WILSON SANCHES MATEUS

MODELO HIERÁRQUICO GENERALIZADO NORMAL ASSIMÉTRICO BAYESIANO APLICADO À ANÁLISE GENÔMICA

LAVRAS – MG

2020

WILSON SANCHES MATEUS

MODELO HIERÁRQUICO GENERALIZADO NORMAL ASSIMÉTRICO BAYESIANO APLICADO À ANÁLISE GENÔMICA

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Dr. Márcio Balestre
Orientador

Prof. Dr. Júlio Sílvio de Sousa Bueno Filho
Coorientador

LAVRAS – MG
2020

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Mateus, Wilson Sanches
Modelo Hierárquico Generalizado Normal Assimétrico Bayesiano Aplicado à Análise Genômica / Wilson Sanches Mateus. – Lavras : UFLA, 2020.
73 p. : il.

Dissertação(Mestrado)–Universidade Federal de Lavras, 2020.
Orientador: Dr. Márcio Balestre.
Bibliografia.

1. Assimetria. 2. Predição. 3. Herdabilidade. 4. Análise Genômica. I. Balestre, Márcio. II. Bueno Filho, Júlio Sílvio de Sousa. III. Título.

CDD-808.066

WILSON SANCHES MATEUS

MODELO HIERÁRQUICO GENERALIZADO NORMAL ASSIMÉTRICO BAYESIANO APLICADO À ANÁLISE GENÔMICA

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 11 de Junho de 2020.

Prof. Dr. Júlio Sílvio de Sousa Bueno Filho	UFLA
Prof. Dr. Moysés Nascimento	UFV
Profa. Dr. Isabela Regina Cardoso de Oliveira	UFLA

Dr. Márcio Balestre
Orientador

Prof. Dr. Júlio Sílvio de Sousa Bueno Filho
Co-Orientador

LAVRAS – MG
2020

Dedico esta dissertação a Deus, que sempre esteve comigo; a minha esposa, que se multiplicou em infinitas mulheres para criar um mundo para eu estudar; a minha filha, por compreender minha ausência e a minha família.

AGRADECIMENTOS

A Deus por guiar meus caminhos.

A minha esposa que sempre acreditou em mim, nunca me deixando desanimar nos momentos mais difíceis, por enfrentar sozinha a criação e educação de nossa filha, por vencer batalhas sozinhas para criar um ambiente favorável aos meus estudos. Por fim, não há palavras para agradecer a esta mulher incrível, que tenho muito orgulho.

Aos meus pais, Amado Antônio Mateus e Maria Claret Mateus, pelo apoio constante, por acreditarem nos meus sonhos e se fazerem diariamente presentes apesar da distância.
A meus irmãos, pelo companheirismo, amizade e por estarem sempre dispostos a me ajudar

Ao Prof. Dr. Márcio Balestre, pela excelente orientação e paciência em ensinar, contribuindo para o desenvolvimento deste trabalho, assim como para o meu desenvolvimento profissional e pessoal. Acima de tudo, por ser exemplo de profissionalismo e dedicação.

Ao Prof. Dr. Júlio Sílvio de Sousa Bueno Filho, por ter me recebido tão bem, me ajudado no encerramento deste trabalho.

Aos demais professores e funcionários do Departamento do DES que contribuíram para a minha formação.

Aos amigos do Curso de Estatística, Henrique, Cristian, Carlos, Luciano, Kelly, e aos demais colegas.

Aos professores Renato e Márcio, pela harmoniosa convivência e por contribuírem imensamente pelo meu amadurecimento profissional e pessoal.

Ao Ernades, por todas as conversas sobre seleção genômica, sempre esclarecendo minhas dúvidas e compartilhando experiências.

A todos os meus familiares, pelo carinho.

Aos amigos de Lavras e da UFLA, por sempre me apoiarem e por entenderem meus momentos de ausência.

Agradeço a Universidade Federal de Lavras por todos espaços concedidos, em especial ao departamento de Estatística por sempre resolver todos os contra-tempos nesta jornada. Por todo o conhecido ensinado pelos professores, aos amigos que sempre estiveram ao meu lado nesta batalha...

À CAPES e CNPQ, pelas bolsas de estudo concedidas.

