	2,720
R				3,676	2,263	2,553
De					2,436	1,776
GtS						3,249

Source: Bussiman (2021). Abbreviations: Di = Dissociation; C = Comfort; S = Style; R = Regularity; De = Development; GtS = Gait total Scores.

The pedigree for (co)variance components estimation contained 3 generations, thus the additive genetic relationship matrix **A** (WRIGHT, 1922; HENDERSON, 1975) had 14,079 horses, born between 1900 and 2013. There were 1,825 stallions with a progeny average of 7.03, and 7,281 mares with a progeny average of 1.76. The average inbreeding was 3.72% (complete pedigree), and the average additive relationship coefficient (excluding self-relationships) was 0.05. In addition, 12,806 horses had both parents known, 29 horses had only sire unknown, and none of the horses had only dam unknown.

6.2.2. Statistical analyses

Following Valente et al. (2010) orientations, the analyses were performed in 3 steps, as follows: 1) a standard MTM was fitted; 2) the IC algorithm was applied to the samples of residual (co)variance matrix, and 3) the SEM was fitted. For all traits the following statistical model was assumed:

$$y_{ijkl} = \mu + \beta_1 \text{Age}_k + \beta_2 \text{Age}_k^2 + \text{CG}_i + \text{Stud}_j + u_k + t_l + e_{ijkl}, \quad (6.1)$$

Where: y_{ijkl} represent the observed phenotypic observation (i.e. Di, C, S, R, De, or GtS); μ is just a constant; β_1 and β_2 are the regression coefficients of covariate age at the measurement of the k^{th} (Age_k) horse, linear, and quadratic respectively; CG_i represent the systematic effect of the i^{th} contemporary group; Stud_j represent the systematic effect of the j^{th} stud of birth (same as herd); u_k represent the random additive genetic effect of the k^{th} horse; t_l represent the uncorrelated random effect of the l^{th} technician (person who gave the scores), and e_{ijkl} represent the random residual terms.

a) Multitrait model

Under matrix notation, (6.1) can be re-written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u} + \mathbf{Z}_2\mathbf{t} + \mathbf{e}, \quad (6.2)$$

where: $\mathbf{y}$ is the vector of phenotypic observations; $\boldsymbol{\beta}$ is the vector of systematic effects, assumed normally distributed ($\boldsymbol{\beta} \sim N\{0, \boldsymbol{\Sigma}_\beta\}$); $\mathbf{u}$ is the vector of random additive individual genetic effects; $\mathbf{t}$ is the vector of uncorrelated random technician effects; $\mathbf{e}$ is the vector of random residual terms; $\mathbf{X}$ and $\mathbf{Z}_i$ are incidence matrices for systematic and random effects, respectively. Under normality assumptions, the joint distribution of random effects was given by:

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{t} \\ \mathbf{e} \end{bmatrix} \sim \text{MVN} \left\{ \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{G}_0 \otimes \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_1 \otimes \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_0 \otimes \mathbf{I} \end{bmatrix} \right\}, \quad (6.3)$$

Where: $\mathbf{u}$, $\mathbf{t}$ and $\mathbf{e}$ are defined above; $\mathbf{G}_0$ is the additive genetic (co)variance matrix among the studied traits (6 x 6); $\mathbf{G}_1$ is the (co)variance matrix among the technician across traits (6 x 6); $\mathbf{R}_0$ is the residual (co)variance matrix (6 x 6), and $\mathbf{I}$ is an identity matrix of proper order.

Given those assumptions in (6.3), the conditional distribution of $\mathbf{y}$ was assumed *a priori* as: $\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \mathbf{t}, \mathbf{G}_0, \mathbf{G}_1, \mathbf{R}_0 \sim \text{MVN}\{\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_2\mathbf{u} + \mathbf{Z}_1\mathbf{t}, \mathbf{R}_0 \otimes \mathbf{I}\}$. Additionally, $\mathbf{G}_0$, $\mathbf{G}_1$ and $\mathbf{R}_0$ were assumed to follow an inverted Wishart distribution with hyperparameters $\mathbf{v}$ and $\mathbf{V}$ (i.e. $\text{WI}^{-1}\{\mathbf{v}, \mathbf{V}\}$), and all hyperparameters were chosen to provide non-informative priors. The (co)variance components and genetic parameters were estimated by Gibbs's sampling algorithm through the GIBBS2F90 software (MISZTAL et al., 2002) with a total of 1,000,000 samples were generated, discarding the first 500,000 samples as burn-in and the remaining samples were stored with a thinning interval of 100. Convergence evidence was checked through graphical inspection of the generated chain and applying the Geweke's test (Z score) (GEWEKE, 1992) and Raftery-Lewis's test (RAFTERY; LEWIS, 1992), implemented in the POSTGIBBSF90 software (MISZTAL et al., 2002).

b) Inductive causation algorithm

The MTM samples $\mathbf{R}_0$ were considered to have sufficient information about the functional relationship among studied traits conditional on genetic effects (VALENTE et al., 2010; ROSA et al., 2011); thus, the search for causal links was made over $p(\mathbf{y}|\mathbf{u})$ instead of $p(\mathbf{y})$; that is, the confounding unobserved correlated genetic effects were removed (VALENTE; ROSA, 2013). The IC algorithm is one of the so-called constraint-based algorithms, performing a search based on causal sufficiency assumption (SPIRTES; GLYMOUR; SCHEINES, 2001; VALENTE; ROSA, 2013); it follows three steps:

- I. Let S be a set of variables, that is $S = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_j, \dots, \mathbf{y}_k\}$. For each pair of traits $\mathbf{y}_j$ and $\mathbf{y}_{j'}$, compute partial correlations conditional on all other traits ($r_{\mathbf{y}_j\mathbf{y}_{j'}|S_{jj'}}^2$, $\forall j \neq j'$), and if it is statically significant, then those traits are considered adjacent and connected with an undirected link, this results in a partially oriented graph (i.e., $\mathbf{y}_j \longleftrightarrow \mathbf{y}_{j'}$);

- II. For non-adjacent pairs of variables y_j and $y_{j'}$, with an adjacent variable in common $y_{j''}$ (that is, $y_j \longleftrightarrow y_{j''} \longleftrightarrow y_{j'}$), compute partial correlations on all other remaining variables. If it is statistically significant, then add directed links pointed to $y_{j''}$ (e.g., $y_j \longrightarrow y_{j''} \longleftarrow y_{j'}$), else continue; and
- III. Every nondirectional edge remaining can be oriented if they do not create a new unshielded collider or cyclic network.

Partial correlations were declared statistically significant (i.e., non-null) when its highest posterior density interval (HPD) did not include zero. Those three steps were carried in *R* (R CORE TEAM, 2020), using the codes provided by Valente and Rosa (2013), for HPD content varying from 70% to 95% (by 5%). From the resulting undirected network, different orientations were proposed to be tested by fitting SEM as in Valente et al. (2010).

c) Structural equation models

The causal structures resulting from the previous step were used to fit the SEM, which can be written as:

$$\mathbf{y} = (\mathbf{\Lambda} \otimes \mathbf{I})\mathbf{y} + \mathbf{X}\boldsymbol{\beta}^* + \mathbf{Z}_1\mathbf{u}^* + \mathbf{Z}_2\mathbf{t}^* + \mathbf{e}^*, \quad (6.4)$$

Where: $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}^*$, $\mathbf{Z}_1$, $\mathbf{u}^*$, $\mathbf{Z}_2$, $\mathbf{t}^*$, and $\mathbf{e}^*$ are analogous as defined in (6.2); and $\mathbf{\Lambda}$ is a square matrix (6 x 6) with causal coefficients of one trait over another trait.

The SEM was fitted under the same assumptions as of the MTM; thus, the joint distribution of the random effects was given by:

$$\begin{bmatrix} \mathbf{u}^* \\ \mathbf{t}^* \\ \mathbf{e}^* \end{bmatrix} \sim \text{MVN} \left\{ \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{G}_0^* \otimes \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_1^* \otimes \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Psi}_0 \otimes \mathbf{I} \end{bmatrix} \right\}, \quad (5)$$

Where: $\mathbf{u}^*$, $\mathbf{t}^*$ and $\mathbf{e}^*$ are defined above; $\mathbf{G}_0^*$ is the SEM additive genetic (co)variance matrix among the studied traits (6 x 6); $\mathbf{G}_1^*$ is the SEM (co)variance matrix among the technician across traits (6 x 6), and $\boldsymbol{\Psi}_0$ is a diagonal matrix with the SEM residual variances. All residual (co)variances were assumed to be zero, guaranteeing identifiability to the causal coefficients in the likelihood function (VALENTE et al., 2010, 2013; ROSA et al., 2011; VALENTE; ROSA, 2013).

The equivalence between SEM and MTM can be shown by solving (4) for $\mathbf{y}$ (VALENTE; ROSA, 2013), which results in: $\mathbf{G}_0 = (\mathbf{I} - \mathbf{\Lambda})^{-1} \mathbf{G}_0^* (\mathbf{I} - \mathbf{\Lambda})^{-T}$; $\mathbf{G}_1 = (\mathbf{I} - \mathbf{\Lambda})^{-1} \mathbf{G}_1^* (\mathbf{I} - \mathbf{\Lambda})^{-T}$; and $\mathbf{R}_0 = (\mathbf{I} - \mathbf{\Lambda})^{-1} \mathbf{\Psi}_0 (\mathbf{I} - \mathbf{\Lambda})^{-T}$. Concerning $\mathbf{\Lambda}$, for a fully recursive model (with 6 traits) it is as follows (GIANOLA; SORENSEN, 2004; VALENTE; ROSA, 2013; ROSA; VALENTE, 2014):

$$\mathbf{\Lambda} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \lambda_{2 \leftarrow 1} & 0 & 0 & 0 & 0 & 0 \\ \lambda_{3 \leftarrow 1} & \lambda_{3 \leftarrow 2} & 0 & 0 & 0 & 0 \\ \lambda_{4 \leftarrow 1} & \lambda_{4 \leftarrow 2} & \lambda_{4 \leftarrow 3} & 0 & 0 & 0 \\ \lambda_{5 \leftarrow 1} & \lambda_{5 \leftarrow 2} & \lambda_{5 \leftarrow 3} & \lambda_{5 \leftarrow 4} & 0 & 0 \\ \lambda_{6 \leftarrow 1} & \lambda_{6 \leftarrow 2} & \lambda_{6 \leftarrow 3} & \lambda_{6 \leftarrow 4} & \lambda_{6 \leftarrow 5} & 0 \end{bmatrix}, \quad (6)$$

Where: $\lambda_{i \leftarrow j}$ represent the causal effect of the j^{th} trait on the i^{th} . Depending on the causal structure $\lambda_{i \leftarrow j}$ can be freely estimated or be assumed as zero, allowing one to compare different SEM to find the best fit of the causal network (VALENTE et al., 2010; VALENTE; ROSA, 2013; ROSA; VALENTE, 2014).

The SEM was fitted using the GIBBS2F90 software (MISZTAL et al., 2002) through Gibbs's sampling algorithm. A single chain of 1,000,000 samples was generated, discarding the first 500,000 samples, and saving the remaining with a thinning interval of 100. Convergence evidence was also checked by graphical inspection of the generated chain, and by Geweke's test (GEWEKE, 1992) and Raftery-Lewis's test (RAFTERY; LEWIS, 1992), implemented in the POSTGIBBSF90 software (MISZTAL et al., 2002). Different SEMs were compared by the Deviance Information Criteria (DIC) (SPIEGELHALTER et al., 2002, 2014). Models with smaller DIC are better supported by the data.

6.3. Results

6.3.1. Convergence diagnoses

The whole set of parameters to be estimated in this study did not present any trend or signs of non-convergency on graphical inspection (data not being shown). The Z-scores from Geweke's test for the posterior distribution of (co)variance components, heritability estimates, and genetic correlations ranged between -1.50 and 1.50. All parameters passed in Raftery-Lewis's test. The autocorrelation (*r-lag50*) varied from -0.04 to 0.53, and all parameters passed

the Heidelberg-Welch stationary test (HEIDELBERGER; WELCH, 1983). The main concern of non-convergence signs was on $\lambda_{i \leftarrow j}$, but for those parameters, Z-scores ranged between -2.68 and 2.75, and $r\text{-lag}50$ varied between -0.13 and 0.67. Additionally, a graphical inspection of chains for $\lambda_{i \leftarrow j}$ showed signs of convergence (data not being shown), and they passed the Raftery-Lewis's test.

6.3.2. Multitrait model

Heritability estimates ranged from 0.08 ± 0.02 (C) to 0.16 ± 0.02 (R), being from low to moderate (Table 3), whereas the proportion of phenotypic variance explained by the technician was much higher, varying between 0.37 ± 0.01 (Di) to 0.46 ± 0.01 (C) (Table 4). Genetic correlations ranged from 0.52 ± 0.13 (between S and De) to 0.99 ± 0.00 (between Di and R) (Table 3), while correlations among technician effects were from 0.31 ± 0.01 (between C and S) to 0.98 ± 0.00 (between R and De) (Table 4). All genetic and technician correlations were positive, as well as the residual correlations, which varied between 0.27 ± 0.03 (between De and C) and 0.77 ± 0.01 (between GtS and S) (Table 3).

Table 3 - Genetic parameter estimates for all studied traits. Heritability (bold diagonal); genetic correlations (above diagonal); and residual correlations (below diagonal)

Traits	Traits					
	Di	C	S	R	De	GtS
Di	0.12 ± 0.02 [0.08;0.15]	0.71 ± 0.08 [0.54;0.87]	0.82 ± 0.06 [0.70;0.93]	0.99 ± 0.00 [0.98;1.00]	0.81 ± 0.07 [0.67;0.94]	0.93 ± 0.03 [0.45;0.88]
C	0.52 ± 0.02 [0.48;0.56]	0.08 ± 0.02 [0.05;0.11]	0.67 ± 0.11 [0.45;0.88]	0.74 ± 0.08 [0.59;0.88]	0.70 ± 0.10 [0.51;0.89]	0.86 ± 0.05 [0.76;0.95]
S	0.61 ± 0.02 [0.58;0.65]	0.39 ± 0.02 [0.35;0.44]	0.13 ± 0.02 [0.08;0.18]	0.84 ± 0.05 [0.74;0.93]	0.52 ± 0.13 [0.28;0.76]	0.91 ± 0.03 [0.85;0.96]
R	0.44 ± 0.02 [0.39;0.48]	0.34 ± 0.02 [0.30;0.39]	0.30 ± 0.03 [0.24;0.35]	0.16 ± 0.02 [0.11;0.20]	0.84 ± 0.06 [0.73;0.95]	0.95 ± 0.02 [0.92;0.98]
De	0.40 ± 0.02 [0.36;0.45]	0.27 ± 0.03 [0.22;0.32]	0.32 ± 0.03 [0.26;0.38]	0.69 ± 0.02 [0.66;0.73]	0.12 ± 0.02 [0.08;0.17]	0.79 ± 0.07 [0.66;0.91]
GtS	0.71 ± 0.01 [0.68;0.73]	0.75 ± 0.01 [0.73;0.78]	0.77 ± 0.01 [0.75;0.79]	0.68 ± 0.01 [0.65;0.71]	0.66 ± 0.02 [0.63;0.70]	0.15 ± 0.02 [0.11;0.19]

Source: Bussiman (2021). Abbreviations: Di = Dissociation; C = Comfort; S = Style; R = Regularity; De = Development; GtS = Gait total Scores.

Table 4 - Estimated technician effects over all studied traits. The proportion of phenotypic variance explained by technician (diagonal bold) and correlation between technician effects among traits (above diagonal)

Traits	Traits					
	Di	C	S	R	De	GtS
Di	0.37 ± 0.01 [0.36;0.39]	0.71 ± 0.00 [0.70;0.71]	0.78 ± 0.00 [0.78;0.79]	0.84 ± 0.00 [0.84;0.85]	0.85 ± 0.00 [0.85;0.86]	0.97 ± 0.00 [0.96;0.97]
C		0.46 ± 0.01 [0.44;0.47]	0.31 ± 0.01 [0.29;0.32]	0.57 ± 0.01 [0.56;0.58]	0.61 ± 0.01 [0.60;0.62]	0.75 ± 0.00 [0.74;0.76]
S			0.41 ± 0.01 [0.40;0.42]	0.55 ± 0.01 [0.54;0.56]	0.54 ± 0.01 [0.53;0.55]	0.79 ± 0.00 [0.78;0.80]
R				0.40 ± 0.01 [0.39;0.41]	0.98 ± 0.00 [0.98;0.98]	0.87 ± 0.00 [0.86;0.87]
De					0.38 ± 0.01 [0.37;0.39]	0.88 ± 0.00 [0.87;0.88]
GtS						0.43 ± 0.01 [0.42;0.44]

Source: Bussiman (2021). Abbreviations: Di = Dissociation; C = Comfort; S = Style; R = Regularity; De = Development; GtS = Gait total Scores.

6.3.3. Inductive causation algorithm

The undirected network obtained by IC did not change for any HPD content tested. Links among Di, S, C, De, and R were identified, as well as causal effects $\lambda_{GtS \leftarrow S}$, $\lambda_{GtS \leftarrow De}$, $\lambda_{GtS \leftarrow R}$ and $\lambda_{GtS \leftarrow C}$ (Figure 1). This undirected network was then oriented in six different acyclic orientations to produce the causal structures in Figure 2. Those gait scores are evaluated altogether at the phenotypic inspection for registration, thus, no temporal information could be used to orient the graphs.

Figure 1 – Links between studied traits by the inductive causation algorithm with highest density interval varying from 70% to 95%. Abbreviations: Di = Dissociation; C = Comfort; S = Style; R = Regularity; De = Development; and GtS = Gait total Scores

Source: Bussiman (2021).

Figure 2 – Tested causal networks among all studied traits. Abbreviations: Di = Dissociation; C = Comfort; S = Style; R = Regularity; De = Development; GtS = Gait total Scores; and SEM = Structural Equation Model

Source: Bussiman (2021).

6.3.4. Structural equation model

The best-fitting model was SEM01 (Figure 2) based on DIC (Table 5). As the (co)variance components from SEM and MTM may differ, because in SEM the genetic effects represent only those directly affecting each trait, whereas in MTM they are overall effects (direct plus indirect) (VALENTE et al., 2013); (co)variance components from SEM01 were rescaled to obtain the predicted MTM (co)variance components by using the equivalence between MTM and SEM provided in section 6.2.2. The overall difference between those matrices was 10.93% (for the additive (co)variance matrix); 0.70% (for the technician (co)variance matrix); and 0.95% (for residual (co)variance matrix).

Table 5 - Deviance Information Criteria (DIC); posterior Deviance ($-2\log p(\mathbf{y}|\hat{\boldsymbol{\theta}})$); the effective number of parameters (p_D); and number of causal coefficients ($N\lambda$), according to each fitted model

Model	$-2\log p(\mathbf{y} \hat{\boldsymbol{\theta}})$	DIC ¹	p_D	$N\lambda$
SEM 01	62.482,16	63.652,63	8.749,52	9
SEM 03	59.725,64	67.406,07	10.687,97	9
SEM 05	61.219,23	68.325,24	10.345,81	9
SEM 04	61.669,47	68.489,05	10.380,17	9
SEM 06	61.893,70	68.577,78	10.259,67	9
SEM 02	62.451,39	68.879,07	9.703,89	9
FRM	69.650,89	74.875,52	7.527,85	15
MTM	90.017,27	92.709,66	4.486,86	--

Source: Bussiman (2021). ¹Models are ranked according to DIC values from smallest to highest. Abbreviations: SEM = Structural Equation Model, from 1 to 6 (Figure 2); FRM = Fully Recursive Model (Λ presented in (6), section 2.2.3); and MTM = Multitrait Model.