RESUMO

MODELO HIERÁRQUICO GENERALIZADO NORMAL ASSIMÉTRICO BAYESIANO APLICADO À ANÁLISE GENÔMICA

Um ponto crucial da análise genômica é a seleção correta dos indivíduos geneticamente superiores para características de importância econômica. Nessa dissertação investigamos a aplicação do Modelo Hierárquico Generalizado (Assimétrico) Bayesiano (MHGB) ao problema da seleção genômica ampla (GWS). Isto porque, em geral, os modelos de seleção genômica assumem que os dados seguem uma distribuição normal, tornando-se pouco confiáveis quando isto não ocorre. Há indícios de que o MHGB pode ser utilizado com vantagens em seleção genômica, sempre que se identificar assimetrias na distribuição dos dados.

Palavras-chave: assimetria, predição, herdabilidade, análise genômica.

ABSTRACT

BAYESIAN ASYMMETRIC GAUSSIAN GENERALIZED HIERARCHICAL MODEL APPLIED TO GENOMIC ANALYSIS

A crucial point in genomic analysis is the correct selection of genetically superior individuals for characters of economic importance. In this dissertation we study the application of the Generalized Hierarchical Bayesian Model (MHGB) using the asymetric gaussian distribution to the Genome wide Selection problem (GWS). The reasoning for this choice of modelling is to challenge current models of GWS when they fail their assumptions and become less reliable. A simulation study was carried to compare reference models to MHGB. Markers of actual SNPs data where used to simulated phenotypes in different scenarios for number of genes and heritability, as well as degrees of asymmetry in the error distribution. In symmetric scenarios MHGB was almost as acurate as main reference methods GBLUP. When asymetry arises, MHGB accuracy overtakes GBLUP and all other considered methods. There are evidence that MHGB should be used with advantages in GWS, whenever asymmetries are identified in the data distribution.

Keywords: asymmetry, prediction, heritability, genetic analysis.

SUMÁRIO

1 INTRODUÇÃO 10
2 OBJETIVOS 12
2.1 Objetivo geral 12
2.2 Objetivos específicos 12
3 JUSTIFICATIVA 13
4 REFERENCIAL TEÓRICO 14
4.1 Inferência Bayesiana 14
4.2 Distribuição a priori 15
4.3 Distribuição a priori não informativa 16
4.4 Distribuição a priori conjugadas 16
4.5 Métodos computacionais 17
4.6 Método de Monte Carlo via Cadeias de Markov 18
4.7 Amostrador de Gibbs 18
4.8 Conceitos de genética 20
4.9 Análise de marcadores 20
4.10 Seleção genômica ampla (GWS) 21
4.10.1 Polimorfismos de Nucleotídeo Único (SNPs) 21
4.11 Modelos de Regressão Empregados em Seleção Genômica 22
4.12 Modelos Lineares Generalizados Mistos 24
4.12.1 Preditor Linear η 24
4.13 Distribuição Normal Assimétrica 26
5 METODOLOGIA 37
5.1 Modelo Hierárquico Generalizado Normal Assimétrico Bayesiano 37
5.1.1 Distribuições a priori para os parâmetros do modelo 39
5.1.2 Distribuições a posteriori condicionais para os parâmetros 41
5.1.3 Distribuições a posteriori condicionais completas para os parâmetros 42
5.2 MCMC 45
5.3 Simulação dos Fenótipos 46
5.4 Avaliação dos Modelos na Seleção Genômica 50
5.5 Validação do Modelo 51
5.5.1 Método k-fold 51

5.5.2 Pacotes estatísticos 52
6 RESULTADOS E DISCUSSÃO 54
6.1 Propriedades das cadeias de Markov 54
6.2 Apresentação dos resultados 55
6.3 Discussão 65
7 CONCLUSÕES 68
REFERÊNCIAS BIBLIOGRÁFICAS 69

LISTA DE FIGURAS

Figura 4.1 – Se δ for zero, tem uma distribuição normal simétrica, se δ for negativo, uma distribuição normal assimétrica a esquerda, se δ for positivo, uma distribuição normal assimétrica a direita. 28