For all causal coefficients estimated from SEM01, none HPD did include zero; thus, those coefficients can be considered different from 0 (Table 6). The posterior mean for those coefficients ranged from -0.042 ± 0.010 ($\lambda_{GtS \leftarrow S}$) to 1.107 ± 0.007 ($\lambda_{GtS \leftarrow C}$). Posterior means of genetic parameters and technician effects for SEM01 are presented in Table 7 and Table 8, respectively. No causal effect of Di over all other traits were found, and the most influential trait over GtS was S, for instance, it had a direct effect over GtS ($\lambda_{GtS \leftarrow S} = -0.042$); an indirect effect through De ($\lambda_{De \leftarrow S} \lambda_{GtS \leftarrow De} = 0.418$); an indirect effect through C ($\lambda_{C \leftarrow S} \lambda_{GtS \leftarrow C} = 0.623$); an indirect effect through De and R ($\lambda_{De \leftarrow S} \lambda_{R \leftarrow De} \lambda_{GtS \leftarrow R} = 0.282$); and finally, an indirect effect through De, R, and C ($\lambda_{De \leftarrow S} \lambda_{R \leftarrow De} \lambda_{C \leftarrow R} \lambda_{GtS \leftarrow C} = 0.062$) (Figure 2 and Table 6).

Table 6 - Causal coefficient estimates from Structural Equation Model 01; Posterior means $\pm$ posterior standard deviation; highest density intervals (HPD) and convergence diagnoses (Geweke's Z-score and r -Lag50)

Causal Coefficient	Mean	HPD	Geweke's Z-score	r -Lag50
$\lambda_{\text{GtS} \leftarrow \text{C}}$	1.107 ± 0.007	[1.094;1.121]	2.433	-0.029
$\lambda_{\text{Di} \leftarrow \text{S}}$	0.345 ± 0.013	[0.319;0.372]	0.298	-0.015
$\lambda_{\text{C} \leftarrow \text{S}}$	0.563 ± 0.021	[0.522;0.604]	-0.491	0.008
$\lambda_{\text{De} \leftarrow \text{S}}$	0.411 ± 0.014	[0.383;0.440]	0.582	0.014
$\lambda_{\text{GtS} \leftarrow \text{S}}$	-0.042 ± 0.010	[-0.062;-0.023]	-1.101	-0.022
$\lambda_{\text{C} \leftarrow \text{R}}$	0.189 ± 0.027	[0.138;0.244]	-0.159	-0.017
$\lambda_{\text{GtS} \leftarrow \text{R}}$	0.947 ± 0.018	[0.911;0.982]	0.069	0.035
$\lambda_{\text{R} \leftarrow \text{De}}$	0.724 ± 0.012	[0.700;0.747]	0.949	-0.005
$\lambda_{\text{GtS} \leftarrow \text{De}}$	1.017 ± 0.017	[0.984;1.049]	0.632	0.039

Source: Bussiman (2021). Abbreviations: Di = Dissociation; C = Comfort; S = Style; R = Regularity; De = Development; GtS = Gait total Scores.

Table 7 - Genetic parameter estimates for all studied traits from Structural Equation Model 01 (Figure 2). Heritability (bold diagonal); genetic correlations (above diagonal); and residual correlations were assumed to be zero

Traits	Traits					
	Di	C	S	R	De	GtS
Di	0.51 ± 0.03 [0.45;0.56]	0.74 ± 0.12 [0.5;0.88]	0.70 ± 0.03 [0.65;0.75]	0.48 ± 0.07 [0.35;0.60]	0.48 ± 0.05 [0.37;0.59]	0.64 ± 0.06 [0.48;0.71]
C		0.23 ± 0.19 [0.09;0.62]	0.20 ± 0.06 [0.10;0.31]	0.17 ± 0.14 [-0.08;0.41]	0.42 ± 0.23 [0.00;0.71]	0.09 ± 0.19 [-0.31;0.29]
S			0.71 ± 0.01 [0.70;0.72]	0.46 ± 0.06 [0.33;0.58]	0.07 ± 0.05 [-0.04;0.18]	0.98 ± 0.03 [0.91;1.00]
R				0.27 ± 0.04 [0.19;0.34]	-0.41 ± 0.12 [-0.62;-0.18]	0.46 ± 0.06 [0.33;0.58]
De					0.25 ± 0.04 [0.17;0.32]	0.06 ± 0.06 [-0.06;0.17]
GtS						0.71 ± 0.01 [0.70;0.73]

Source: Bussiman (2021). Abbreviations: Di = Dissociation; C = Comfort; S = Style; R = Regularity; De = Development; GtS = Gait total Scores.

Table 8 - Estimated technician effects over all studied traits from Structural Equation Model 01 (Figure 2). The proportion of phenotypic variance explained by technician (diagonal bold) and correlation between technician effects among traits (above diagonal)

Traits	Traits					
	Di	C	S	R	De	GtS
Di	0.24 ± 0.01 [0.22;0.25]	0.33 ± 0.01 [0.32;0.35]	0.83 ± 0.00 [0.82;0.83]	0.67 ± 0.00 [0.66;0.68]	0.58 ± 0.01 [0.57;0.59]	0.76 ± 0.00 [0.75;0.77]
C		0.37 ± 0.03 [0.31;0.40]	0.01 ± 0.01 [-0.01;0.02]	0.15 ± 0.01 [0.13;0.16]	0.12 ± 0.01 [0.10;0.13]	-0.09 ± 0.01 [-0.10;-0.07]
S			0.28 ± 0.01 [0.27;0.29]	0.40 ± 0.01 [0.39;0.41]	0.37 ± 0.01 [0.35;0.38]	0.98 ± 0.00 [0.98;0.98]
R				0.13 ± 0.00 [0.12;0.13]	0.78 ± 0.00 [0.78;0.79]	0.36 ± 0.01 [0.34;0.37]
De					0.25 ± 0.01 [0.24;0.26]	0.34 ± 0.01 [0.32;0.35]
GtS						0.28 ± 0.01 [0.27;0.29]

Source: Bussiman (2021). Abbreviations: Di = Dissociation; C = Comfort; S = Style; R = Regularity; De = Development; GtS = Gait total Scores.

6.4. Discussion

The found (co)variance components and genetic parameters suggest the Campolina population could be selected for gait scores, even under such high subjectivity of the phenotypic process (as evidenced by the high technician effects). Despite subjectivity, functional traits are commonly highly affected by environmental forces in horses (ARNASON; VAN VLECK, 2000; ÁRNASON, 2013), such environmental effects can be attributed to the rider (VELIE; HAMILTON; WADE, 2015; BARTOLOMÉ et al., 2018; CERVANTES et al., 2020); appraiser/technician (RUSTIN et al., 2009; DE OLIVEIRA BUSSIMAN et al., 2018); competition/event (RICARD; LEGARRA, 2010; CERVANTES et al., 2020); or even other horses if the trait is a competition trait (ARNASON; VAN VLECK, 2000; RICARD; LEGARRA, 2010; CERVANTES et al., 2020; VARONA; LEGARRA, 2020).

The challenge for gait/performance selection on horses is then to overcome such subjectivity or environmental impacts. Therefore, new phenotyping strategies, such as kinematic (MOLINA et al., 2008; SOLÉ et al., 2014); accelerometry (ROBILLIARD; PFAU; WILSON, 2007); or inertial measurement units (SERRA BRAGANÇA et al., 2020); for example, could benefit the process. De Oliveira Bussiman et al. (2018) first addressed the problem of high technician effect over GtS. Those authors stated that frequent training could help to decrease the proportion of phenotypic variance explained by the technician, which was

further briefly investigated by de Oliveira Bussiman et al. (2020), who estimated correlations between technician effects across the same traits as in this study. Those authors found a negative correlation between technician effect over C and S, and C and Di. In our study all technician correlations were positive, differing from the previous analysis (DE OLIVEIRA BUSSIMAN et al., 2020), but those authors did not include GtS in the MTM, which can explain the different (co)variance components.

The technician correlations ranged from 0.31 (C with S) to 0.98 (R with De), showing that the way scores are given is not uniformly across traits. Such reasoning can be pointed because the Campolina horse can perform two types of gaits: *marcha batida* or *marcha picada* (ABCCCAMPOLINA, 2018; NASCIMENTO et al., 2019). The scores may then be conditioned to the gait type, and the sequence in which each technician gave the scores may influence the direction of those correlations. Not surprising, the correlations between Di, C, S, R, and De with GtS were among the highest ones, indicating the technician effect over the final scores is highly dependent on this effect over each partial score. This also makes sense because GtS is obtained by summing all other traits.

Studying how those traits (Di, C, S, R, and De) could be better combined, de Oliveira Bussiman et al. (2020) found that combination by principal component analysis, as proposed by Buzanskas et al. (2013) and Solar Diaz et al. (2020), could lead to correlated response on GtS higher than direct response (1.39 against 1.37, respectively). This supports the causal network found in this study because the overall genetic merit of GtS can be predicted by a combination of all other traits as stated by Valente et al. (2013) under the SEM approach. Although, for selection purposes, it is not necessary to know the causal links between phenotypes (VALENTE et al., 2011, 2013).

SEM is very powerful in cases where interventions are made in the selection process (GIANOLA; SORENSEN, 2004; VALENTE et al., 2013; VALENTE; ROSA, 2013), they can better accommodate selection decisions (DE LOS CAMPOS; GIANOLA; HERINGSTAD, 2006; ROSA et al., 2011) because traditional MTM resolves the selection bias only when there are traits measured before and after selection in a given trait (MEYER, 1995; BOLIGON et al., 2009), while under an SEM the selection over a given trait directly affects other traits through the causal effects. Thus, SEM is the only model that can solve the prediction when a given trait is fixed (VALENTE et al., 2013), which can be related to the restriction made by ABCCCampolina for registration, in which there is an acceptable range for each partial score (e.g., if a horse has 0, it is not registered and not included in the dataset, so the phenotypic range for GtS is completely dependent of it).

It must be noticed that despite changes in the HPD content (between 70% and 95%) the result from the IC algorithm was the same. This result is unexpected because when a short interval is applied more links are recovered (VALENTE et al., 2010, 2011; INOUE et al., 2016), but in our analysis, the presence of GtS being the sum of all other traits can make the network overidentified, even though most of the links were undirected. Cavani et al. (2020) stated that is essential to find unshielded colliders for edge orientation; thus, the IC algorithm resulted in oriented edges towards GtS that is caused by other traits, which has mathematical support, since GtS is the sum of Di, S, C, R, and De.

Posterior means from (co)variance components from MTM and SEM were different, even for the best model (SEM01) which may be related to the fact that in SEM genetic effects represent only the direct effects while in MTM those effects are overall (direct and indirect) (VALENTE et al., 2013; CAVANI et al., 2020) therefore they are not directly comparable (BOUWMAN et al., 2014; INOUE et al., 2016). Taking the equivalence between SEM and MTM (GIANOLA; SORENSEN, 2004; VALENTE et al., 2010, 2013; VALENTE; ROSA, 2013) one can estimate the (co)variance matrices in the MTM scale and compare them. In this study, the additive genetic variance increased when the SEM was fitted, mainly because the technician effect decreased, and residual variances were similar. This may be related to indirect technician effects in a given trait passing to another through the causal network, reducing the variability originally captured by the MTM. In addition, since the IC algorithm was applied to the residual (co)variance matrix, SEM and MTM may be equivalent at the residual level (VALENTE et al., 2010; CHITAKASEMPORNKUL et al., 2019).

Commonly, horse gaits are classified based on kinematics (CLAYTON; HOBBS, 2019) which is performed by Video analysis or visual evaluation. There are several different types of intermediate gaits which differ in terms of support types; steps frequency; speed or even at metabolic level (NICODEMUS; CLAYTON, 2003; WANDERLEY et al., 2010). This study is not concerned with gait classification but with selection for quality attributes of gait. Quality is a very subjective and specific aspect of those traits because different breeders may have different preferences also, thus the use of the found causal network should allow breeders to better select under every specific condition. For example, suppose a breeder wants a very specific gait, let's say with maximum values for C. The traditional MTM only says that given the genetic correlation between C and GtS, horses with high values for GtS are expected to have also high values for C. But under the SEM approach, this breeder can easily compute expected phenotypes for GtS, given the causal structure, when C is at a fixed value, and exactly rank the animals, finding those that should have maximum values for C.

At least in terms of kinematics, dissociation arose from the asynchrony between timing limb supports (HOBBS; BERTRAM; CLAYTON, 2016). At moderate speeds, trotting horses use temporal dissociation; adjustment of hoof contact position (relative to the center of mass); and alteration of the distribution of vertical forces between fore- and hindlimbs concurrently loaded, to achieve balance (HOBBS; BERTRAM; CLAYTON, 2016; CLAYTON; HOBBS, 2017), which may apply to gaited horses by analogy. In addition, different dissociation patterns were attributed to differences in body motions, that are resultant of body/limb proportion and mass distribution (HOBBS; BERTRAM; CLAYTON, 2016). Thus, it is perfectly plausible that different posture, balance, and harmony of movements (S) will cause more, or less dissociation, as depicted in SEM01 (Figure 2 and Table 6).

In terms of energy, more steps per stride are more efficient (RUINA; BERTRAM; SRINIVASAN, 2005). An increase in steps frequency would decrease the ability of the horse to cover long distances with fewer movements; thus, impacting its rhythm. The rhythm of a gait is influenced by the sequence of steps, which would be influenced by its frequency (NICODEMUS; CLAYTON, 2003; CLAYTON; HOBBS, 2019). This can explain the causal effect of De over R because De reflects step frequency while R reflects stability and rhythm (ABCCCAMPOLINA, 2018). Concerning stability, the trunk will be stable when the center of mass is stable (HOBBS; BERTRAM; CLAYTON, 2016) and the center of mass is more sensitive to vertical forces (HOBBS; RICHARDS; CLAYTON, 2014); thus, it is plausible that oscillations on gait stability/rhythm will affect the comfort (C) to the rider because the vertical forces will be transferred from the center of mass to the rider.

The force balance during movement is highly dependent on the horse posture (ALEXANDER, 1984; CLAYTON; HOBBS, 2019), for example, head and trunk work together to maintain spatial body orientation (DUNBAR et al., 2008); *longissimus dorsi* and *rectus abdominis* muscles activation increase spinal stiffness on trotting horses (ROBERT et al., 2002), which may occur on gaited horses as well. The most commercial appeal of gaited horses in Brazil is the ability of the horse to carry the rider comfortably, thus trunk stability is a must for Brazilian gaited horses. Trunk stability in terms of vertical, lateral, and frontal oscillations is related to the center of mass movements (CLAYTON; HOBBS, 2017), which is influenced by horse posture (HOBBS; BERTRAM; CLAYTON, 2016), conformation (HOBBS; RICHARDS; CLAYTON, 2014), and support patterns (MCGHEE; FRANK, 1968). Finally, support patterns are influenced by the steps frequency (CRUZ et al., 2018) as well as the speed of movement (NICODEMUS; CLAYTON, 2003); thus, the causal path among S, De,

R, and C, can be explained by those complex mechanisms of body-posture-movement control during the gait execution.

The IC algorithm performs the search for causal structures using partial correlations (VALENTE et al., 2010); thus, the causal meaning of resulted graphs is limited to their values. If two variables are correlated, this does not mean one causes another, but they may share a “parent”; for instance, a third variable affects them both (PEARL, 2009). This can explain the negative direct effect of S on GtS because S affects Di directly, and the partial correlation between Di and GtS is substituted by those effects together. This suggests that on the genetic level those traits are correlated, but conditional on those genetic effects they are d-separated (PEARL, 1988, 2009; VALENTE et al., 2010).

Causal statistical models are an unperfect translation of causal hypothesis (PEÑAGARICANO et al., 2015), and the imperfection arose from the replacement of asymmetrical relationships by symmetrical, involving probability distributions (SHIPLEY, 2003). Consequently, different models can be equivalent, implying the same (co)variance matrices and providing similar goodness-of-fit values (BROWN, 2015; PEÑAGARICANO et al., 2015; SCHUMACKER; LOMAX, 2015). It is important to address that none of the tested SEM were statistically equivalent, thus, the final model was chosen only based on DIC value, this allows to better rank non-nested models by balancing model adequacy and complexity (SPIEGELHALTER et al., 2002).

Here it was applied the method developed by Valente et al. (2010) to search for the causal relationship among studied traits. Such a method was considered a powerful tool without proper prior knowledge or if there is an uncertain relationship among variables (INOUE, 2020). But there are some limitations, mainly about the assumptions made to search for causal links: Markov condition, faithfulness, and causal sufficiency (PEARL, 2009); those are not testable from data and are paramount even under randomized experiments (CHITAKASEMPORNKUL et al., 2020), thus by working with observational data, as in this study, one should be aware that causal assumptions required to allow causal interpretation to estimated structural coefficients regardless observational or experimental data (BELLO; RENTER, 2018).

6.5. Conclusion

To the best of our knowledge, it is the first application of the inductive causation algorithm in horse gait traits. It was demonstrated how gait visual scores may interact to cause de overall gait in Campolina horses. It seems, from the results, that Style has a major effect on

gait, positively and directly affecting Development and Comfort; and positively and indirectly affecting Comfort through the causal effects of Development on Regularity, and of Regularity on Comfort. There is also a positive direct causal effect of Style on Dissociation, which can explain how Dissociation and Gait total Scores are correlated on the genetic level, but dissociated conditional on genetic effects. Finally, the negative direct causal effect of Style on Gait total Scores can be considered a side effect since the total effect of Style on Gait total Scores is positive. Those paths are supported by biological processes of biomechanics and kinematics of horse movements and may serve both scientists and breeders, to better understand gait on gaited horses.

6.6. References

- ABCCCAMPOLINA, Associação Brasileira dos Criadores do Cavalo Campolina. **Regulamento do Serviço de Registro Genealógico do cavalo Campolina – SRGCC**. Brazil: Ministério da Agricultura, Pesca e Abastecimento, 2018.
- ALEXANDER, R. M. The Gaits of Bipedal and Quadrupedal Animals. **The International Journal of Robotics Research**, v. 3, n. 2, p. 49–59, 1984.
- ÁRNASON, T. Breeding in horses. In: **Sustainable Food Production**. New York, NY: Springer New York, 2013. p. 401–416.
- ARNASON, T.; VAN VLECK, L. D. Genetic Improvement of the Horse. In: BOWLING, A. T.; RUVINSKY, A. (Eds.). **The genetics of the horse**. 1. ed. Oxford: CABI Publishing, 2000. p. 473–498.
- BARTOLOMÉ, E. et al. Plasticity effect of rider-horse interaction on genetic evaluations for Show Jumping discipline in sport horses. **Journal of Animal Breeding and Genetics**, v. 135, n. 2, p. 138–148, 2018.
- BELLO, N. M.; RENTER, D. G. Invited review: Reproducible research from noisy data: Revisiting key statistical principles for the animal sciences. **Journal of Dairy Science**, v. 101, n. 7, p. 5679–5701, 2018.
- BOLIGON, A. A. et al. Herdabilidades e correlações entre pesos do nascimento à idade adulta em rebanhos da raça Nelore. **Revista Brasileira de Zootecnia**, v. 38, n. 12, p. 2320–2326, 2009.
- BOUWMAN, A. C. et al. Exploring causal networks of bovine milk fatty acids in a multivariate mixed model context. **Genetics Selection Evolution**, v. 46, n. 1, p. 2, 2014.
- BROWN, T. A. **Confirmatory Factor Analysis for Applied Research**. 2. ed. New York: The Guilford Press, 2015.
- BUZANSKAS, M. E. et al. Genetic parameter estimates and principal component analysis of

breeding values of reproduction and growth traits in female Canchim cattle. **Reproduction, Fertility and Development**, v. 25, n. 5, p. 775, 2013.