Figura 4.2 – Distribuição $SN_2((0,0)^T, diag(1,1), diag(0,0))$ 29

Figura 4.3 – Distribuição $SN_2((0,0)^T, diag(1,1), diag(0,-5))$ 29

Figura 4.4 – Distribuição $SN_2((0,0)^T, diag(1,1), diag(5,0))$ 29

Figura 4.5 – Distribuição $SN_2((0,0)^T, diag(1,1), diag(5,5))$ 30

Figura 4.6 – Distribuição $SN_2((0,0)^T, diag(1,1), diag(-5,-5))$ 30

Figura 5.1 – Fenótipo oligogênico,herdabilidade 0,2 e $\lambda = 0$ 47

Figura 5.2 – Cenário Oligogênico com herdabilidade 0,2 e λ = 0,9996628, com validação cruzada 48

Figura 5.3 – Cenário Oligogênico com herdabilidade 0,2 e δ = 0,9999865, com validação cruzada 48

Figura 5.4 – Cenário poligênico com herdabilidade 0,2 $\lambda = 0$, sem validação cruzada . . 49

Figura 5.5 – Cenário poligênico com herdabilidade 0,2 e λ = 0,9999955 , com validação cruzada 49

Figura 5.6 – Cenário poligênico com herdabilidade 0,2 e λ = 0,99999982 , com validação cruzada 50

Figura 5.7 – Diagrama de k-fold validação cruzada. 51

Figura 6.1 – Amostras da distribuição a posteriori para os parâmetros do modelo hierárquico generalizado normal assimétrico Bayesiano. Os gráficos de traço e densidades representam, respectivamente: (A) Variância do Erro, (B) Variância de Genótipos, (C) Valores preditos, (D) efeitos de genótipos, (E) Parâmetro de assimetria e (F) Média Geral. 54

Figura 6.2 – Os gráficos ilustram os ajustes dos modelos MHGB (esquerda) e Bayes L (direita). Cenário oligogênico com herdabilidade $h^2 = 0,2$ e sem assimetria $\Delta = 0$, sem validação cruzada 64

Figura 6.3 – Os gráficos ilustram os ajustes dos modelos MHGB (esquerda) e Bayes L (direita). Cenário poligênico com herdabilidade $h^2 = 0,8$ com parâmetro de assimetria $\Delta = 10$, com validação cruzada. 65

LISTA DE TABELAS

Tabela 6.1 – Médias da distribuição *a posteriori* marginal para cada um dos parâmetros comuns aos diversos modelos de ajuste. Cenários com diferentes herdabilidades e números de genes controladores da expressão do fenótipo e sem parâmetro de assimetria ($\Delta = 0$) 58

Tabela 6.2 – Propriedades preditivas dos diferentes modelos de ajuste. Cenários com diferentes herdabilidades e números de genes controladores da expressão do fenótipo e com parâmetro de assimetria ($\Delta = 0$) e dados apresentados como porcentagens (%) para coeficientes de herdabilidades, correlações e determinações em em notação científica indicada na linha para o MSE). . . 59

Tabela 6.3 – Médias da distribuição *a posteriori* marginal para cada um dos parâmetros comuns aos diversos modelos de ajuste. Cenários com diferentes herdabilidades e números de genes controladores da expressão do fenótipo e com parâmetro de assimetria ($\Delta = 2$) 60

Tabela 6.4 – Propriedades preditivas dos diferentes modelos de ajuste. Cenários com diferentes herdabilidades e números de genes controladores da expressão do fenótipo e com parâmetro de assimetria ($\Delta = 2$) e dados apresentados como porcentagens (%) para coeficientes de herdabilidades, correlações e determinações em em notação científica indicada na linha para o MSE). . . 61

Tabela 6.5 – Médias da distribuição *a posteriori* marginal para cada um dos parâmetros comuns aos diversos modelos de ajuste. Cenários com diferentes herdabilidades e números de genes controladores da expressão do fenótipo e com parâmetro de assimetria ($\Delta = 10$) 62

Tabela 6.6 – Propriedades preditivas dos diferentes modelos de ajuste. Cenários com diferentes herdabilidades e números de genes controladores da expressão do fenótipo e com parâmetro de assimetria ($\Delta = 10$) e dados apresentados como porcentagens (%) para coeficientes de herdabilidades, correlações e determinações em em notação científica indicada na linha para o MSE). . . 63