CAVANI, L. et al. Inferring phenotypic causal networks for tick infestation, *Babesia bovis* infection, and weight gain in Hereford and Braford cattle using structural equation models. **Livestock Science**, v. 238, p. 104032, 2020.

CERVANTES, I. et al. Combining Threshold, Thurstonian and Classical Linear Models in Horse Genetic Evaluations for Endurance Competitions. **Animals**, v. 10, n. 6, p. 1075, 2020.

CHITAKASEMPORNKUL, K. et al. Investigating causal biological relationships between reproductive performance traits in high-performing gilts and sows. **Journal of Animal Science**, v. 97, n. 6, p. 2385–2401, 2019.

CHITAKASEMPORNKUL, K. et al. Hierarchical Modeling of Structural Coefficients for Heterogeneous Networks with an Application to Animal Production Systems. **Journal of Agricultural, Biological and Environmental Statistics**, v. 25, n. 4, p. 1–22, 2020.

CLAYTON, H. M.; HOBBS, S. J. An exploration of strategies used by dressage horses to control moments around the center of mass when performing passage. **PeerJ**, v. 5, p. e3866, 2017.

CLAYTON; HOBBS. A Review of Biomechanical Gait Classification with Reference to Collected Trot, Passage and Piaffe in Dressage Horses. **Animals**, v. 9, n. 10, p. 763, 2019.

CRUZ, A. M. et al. Effect of trotting speed on kinematic variables measured by use of extremity-mounted inertial measurement units in nonlame horses performing controlled treadmill exercise. **American Journal of Veterinary Research**, v. 79, n. 2, p. 211–218, 2018.

DE LOS CAMPOS, G.; GIANOLA, D.; HERINGSTAD, B. A Structural Equation Model for Describing Relationships Between Somatic Cell Score and Milk Yield in First-Lactation Dairy Cows. **Journal of Dairy Science**, v. 89, n. 11, p. 4445–4455, 2006.

DE OLIVEIRA BUSSIMAN, F. et al. Genetic analysis of morphological and functional traits in Campolina horses using Bayesian multi-trait model. **Livestock Science**, v. 216, n. April, p. 119–129, 2018.

DE OLIVEIRA BUSSIMAN, F. et al. Análise de Componentes Principais e Resposta à Seleção em características de andamento em cavalos Campolina. In: XIV SIMPÓSIO DE PÓS-GRADUAÇÃO E PESQUISA EM NUTRIÇÃO E PRODUÇÃO ANIMAL, 14. **Anais...** Pirassununga: 5D Editora, 2020.

DUNBAR, D. C. et al. Stabilization and mobility of the head, neck and trunk in horses during overground locomotion: comparisons with humans and other primates. **Journal of Experimental Biology**, v. 211, n. 24, p. 3889–3907, 2008.

GEWEKE, J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. **Bayesian Statistics 4**, p. 169–193, 1992.

GIANOLA, D.; SORENSEN, D. Quantitative Genetic Models for Describing Simultaneous and Recursive Relationships Between Phenotypes. **Genetics**, v. 167, n. 3, p. 1407–1424,

2004.

HAAVELMO, T. The Statistical Implications of a System of Simultaneous Equations. **Econometrica**, v. 11, n. 1, p. 1, 1943.

HEIDELBERGER, P.; WELCH, P. D. Simulation Run Length Control in the Presence of an Initial Transient. **Operations Research**, v. 31, n. 6, p. 1109–1144, 1983.

HENDERSON, C. R. Rapid method for computing the inverse of a relationship matrix. **Journal of Dairy Science**, v. 58, n. 11, p. 1727–1730, 1975.

HOBBS, S. J.; BERTRAM, J. E. A.; CLAYTON, H. M. An exploration of the influence of diagonal dissociation and moderate changes in speed on locomotor parameters in trotting horses. **PeerJ**, v. 4, p. e2190, 2016.

HOBBS, S. J.; RICHARDS, J.; CLAYTON, H. M. The effect of centre of mass location on sagittal plane moments around the centre of mass in trotting horses. **Journal of Biomechanics**, v. 47, n. 6, p. 1278–1286, 2014.

INOUE, K. et al. Inferring phenotypic causal structures among meat quality traits and the application of a structural equation model in Japanese Black cattle1. **Journal of Animal Science**, v. 94, n. 10, p. 4133–4142, 2016.

INOUE, K. Application of Bayesian causal inference and structural equation model to animal breeding. **Animal Science Journal**, v. 91, n. 1, 2020.

MCGHEE, R.; FRANK, A. On the stability properties of quadruped creeping gaits. **Bellman Prize in Mathematical Biosciences**, v. 3, p. 331–351, 1968.

MEYER, K. Estimates of genetic parameters for mature weight of Australian beef cows and its relationship to early growth and skeletal measures. **Livestock Production Science**, v. 44, n. 2, p. 125–137, 1995.

MEYER, K. Genetic principal components for live ultrasound scan traits of Angus cattle. **Animal Science**, v. 81, n. 3, p. 337–345, 2005.

MISZTAL, I. et al. BLUPF90 and related programs (BGF90). In: 7th WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 7. **Proceeding...** Montpellier: Wageningen University, 2002.

MOLINA, A. et al. Genetic parameters of morphofunctional traits in Andalusian horse. **Livestock Production Science**, v. 60, n. 2–3, p. 295–303, 1999.

MOLINA, A. et al. Genetic parameters of biokinematic variables at walk in the Spanish Purebred (Andalusian) horse using experimental treadmill records. **Livestock Science**, v. 116, n. 1–3, p. 137–145, 2008.

NASCIMENTO, C. A. DE M. S. et al. Participation frequency and performance of horses in national shows of Campolina and Mangalarga Marchador breeds. **Revista Brasileira de Zootecnia**, v. 48, 2019.

NICODEMUS, M. C.; CLAYTON, H. M. Temporal variables of four-beat, stepping gaits of

gaited horses. **Applied Animal Behaviour Science**, v. 80, n. 1, p. 133–142, 2003.

PEARL, J. **Probabilistic Reasoning in Intelligent Systems: networks of plausible inference**. 2. ed. San Francisco: Morgan Kaufmann Publishers, INC., 1988.

PEARL, J. **Causality: Models, Reasoning and Inference**. 2. ed. Cambridge: Cambridge University Press, 2009.

PEÑAGARICANO, F. et al. Searching for causal networks involving latent variables in complex traits: Application to growth, carcass, and meat quality traits in pigs¹. **Journal of Animal Science**, v. 93, n. 10, p. 4617–4623, 2015.

PROCÓPIO, A. M. M.; BERGMANN, J. A. G. A. G.; COSTA, M. D. D. Formação e demografia da raça Campolina. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 55, n. 3, p. 361–365, 2003.

R CORE TEAM. **R: A Language and Environment for Statistical Computing** (R. D. C. Team, Ed.) R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. 2020.

RAFTERY, A. E.; LEWIS, S. How many iterations in the Gibbs sampler? **Bayesian Statistics**, p. 763--773, 1992.

RICARD, A.; LEGARRA, A. Validation of models for analysis of ranks in horse breeding evaluation. **Genetics Selection Evolution**, v. 42, n. 1, p. 3, 2010.

ROBERT, C. et al. Effects of trotting speed on muscle activity and kinematics in saddlehorses. **Equine Veterinary Journal**, v. 34, n. S34, p. 295–301, 2002.

ROBILLIARD, J. J.; PFAU, T.; WILSON, A. M. Gait characterisation and classification in horses. **Journal of Experimental Biology**, v. 210, n. 2, p. 187–197, 2007.

ROSA, G. J. M. et al. Inferring causal phenotype networks using structural equation models. **Genetics Selection Evolution**, v. 43, n. 1, p. 6, 2011.

ROSA, G. J. M.; VALENTE, B. D. Structural Equation Models for Studying Causal Phenotype Networks in Quantitative Genetics. In: **Probabilistic Graphical Models for Genetics, Genomics, and Postgenomics**. Oxford University Press, 2014. p. 196–214.

RUINA, A.; BERTRAM, J. E. A.; SRINIVASAN, M. A collisional model of the energetic cost of support work qualitatively explains leg sequencing in walking and galloping, pseudo-elastic leg behavior in running and the walk-to-run transition. **Journal of Theoretical Biology**, v. 237, n. 2, p. 170–192, 2005.

RUSTIN, M. et al. Multi-trait animal model estimation of genetic parameters for linear type and gait traits in the Belgian warmblood horse. **Journal of Animal Breeding and Genetics**, v. 126, n. 5, p. 378–386, 2009.

SCHUMACKER, R. E.; LOMAX, R. G. **A Beginner's guide to Structural Equation Modeling**. 4. ed. New York: Routledge, 2015.

SERRA BRAGANÇA, F. M. et al. Improving gait classification in horses by using inertial

measurement unit (IMU) generated data and machine learning. **Scientific Reports**, v. 10, n. 1, p. 17785, 2020.

SHIPLEY, B. From biological hypotheses to structural equation models: the imperfection of causal translation. In: PUGESEK, B.; TOMER, A.; VON EYE, A. (Eds.) **Structural Equation Modeling**. Cambridge: Cambridge University Press, 2003. p. 194–211.

SOLAR DIAZ, I. D. P. et al. Principal components for morphometric traits in Campolina horses. **Journal of Animal Breeding and Genetics**, v. 00, p. 1–9, 2020.

SOLÉ, M. et al. Genetic analysis of kinematic traits at the trot in Lusitano horse subpopulations with different types of training. **Animal**, v. 8, n. 2, p. 192–199, 2014.

SPIEGELHALTER, D. J. et al. Bayesian measures of model complexity and fit. **Journal of the Royal Statistical Society. Series B: Statistical Methodology**, v. 64, n. 4, p. 583–616, 2002.

SPIEGELHALTER, D. J. et al. The deviance information criterion: 12 years on. **Journal of the Royal Statistical Society. Series B: Statistical Methodology**, v. 76, n. 3, p. 485–493, 2014.

SPIRITES, P.; GLYMOUR, C.; SCHEINES, R. **Causation, Prediction, and Search**. 2. ed. Cambridge, Massachusetts, EUA: MIT Press, 2001.

THOMPSON, R.; MEYER, K. A review of theoretical aspects in the estimation of breeding values for multi-trait selection. **Livestock Production Science**, v. 15, n. 4, p. 299–313, 1986.

VALENTE, B. D. et al. Searching for Recursive Causal Structures in Multivariate Quantitative Genetics Mixed Models. **Genetics**, v. 185, n. 2, p. 633 LP – 644, 2010.

VALENTE, B. D. et al. Searching for phenotypic causal networks involving complex traits: an application to European quail. **Genetics Selection Evolution**, v. 43, n. 1, p. 37, 2011.

VALENTE, B. D. et al. Is Structural Equation Modeling Advantageous for the Genetic Improvement of Multiple Traits? **Genetics**, v. 194, n. 3, p. 561–572, 2013.

VALENTE, B. D.; ROSA, G. J. DE M. Mixed Effects Structural Equation Models and Phenotypic Causal Networks. In: GONDRO, C.; VAN DER WERF, J.; HAYES, B. (Eds.) **Genome-Wide Association Studies and Genomic Prediction**. Totowa, NJ: Humana Press, 2013. p. 449–464.

VAN DER WERF, J. H. J.; VAN ARENDONK, J. A. M.; DE VRIES, A. G. Improving selection of pigs using correlated characters. In: 43th ANNUAL MEETING OF EUROPEAN FEDERATION OF ANIMAL SCIENCE, 43. **Proceedings...** Madrid: Wageningen Academic Publishers, 1992.

VARONA, L.; LEGARRA, A. GIBBSTHUR: Software for Estimating Variance Components and Predicting Breeding Values for Ranking Traits Based on a Thurstonian Model. **Animals**, v. 10, n. 6, p. 1001, 2020.

VELIE, B. D.; HAMILTON, N. A.; WADE, C. M. Heritability of racing performance in the Australian Thoroughbred racing population. **Animal Genetics**, v. 46, n. 1, p. 23–29, 2015.

VERMA, T. S.; PEARL, J. **Equivalence and synthesis of causal models**. In: 6th CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE, 6. **Proceedings...** Cambridge: Elsevier Science Inc., 1990.

VICENTE, A. A. et al. Selection for morphology, gaits and functional traits in Lusitano horses: I. Genetic parameter estimates. **Livestock Science**, v. 164, n. 1, p. 1–12, 2014a.

VICENTE, A. A. et al. Selection for morphology, gaits and functional traits in Lusitano horses: II. Fixed effects, genetic trends and selection in retrospect. **Livestock Science**, v. 164, n. 1, p. 13–25, 2014b.

VOSTRY, L. et al. Genetic Parameters for Linear Type Traits in Three Czech Draught Horse Breeds. **Agriculturae Conspectus Scientificus**, v. 82, n. 2, p. 111–115, 2017.

WANDERLEY, E. K. et al. Metabolic changes in four beat gaited horses after field marcha simulation. **Equine Veterinary Journal**, v. 42, n. SUPPL. 38, p. 105–109, 2010.

WRIGHT, S. Correlation and Causation. **Journal of Agriculture Research**, v. 201, p. 557–585, 1921.

WRIGHT, S. Coefficients of Inbreeding and Relationship. **The American Naturalist**, v. 56, n. 645, p. 330–338, 1922.

7. EVALUATING HOW TO IMPLEMENT GENOMIC SELECTION IN A NON-GENOTYPED POPULATION USING SINGLE-STEP METHOD COMBINED WITH KERNEL REGRESSIONS: A SIMULATION STUDY

*“Do you know calculus? [...] You had better learn it [...]. It's the language God talks”
(Richard Feynman)¹⁵.*

F. de Oliveira Bussiman¹; D.A.L. Lourenço²; R.V. Ventura¹; F.F. Silva (*in memoriam*)³;
R.S.B. Carvalho⁴; E.C. Mattos⁵; J.B.S. Ferraz⁵; J.P. Eler⁵; J.C.C. Balieiro¹

¹*Bioinformatic and Animal Breeding Lab., Department of Animal Nutrition and Production, College of Veterinary Medicine and Animal Science University of São Paulo (BIOMA-VNP/FMVZ-USP), Av. Duque de Caxias Norte, 225, 13.635-900, Pirassununga-SP, Brazil;*

²*Animal Breeding and Genetics Group, Department of Animal and Dairy Science, University of Georgia, E.L. Rhodes Center for Animal and Dairy Science, 425 River Road, Athens, GA 30602 0000;*

³*Department of Animal Science, Federal University of Viçosa (DZO/UFV), Av. P H Rolfs, s/n, 36.750-900, Viçosa-MG, Brazil;*

⁴*Department of Basic Sciences, College of Animal Science and Food Engineering, University of São Paulo (ZAB/FZEA-USP), Av. Duque de Caxias Norte, 225, 13.635-900, Pirassununga-SP, Brazil;*

⁵*Group of Animal Breeding and Biotechnology, Department of Veterinary Medicine, College of Animal Science and Food Engineering, University of São Paulo (GMAB-ZMV/FZEA-USP), Av. Duque de Caxias Norte, 225, 13.635-900, Pirassununga-SP, Brazil.*

RESUMO: Diferentes estratégias de genotipagem resultam em diferentes matrizes de relacionamento genômico e por isso talvez afetem as predições genômicas. Antes de investir dinheiro para genotipar os indivíduos, é melhor saber, se possível, qual a melhor estratégia para amostrarem-se os candidatos à genotipagem. Este estudo foi concebido para avaliar estratégias de genotipagem e modelos estatísticos para seleção genômica usando o método de passo único em uma população não genotipada de cavalos. Três cenários diferentes em termos de estrutura populacional e endogamia foram simulados, assim como, quatro características quantitativas, adotando cinco replicatas. O genoma simulado foi composto de 461.523 marcadores designados aleatoriamente a 31 cromossomos autossomos. O desequilíbrio de ligação simulado está de acordo com outras populações de cavalos, assim como, a média da frequência alélica do alelo menor por cromossomo. Foram testadas duas proporções de animais genotipados, 10% e 50% dos candidatos à genotipagem disponíveis. No que tange as estratégias de genotipagem, foram testadas seis formas de se amostrar os animais: aleatoriamente; maior e menor valor genético verdadeiro (TBV); maior e menor valor fenotípico (P); maior TBV; maior P; e maior número de progênes (Top Pr). Os modelos testados foram GBLUP; ssGBLUP; RKHS (com dois *kernels*) e, a combinação do método de passo único com RKHS (com dois *kernels*). Houve um efeito significativo da interação entre característica e cenário, assim como da interação entre característica e estratégia de genotipagem sobre a acurácia de predição, enquanto sobre a proporção de viés, o único efeito significativo foi a estratégia de genotipagem. A melhor estratégia de genotipagem foi Top Pr, na qual foi encontrada a maior acurácia de predição

¹⁵ WOUK, H. **The language God talks:** on science and religion. 1. ed. New York: Little, Brown and Company, 2010.

juntamente com os menores viés e raiz do erro quadrático médio. No tocante aos modelos, os métodos RKHS foram os mais acurados e a combinação ssRKHS funcionou melhor para características categóricas. Para implementar a seleção genômica no futuro, os criadores de cavalos Campolina devem genotipar os cavalos com o maior número de progênies, o que beneficiará as predições com alta acurácia e baixo viés, aplicando métodos ssRKHS, quando possível, para as predições.

Palavras-chave: estratégias de genotipagem; seleção genômica; melhoramento de cavalos; genotipagem seletiva; genômica.

ABSTRACT: Different genotyping strategies result in different genomic relationship matrices and thus may affect genomic predictions. Before investing money to genotype individuals it is better to know, if possible, the best way to sample genotype candidates. This study was designed to evaluate genotyping strategies and statistical models for genomic selection using the single-step method in a non-genotyped horse population. Three different scenarios in terms of population structure and inbreeding were simulated, as well as four quantitative traits, adopting five replicates. The simulated genome was comprised of 461,523 markers randomly assigned to 31 autosomes. Simulated linkage disequilibrium was in the range for horse populations as well as the average minor allele frequencies per chromosome. Were tested two proportions of genotyped animals, 10% and 50% percent of available genotyping candidates. Concerning genotyping strategy there were tested six sampling ways: randomly; top and bottom true breeding value (TBV); top and bottom phenotypic value (P); top TBV; top P; and highest progeny number. The tested models were GBLUP; ssGBLUP; RKHS (with two kernels); and the combination of the single-step method with RKHS (ssRKHS, with two kernels). There was an effect of the interaction between trait and scenario, as well as the interaction between trait and genotyping strategy over the prediction reliability, while over bias proportion the only significant effect was genotyping strategy. The best genotyping strategy was Top Pr, which was found the highest reliability along with the lowest bias and root mean squared error. Concerning models, the most accurate predictions were found by RKHS methods, and the ssRKHS method worked better for the categorical outcome. To perform genomic selection in the future, the Campolina breeders should sample the horses with the highest progeny number, which will benefit predictions by high reliability and low bias, applying ssRKHS methods, when possible, for predictions.

Keywords: genotyping strategy; genomic selection; horse breeding; selective genotyping; genomics.

7.1. Introduction

Prediction of breeding values (BV) depends on knowing genetic relationships among individuals, which allows estimation of the heritable proportion of phenotypic variation (CALUS, 2010). Since the derivation of the inverse relationship matrix, $\mathbf{A}^{-1}$ (HENDERSON, 1975), traditional additive kinship coefficients (WRIGHT, 1922) was widely used by animal breeders to predict BV. Aiming to better account for the mendelian sampling those coefficients were replaced by genomic relationships (NEJATI-JAVAREMI; SMITH; GIBSON, 1997;

VANRADEN, 2008) which culminated with the development of a new method combining genotyped and non-genotyped animals known as single-step (LEGARRA; AGUILAR; MISZTAL, 2009; MISZTAL; LEGARRA; AGUILAR, 2009).