Tabela 6.7 – Propriedades assimétricas dos diferentes modelos de ajuste. Cenários com diferentes herdabilidades e números de genes controladores da expressão do fenótipo e com parâmetro de assimetria ($\Delta = 0$), ($\Delta = 2$) e ($\Delta = 10$) . . 64

1 INTRODUÇÃO

O Método de Seleção Genômica Ampla (*GWS – Genome Wide Selection*) proposto por Meuwissen, Hayes e Goddard (2001) embasa-se na aplicação de marcadores moleculares com intuito de fazer predições dos Valores Genéticos Genômicos (*GEBV- Genomic Estimated Breeding Value*) dos indivíduos candidatos à seleção. Os marcadores moleculares mais adotados na GWS são os marcadores codominantes SNPs (*Single Nucleotide Polymorphisms*) e os marcadores dominantes DArTs (*Diversity Array Technology*). O uso de informações de marcadores possibilita agregar informações de DNA na seleção de genótipos superiores, além de proporcionar maiores ganhos genéticos com maior eficácia e menor custo, assim reduzindo o intervalo de gerações (BORÉM, 1997; RESENDE *et al.*, 2008).

Um ponto crucial da análise genômica é auxiliar na seleção dos indivíduos geneticamente superiores para características de importância econômica. Os métodos usados na identificação desses indivíduos têm evoluído continuamente, chegando-se às atuais avaliações genéticas que, usando modelos estatísticos combinam informações fenotípicas e de genealogia, na predição do valor genético dos indivíduos para as diferentes características. Quer dizer que, quanto mais próximo for o valor estimado em relação ao valor genético verdadeiro, com isto aumentando a eficiência no critério de decisão sobre a característica avaliada, podendo melhorar mais o ganho genético.

Vários métodos de predição dos efeitos dos marcadores foram propostos: G-BLUP, Bayes A e Bayes B (MEUWISSEN; HAYES; GODDARD, 2001), regressão de cumeeira Bayesiana (GIANOLA; PEREZ-ENCISO, TORO,2003), regressão kernel não paramétrica via modelos aditivos generalizados (GIANOLA; FERNANDO; STELLA, 2006), aprendizado de máquina (LONG et al., 2007), regressão stepwise (HABIER; FERNANDO; DEKKERS, 2007), regressão RKHS (Reproducing Kernel Hilbert Spaces, GIANOLA; VAN KAAM, 2008), LASSO Bayesiano (PARK; CASELLA, 2008; DE LOS CAMPOS et al., 2009), Bayes B acelerado (MEUWISSEN et al., 2009), regressão via quadrados mínimos parciais e via componentes principais (SOLBERG et al., 2009) e Bayes C, Bayes Cπ, Bayes D e Bayes Dπ (HABIER et al., 2011).

Essas metodologias para predição do GEBV em geral assumem que a distribuição dos dados, os efeitos aleatórios e os resíduos são independentes e que seguem uma distribuição normal. No entanto, há conjuntos de dados que apresentam problemas como a falta de norma-

lidade, em áreas de estudo como a economia, as ciências agrárias, as ciências biológicas e em especial, na área de genética.

Geralmente, esses casos citados necessitam da utilização de metodologias que manipulam a falta de normalidade para a variável resposta (fenótipo) e/ou efeitos aleatórios e o resíduo do modelo estatístico.

Conjuntos de dados que são não normais veem sendo estudados há algum tempo em diversas áreas, e uma alternativa para contornar o problema da não normalidade da variável resposta pode ser obtido nos seguintes modelos: Modelo Threshold ou Modelo de limiar (GIANOLA, 1982; GIANOLA e FOULLEY, 1983; FOULLEY et al., 1987), Regressão Quantílica (RQ) (KOENKER e BASSET, 1978); (NASCIMENTO *et al.*, 2017), SILVA; JÚNIOR; SILVA 2006; SAMPAIO 2009; OLIVEIRA 2019) e Modelo Misto Normal Assimétrico (BALDONI *et al.*, 2014); (BOLKER *et al.*, 2009). (LEE; NELDER, 1996), mas podem existir outros modelos que lidem com este problema.

Os modelos Threshold e o Modelo Misto Normal Assimétrico pertencem à uma classe de modelos, conhecida como Modelos Lineares Generalizados Mistos (*GLMM- Generalized Linear Model Mixed*) e podem contribuir muito para a GWS (BISCARINI et al., 2014; MONTESINOS-LÓPEZ et al., 2015a; 2015b).