The single-step method is an extension of traditional best linear unbiased prediction (BLUP - HENDERSON, 1949) replaced $\mathbf{A}^{-1}$ by $\mathbf{H}^{-1}$ (AGUILAR et al., 2010; CHRISTENSEN; LUND, 2010): thus, this method is often called single-step genomic best linear unbiased prediction (ssGBLUP). $\mathbf{H}^{-1}$ is the inverse of a combined relationship matrix in which non-genotyped animals have genomic relationships with others imputed through their relationship with genotyped animals (CHRISTENSEN; LUND, 2010; HSU; GARRICK; FERNANDO, 2017; LEGARRA; AGUILAR; MISZTAL, 2009).

This imputation can result in some bias on (co)variance components estimation (GAO et al., 2019) or the BV prediction (HSU; GARRICK; FERNANDO, 2017; VITEZICA et al., 2011) because models for genomic selection (MEUWISSEN; HAYES; GODDARD, 2001) assume implicitly a non-selected genotyped population (HAYES; VISSCHER; GODDARD, 2009), as well as traditional BLUP, assumes an unselected base population (VITEZICA et al., 2011), but genotyped animals are often highly selected, thus genomic and pedigree-based relationships are not directly compatible if there is selection undergoing.

Legarra et al. (2014) showed ssGBLUP is resistant to the preselection of genotyped animals, however, the BV accuracy still depends on the reference population (LOURENCO et al., 2015) because different genotyping strategies will result in distinct reference populations when ssGBLUP is performed since the genomic relationship matrix ($\mathbf{G}$; VANRADEN, 2008) will differ for each set of genotyped animals (CHEN et al., 2011). Thus, it is expected that genotyping different sets of animals will produce different prediction accuracy and level of bias (ANDONOV et al., 2017; PSZCZOLA et al., 2012).

Concerning horses, despite the number of genotyped animals is increasing, it still small; thus, ssGBLUP could be a powerful tool to implement genomic selection on this species (STOCK et al., 2016), since this method is considered to have optimal use of available information even when there is a small, genotyped part of the population (AGUILAR et al., 2010). Genotyping costs are decreasing nowadays, but it still is an obstacle to implementing genomic selection on horses (SITZENSTOCK et al., 2010). In general, most genomic studies on horses are mapping such as genome-wide association and fewer were applying single-step methods (DE OLIVEIRA BUSSIMAN et al., 2020a; STOCK et al., 2016).

Despite all promises of genomic selection, there is a lack of its application in the horse industry mainly concerning gaited horses in which recent studies have focused on mapping (ANDERSSON et al., 2012; DE OLIVEIRA BUSSIMAN et al., 2020a; FONSECA et al., 2017; JÄDERKVIST FEGRAEUS et al., 2017; MCCOY et al., 2019); population structure (DOS SANTOS et al., 2019; NOVOA-BRAVO; BERNAL-PINILLA; GARCÍA, 2021; SCHURINK et al., 2019); and selection signatures (LIU et al., 2018; SANTOS et al., 2021). Among those gaited horse breeds, there are at least two Brazilian horses: the Mangalarga Marchador and the Campolina, both known by the ability to perform four-beat gaits in either lateral (marcha picada gait) or diagonal (marcha batida gait) predominance of supports (DE OLIVEIRA BUSSIMAN et al., 2018b, 2019; MANSO FILHO et al., 2015; PATTERSON; STAIGER; BROOKS, 2015).

The Campolina breed was developed in Minas Gerais State (MG) around the year 1870 by using European stallions (such as Anglo-Norman; Clydesdale; Holsteiner; Percheron; and Thoroughbred) on crossed mares (Andalusian; Lusitano; and Berber) (PROCÓPIO; BERGMANN; COSTA, 2003). In 1951 the Campolina Horse Breeders Association was founded (ABCCCampolina; <http://www.campolina.org.br/portal/>) and the Campolina breeders started to select their horses based on competition results and phenotypic aspects. This breed is related to the historic-economic development of MG (VIEIRA et al., 2015) and has been exported to various countries receiving international horse industry attention.

However, this breed was never under a formal breeding program; and no study on genomic selection was carried out for BV prediction on Campolina horses. Given this lack of information, this study aimed to use simulation techniques to evaluate the better genotyping strategy to achieve high accuracy and low bias on genomic predictions through ssGBLUP on this population, as well as to evaluate statistical models (trait-dependent) to perform the genomic selection.

7.2. Material and Methods

First, the real pedigree of the Campolina breed (provided by the ABCCCampolina) was assessed through the CFC software (SARGOLZAEI; IWASAKI; COLLEAU, 2006). Based on the generation interval of this breed (9.00 ± 0.09 years - DE OLIVEIRA BUSSIMAN et al., 2018b) animals born after 1990 (which represents three times the generation interval) were considered as genotyping candidates. For those animals, the complete pedigree was extracted, and generation numbers were assigned based on birth order by using the “countGen” function from the R (R CORE TEAM, 2020) package “pedigree” (COSTER, 2012). For each generation

was calculated the number of founders (males and females) coming from all precedent generations and the number of foals per mare, as well as the average inbreeding coefficient. Then the effective population size was calculated as follows (GUTIÉRREZ et al., 2008; GUTIÉRREZ; CERVANTES; GOYACHE, 2009):

$$\overline{N_e} = \frac{1}{2\overline{\Delta F}}, \quad (7.1)$$

Where: $\overline{N_e}$ is the average effective population size across generations, and $\overline{\Delta F}$ is the average rate of inbreeding among generations. The rate of inbreeding for i^{th} generation can be calculated as:

$$\Delta F_i = \frac{\overline{F}_i - \overline{F}_{i-1}}{1 - \overline{F}_{i-1}}, \quad (7.2)$$

Where: $\overline{F}_i$ is the average inbreeding coefficient at i^{th} generation and $\overline{F}_{i-1}$ is the average inbreeding at the previous generation. Those parameters were used to simulate three scenarios using version 2.0 of the QMSim software (SARGOLZAEI; SCHENKEL, 2009); there were simulated four traits (Table 1) and three scenarios (depicted later).

Table 1 - Assumed variance components and genetic parameters of simulated traits. Additive genetic variance (σ_u^2); variance due to quantitative trait loci (σ_a^2); residual variance (σ_e^2); phenotypic variance (σ_p^2); heritability ($h^2 = (\sigma_u^2 + \sigma_a^2) / \sigma_p^2$); and quantitative trait loci heritability ($h_{QTL}^2 = \sigma_a^2 / \sigma_p^2$)

Trait	Assumed variance component				h^2	h_{QTL}^2
	σ_u^2	σ_a^2	σ_e^2	σ_p^2		
T1	1.00	1.57	4.58	7.15	0.36	0.22
T2	0.10	0.14	0.91	1.15	0.21	0.12
T3	0.02	0.01	1.09	1.12	0.03	0.01
T4	3.11	2.28	0.95	6.34	0.85	0.36

Source: Bussiman (2021).

7.2.1. Simulated scenarios

A historical population was created with a fixed effective population size of 1,000 animals by 500 generations. Hereafter, 500 historical generations were simulated, and the number of animals was gradually reduced to 186 animals. Finally, 20 historical generations were simulated to reduce the number of animals from 186 to 93, ensuring that mutation-drift equilibrium was established, and initial linkage disequilibrium (LD) was created. The number

of males and females was equal for the first 500 generations, then the number of males was gradually reduced until 46 at last historical generation (Figure 1).

Figure 1 - Graphical representation of the used simulation scheme

Source: Bussiman (2021).

The population was expanded by selecting all males and females in the last historical generation, through nine subsequent generations with, on average, five foals per mare and exponential growth of the number of mares, considering as mating design random union of gametes without selection. After it was selected, at last, expansion generation and with highest phenotypic value, 62 males and 411 females to form the founder population of the three simulated scenarios (Figure 1).

For the first scenario, it was considered each generation from the real Campolina pedigree as a single simulated population, which allowed to pick the exact number of males and females from previous populations as in the real situation (Table 2). This was used to create a very complex pedigree structure. Every single simulated population was mated by two generations, and only animals from the second generation were selected to be founders of the next population.

Table 2 - Number of founders coming from each generation to form every single simulated population (representing generation on the real pedigree)

Simulated Population	Sex	Founder Population												
		0	1	2	3	4	5	6	7	8	9	10	11	12
1	M	62												
2		13	114											
3		2	2	287										
4		1	2	4	397									
5		6	2	4	10	459								
6		6		2	6	11	520							
7		4		3	2	6	15	519						
8		1			1	2	5	18	455					
9		2					5	10	36	396				
10						1	1	2	8	9	15	331		
11									1	4	7	8	217	
12										2	4	3	1	62
13											1	1	1	3
1	F	411												
2		22	927											
3		14	153	2169										
4		9	117	349	2547									
5		5	57	209	448	2536								
6		5	27	142	319	648	2267							
7		5	16	100	260	469	552	1747						
8		1	15	27	121	310	311	522	1405					
9				8	38	187	163	359	485	929				
10					8	48	82	202	247	297	719			
11					3	17	24	61	101	153	179	488		
12						4	7	18	20	48	50	59	74	
13									1	3	2	1	7	3

Source: Bussiman (2021).

Concerning founders, those were selected through the highest phenotypic value and the mating design was based on the maximum inbreeding coefficient of the foal. After 14 populations the first scenario was concluded (Figure 1), and animals from the first and second generation within each population were selected to form the final simulated population, for which the pedigree was then compared with the real Campolina pedigree (Table 3). In addition, the distribution of horses within generations and per inbreeding class is presented in Table 4.

Table 3 - The pedigree structure of the simulated population compared to the real Campolina population. Total number of animals (N); the number of founders (NF); the number of stallions (NS); the number of mares (NM); average inbreeding coefficient ($\bar{F}$; %); average relationship coefficient ($\bar{f}$; excluding self-relationships); the number of full-sib groups (FS groups); average family size (Family size); and effective population size, based on the rate of inbreeding ($N_e \Delta F$)

Populational Parameter	Population			
	Real	S1 ¹	S2 ¹	S3 ¹
N	81,803	82,643.05 $\pm$ 379.09	9,162.30 $\pm$ 62.11	9148.25 $\pm$ 61.93
NF	7,853	2,043.65 $\pm$ 58.82	499.25 $\pm$ 9.83	499.25 $\pm$ 9.83
Progeny	17,239	2,258.00 $\pm$ 0.00	473.00 $\pm$ 0.00	473.00 $\pm$ 0.00
NS	4,061	3,847.85 $\pm$ 0.49	352.55 $\pm$ 3.61	352.55 $\pm$ 3.61
Progeny	73,950	80,599.40 $\pm$ 364.15	8,648.05 $\pm$ 72.64	8632.50 $\pm$ 73.81
NM	24,338	16,783.80 $\pm$ 58.53	1,399.70 $\pm$ 6.84	1399.70 $\pm$ 6.84
Progeny	73,950	80,599.40 $\pm$ 364.15	12,363.05 $\pm$ 61.70	86,29.00 $\pm$ 77.57
$\bar{F}$	2.97	0.69 $\pm$ 0.29	10.89 $\pm$ 5.13	0.63 $\pm$ 0.38
$\bar{f}$	0.03	0.01 $\pm$ 0.01	0.03 $\pm$ 0.02	0.02 $\pm$ 0.01
FS groups	13,259	17,313.50 $\pm$ 76.69	1,553.15 $\pm$ 85.71	1,524.55 $\pm$ 33.32
Family size	2.53	3.27 $\pm$ 0.01	3.20 $\pm$ 0.16	2.81 $\pm$ 0.04
$N_e \Delta F$	81.67	294.62 $\pm$ 213.46	36.46 $\pm$ 43.13	7,366.31 $\pm$ 1,174.91

Source: Bussiman (2021). ¹Values are averages across replicates presented as average $\pm$ standard deviation.

Table 4 - Distribution of animals within generations and per class of inbreeding, according to the simulated scenario

Value	Real	S1	S2	S3
Number of animals per generation				
0	7853	2043.65 ± 58.82	499.25 ± 9.83	499.25 ± 9.83
1	535	2258.00 ± 0.00	473.00 ± 0.00	473.00 ± 0.00
2	1598	967.80 ± 26.97	2751.20 ± 1441.74	2716.45 ± 1399.17
3	4154	3005.45 ± 103.30	1320.45 ± 516.50	1257.50 ± 471.81
4	6256	6996.20 ± 332.84	890.20 ± 140.77	838.35 ± 122.21
5	8520	10289.20 ± 648.73	652.35 ± 182.69	667.85 ± 145.78
6	10912	12363.60 ± 787.85	572.55 ± 272.45	612.10 ± 277.69
7	11800	12877.45 ± 814.06	506.75 ± 285.91	615.35 ± 383.45
8	10823	11849.45 ± 266.74	473.50 ± 321.66	514.65 ± 382.32
9	8555	9150.95 ± 421.14	386.45 ± 285.87	415.30 ± 357.71
10	6151	6174.00 ± 792.02	333.70 ± 293.22	281.95 ± 294.18
11	3887	3262.00 ± 770.81	220.25 ± 196.78	165.65 ± 213.27
12	769	1206.40 ± 493.18	146.35 ± 157.41	68.45 ± 105.11
13	40	192.20 ± 115.24	69.55 ± 84.68	19.40 ± 34.40
Number of animals per inbreeding class				
0.00 -- 0.05	28726	51172.05 ± 5069.26	426.25 ± 329.92	2311.55 ± 1909.66
0.05 -- 0.10	8493	934.85 ± 360.81	320.80 ± 151.07	168.30 ± 67.59
0.10 -- 0.15	3891	167.35 ± 52.34	414.00 ± 59.56	109.95 ± 23.94
0.15 -- 0.20	1647	5.60 ± 9.79	192.85 ± 98.64	--
0.20 -- 0.25	1036	16.65 ± 7.13	786.70 ± 161.86	--
0.25 -- 0.30	650	32.85 ± 15.83	182.15 ± 121.00	--
0.30 -- 0.35	136	0.15 ± 0.37	244.00 ± 121.12	--
0.35 -- 0.40	36	--	325.90 ± 131.53	--
0.40 -- 0.45	15	--	190.05 ± 121.07	--
0.45 -- 1.00	0	--	594.55 ± 516.17	--

Source: Bussiman (2021).

The second and third scenarios were comprised of 14 generations, in which the number of foals per mare varied from 1 to 5 (with probabilities from the frequency of mares with 1 to 5 foals from the real Campolina pedigree). The selection design was based on the highest phenotypic value and the culling design on the lowest. Stallion and mare replacement rate was constant over generations at 30% and 10%, respectively. Concerning mating design, for the second scenario, it was used to maximize inbreeding while for the third was to minimize it (Figure 1).

The simulated genome consisted of 4,615,230 markers and 1,000 quantitative trait loci (QTL) randomly distributed over 31 autosomal chromosomes, with length varying from 26 cM to 188 cM, thus identical to the real equine genome based on EquCab3 (KALBFLEISCH et al.,

2018) totaling 2,280 cM. Here was assumed different proportions of variance explained by the QTL for each trait, which was 22% (T1); 12 % (T2); 1% (T3); and 36% (T4). The number of makers and QTL per chromosome ranged between 53,080 and 449,840 and between 12 and 97, respectively. Markers were simulated bi-allelic while the QTL number of alleles per locus was randomly from 2 to 4, considering allele effects sampled from a gamma distribution with shape parameter equal to 0.4 (HAYES; GODDARD, 2001).

To establish mutation-drift equilibrium in historical generations, the genotyping error was assumed to be 0 while QTL and marker recurrent mutation rate were 10^{-4} . To generate genotypic data, after the historical population was created 461,523 segregating markers with minor allele frequency (MAF) higher or equal to 1% were randomly selected from the last historical generation. TBV every individual was the sum of QTL effects plus a polygenic residual effect (since QTL did not explain all phenotypic variance). True phenotypes were then generated by adding to the TBV residuals from a normal distribution with mean zero and variance one. All the simulation process was replicated five times. The simulated genome for each scenario is presented in Table 5.

Table 5 - Average number of markers (N); minor allele frequency (MAF); and linkage disequilibrium (LD), per chromosome (ECA) of the simulated genome across all replicates and scenarios

ECA	N	MAF	LD
1	44,879.80 $\pm$ 860.52	0.24 $\pm$ 0.15	0.24 $\pm$ 0.13
2	37,382.05 $\pm$ 1240.21	0.24 $\pm$ 0.15	0.24 $\pm$ 0.12
3	20,578.65 $\pm$ 434.21	0.24 $\pm$ 0.15	0.24 $\pm$ 0.13
4	18,801.50 $\pm$ 546.31	0.24 $\pm$ 0.15	0.24 $\pm$ 0.15
5	17,449.65 $\pm$ 435.86	0.24 $\pm$ 0.15	0.25 $\pm$ 0.16
6	31,950.30 $\pm$ 1298.27	0.24 $\pm$ 0.15	0.25 $\pm$ 0.19
7	16,576.00 $\pm$ 334.41	0.24 $\pm$ 0.15	0.25 $\pm$ 0.10
8	16,377.35 $\pm$ 560.33	0.24 $\pm$ 0.15	0.24 $\pm$ 0.17
9	14,736.50 $\pm$ 455.60	0.23 $\pm$ 0.15	0.25 $\pm$ 0.20
10	15,221.25 $\pm$ 437.49	0.23 $\pm$ 0.15	0.25 $\pm$ 0.19
11	10,953.05 $\pm$ 278.73	0.24 $\pm$ 0.15	0.25 $\pm$ 0.18
12	7,169.50 $\pm$ 332.41	0.24 $\pm$ 0.15	0.26 $\pm$ 0.19
13	8,339.45 $\pm$ 244.54	0.23 $\pm$ 0.15	0.26 $\pm$ 0.18
14	16,605.25 $\pm$ 474.05	0.24 $\pm$ 0.15	0.25 $\pm$ 0.14
15	16,176.45 $\pm$ 562.66	0.24 $\pm$ 0.15	0.25 $\pm$ 0.18
16	15,915.80 $\pm$ 655.62	0.24 $\pm$ 0.15	0.25 $\pm$ 0.15
17	14,061.65 $\pm$ 390.87	0.24 $\pm$ 0.15	0.25 $\pm$ 0.19
18	14,094.75 $\pm$ 374.48	0.24 $\pm$ 0.15	0.25 $\pm$ 0.25
19	10,826.50 $\pm$ 354.23	0.24 $\pm$ 0.15	0.25 $\pm$ 0.19
20	26,683.70 $\pm$ 695.31	0.24 $\pm$ 0.15	0.25 $\pm$ 0.21
21	10,233.15 $\pm$ 347.35	0.24 $\pm$ 0.15	0.25 $\pm$ 0.17
22	9,235.45 $\pm$ 284.76	0.24 $\pm$ 0.15	0.25 $\pm$ 0.21
23	9,390.60 $\pm$ 331.14	0.24 $\pm$ 0.15	0.25 $\pm$ 0.21
24	8,535.00 $\pm$ 273.06	0.23 $\pm$ 0.15	0.26 $\pm$ 0.27
25	7,375.35 $\pm$ 321.94	0.23 $\pm$ 0.15	0.26 $\pm$ 0.31
26	7,804.40 $\pm$ 195.71	0.24 $\pm$ 0.15	0.26 $\pm$ 0.22
27	7,720.05 $\pm$ 324.35	0.24 $\pm$ 0.15	0.26 $\pm$ 0.16
28	8,385.10 $\pm$ 176.36	0.24 $\pm$ 0.15	0.25 $\pm$ 0.24
29	6,734.65 $\pm$ 230.22	0.24 $\pm$ 0.15	0.26 $\pm$ 0.30
30	6,002.65 $\pm$ 345.31	0.24 $\pm$ 0.15	0.26 $\pm$ 0.20
31	5,327.45 $\pm$ 242.41	0.23 $\pm$ 0.15	0.27 $\pm$ 0.33

Source: Bussiman (2021).

7.2.2. Genotyping strategies

From the total of genotyped animals saved in each replicate within each scenario for every trait, it was sampled 10% (n1) or 50% (n2) of the animals according to the following rules:

- I. The highest number of Progenies (Top Pr); animals were ranked according to the number of progenies from highest to lowest, and then the n1 or n2 animals were chosen;
- II. Highest simulated TBV (Top TBV); animals were ranked according to TBV from highest to lowest, and then n1 or n2 animals were chosen;
- III. Highest simulated phenotype (Top P); animals were ranked according to phenotypic value from highest to lowest, and then n1 or n2 animals were chosen;
- IV. Randomly (R); animals were randomly sampled for n1 or n2;
- V. Highest and Lowest simulated TBV (TB TBV); from the total n1 or n2, 50% of the animals were chosen based on highest TBV and the remaining 50% based on lowest; and
- VI. Highest and Lowest simulated phenotype (TB P); from the total n1 or n2, 50% of the animals were chosen based on the highest phenotypic value and the remaining 50% based on the lowest.

It must be noticed those percentages of genotyped animals sampled were arbitrarily chosen because at the present there are few genotyped horses on published studies, thus it was intended to better represent the reality of the species. In addition, the average ($\pm$ SD) number of sampled animals according to each rule of sampling across replicates within the scenario is presented in Table 6.

Table 6 - Average number of genotyped animals according to percentage of genotyping candidate sampled (p; n1 = 10%; n2 = 50%) for every trait across each scenario

S	p	Traits			
		T1	T2	T3	T4
S1	n1	3,928.00 $\pm$ 42.97	3,916.40 $\pm$ 32.91	3,922.20 $\pm$ 39.95	3,914.80 $\pm$ 35.99
	n2	19,638.00 $\pm$ 215.00	19,582.80 $\pm$ 166.18	19,612.00 $\pm$ 200.20	19,574.80 $\pm$ 178.73
S2	n1	231.60 $\pm$ 5.18	236.60 $\pm$ 2.97	232.80 $\pm$ 2.28	232.00 $\pm$ 2.45
	n2	1,177.20 $\pm$ 28.45	1,168.40 $\pm$ 17.85	1,162.40 $\pm$ 9.84	1,158.80 $\pm$ 10.55
S3	n1	235.60 $\pm$ 0.89	234.00 $\pm$ 4.32	236.40 $\pm$ 4.34	234.00 $\pm$ 2.00
	n2	1,176.80 $\pm$ 5.40	1,169.60 $\pm$ 29.38	1,182.00 $\pm$ 21.35	1,169.60 $\pm$ 10.81

Source: Bussiman (2021). The number of genotyped animals was the same for every genotyping strategy, thus it was omitted from the table.

7.2.3. Breeding value prediction

Across all traits and scenarios, estimated BV (EBV) was predicted by the following model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad (7.3)$$

Where: $\mathbf{y}$ is the vector of simulated phenotypes; μ is the estimated fixed effect of the overall mean; $\mathbf{u}$ is the random additive genetic effect; $\mathbf{e}$ is the vector of random residual terms; $\mathbf{1}$ is a vector of ones; and $\mathbf{Z}$ is the incidence matrix for the random additive genetic effect. Under multivariate normality assumption, the distribution of $\mathbf{y}$ was given by $\mathbf{y} \sim \text{MVN}(\mathbf{1}\mu, \mathbf{V})$; and for our analysis $\mathbf{V}$ was as follows:

$$\mathbf{V} = \text{Var} \begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}\mathbf{H}\mathbf{Z}^T\sigma_u^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_e^2 \end{bmatrix}, \quad (7.4)$$

Where: $\mathbf{u}$; $\mathbf{e}$; and $\mathbf{Z}$ were defined in (7.3); $\mathbf{I}$ is an identity matrix of proper order; σ_u^2 is the additive genetic variance; σ_e^2 is the residual variance; and $\mathbf{H}$ is the combined relationship matrix (CHRISTENSEN; LUND, 2010; LEGARRA; AGUILAR; MISZTAL, 2009). Since additive and residual variances were known (Table 1); the solutions for μ and $\mathbf{u}$ were given by:

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{1} & \mathbf{Z}^T\mathbf{Z} + \frac{\sigma_e^2}{\sigma_u^2}\mathbf{H}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{bmatrix}, \quad (5)$$

where: $\hat{\mu}$ and $\hat{\mathbf{u}}$ are the estimated solution for overall mean fixed effect and predicted solutions for the random additive genetic effect (EBV), respectively; $\mathbf{1}$; $\mathbf{Z}$; σ_u^2 ; and σ_e^2 were defined above; and $\mathbf{H}^{-1}$ is the inverse of additive combined relationship matrix, given by (AGUILAR et al., 2010; CHRISTENSEN; LUND, 2010):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{A}_{12}^{-1} \\ \mathbf{A}_{21}^{-1} & \mathbf{G}^{-1} \end{bmatrix}, \quad (7.6)$$

Where: $\mathbf{A}^{-1}$ is the inverse of additive relationship matrix (HENDERSON, 1975); $\mathbf{G}^{-1}$ is the inverse of genomic relationship matrix (VANRADEN, 2008); and $\mathbf{A}_{11}^{-1}$; $\mathbf{A}_{12}^{-1}$; and $\mathbf{A}_{21}^{-1}$ are the first, second, and the third quadrant from $\mathbf{A}^{-1}$ (non-genotyped animals), respectively; finally $\mathbf{A}_{22}^{-1}$ is the fourth quadrant from $\mathbf{A}^{-1}$ (genotyped animals).

All this procedure was carried out using the BLUPF90 software (MISZTAL et al., 2002) which has included the PREGSF90 software for genomic processing data (AGUILAR; MISZTAL; TSURUTA, 2014), considering default parameters for convergence. The EBV prediction accuracy was calculated by the square correlation coefficient (r^2) between TBV and

EBV; and the prediction bias was assessed by $1 - \beta_{\text{TBV,EBV}}$ (HSU; GARRICK; FERNANDO, 2017) in which $\beta_{\text{TBV,EBV}}$ is the regression coefficient of TBV on EBV, estimated as follows:

$$\beta_{\text{TBV,EBV}} = \frac{E(\mathbf{u}\hat{\mathbf{u}}^T)}{E(\mathbf{u}\mathbf{u}^T)} = \frac{\sigma_{\mathbf{u},\hat{\mathbf{u}}}}{\sigma_{\mathbf{u}}^2}, \quad (7.7)$$

Where: $\mathbf{u}$ is the vector of TBV; $\hat{\mathbf{u}}$ is the vector of EBV; $\sigma_{\mathbf{u},\hat{\mathbf{u}}}$ is the covariance between TBV and EBV, and $\sigma_{\mathbf{u}}^2$ is the variance of TBV.

7.2.4. Statistical analyses

After estimation of $\beta_{\text{TBV,EBV}}$ and r^2 , those coefficients were transformed to achieve normality by taking the logarithm in base 10 of them. Then, those two parameters were analyzed through a complete unrandomized design with a factorial treatment arrangement. For instance, 4 traits; 3 scenarios; 2 percentages of selected genotyping candidates; and 6 sets of genotyped animals; the model for this included each principal effect as well as all possible interactions among them.

Analysis was conducted using the GLM procedure from SAS[®] version 9.4 (SAS INSTITUTE INC., 2013); normality of residuals was checked by the Shapiro-Wilk's test (SHAPIRO; WILK, 1965) implemented at the UNIVARIATE procedure from SAS[®] version 9.4 (SAS INSTITUTE INC., 2016); and the homoscedasticity of residuals for principal effects were checked by Levene's test (LEVENE, 1960) implemented at GLM procedure. When appropriate least-square means were compared by Tukey's test, and the significance level was fixed at 5% (i.e., $\alpha = 0.05$).

7.2.5. Model comparison

Using the best genotyping strategy, found in the previous step there were tested several models to determine the best suitable for each simulated trait. For the fourth trait, it was assumed that simulated values are the liability of a categorical trait, thus it was categorized on two phenotypic levels: 0 if simulated phenotype was negative, and 1 if it was positive. This was conducted to mimic the Gait Type as analyzed by de Oliveira Bussiman et al. (2020a, 2020b).

For the first three traits it was applied the genomic best linear unbiased prediction (GBLUP; VANRADEN, 2008), which is equivalent to using the BLUP method

(HENDERSON, 1949, 1950) but fitting marker effects with common marker variance (KOLBEHDARI; SCHAEFFER; ROBINSON, 2007; MEUWISSEN; HAYES; GODDARD, 2001; MUIR, 2007; STRANDÉN; GARRICK, 2009; VANRADEN, 2008), known as SNP-BLUP. This equivalence was explored to compute a direct genomic value (DGV) from GBLUP as it would be from SNP-BLUP. In addition, it was fitted the ssGBLUP again, as previously described, and a semi-parametric method (defined later).

The equivalence between SNP-BLUP and GBLUP can be done by revisiting Falconer and Mackay (1996) from which the additive genetic variance at locus i is given by $2p_i(1-p_i)a_i^2$ where: p_i is the allele frequency, and a_i^2 is the squared allele substitution effect. Under the assumption of common variance and $E[a_i]=0$; the total marker variance is then (STRANDÉN; GARRICK, 2009; VANRADEN, 2008):

$$\sigma_a^2 = \frac{\sigma_u^2}{2 \sum_{i=1}^m p_i(1-p_i)}, \quad (7.8)$$

Where: σ_a^2 is the common marker variance; σ_u^2 is the trait additive genetic variance; p_i is the allele frequency of i^{th} marker, and m is the total number of markers. Then marker effects can be converted in BV by $\mathbf{u} = \mathbf{Z}^* \mathbf{a}$ (MEUWISSEN; HAYES; GODDARD, 2001; STRANDÉN; GARRICK, 2009), also known as DGV; where $\mathbf{Z}^*$ is the matrix of marker content.

The model for GBLUP and ssGBLUP was the same as described in (3), except that for GBLUP, the assumption about the variance of random terms was:

$$\mathbf{V} = \text{Var} \begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{ZGZ}^T \sigma_u^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \sigma_e^2 \end{bmatrix}, \quad (7.9)$$

Where: $\mathbf{u}$; $\mathbf{e}$; $\mathbf{Z}$; $\mathbf{I}$; σ_u^2 ; and σ_e^2 were defined in (7.3); and $\mathbf{G}$ is the genomic relationship matrix among individuals, computed as in VanRaden (2008).

The semi-parametric approach was the reproducible kernel Hilbert spaces (RKHS) regression as described by Gianola, Fernando, and Stella (2006) and Gianola and Van Kaam (2008). This method consists of a non-parametric way (Euclidean distance) to compute similarity among individuals and then to use it in the mixed model equations (WINKELMAN; JOHNSON; HARRIS, 2015). The model for this analysis can be written as:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{K}_0 \boldsymbol{\alpha} + \mathbf{e}, \quad (7.10)$$

Where: $\mathbf{y}$; $\mathbf{1}$; and μ were defined in (7.3); $\mathbf{K}_\theta$ is a positive kernel matrix (defined above), dependent on a bandwidth parameter θ ; $\boldsymbol{\alpha}$ is a vector of non-parametric coefficients, and $\mathbf{e}$ is a vector of random residual terms.

Concerning $\mathbf{K}_\theta$, let $\mathbf{M}$ be a matrix of markers as: $\mathbf{M} = [\mathbf{m}_1 \ \mathbf{m}_2 \ \cdots \ \mathbf{m}_m]$, then $\mathbf{K}_\theta$ can be calculated by taking the exponential of the scaled Euclidean distance between each pair of individuals divided by θ (WINKELMAN; JOHNSON; HARRIS, 2015); i.e., $\mathbf{K}_\theta = \exp(-\bar{\mathbf{E}}/\theta)$ (hereafter referred as RKHSW), or by taking the exponential of the scaled Euclidean distance matrix pre-multiplied by $-\theta$ (GIANOLA; VAN KAAM, 2008); that is, $\mathbf{K}_\theta = \exp(-\theta\bar{\mathbf{E}})$ (hereafter referred as RKHSG); where: θ is the bandwidth parameter; and $\bar{\mathbf{E}}$ is defined as (WINKELMAN; JOHNSON; HARRIS, 2015):

$$\bar{\mathbf{E}} = [\bar{e}_{ij}] \quad \forall \quad \bar{e}_{ij} = \frac{\text{Diag}[\mathbf{M}\mathbf{M}^T]_i + \text{Diag}[\mathbf{M}\mathbf{M}^T]_j - 2\mathbf{M}_i\mathbf{M}_j^T}{\sum_{i=1}^n \sum_{j=1}^n (\text{Diag}[\mathbf{M}\mathbf{M}^T]_i + \text{Diag}[\mathbf{M}\mathbf{M}^T]_j - 2\mathbf{M}_i\mathbf{M}_j^T)}, \quad (7.11)$$

$$n^2$$

Where: n is the number of genotyped animals, and other parameters were defined above. For our analysis, θ assumed the value of 1.333 as in Winkelman, Johnson, and Harris (2015) following the recommendation of Pérez-Elizalde et al. (2015) (i.e., $\theta \leq 2$). For this case, the DGV can be predicted as $\text{DGV} = \mathbf{K}_\theta \hat{\boldsymbol{\alpha}}$ (MOROTA et al., 2014).

After categorization, variance components (Table 1) for the fourth trait were transformed from the underlying scale to the categorical scale, following Dempster and Lerner (1950) and Gianola (1979) to given $\sigma_{u_0}^2 = 0.14$ and $\sigma_{e_0}^2 = 0.11$. Then, the fourth trait was analyzed through a generalized linear mixed model (defined later) and also by the RKHS (defined above). Let the categorial response for T4 be a binomially distributed variable, then it can be modeled by assuming a probit link function, and the liability, given the model presented in (7.3), can be assumed as:

$$\mathbf{l} | \mu, \mathbf{u}, \sigma_u^2, \sigma_e^2 \sim \text{MVN} \{ \mathbf{1}\mu + \mathbf{Z}\mathbf{u}, \mathbf{I}\sigma_e^2 \}, \quad (7.12)$$

where: $\mathbf{l}$ is the liability of T4; μ ; $\mathbf{u}$; σ_u^2 ; σ_e^2 ; $\mathbf{Z}$; and $\mathbf{I}$ are as defined in (7.3). And the joint distribution of random effects ($\mathbf{u}$ and $\mathbf{e}$) was assumed as in (7.9). If there is a threshold point on this distribution, let say $t = 0$, from there the observed response is changed the model is called a threshold model (hereafter called THR), if not it is simply a generalized model (hereafter called GLM).

All models, GBLUP; ssGBLUP; RKHS; THR-GBLUP; and G-GBLUP were compared within trait by the prediction accuracy; prediction bias; and root mean squared error (RMSE). Prediction accuracy was calculated as the square correlation between TBV and DGV; the prediction bias was calculated as $1 - \beta_{\text{TBV,DGV}}$ in which $\beta_{\text{TBV,DGV}}$ was calculated as in (7.7) but using DGV despite EBV. The MSE was calculated as the square mean of the difference between TBV and DGV:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (\text{TBV}_i - \text{DGV}_i)^2}{n}}, \quad (7.13)$$

Where: TBV_i is the true breeding value for the i^{th} animal; DGV_i is the direct genomic value for the i^{th} animal, and n is the number of genotyped animals.

For the first three traits, all models were fitted using the BLUPF90 software (MISZTAL et al., 2002) and DGV was calculated by PREDF90 software (MISZTAL et al., 2002). For the fourth trait, all models were implemented as DMU package (JENSEN; MADSEN, 1994) and the DGV was calculated by homemade R (R CORE TEAM, 2020) scripts. Concerning scenario, the model comparisons were performed only within the first scenario, since the different scenarios were used to improve the best genotyping strategy selection, as well as the first scenario, which is the most realistic one.

7.2.6. Validation scheme

Concerning the implemented validation scheme, it is depicted in Figure 2. To validate genomic predictions the “saved” genotyped animals at the end of the simulation were considered as “genotyping candidates”. Those animals were divided into two groups: first a training set, which was based on each genotyping strategy being tested, and second a validation set, the remaining animals. After that, two approaches were conducted as follows: 1) all simulated pedigrees along with the training genotypes and all available phenotypes were used, and 2) the phenotypes for validation animals were also masked along with their genotypes. Our results are being presented based on the second validation approach, because in a previous analysis we verified that inclusion of phenotypes from validation animals on the ssGBLUP increased r^2 values, not allowing the marked difference compared with the RKHS method.

Figure 2 - Graphical scheme of the validation implemented in this study

Source: Bussiman (2021).

7.3. Results and Discussion

7.3.1. Linkage disequilibrium and population structure

To the date and to the best of our knowledge, there is no study with LD measures in Campolina horses. Thus, our simulations cannot be compared to the real Campolina genome. Given its origins, it is plausible to expect that the Campolina horse genome is somehow like the Mangalarga Marchador genome. By comparing LD estimates from dos Santos et al. (2019) with those in this study (Table 5 and Table 7), despite similarity among minor allele frequencies (MAF) per ECA (Table 5), substantially more LD was simulated, as well as was observed higher LD in longer distances (Table 7). However, our LD estimates fluctuate among several previous estimates on real horse populations (CORBIN et al., 2010; KIM et al., 2018; MCCUE et al., 2012; WADE et al., 2009) which turns them in valid for horses in general.

Table 7 - Linkage disequilibrium decay on simulated genotype according to five intervals of physical distance between pairs of markers

Distance (Kb)	LD
00 - 25	0.32 ± 0.36
25 - 100	0.30 ± 0.32
100 - 200	0.27 ± 0.31
200 - 500	0.23 ± 0.30
500 - 1,000	0.21 ± 0.28

Source: Bussiman (2021).

LD plays a very important role in the allele effects estimation (LEGARRA; RICARD; VARONA, 2018) but seems to have a small effect over genomic prediction accuracy (WIENTJES; VEERKAMP; CALUS, 2013), even though Sadeghi, Rafat, and Alijani (2018) found medium effects of LD over reliability. Despite that the increase in marker density tends to increase LD between QTL and SNP, which can have an impact on reliability, considering that genomic selection methods are based on the indirect effect of the QTL over the traits through the SNP (CALUS, 2010). In our study, there was simulated a high-density panel, thus minor effects of LD are not expected to affect reliabilities.

Boison et al. (2014) by comparing simulated and real Nellore genome found that even the simulation does mimic the population structure of real data, the simulated LD was higher than real in long distances. For our analysis, in addition to that, the simulation tends to give a more uniform distribution of LD and MAF across the chromosome, which in fact can be handled by very specific simulation parameters but should have minor effects on reliability. Further, when considering only genomic information, if the panel density is high and LD from the simulation is comparable to commercially available SNP chips (at same density), differences in reliability are negligible (CALUS, 2010; CALUS et al., 2008).

In addition, as the level of LD was high for our simulations, and the LD decay resulted in a long-LD breed (STOCK et al., 2016), our reliability estimates can be inflated for model comparisons and genotyping strategy determination when compared with the real Campolina genome when it became available, however, our results should apply not only for this horse breed but to others too, given different population structures were simulated for a level of LD. Besides, when genomic information is not available, stochastic simulation is the best we can have to prospect what the results of genomic selection would be (HOWARD et al., 2017).