A análise que considera os dados assimétricos permite descrever a influência relativa de um conjunto de variáveis explicativas na variável resposta, sem que haja a pressuposição de normalidade na distribuição dos dados, resíduo e efeitos aleatórios do modelo. Desta forma, pode-se admitir outras pressuposições para a distribuição dos dados e efeitos do modelo.

Uma das formas possíveis de implementar generalizações de modelos lineares é a construção de modelos hierárquicos (generalizados) Bayesianos (MHGB), uma alternativa que relaxa a pressuposição de normalidade, em geral substituindo esta por uma pressuposição de distribuição assimétrica, de forma a que se evite transformações nos dados. Nesta classe de modelos, as observações, os resíduos e, ou, os efeitos aleatórios podem assumir distribuição normal assimétrica.

2 OBJETIVOS

2.1 Objetivo geral

O objetivo deste trabalho foi investigar a aplicação do Modelo Hierárquico Generalizado Normal Assimétrico Bayesiano utilizando uma distribuição normal assimétrica e compará-lo com o Modelos não assimétricos correspondentes, utilizando para isso, medidas de acurácia de estimativas e de previsões de valores genéticos genômicos.

2.2 Objetivos específicos

- Implementar a análise do Modelo Hierárquico Generalizado Normal Assimétrico Bayesiano para a seleção genômica;
- Verificar as propriedades do ajuste do modelo proposto em diferentes cenários de assimetria e herdabilidade;
- Verificar o ganho nas estimativas dos parâmetros genéticos e na capacidade preditiva;
- Inferir sobre os parâmetros dos modelos gaussianos de seleção genômica;
- Comparar o Modelo Hierárquico Generalizado Normal Assimétrico Bayesiano aos Modelos Gaussianos (simétricos).

3 JUSTIFICATIVA

Os métodos tradicionais de Seleção Genômica Ampla (*GWS – Genome Wide Selection*) para predição dos efeitos dos marcadores utilizam modelos gaussianos. Contudo, esses modelos não levam em consideração à assimetria dos dados, que muitas vezes é um fenômeno típico dos caracteres em estudo.

Modelos Lineares Generalizados Mistos (GLMM) Bolker *et al.* (2009), são generalizações dos Modelos Lineares que permitem ajustar respostas com outra distribuição; que não a normal, em geral da família exponencial. Modelos Hierárquicos (Generalizados) Bayesianos são uma classe análoga de modelos, mas permitem maior flexibilidade, sendo que a distribuição a ser considerada não precisa ser da família exponencial. Tal classe de modelos pode proporcionar uma metodologia mais ampla e flexível para ajustar a complexidade de fatores genéticos e ambientais, que afetam o desempenho da seleção genômica em animais e plantas em características biológicas complexas.

Neste trabalho apresentamos a aplicação de um modelo estatístico que apresenta parâmetro de locação, escala e também um parâmetro de assimetria para modelar regressões em marcadores típicas da GWS.

Será utilizada a distribuição normal assimétrica, que faz parte de uma classe de generalizações da distribuição normal que poderiam ser empregadas. Esta classe envolve também outras distribuições como a de Laplace, a log-normal e a uniforme, podendo ser considerada uma ferramenta para reduzir os efeitos de observações atípicas e obter estimativas robustas (Box e Tiao, 1973; Walker e Gutiérrez-Peña, 1999; Liang et al., 2007). Escolhemos a distribuição normal assimétrica por ser especificamente voltada a problemas de assimetria.

O modelo é capaz de lidar com dados proveniente de distribuições com caudas mais pesadas do que a normal, com dados originados de distribuições assimétricas e com presença de *outliers*, de forma a obter estimativas robustas.

Para a implementação computacional será utilizado o *software* ***R*** (R Development Core Team, 2019), pelo fato de ser um software gratuito, prático e processamento satisfatório com modelos complexos.

4 REFERENCIAL TEÓRICO

A classe dos Modelos Lineares Generalizados Mistos (GLMM) é uma extensão dos Modelos Lineares Mistos (MLM) e dos Modelos Lineares Generalizados (GLM) McCullagh e Nelder (1989). Como tais, os GLMM são de grande importância e possuem diversas aplicações, dada a sua capacidade de modelar a super dispersão dos dados Williams (1982) e a dependência entre observações em estudos longitudinais Stiratelli, Laird e Ware (1984) ou em dados com medidas repetidas Breslow (1984), quando incorporamos efeitos aleatórios.