The extension of LD is related to the N_e (BOHLOULI et al., 2017; SADEGHI; RAFAT; ALIJANI, 2018; WANG et al., 2017), but it is also affected by the inbreeding since an increase in it led to a decrease in heterozygosity, which changes the allele frequencies and affects both, LD and N_e . Inbreeding is a key in simulation studies. In general, simulated matings are very

well-balanced causing homogenous inbreeding effects and reducing it in populational terms. Comparing three simulated populations with the real Campolina (Table 3 and Table 4) the real population falls between scenarios S1 and S2 and is very distinct from S3.

Although population S1 was designed to represent the real Campolina population, differences were found comparing the inbreeding distribution (Table 4); N_e (Table 3); and the number of founders (Table 3). Despite those differences, principal component analysis from the additive relationship matrices of each population (S1 and Real), showed good agreement between those populations (data not being shown). Given the range of N_e and average inbreeding coefficient for simulated scenarios the conclusions, based on simulated population structures are valid for real horse populations.

Working with the entire population, de Oliveira Bussiman et al. (2018a) presented a very deep description of the Campolina pedigree. The N_e estimate was 93 which is closer to the found in the present study (81.67) for the recent population (animals born after 1990). The N_e from population S2 is far too small to be viable in long term (FRANKHAM et al., 2002), then this population can represent an endangered horse population, while S3 can be interpreted as a wild horse population. Population S1 is the most realistic scenario, even with the differences compared to the real population.

7.3.2. Genotyping strategy

The interaction between simulated trait and scenario was statistically significant ($P < 0.0001$) (Figure 3A), as well as the interaction between simulated trait and genotyping strategy ($P = 0.0266$) over the reliability (Figure 3B). While for bias, the only significant effect was genotyping strategy ($P < 0.0001$) (Figure 3C). T3 showed the lowest mean of reliability: 0.03 ± 0.01 ; 0.62 ± 0.01 ; and 0.57 ± 0.01 ; for S1; S2; and S3; respectively. S2 and S3 were not statistically different ($P = 0.1195$) but were both different ($P < 0.0001$) from S1. Across scenario, for T1 and T2 all reliability averages were different ($P < 0.0001$). Those means were, for T1: 0.84 ± 0.00 (S1); 0.96 ± 0.00 (S2); and 0.93 ± 0.00 (S3). For T2 reliability estimates were 0.70 ± 0.01 ; 0.93 ± 0.00 ; and 0.83 ± 0.00 ; for S1; S2; and S3; respectively. There was no reliability difference for T4 across all simulated scenario, which was 0.99 ± 0.00 .

Those effects are related to differences in heritability (BRITO et al., 2011; DE LOS CAMPOS et al., 2013; HAYES et al., 2009; MORGANTE et al., 2018); while T1 and T2 have moderate heritability (0.36, and 0.21, respectively), T4 had a very high heritability coefficient (0.85), and remarkably, T3 had $h^2 = 0.03$. Thus, for T3 were expected low values of r^2 , and a

high proportion of bias, which was confirmed (Figure 3). The upper bound on r^2 is, assuming perfect LD between SNP and QTL, the heritability of a trait (DE LOS CAMPOS et al., 2013), which explains low values for T2 and high values for T4. Within trait, there was no difference ($P = 1.0000$) among genotype strategies, while the averages of r^2 were all different ($P < 0.0001$) within genotype strategy for different traits (Figure 3B).

Figure 3 - Statistical results of genotyping strategy. A: interaction between trait and scenario over reliability; B: interaction between trait and genotyping strategy over reliability; and C: bias percentage

Source: Bussiman (2021).

In practical terms, Figure 3B clearly shows that prediction accuracy is a function of heritability, while the bias is influenced by the genotyping strategy (Figure 3C). As population S1 was simulated by declaring each generation on the pedigree as a single population in the simulation process, the more complex population structure affected predictions, while populations S2 and S3 were simulated in a “conventional” way, increasing prediction accuracy. Inbreeding was an issue, our results show that inbred populations have higher prediction accuracy, but only when the inbreeding coefficient is ignored in the model (data not being shown). That is, the inbreeding is used to build $\mathbf{H}$, and this controls prediction error variance, which is also influenced by heritability (AGUILAR et al., 2020), but for our analyses when the inbreeding was used as a covariate in the model for S2 the reliability decrease, as well as the proportion of bias.

Cesarani et al. (2019) and Lourenço et al. (2015) investigated different genotyping strategies applying ssGBLUP, those authors found improvement in reliability by increasing the number of genotyped animals and sampling for genotyping those animals with high EBV accuracy. Despite there being no difference in reliability for the applied genotyping strategy in the present study, it must be noticed that sampling animals with a high number of progenies for genotyping dramatically reduced bias ($P < 0.0001$). As the progeny number grows, the

prediction accuracy does so too, thus even under single-step methods there is no explicit reference population (as in multistep methods), when the genotyped animals are more important to the population, the imputation of genomic relationships occurred by building $\mathbf{H}$ (FERNANDO et al., 2016; HSU; GARRICK; FERNANDO, 2017) is more accurate resulting in less biased predictions.

By using graph theory, Perez et al. (2018) studied genotyping strategies for dairy cattle populations, those authors argued that increasing the number of genotyped animals in the reference population increases the reliability, they showed that sampling the most connected animals in the communities formed based on pedigree relationship is the most promising strategy to found genotyping candidates. This supports the bias reductions when genotyping animals with more progeny, because those animals will have more connections with others, and the progeny number is a scaling method, that is when a given animal has a large offspring number and one of those progenies has also a large number of offspring, more connected the previous animal will be with the whole community formed.

Bias may arise from different baseline populations (POWELL; VISSCHER; GODDARD, 2010; VANRADEN, 2008; VITEZICA et al., 2011): while the pedigree base population is the founder population, for the G it is assumed the reference population (or the current genotyped population) (CHRISTENSEN et al., 2012; GAO et al., 2019; LEGARRA et al., 2015). This leads to biased (co)variance components estimation (GAO et al., 2019; JENSEN, 2016) and breeding value prediction (MISZTAL et al., 2021; MISZTAL; LOURENCO; LEGARRA, 2020), however, the use of ssGBLUP should decrease bias since it provides optimal use of all available information (LEGARRA et al., 2014).

Another source of bias could be the ability of genotyped animals to represent the population structure captured by pedigree, i.e., the more selected or different it is, the more biased genomic evaluations will be because of the large impact on the accuracy of genotype relationship imputation on single-step (GAO et al., 2019; HSU; GARRICK; FERNANDO, 2017; VITEZICA et al., 2011). Even under a Bayesian approach, in which there are explicit imputed genotype covariates for ungenotyped animals, the accuracy depends on the genetic relationship between genotyped and non-genotyped animals (FERNANDO et al., 2016; FERNANDO; DEKKERS; GARRICK, 2014; GAO et al., 2019).

Furthermore, as no statistical difference was found between reliability estimates from different genotyping strategies within traits, future genotyping strategies should be implemented to avoid bias. By that, to sample the animals with high progeny numbers (to be genotyped) is the best approach. This strategy provides high accuracy with the lowest percent

of bias as possible for ssGBLUP evaluations in practical conditions, such as those applied in this study. We did run a forward validation, by using allele effects from each genotyping strategy tested to predict DGV for young animals (data not being shown) but this resulted in the same conclusions as depicted in Figure 2. This may be related to the fact that although indirect predictions via ssGBLUP are calculated using a similar concept of multistep predictions, they usually are more accurate because on multistep methods those predictions involve approximations and possible double counting of phenotypic information (LOURENÇO et al., 2015).

It must be noticed that the proportion of genotyped animals (n_1 , or n_2 : 10% and 50%, respectively) did not significantly affect the reliability or the bias ($P = 0.3360$ and $P = 1.0000$). That is an unexpected result once the increase on N is related to an increase in reliability and a decrease in bias. Despite it, there was a quadratic effect ($P < 0.0001$) of the genotyped number of animals over the reliability. Thus, no matter what proportion of the population is genotyped, it should be genotyped as many animals as possible to allow unbiased and reliable genomic predictions. By taking the first derivative of the quadratic function of reliability (data not being shown) over the number of genotyped animals and equating it equals to zero, the number of genotyped animals to achieve the maximum reliability should be around 10,733 horses (average over simulated traits).

7.3.3. Model comparison

The more heritable the trait is, the higher reliability of genomic predictions are (Figure 3). As expected, the model which achieved higher r^2 was the semi-parametric one (RKHS), and between two different kernels, the RKHSG was the best (Figure 4). Those results are supported by previous one (CROSSA et al., 2013; CUEVAS et al., 2016; DE ALMEIDA FILHO et al., 2019; DE LOS CAMPOS et al., 2010; JACQUIN; CAO; AHMADI, 2016; MOMEN et al., 2018; MOROTA et al., 2014; WINKELMAN; JOHNSON; HARRIS, 2015) that found semi-parametric or non-parametric models are better for genomic predictions. Following the approach of Winkelman, Johnson, and Harris (2015) we combined ssGBLUP with RKHS, by replacing $\mathbf{G}$ with the $\mathbf{K}_0$ (hereafter called ssRKHS). ssRKHS resulted in 1) high r^2 (not higher than RKHS); 2) smaller bias than RKHS (but higher than ssGBLUP); and 3) highest predictive power, as those ssRKHS models had the smallest RMSE values.

Figure 4 - Statistical results of model comparison for linear traits. A: reliability estimates; B: percentage of bias; and C: root mean squared error

Source: Bussiman (2021).

The advantages of RKHS over GBLUP or ssGBLUP are often attributed to the genetic architecture of the traits (CROSSA et al., 2013; JACQUIN; CAO; AHMADI, 2016; MOMEN et al., 2018; MOROTA et al., 2014) because $\mathbf{K}_0$ captures not only the additive gene interactions but epistatic too. By that, the solutions are more realistic, and the predictive power is high (GIANOLA; FERNANDO; STELLA, 2006; GIANOLA; VAN KAAM, 2008; JACQUIN; CAO; AHMADI, 2016). On other hand, the bias is more affected by the number of genotyped animals, since both GBLUP and RKHS can be performed only for genotyped animals. For our analysis, it was considered as reference population (to estimate allele effects) the best-genotyping strategy found on the previous step (Top Pr), thus there were fewer animals on the training set compared to the validation set, which can explain the increase in bias on GBLUP and RKHS, but also the bias reduction on ssGBLUP and ssRKHS.

For our analysis, the RKHS outperformed GBLUP and ssGBLUP in terms of r^2 even for the lowest heritable trait (T3, $h^2 = 0.03$), and concerning the bias, this trait resulted to be the more biased one. This increase in prediction reliability should be interpreted with caution because of the high percentage of bias. Since the trait presents very low heritability it should have very low values of r^2 which is confirmed by ssGBLUP or ssRKHS when all available phenotypes are used along with the Top Pr genotypes. The predictive ability reflected by the RMSE was high for GBLUP and both RKHS kernels. This may be related to the use of only Top Pr genotypes and then to validate prediction with DGV for all genotypes available. On the other hand, comparing methods, ssGBLUP and ssRKHS showed the highest predictive ability (lowest RMSE).

As stated before, the ssGBLUP is less sensitive to bias (LEGARRA et al., 2014) and that is a result of combining at once all phenotypes, pedigree, and genotypes. This is also the case when we combined RKHS and single-step method; as the number of animals increase because of pedigree and phenotypes, the bias reduced and predictive power increased along with reliability, by the more accurate genomic relationship used when $\mathbf{K}_0$ replaced $\mathbf{G}$. Concerning traits T1 ($h^2 = 0.36$) and T2 ($h^2 = 0.21$), the higher heritability is, the higher prediction accuracy and the smaller bias and RMSE are, but for trait T3 ($h^2 = 0.03$) the RMSE was very low across all tested models. This result can be a statistical artifact since the trait is almost not heritable, predictions are done randomly and by chance the differences between TBV and DGV were low. Those results allow the recommendation of ssRKHS for genomic evaluations since it will bring together high accuracy and predictive ability along with small bias, but also more studies are required to confirm this hypothesis.

Concerning T4, it was treated as categorical for model comparisons, and several models were compared using or not a threshold on the underlying distribution (Figure 5). As this trait had the highest heritability, it was the one with a more accurate prediction across all tested models. The RKHS outperformed GBLUP and ssGBLUP, except when using a threshold model in which the best option is the ssGBLUP (Figure 5A). When the ssRKHS was tested it turned to be equivalent to RKHS, but with the benefiting of small bias and RMSE (Figures 4B, and 5C). By comparing the use or not of a threshold model (THR), for GBLUP, RKHSW, RKHSG, and ssGBLUP, the THR was better than GLM, but comparing the three single-step methods it appears that GLM is better than THR (Figure 5).

Figure 5 - Statistical results of model comparison for the binary trait. A: reliability estimates; B: percentage of bias; and C: root mean squared error

Source: Bussiman (2021).

Generalized linear models were first proposed aiming to unify linear models' theory (NELDER; WEDDERBURN, 1972), those are powerful models when the normality assumptions do not hold. With or without a threshold in the liability distribution those models linearize nonlinear predictors through the link function (ENGEL; BUIST; VISSCHER, 1995; NELDER; WEDDERBURN, 1972). Several desired properties of threshold models in the context of animal breeding were presented by Dempster and Lerner (1950); Gianola (1982); Gianola and Foulley (1983); and Meijering and Gianola (1985), however, threshold models are just a class of generalized models with probit link function and the presence of a threshold in the underlying distribution (DEMPSTER; LERNER, 1950; ENGEL; BUIST; VISSCHER, 1995).

For our analyses, the presence of the threshold increases reliability only for ssGBLUP while for other methods r^2 favors choosing models without the threshold; and the same pattern was verified on bias and RMSE. However, when we compare ssRKHS with other methods, those outperformed in terms of r^2 and predictive ability. The best option was ssRKHSW (Figure 4), which is an ssRKHS method with the kernel calculated as in Winkelman, Johnson, and Harris (2015) but not assuming the occurrence of a threshold in the liability distribution. While for GBLUP, RKHSW, RKHSG and ssGBLUP the smaller percentage of bias were by assuming the threshold, for ssRKHS methods the better option is not to assume it. This may be related to a better matrix conditioning during the application of the probit link function in the linear predictor once the relationship with individuals was nonlinear based.

Working with mortality in broilers, González-Recio et al. (2008) found an increase in predictive power when fitting RKHS regression in a binary variable, and Mallick, Ghosh, and Ghosh (2005) used the RKHS kernel to conduct Bayesian classification of tumors and found this method performed better than other classification methods. The high predictive power of the RKHS regression may be related to classification; in fact, de los Campos et al. (2010) and Vapnik (1998) showed how estimates in RKHS regression can be viewed as penalized optimization problem allowing regression with any kind of information set. Thus, conditional on the input, the output will be treated as a classification problem through RKHS regression, which can explain the increase in reliability and decrease in RMSE.

When the assumption of the threshold holds (FALCONER, 1965; GIANOLA, 1982; GIANOLA; FOULLEY, 1983; WRIGHT, 1934) a threshold model must be always applied. Our results may suggest for T4 the assumption of a threshold-liability is wrong and thus a conventional generalized linear model performed better considering the ssRKHS methods. Furthermore, there are several advantages of using semi-parametric methods to do genomic

predictions (DE LOS CAMPOS et al., 2010), those are equivalent to doing an animal model in BLUP (DE LOS CAMPOS; GIANOLA; ROSA, 2009) but using kernel matrices instead of numerator relationship matrix (MOROTA et al., 2014; MOROTA; GIANOLA, 2014); and de los Campos, Gianola, and Rosa (2009) argued that $\mathbf{A}$ is a special case of RKHS matrices. Under this reasoning, the combination of $\mathbf{K}_0$ with $\mathbf{A}_{22}$ (pedigree-based relationship matrix for genotype animals) should provide a better version of ssGBLUP in a semi-parametric way. By that, the implicit imputation of genomic relationship made when $\mathbf{H}$ is built, is performed with the kernel distance matrix instead of traditional $\mathbf{G}$.

In addition, in the present study the bandwidth parameter was set, empirically, to 1.333, and the proportion of $\mathbf{A}_{22}$ included in $\mathbf{K}_0$ to build $\mathbf{G}_0$ and then to create $\mathbf{H}$ was set to 0.02 (WINKELMAN; JOHNSON; HARRIS, 2015). Pérez-Elizalde et al. (2015) suggested that better results are faced when the bandwidth parameter is smaller than 2, but Mallick, Ghosh, and Ghosh (2005) suggested a Bayesian estimation of it while de los Campos et al. (2010) argued in favor of cross-validation to do it. Thus, different results may be found with other bandwidth parameters and with other proportions of $\mathbf{A}_{22}$ included $\mathbf{G}_0$. Those values could be better investigated in other approaches to determine the optimal way to combine distance-based matrix ($\mathbf{K}$) with pedigree-based relationships allowing a new semi-parametric single-step method.

Finally, it must be noticed that for our analysis the apparent superiority of GBLUP over ssGBLUP could be explained by the population structure. Since the addition of the genotypes and the phenotypes from validation animals on the model increased substantially the r^2 values (data not being shown). By doing predictions and performing validation for scenarios S2 and S3 (data not shown), even masking the phenotypes and genotypes from validation animals did not reduce r^2 from ssGBLUP compared with GBLUP. In addition, for this situation, the ssGBLUP was not as different from RKHS methods as it was in the presented validation schemes. This should indicate that for strong complexes population structures, impacts on r^2 values should be expected.

7.4. Conclusion

The decision to implement genomic selection will always be a tradeoff between cost and benefit. Aiming to achieve higher reliability and low bias in predictions, one must genotype as many as possible animals, but genotyping costs for horses still being difficult to do so.

Presented results attempt to solve partially the cost problem by using simulation to prospect how genomic selection can be better implemented on horse populations. As a general recommendation, it should be genotyped the most important animals, that is, the animals with the highest progeny number. Those animals will have more connections among the populations, higher prediction accuracy, and will somehow better represent the pedigreed population. In addition, if the number of animals with high progeny numbers is small, the combination of sampling animals with the highest progeny number along with the remaining genotype candidates sampled randomly, based on our results, should not impact the reliability and would still present low values of bias. In terms of models, we presented a new class of semi-parametric methods, by combining kernel method with conventional single-step genomic best linear prediction. Those performed better for both continuous and binary outcomes, they should be better researched in the future along with other genotyping strategies.

7.5. References

- AGUILAR, I. et al. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Dairy Science**, v. 93, n. 2, p. 743–752, 2010.
- AGUILAR, I. et al. Effects of ignoring inbreeding in model-based accuracy for BLUP and SSGBLUP. **Journal of Animal Breeding and Genetics**, v. 137, n. 4, p. 356–364, 2020.
- AGUILAR, I.; MISZTAL, I.; TSURUTA, S. PREGSF90–POSTGSF90: Computational tools for the implementation of single-step genomic selection and genome-wide association with ungenotyped. In: 10th WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 10. **Proceedings...** Vancouver: American Society of Animal Science, 2014.
- ANDERSSON, L. S. et al. Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. **Nature**, v. 488, n. 7413, p. 642–646, 2012.
- ANDONOV, S. et al. Accuracy of breeding values in small genotyped populations using different sources of external information-A simulation study. **Journal of Dairy Science**, v. 100, n. 1, p. 395–401, 2017.
- BOHLOULI, M. et al. Genomic prediction by considering genotype $\times$ environment interaction using different genomic architectures. **Annals of Animal Science**, v. 17, n. 3, p. 683–701, 2017.
- BOISON, S. A. et al. Imputation of non-genotyped individuals using genotyped progeny in Nellore, a Bos indicus cattle breed. **Livestock Science**, v. 166, p. 176–189, 2014.
- BRITO, F. V et al. Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. **BMC Genetics**, v. 12, n. 1, p. 80, 2011.