Neste trabalho utilizaremos o Modelo Hierárquico Generalizado Normal Assimétrico Bayesiano (MHGB), que consiste em outra formalização, potencialmente mais geral que os GLMM. O foco deste estudo é avaliar os aspectos inferenciais desta classe de modelos, e seu impacto na GWS para caracteres assimétricos ou não de potencial interesse.

Quando se deseja quantificar a taxa de transferência genética de uma população, entre uma geração e sua próxima geração pode-se utilizar modelos estatísticos, em virtude da grande complexidade inerente ao processo de herdabilidade, que é um coeficiente genético que expressa a relação entre a variância genotípica e a variância fenotípica, ou seja, mede o nível da correspondência entre o fenótipo e o valor genético.

4.1 Inferência Bayesiana

Um dos principais objetivos da estatística é fazer inferências ou predições acerca dos parâmetros de interesse seja na abordagem clássica ou Bayesiana. Uma forma sutil de diferenciar as duas abordagem é que a abordagem clássica (frequentista) o parâmetro θ, ainda que desconhecido, é tratado como uma constante (fixo) ao invés de aleatório como faz a abordagem Bayesiana.

A inferência clássica, assim como a Bayesiana, trabalha na presença de observações (y), que podem ser descritas por uma distribuição de probabilidades $f(y|\theta)$, sendo θ uma quantidade desconhecida e necessária para descrever a distribuição das observações (y). A inferência Bayesiana modela a incerteza associada aos parâmetros levando em conta informação a priori, através das distribuições de probabilidades chamadas de distribuições a priori.

A metodologia Bayesiana leva em conta duas coisas, a informação vinda da distribuição de verossimilhança e informação vinda da distribuição a priori $\pi(\boldsymbol{\theta})$. Juntando essas duas distribuições e fazendo uso do teorema de Bayes podemos obter a distribuição a posteriori para

$\boldsymbol{\theta}$, denominado por $\pi(\boldsymbol{\theta}|y)$;

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{L(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{\theta})}{f(\boldsymbol{y})} \tag{4.1}$$

em que:

$$f(\boldsymbol{y}) = \int_{\boldsymbol{\theta}} f(\boldsymbol{y},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \text{ para o caso contínuo.} \tag{4.2}$$

$$f(\boldsymbol{y}) = \sum_{\boldsymbol{\theta}} f(\boldsymbol{y},\boldsymbol{\theta}) = \sum_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{\theta}) \text{ para o caso discreto.} \tag{4.3}$$

A função $f(\mathbf{y})$, no denominador não depende de θ e, portanto, para a determinação da distribuição de interesse $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, representa apenas uma constante normalizadora, então podemos dizer que a equação 4.1 pode ser reescrita como:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = kL(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{\theta}) \tag{4.4}$$

em que $k = \frac{1}{f(y)}$ é a constante normalizadora, assim obtemos a seguinte expressão:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto L(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{\theta}) \tag{4.5}$$

o símbolo $\propto$ representa a proporcionalidade. A proporcionalidade pode ser usada porque quando se multiplica a função de verossimilhança por uma constante não se altera a informação relativa ao parâmetro $\boldsymbol{\theta}$, a distribuição a posteriori não será alterada (BOX; TIAO, 1992).

4.2 Distribuição a priori

A distribuição a priori representa informação relativa ao parâmetro antes da observação dos dados, isto é, resume a informação pertinente ao parâmetro $\boldsymbol{\theta}$, desconhecido antes da realização do experimento. Essa distribuição tem uma função importante na inferência Bayesiana, pois pode representar a probabilidade do conhecimento prévio ou ignorância sobre a quantidade desconhecida de $\boldsymbol{\theta}$ antes dos dados serem obtidos (BOX; TIAO, 1992).

A distribuição a priori assumida para o parâmetro $\boldsymbol{\theta}$ pode divergir de pesquisador para pesquisador. Existe a distribuição a priori informativa, que traz algum conhecimento empírico sobre o parâmetro desconhecido e a distribuição a priori não informativa, quando não existe informação a priori ou que o conhecimento a priori é pouco significativo (o estado de conheci-

mento ‘vago’), que não traz conhecimento nenhum sobre o parâmetro; assim, toda informação sobre o parâmetro $\boldsymbol{\theta}$ se encontrará nos dados observados.