- CALUS, M. P. L. et al. Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. **Genetics**, v. 178, n. 1, p. 553–561, 2008.
- CALUS, M. P. L. Genomic breeding value prediction: methods and procedures. **Animal**, v. 4, n. 2, p. 157–164, 2010.
- CESARANI, A. et al. Bias in heritability estimates from genomic restricted maximum likelihood methods under different genotyping strategies. **Journal of Animal Breeding and Genetics**, v. 136, n. 1, p. 40–50, 2019.
- CHEN, C. Y. et al. Effect of different genomic relationship matrices on accuracy and scale1. **Journal of Animal Science**, v. 89, n. 9, p. 2673–2679, 2011.
- CHRISTENSEN, O. F. et al. Single-step methods for genomic evaluation in pigs. **Animal**, v. 6, n. 10, p. 1565–1571, 2012.
- CHRISTENSEN, O. F.; LUND, M. S. Genomic prediction when some animals are not genotyped. **Genetics Selection Evolution**, v. 42, n. 1, p. 2, 2010.
- CORBIN, L. J. et al. Linkage disequilibrium and historical effective population size in the Thoroughbred horse. **Animal Genetics**, v. 41, p. 8–15, 2010.
- COSTER, A. **pedigree**: Pedigree Functions. R Foundation for Statistical Computing, 2012. Disponível em: <<https://cran.r-project.org/package=pedigree>>.
- CROSSA, J. et al. Genomic Prediction in Maize Breeding Populations with Genotyping-by-Sequencing. **G3 Genes|Genomes|Genetics**, v. 3, n. 11, p. 1903–1926, 2013.
- CUEVAS, J. et al. Genomic Prediction of Genotype × Environment Interaction Kernel Regression Models. **The Plant Genome**, v. 9, n. 3, 2016.
- DE ALMEIDA FILHO, J. E. et al. Genomic Prediction of Additive and Non-additive Effects Using Genetic Markers and Pedigrees. **G3: Genes, Genomes, Genetics**, v. 9, n. 8, p. 2739–2748, 2019.
- DE LOS CAMPOS, G. et al. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. **Genetics Research**, v. 92, n. 4, p. 295–308, 2010.
- DE LOS CAMPOS, G. et al. Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. **PLoS Genetics**, v. 9, n. 7, p. e1003608, 2013.
- DE LOS CAMPOS, G.; GIANOLA, D.; ROSA, G. J. M. Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation1. **Journal of Animal Science**, v. 87, n. 6, p. 1883–1887, 2009.
- DE OLIVEIRA BUSSIMAN, F. et al. Pedigree analysis and inbreeding effects over morphological traits in Campolina horse population. **Animal**, v. 12, n. 11, p. 2246–2255, 2018a.
- DE OLIVEIRA BUSSIMAN, F. et al. Genetic analysis of morphological and functional traits in Campolina horses using Bayesian multi-trait model. **Livestock Science**, v. 216, n. April, p.

119–129, 2018b.

DE OLIVEIRA BUSSIMAN, F. et al. Allelic and genotypic frequencies of the DMRT3 gene in the Brazilian horse breed Mangalarga Marchador and their association with types of gait. **Genetics and Molecular Research**, v. 18, n. 1, 2019.

DE OLIVEIRA BUSSIMAN, F. et al. Genome-wide association study: Understanding the genetic basis of the gait type in Brazilian Mangalarga Marchador horses, a preliminary study. **Livestock Science**, v. 231, p. 103867, 2020a.

DE OLIVEIRA BUSSIMAN, F. et al. Model comparisons for genetic evaluation of gait type in Mangalarga Marchador horses. **Livestock Science**, v. 239, p. 104168, 2020b.

DEMPSTER, E. R.; LERNER, I. M. Heritability of Threshold Characters. **Genetics**, v. 35, n. 2, p. 212–236, 1950.

DOS SANTOS, B. A. et al. Genomic analysis of the population structure in horses of the Brazilian Mangalarga Marchador breed. **Livestock Science**, v. 229, p. 49–55, 2019.

ENGEL, B.; BUIST, W.; VISSCHER, A. Inference for threshold models with variance components from the generalized linear mixed model perspective. **Genetics Selection Evolution**, v. 27, n. 1, p. 15, 1995.

FALCONER, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. **Annals of Human Genetics**, v. 29, n. 1, p. 51–76, 1965.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to Quantitative Genetics**. 4. ed. Essex: Pearson Education Ltda, 1996.

FERNANDO, R. L. et al. Computational strategies for alternative single-step Bayesian regression models with large numbers of genotyped and non-genotyped animals. **Genetics Selection Evolution**, v. 48, n. 1, p. 96, 2016.

FERNANDO, R. L.; DEKKERS, J. C.; GARRICK, D. J. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. **Genetics Selection Evolution**, v. 46, n. 1, p. 50, 2014.

FONSECA, M. G. et al. A Genome-Wide Association Study Reveals Differences in the Genetic Mechanism of Control of the Two Gait Patterns of the Brazilian Mangalarga Marchador Breed. **Journal of Equine Veterinary Science**, v. 53, p. 64–67, 2017.

FRANKHAM, R. et al. **Introduction to Conservation Genetics**. 1. ed. Cambridge: Cambridge University Press, 2002.

GAO, H. et al. Bias in estimates of variance components in populations undergoing genomic selection: a simulation study. **BMC Genomics**, v. 20, n. 1, p. 956, 2019.

GIANOLA, D. Heritability of polychotomous characters. **Genetics**, v. 93, n. 4, p. 1051–1055, 1979.

GIANOLA, D. Theory and Analysis of Threshold Characters. **Journal of Animal Science**, v. 54, n. 5, p. 1079, 1982.

GIANOLA, D.; FERNANDO, R. L.; STELLA, A. Genomic-Assisted Prediction of Genetic Value with Semiparametric Procedures. **Genetics**, v. 173, n. 3, p. 1761–1776, 2006.

GIANOLA, D.; FOULLEY, J. L. Sire evaluation for ordered categorical data with a threshold model. **Genetics Selection Evolution**, v. 15, n. 2, p. 201–224, 1983.

GIANOLA, D.; VAN KAAM, J. B. C. H. M. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. **Genetics**, v. 178, n. 4, p. 2289–2303, 2008.

GONZÁLEZ-RECIO, O. et al. Nonparametric Methods for Incorporating Genomic Information Into Genetic Evaluations: An Application to Mortality in Broilers. **Genetics**, v. 178, n. 4, p. 2305–2313, 2008.

GUTIÉRREZ, J. et al. Individual increase in inbreeding allows estimating effective sizes from pedigrees. **Genetics Selection Evolution**, v. 40, n. 4, p. 359, 2008.

GUTIÉRREZ, J. P.; CERVANTES, I.; GOYACHE, F. Improving the estimation of realized effective population sizes in farm animals. **Journal of Animal Breeding and Genetics**, v. 126, n. 4, p. 327–332, 2009.

HAYES, B.; GODDARD, M. E. The distribution of the effects of genes affecting quantitative traits in livestock. **Genetics Selection Evolution**, v. 33, n. 3, p. 209, 2001.

HAYES, B. J. et al. Invited review: Genomic selection in dairy cattle: Progress and challenges. **Journal of Dairy Science**, v. 92, n. 2, p. 433–443, 2009.

HAYES, B. J.; VISSCHER, P. M.; GODDARD, M. E. Increased accuracy of artificial selection by using the realized relationship matrix. **Genetics Research**, v. 91, n. 1, p. 47–60, 2009.

HENDERSON, C. R. Estimation of Changes in Herd Environment. In: FORTY-FOURTH ANNUAL MEETING OF THE AMERICAN DAIRY SCIENCE ASSOCIATION, 44. **Book of Abstracts...** Minneapolis, Minnesota: Elsevier, 1949.

HENDERSON, C. R. Estimation of Genetic Parameters. In: THE ANNALS OF MATHEMATICAL STATISTICS. **Anais...** Chapel Hill, North Carolina: The Annals of Mathematical Statistics, 1950.

HENDERSON, C. R. Rapid method for computing the inverse of a relationship matrix. **Journal of Dairy Science**, v. 58, n. 11, p. 1727–1730, 1975.

HOWARD, J. T. et al. Geno-Diver: A combined coalescence and forward-in-time simulator for populations undergoing selection for complex traits. **Journal of Animal Breeding and Genetics**, v. 134, n. 6, p. 553–563, 2017.

HSU, W.-L.; GARRICK, D. J.; FERNANDO, R. L. The Accuracy and Bias of Single-Step Genomic Prediction for Populations Under Selection. **G3: Genes, Genomes, Genetics**, v. 7, n. 8, p. 2685–2694, 2017.

JACQUIN, L.; CAO, T.-V.; AHMADI, N. A Unified and Comprehensible View of Parametric and Kernel Methods for Genomic Prediction with Application to Rice. **Frontiers**

in **Genetics**, v. 7, 2016.

JÄDERKVIST FEGRAEUS, K. et al. To pace or not to pace: a pilot study of four- and five-gaited Icelandic horses homozygous for the DMRT3 ‘Gait Keeper’ mutation. **Animal Genetics**, v. 48, n. 6, p. 694–697, 2017.

JENSEN, J. Estimation of genetic variance in the age of genomics. **Journal of Animal Breeding and Genetics**, v. 133, n. 5, p. 333–333, 2016.

JENSEN, J.; MADSEN, P. DMU: A package for the analysis of multivariate mixed models. In: 5th CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION. **Proceedings...** Guelph: Wageningen University, 1994.

KALBFLEISCH, T. S. et al. Improved reference genome for the domestic horse increases assembly contiguity and composition. **Communications Biology**, v. 1, n. 1, p. 197, 2018.

KIM, N. Y. et al. Genome-wide analyses of the Jeju, Thoroughbred, and Jeju crossbred horse populations using the high density SNP array. **Genes & Genomics**, v. 40, n. 11, p. 1249–1258, 2018.

KOLBEHDARI, D.; SCHAEFFER, L. R.; ROBINSON, J. A. B. Estimation of genome-wide haplotype effects in half-sib designs. **Journal of Animal Breeding and Genetics**, v. 124, n. 6, p. 356–361, 2007.

LEGARRA, A. et al. Single Step, a general approach for genomic selection. **Livestock Science**, v. 166, n. 1, p. 54–65, 2014.

LEGARRA, A. et al. Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships. **Genetics**, v. 200, n. 2, p. 455–468, 2015.

LEGARRA, A.; AGUILAR, I.; MISZTAL, I. A relationship matrix including full pedigree and genomic information. **Journal of Dairy Science**, v. 92, p. 4656–4663, 2009.

LEGARRA, A.; RICARD, A.; VARONA, L. GWAS by GBLUP: Single and Multimarker EMMAX and Bayes Factors, with an Example in Detection of a Major Gene for Horse Gait. **G3: Genes, Genomes, Genetics**, v. 8, n. 7, p. 2301–2308, 2018.

LEVENE, H. Robust tests for equality of variances. In: OLKIN, I. et al. (Eds.) **Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling**. 1. ed. Palo Alto: Stanford University Press, 1960. p. 278–292.

LIU, L.-L. et al. Genome-wide analysis reveals signatures of selection for gait traits in Yili horse. **bioRxiv**, p. 471797, 2018.

LOURENÇO, D. A. L. et al. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. **Journal of Animal Science**, v. 93, n. 6, p. 2653–2662, 2015.

MALLICK, B. K.; GHOSH, D.; GHOSH, M. Bayesian classification of tumours by using gene expression data. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 67, n. 2, p. 219–234, 2005.

MANSO FILHO, H. C. et al. Alelo DMRT3 mutante em equinos de marcha batida e picada das raças Campolina e Mangalarga Marchador. **Ciência Veterinária nos Trópicos**, v. 18, n. 1, p. 6–11, 2015.

MCCOY, A. M. et al. Identification and validation of genetic variants predictive of gait in standardbred horses. **PLoS Genetics**, v. 15, n. 5, p. 1–16, 2019.

MCCUE, M. E. et al. A High Density SNP Array for the Domestic Horse and Extant Perissodactyla: Utility for Association Mapping, Genetic Diversity, and Phylogeny Studies. **PLoS Genetics**, v. 8, n. 1, p. e1002451, 2012.

MEIJERING, A.; GIANOLA, D. Linear versus nonlinear methods of sire evaluation for categorical traits: a simulation study. **Genetics Selection Evolution**, v. 17, p. 115–132, 1985.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819–1829, 2001.

MISZTAL, I. et al. BLUPF90 and related programs (BGF90). In: 7th WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 7. **Proceeding...** Montpellier: Wageningen University, 2002.

MISZTAL, I. et al. Emerging issues in genomic selection. **Journal of Animal Science**, v. 99, n. 6, 2021.

MISZTAL, I.; LEGARRA, A.; AGUILAR, I. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. **Journal of Dairy Science**, v. 92, n. 9, p. 4648–4655, 2009.

MISZTAL, I.; LOURENCO, D.; LEGARRA, A. Current status of genomic evaluation. **Journal of Animal Science**, v. 98, n. 4, 2020.

MOMEN, M. et al. Predictive ability of genome-assisted statistical models under various forms of gene action. **Scientific Reports**, v. 8, n. 1, p. 12309, 2018.

MORGANTE, F. et al. Effect of genetic architecture on the prediction accuracy of quantitative traits in samples of unrelated individuals. **Heredity**, v. 120, n. 6, p. 500–514, 2018.

MOROTA, G. et al. Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. **Frontiers in Genetics**, v. 5, 2014.

MOROTA, G.; GIANOLA, D. Kernel-based whole-genome prediction of complex traits: a review. **Frontiers in Genetics**, v. 5, 2014.

MUIR, W. M. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. **Journal of Animal Breeding and Genetics**, v. 124, n. 6, p. 342–355, 2007.

NEJATI-JAVAREMI, A.; SMITH, C.; GIBSON, J. P. Effect of total allelic relationship on accuracy of evaluation and response to selection. **Journal of Animal Science**, v. 75, n. 7, p. 1738, 1997.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **Journal of the Royal Statistical Society. Series A: General**, v. 135, n. 3, p. 370–384, 1972.

NOVOA-BRAVO, M.; BERNAL-PINILLA, E.; GARCÍA, L. F. Microevolution operating in domestic animals: evidence from the Colombian Paso horses. **Mammalian Biology**, v. 101, n. 2, p. 181–192, 2021.

PATTERSON, L.; STAIGER, E. A.; BROOKS, S. A. DMRT3 is associated with gait type in Mangalarga Marchador horses, but does not control gait ability. **Animal Genetics**, v. 46, n. 2, p. 213–215, 2015.

PÉREZ-ELIZALDE, S. et al. Selection of the Bandwidth Parameter in a Bayesian Kernel Regression Model for Genomic-Enabled Prediction. **Journal of Agricultural, Biological, and Environmental Statistics**, v. 20, n. 4, p. 512–532, 2015.

PEREZ, B. DA C. et al. Accounting for population structure in selective cow genotyping strategies. **Journal of Animal Breeding and Genetics**, v. 00, p. 1–17, 2018.

POWELL, J. E.; VISSCHER, P. M.; GODDARD, M. E. Reconciling the analysis of IBD and IBS in complex trait studies. **Nature Reviews Genetics**, v. 11, n. 11, p. 800–805, 2010.

PROCÓPIO, A. M. M.; BERGMANN, J. A. G. A. G.; COSTA, M. D. D. Formação e demografia da raça Campolina. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 55, n. 3, p. 361–365, 2003.

PSZCZOLA, M. et al. The impact of genotyping different groups of animals on accuracy when moving from traditional to genomic selection. **Journal of Dairy Science**, v. 95, n. 9, p. 5412–5421, 2012.

R CORE TEAM. **R: A Language and Environment for Statistical Computing** (R. D. C. Team, Ed.) R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. 2020.

SADEGHI, S.; RAFAT, S. A.; ALIJANI, S. Evaluation of imputed genomic data in discrete traits using Random forest and Bayesian threshold methods. **Acta Scientiarum. Animal Sciences**, v. 40, n. 1, p. 39007, 2018.

SANTOS, W. et al. Fine-scale estimation of inbreeding rates, runs of homozygosity and genome-wide heterozygosity levels in the Mangalarga Marchador horse breed. **Journal of Animal Breeding and Genetics**, v. 138, n. 2, p. 161–173, 2021.

SARGOLZAEI, M.; IWASAKI, H.; COLLEAU, J. J. CFC: A tool for Monitoring genetic diversity. In: 8th WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 8. **Proceedings...** Belo Horizonte: 2006.

SARGOLZAEI, M.; SCHENKEL, F. S. QMSim: A large-scale genome simulator for livestock. **Bioinformatics**, v. 25, n. 5, p. 680–681, 2009.

SAS INSTITUTE INC. The GLM Procedure. In: SAS INSTITUTE INC. (Ed.) **SAS/STAT 13.1 User's Guide**. 13.1 ed. Cary, North Carolina: SAS Institute Inc., SAS Campus Drive, 2013. p. 206.

SAS INSTITUTE INC. The UNIVARIATE Procedure. In: SAS INSTITUTE INC. (Ed.) **Base SAS 9.4 Procedures Guide: Statistical Procedures**. 6. ed. Cary, North Carolina: SAS Institute Inc., SAS Campus Drive, 2016. p. 287–551.

SCHURINK, A. et al. The Genomic Makeup of Nine Horse Populations Sampled in the Netherlands. **Genes**, v. 10, n. 6, p. 480, 2019.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, v. 52, n. 3–4, p. 591–611, 1965.

SITZENSTOCK, F. et al. Evaluation of genomic selection for functional traits in horse breeding programs. In: 9th WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 9. **Proceedings...** Leipzig: Wageningen University, 2010.

STOCK, K. F. et al. Genomic applications in horse breeding. **Animal Frontiers**, v. 6, n. 1, p. 45–52, 2016.

STRANDÉN, I.; GARRICK, D. J. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. **Journal of Dairy Science**, v. 92, n. 6, p. 2971–2975, 2009.

VANRADEN, P. M. Efficient Methods to Compute Genomic Predictions. **Journal of Dairy Science**, v. 91, n. 11, p. 4414–4423, 2008.

VAPNIK, V. N. **Statistical learning theory**. 1. ed. New York: John Wiley & Sons, Inc., 1998.

VIEIRA, E. R. et al. Caracterização da equideocultura no estado de Minas Gerais. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 67, n. 1, p. 319–323, 2015.

VITEZICA, Z. G. G. et al. Bias in genomic predictions for populations under selection. **Genetics Research**, v. 93, n. 5, p. 357–366, 2011.

WADE, C. M. et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. **Science**, v. 326, n. 5954, p. 865–867, 2009.

WANG, Q. et al. Effects of marker density and population structure on the genomic prediction accuracy for growth trait in Pacific white shrimp *Litopenaeus vannamei*. **BMC Genetics**, v. 18, n. 1, p. 45, 2017.

WIENTJES, Y. C. J.; VEERKAMP, R. F.; CALUS, M. P. L. The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. **Genetics**, v. 193, n. 2, p. 621–631, 2013.

WINKELMAN, A. M.; JOHNSON, D. L.; HARRIS, B. L. Application of genomic evaluation to dairy cattle in New Zealand. **Journal of Dairy Science**, v. 98, n. 1, p. 659–675, 2015.

WRIGHT, S. Coefficients of Inbreeding and Relationship. **The American Naturalist**, v. 56, n. 645, p. 330–338, 1922.

WRIGHT, S. The Results of Crosses between Inbred Strains of Guinea Pigs, Differing in Number of Digits. **Genetics**, v. 19, n. 6, p. 537–551, 1934.

8. CONSIDERAÇÕES FINAIS

“All models are wrong, but some are useful”
(BOX, 1979, p. 202)¹⁶.

“All models are useful, some are wrong”
(Aike Kagada, unpublished)¹⁷.

Durante o desenvolvimento de minha dissertação de mestrado pude sentir a dificuldade de se trabalhar com poucos animais genotipados, bem como alguns problemas quanto a habilidade dos animais genotipados em representar, de fato, a população em estudo. Na época contornamos a maioria dos problemas ao adotar o método de passo único, uma vez que a incorporação da matriz **H** nas equações de Henderson, leva a uma correção natural para a estrutura da população. Ainda assim, surgiu um interesse pelo tema da genotipagem seletiva que, posteriormente, acabou por virar meu pré-projeto de doutorado. Neste meio tempo, tive contato também com um ramo novo do melhoramento genético, mas bastante sedimentado em outras áreas, principalmente na estatística: a inferência causal. Achei por bem então unir as coisas: desenvolver um estudo bastante aprofundado do ponto de vista genético-fenotípico, bem como prospectar a possibilidade de seleção genômica em cavalos Campolina.

Inicialmente, apresentamos um estudo sobre a população fundadora da raça de cavalos Campolina, avaliamos o atual status de diversidade genética e a formação de linhagens fundadoras no pedigree da população mais recente, por meio da técnica de análise de componentes principais (PCA). Estudos sobre a formação de linhagens não são um tema novo, mas volta e meia, ressurgem na literatura objetivando entender como a variabilidade genética está distribuída nos rebanhos de interesse. Nosso trabalho apresenta uma aplicação bastante simples de PCA para identificar linhagens, mas estendemos o estudo para avaliar se de fato as linhagens identificadas impactaram nas predições genéticas nessa população. Nosso principal, e mais relevante resultado, é que a constituição genética dos fundadores permanece importante na população mesmo após muitas gerações. Esse resultado, se empregado por futuros pesquisadores e criadores, auxiliará na formação de novos rebanhos, mesmo os experimentais, uma vez que pode nortear a escolha correta da população fundadora.

¹⁶ BOX, G. E. P. Robustness in the Strategy of Scientific Model Building. In: **Robustness in Statistics**. Elsevier, 1979. p. 201–236.

¹⁷ Informação fornecida por Gianola, D. durante o curso “Prediction of Complex Traits” ministrado na Universitat Politècnica de València, Valencia, 2017.

Uma vez conhecida a dinâmica populacional dos cavalos Campolina, resultante do primeiro trabalho conduzido nesta tese, bem como de meu primeiro artigo, publicado em 2018 na *Animal: Pedigree analysis and inbreeding effects over morphological traits in Campolina horse population*. Era necessário agora, que conhecêssemos as características que, de fato, precisariam ser simuladas para estudar a possibilidade de seleção genômica em cavalos Campolina. No banco de dados fenotípicos, que nos foi gentilmente cedido pela ABCCCampolina, constam mais 20 características, das quais, utilizamos boa parte, mas deixamos de lado avaliações visuais para atributos morfológicos. Portanto, precisávamos agora, saber qual das várias características deveria ser, de fato, simulada para estudar a possibilidade de seleção genômica nesta população. Foi então que resolvemos conduzir algumas análises de redução de posto, pois, se fosse possível através de PCA e da análise de fatores (FA) reduzir o número de características em análise, talvez pudéssemos simular menos características, mas que representassem todas as outras.

Apresentamos então nosso segundo estudo nesta tese, comparando modelos de PCA e FA com análise multicaracterística padrão. O resultado foi bastante promissor: é possível ajustar um modelo que pressupõe que as 19 características de interesse em cavalos Campolina, são na verdade resultado de um fenômeno subjacente, complexo e não correlacionado, capturado por cinco fatores comuns. Apesar de muito promissor, resolvemos dar um passo atrás, criamos características e tentamos entender como este fenômeno subjacente é formado, foi assim que surgiu nosso terceiro trabalho, aplicando análise confirmatória de fatores e modelos de equações estruturais para estudar de forma mais profunda os resultados previamente obtidos. Foi neste ponto que tivemos o resultado mais interessante, as 28 características que utilizamos eram o resultado observável de três características não mensuráveis em essência. Fizemos então a primeira proposição de que na verdade não medimos morfologia e andamento, mas sim atributos desses dois “fenômenos”. Nossos resultados ainda permitiram concluir que a qualidade morfológica cria a morfologia que, por sua vez, cria o andamento.

Pois bem, agora nos cabia estudar o que acontecia dentro do fenômeno andamento. Aplicamos então modelos recursivos e estruturais para tentar entender como os escores visuais se unem criando o escore de andamento. O resultado foi surpreendente, apesar de se tratar de uma técnica puramente estatística, a rede causal que encontramos possui respaldo biológico do ponto de vista biomecânico e, pela primeira vez, é apresentado um coeficiente causal entre atributos de andamento. Vale acrescentar que, neste mesmo estudo tentamos viabilizar a seleção para andamento através da utilização de pseudo-fenótipos para predição de valores genéticos, mas seja por desregressão, seja pelo índice de seleção a partir dos coeficientes causais, os

ganhos em acurácia, quando comparados com a utilização do fenótipo original são pífios frente a necessidade de análise para se obter os pseudo-fenótipos.

Agora possuíamos todos os principais elementos para conduzir nossa simulação. Sabíamos como o pedigree “se comportava”; sabíamos quais características deveríamos simular e conhecíamos seus componentes de variância; entendíamos como os escores de andamento se uniam para formar a pontuação de andamento... Fomos então a simulação, fizemos testes e implementamos da melhor maneira que pudemos os resultados obtidos previamente. Conseguimos via PCA comparar as estruturas populacionais real e simulada, o que revelou uma boa adequação entre elas. Conseguimos simular um desequilíbrio de ligação aplicável a várias populações equinas e, 720 análises de predição; 720 validações; 60 predições com diferentes modelos; e 60 validações depois conseguimos os resultados para prospectar a seleção genômica em cavalos Campolina. Coloco aqui os números, para que, no futuro, quem resolva aplicar nossa abordagem esteja ciente da demanda de análises. Nossos resultados permitem concluir que a melhor opção para amostrar os animais à genotipagem é baseado no número de progênes. Ainda pudemos calcular qual deve ser o número de cavalos genotipados para que a acurácia de predição atinja seu máximo e, comparamos diversos modelos quanto a sua habilidade preditiva.

O principal resultado aqui é a possibilidade de utilizarmos o método de passo único com uma forma semi-paramétrica de se calcular a similaridade entre os animais, que chamamos de ssRKHS. Tal método parece bastante promissor para predição em características com distribuição contínua e categóricas binárias. Não fomos os primeiros a propor este método, mas fomos os primeiros a compará-los com os métodos isolados (passo único e RKHS), bem como com métodos tradicionais e, aplicamos o ssRKHS sob a ótica dos modelos generalizados. Nossos resultados permitem recomendar que, no futuro, aliando a genotipagem do maior número possível de animais, com a população de referência formada pelos animais com maior número de progênes, bem como utilizando o ssRKHS, a seleção genômica em cavalos Campolina será bem-sucedida e exequível. Quebrando o protocolo novamente, usando a primeira pessoa, quero aqui pedir desculpas pelo tamanho de minha tese, mas como espero ter conseguido mostrar nestas considerações finais, foi preciso que cada passo fosse dado para termos uma visão mais holística do melhoramento genético-genômico em cavalos Campolina.

APÊNDICE A - Prova de que $\text{Var}[\mathbf{y}] \equiv \text{Var}[\mathbf{y} - \mathbf{X}\boldsymbol{\beta}]$ nos modelos de posto reduzido.

Seja o modelo $\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}$, em que: $\mathbf{y} = [\mathbf{y}_1^T \quad \mathbf{y}_2^T]^T$

é o vetor de observações fenotípicas, ordenado por animal dentro de característica;

$\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T \quad \boldsymbol{\beta}_2^T]^T$ é o vetor de parâmetros desconhecidos de efeito fixo; $\mathbf{u} = [\mathbf{u}_1^T \quad \mathbf{u}_2^T]^T$ é o vetor de

efeitos genéticos aditivos diretos de animal; e $\mathbf{e} = [\mathbf{e}_1^T \quad \mathbf{e}_2^T]^T$ é o vetor de termos aleatórios

residuais. $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix}$ e $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix}$ são as matrizes de incidência para os efeitos fixos e

aleatórios respectivamente.

Assumindo distribuição normal multivariada, com média zero, e independência entre os efeitos genéticos e residuais, isto é $E[\mathbf{u}\mathbf{e}^T] = 0$ e, fazendo-se $\text{Var}[\mathbf{u}] = \mathbf{A} \otimes \mathbf{G}_0$ e

$\text{Var}[\mathbf{e}] = \mathbf{I} \otimes \mathbf{R}_0$, os primeiro e segundo momentos da distribuição de $\mathbf{y}$ são dados por:

$$\begin{aligned} E[\mathbf{y}] &= E[\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}] \\ &= E[\mathbf{X}\boldsymbol{\beta}] + 0 + 0 \\ &= \mathbf{X}\boldsymbol{\beta} \\ \text{Var}[\mathbf{y}] &= \text{Var}[\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}] \\ &= \text{Var}[\mathbf{Z}\mathbf{u} + \mathbf{e}] \\ &= \mathbf{Z}(\mathbf{A} \otimes \mathbf{G}_0)\mathbf{Z}^T + \mathbf{I} \otimes \mathbf{R}_0 \end{aligned} \quad (9.1)$$

Se obtemos as estimativas de $\mathbf{X}\boldsymbol{\beta}$ no modelo unicaracterística e estimamos os componentes de variância em um modelo bicaracterística reduzido, isto é

$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} - \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}$, os primeiro e segundo momentos da

distribuição de $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ são dados por:

$$\begin{aligned} E[\mathbf{y} - \mathbf{X}\boldsymbol{\beta}] &= E[\mathbf{Z}\mathbf{u} + \mathbf{e}] \\ &= 0 \\ \text{Var}[\mathbf{y} - \mathbf{X}\boldsymbol{\beta}] &= \text{Var}[\mathbf{Z}\mathbf{u} + \mathbf{e}] \\ &= \mathbf{Z}(\mathbf{A} \otimes \mathbf{G}_0)\mathbf{Z}^T + \mathbf{I} \otimes \mathbf{R}_0 \end{aligned} \quad (9.2)$$

Portanto, ambos os modelos são equivalentes, pois, modelam a mesma matriz de (co)variâncias. Se escrevermos $\mathbf{G}_0$ em termos de sua decomposição espectral, isto é

$\mathbf{G}_0 = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^T$, e aplicarmos um modelo de componentes principais obtemos $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}^* + \mathbf{e}$,

em que: $\mathbf{Z}^* = \mathbf{Z}(\mathbf{E} \otimes \mathbf{I})$ é a matriz de incidência para os efeitos aleatórios genéticos aditivos de animal, tomada em função dos autovetores de $\mathbf{G}_0$; $\mathbf{u}^* = (\mathbf{E}^T \otimes \mathbf{I})$ é o novo vetor de efeitos aleatórios; com variância igual a $\text{Var}(\mathbf{u}^*) = \mathbf{\Lambda} \otimes \mathbf{A}$. Se nenhuma redução é ajustada, ou $\mathbf{y}$ seja, no modelo de componentes principais entram os 2 componentes principais do nosso exemplo, a distribuição de $\mathbf{y}$ é a mesma que no modelo convencional, isto é:

$$\begin{aligned}
 E[\mathbf{y}] &= E[\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^*\mathbf{u}^* + \mathbf{e}] \\
 &= E[\mathbf{X}\boldsymbol{\beta}] + 0 + 0 \\
 &= \mathbf{X}\boldsymbol{\beta} \\
 \text{Var}[\mathbf{y}] &= \text{Var}[\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^*\mathbf{u}^* + \mathbf{e}] \\
 &= \text{Var}[\mathbf{Z}^*\mathbf{u}^* + \mathbf{e}] \quad , \\
 &= \mathbf{Z}(\mathbf{E} \otimes \mathbf{I})(\mathbf{\Lambda} \otimes \mathbf{A})(\mathbf{E} \otimes \mathbf{I})^T \mathbf{Z}^T + \mathbf{I} \otimes \mathbf{R}_0 \\
 &= \mathbf{Z}(\mathbf{E}\mathbf{\Lambda} \otimes \mathbf{A})(\mathbf{E}^T \otimes \mathbf{I})\mathbf{Z}^T + \mathbf{I} \otimes \mathbf{R}_0 \\
 &= \mathbf{Z}(\mathbf{E}\mathbf{\Lambda}\mathbf{E}^T \otimes \mathbf{A})\mathbf{Z}^T + \mathbf{I} \otimes \mathbf{R}_0 \\
 &= \mathbf{Z}(\mathbf{G}_0 \otimes \mathbf{A})\mathbf{Z}^T + \mathbf{I} \otimes \mathbf{R}_0
 \end{aligned} \tag{9.3}$$

De forma análoga podemos trabalhar com $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ no modelo de componentes principais, pois, a variância nos modelos mistos independe dos efeitos fixos, desde que sejam os mesmos efeitos em ambos os modelos. Assim:

$$\begin{aligned}
 E[\mathbf{y} - \mathbf{X}\boldsymbol{\beta}] &= E[\mathbf{Z}^*\mathbf{u}^* + \mathbf{e}] \\
 &= 0 \\
 \text{Var}[\mathbf{y} - \mathbf{X}\boldsymbol{\beta}] &= \text{Var}[\mathbf{Z}^*\mathbf{u}^* + \mathbf{e}] \\
 &= \mathbf{Z}(\mathbf{E} \otimes \mathbf{I})(\mathbf{\Lambda} \otimes \mathbf{A})(\mathbf{E} \otimes \mathbf{I})^T \mathbf{Z}^T + \mathbf{I} \otimes \mathbf{R}_0 \quad , \\
 &= \mathbf{Z}(\mathbf{E}\mathbf{\Lambda} \otimes \mathbf{A})(\mathbf{E}^T \otimes \mathbf{I})\mathbf{Z}^T + \mathbf{I} \otimes \mathbf{R}_0 \\
 &= \mathbf{Z}(\mathbf{E}\mathbf{\Lambda}\mathbf{E}^T \otimes \mathbf{A})\mathbf{Z}^T + \mathbf{I} \otimes \mathbf{R}_0 \\
 &= \mathbf{Z}(\mathbf{G}_0 \otimes \mathbf{A})\mathbf{Z}^T + \mathbf{I} \otimes \mathbf{R}_0
 \end{aligned} \tag{9.4}$$

Portanto tanto para modelos fatores analíticos como para modelos de componentes principais, é possível utilizar $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ ao invés de $\mathbf{y}$. desde que as estimativas de $\mathbf{X}\boldsymbol{\beta}$ venham de modelos unicaracterística, de tal sorte que ao ajustarmos um modelo apenas com a média geral como efeito fixo, pequenas diferenças nos efeitos fixos, devido a covariância entre as características, passam a ser levadas em conta por este efeito. Este procedimento reduz a demanda computacional drasticamente, bem como acelera o processo de convergência.

APÊNDICE B - Prova de que os modelos GBLUP, SNP-BLUP, Ridge-BLUP e RR-BLUP são equivalentes sob pressuposição de variância comum para os marcadores.

Seja o modelo misto $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u} + \mathbf{e}$, em que: $\mathbf{y}$ é o vetor de observações fenotípicas; $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos de efeito fixo; $\mathbf{u}$ é o vetor de efeito aleatório genético aditivo de animal; $\mathbf{e}$ é o vetor de termos residuais aleatórios; $\mathbf{X}$ e $\mathbf{Z}_1$ são as matrizes de incidência para os efeitos fixos e aleatórios, respectivamente. As equações de modelos mistos para este modelo são dadas por:

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z}_1 \\ \mathbf{Z}_1^T\mathbf{X} & \mathbf{Z}_1^T\mathbf{Z}_1 + \lambda\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}_1^T\mathbf{y} \end{bmatrix}, \quad (10.1)$$

Em que: $\lambda = \sigma_e^2 / \sigma_u^2$.

Seja ainda outro modelo misto, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_2\mathbf{a} + \mathbf{e}$, em que: $\mathbf{y}$, $\boldsymbol{\beta}$, $\mathbf{e}$ e $\mathbf{X}$ são conforme definidos acima; e $\mathbf{a}$ é o vetor de efeitos de substituição alélica dos marcadores incluídos em análise; portanto, $\mathbf{Z}_2 \neq \mathbf{Z}_1$ é a matriz de marcadores com dimensões $n \times m$, isto é:

$$\mathbf{Z}_2 = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{bmatrix} \quad (\text{sendo } n \text{ o número de animais genotipados e } m \text{ o número de}$$

marcadores). As equações de modelos mistos para este modelo são dadas por:

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z}_2 \\ \mathbf{Z}_2^T\mathbf{X} & \mathbf{Z}_2^T\mathbf{Z}_2 + \gamma\mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}_2^T\mathbf{y} \end{bmatrix}, \quad (10.2)$$

Em que: $\gamma = \sigma_e^2 / \sigma_a^2$.

Assumindo que os m marcadores estão em perfeito equilíbrio de ligação, $\mathbf{Z}_2^T\mathbf{Z}_2$ é uma matriz diagonal e as várias soluções a_i são independentes. Entretanto, este não é o caso na maioria das vezes, e $\mathbf{Z}_2^T\mathbf{Z}_2$ possui elementos não-nulos fora da diagonal. Desta forma, γ atua como um parâmetro de cumeeira se, e somente se, $E[a_i] = 0$ e $E[a_i^2] = \sigma_a^2$. Assim os modelos SNP-BLUP, Ridge-BLUP e RR-BLUP são equivalentes quando os marcadores possuem variância comum e são assumidos de efeito aleatório. Além disso, são equivalentes ao G-BLUP, pois:

$$\begin{aligned}
\mathbf{u} &= \mathbf{Z}_2 \mathbf{a} \\
u_i &= z_{2i} \mathbf{a}, i = \{1, 2, \dots, n\} \\
u_i &= z_{21i} a_1 + z_{22i} a_2 + \dots + z_{2mi} a_m \\
\text{Var}(u_i) &= \text{Var}(z_{21i} a_1 + z_{22i} a_2 + \dots + z_{2mi} a_m) \\
\sigma_u^2 &= \text{Var}(z_{21i} a_1) + \text{Var}(z_{22i} a_2) + \dots + \text{Var}(z_{2mi} a_m) \quad , \\
&= \text{Var}(z_{21i}) \sigma_a^2 + \text{Var}(z_{22i}) \sigma_a^2 + \dots + \text{Var}(z_{2mi}) \sigma_a^2 \\
&= 2p_1 q_1 \sigma_a^2 + 2p_2 q_2 \sigma_a^2 + \dots + 2p_m q_m \sigma_a^2 \\
&= 2 \sum_{i=1}^m p_i q_i \sigma_a^2
\end{aligned} \tag{10.3}$$

Em que: z_{2i} representa a i -ésima linha de $\mathbf{Z}_2$; p_i é a frequência alélica do alelo 1, no i -ésimo marcador; e $q_i = 1 - p_i$ é a frequência alélica do alelo 2, no i -ésimo marcador.

A partir desta equivalência podemos entender então como os métodos SNP-BLUP, Ridge-BLUP, RR-BLUP e G-BLUP particionam a variância genética aditiva σ_u^2 para cada marcador:

$$\sigma_a^2 = \frac{\sigma_u^2}{2 \sum_{i=1}^m p_i q_i} \tag{10.4}$$