4.3 Distribuição a priori não informativa

O emprego de uma distribuição a priori não informativa pressupõe que a informação contida nos dados seja mais relevante que a informação contida na priori. No intuito de que o conhecimento a priori seja vago, não há informações a priori referente ao parâmetro de interesse ou que a informação trazida pela priori seja pouco expressiva (GELMAN *et al.*, 2003). Podemos perceber que uma forma de obter uma distribuição a priori não informativa é pensar que todos os possíveis valores para um dado parâmetro tenham a mesma chance de ocorrer, isto é, como uma distribuição a priori uniforme,$(\pi(\theta) \propto k)$. No entanto, podem surgir algumas dificuldades com esta escolha: $\pi(\theta)$ é imprópria, isto é, $\int \pi(\theta) \to \infty$.

Gamerman e Migon (1993) apresentam algumas dificuldades inerentes a esta escolha, a saber:
(i) $\pi(\theta)$ é imprópria, ou seja, a integral sobre todos os possíveis valores de θ não converge;
(ii) se $\phi = \phi(\theta)$ é uma transformação biunívoca de θ, e se θ tem distribuição uniforme, então pelo teorema de transformações de variáveis, a densidade de ϕ é dada por:

$$\pi(\phi) = \pi(\theta(\phi)) \left| \frac{d\theta}{d\phi} \right| \propto \left| \frac{d\theta}{d\phi} \right| \tag{4.6}$$

Assim, o raciocínio que conduz à especificação de que $\pi(\theta)$ é uma constante, deveria levar também a $\pi(\phi)$ a uma constante, o que não é verdade. O ideal seria estabelecer uma regra que fosse invariante e que $\pi(\phi)$ não fosse imprópria. Como na realidade, o interesse principal está na distribuição a posteriori, e como esta é em geral, própria, mesmo quando a priori não é; a eventual impropriedade das distribuições a priori não é importante (GAMERMAN; MIGON, 1993).

4.4 Distribuição a priori conjugadas

Podemos dizer que temos uma família de distribuições a priori conjugada se as distribuições a priori e a posteriori pertencerem à mesma classe de distribuições e, dessa maneira, a atualização do conhecimento que se tem sobre o parâmetro θ abrange uma mudança nos parâmetros indexadores da família de distribuição a priori, identificados como hiperparâmetros,

que difere dos parâmetros de θ. Se temos $\boldsymbol{y}^T = (y_1, \ldots, y_n)$ um vetor de observações de variáveis aleatórias independentes e identicamente distribuídas na família exponencial, a função de distribuição conjunta é dada por:

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \exp\{a(\theta)b(y_i) + c(\theta) + d(y_i)\} \tag{4.7}$$

e sua função de verossimilhança pode ser dada por:

$$L(\boldsymbol{\theta}|y) \propto \exp\left\{a(\theta)\sum_{i=1}^{n} b(y_i) + nc(\theta)\right\} \tag{4.8}$$

em que $a(\theta)$ e $c(\theta)$ são funções reais de θ, $b(y_i)$ e $d(y_i)$ são funções reais de y, agora suponhamos que a distribuição a priori conjugada para θ dada por:

$$\pi(\boldsymbol{\theta}; k_1, k_2) \propto \exp\{k_1 a(\theta) + k_2 c(\theta)\} \tag{4.9}$$

obtém-se a seguinte distribuição a posteriori

$$\pi(\boldsymbol{\theta}|y) \propto \exp\left\{a(\theta)\left[\sum_{i=1}^{n} b(y_i) + k_1\right] + c(\theta)[n + k_2]\right\} \tag{4.10}$$

O uso de distribuições a priori conjugados é muito importante na estatística Bayesiana Gamerman e Lopes (2006) ressaltam que a distribuição a priori conjugada deve ser usada com certo cuidado, pois sua utilização está muitas vezes associadas as facilidades analíticas e nem sempre é uma representação adequada do conhecimento prévio do parâmetro.

4.5 Métodos computacionais

Para se inferir em relação a qualquer parâmetro unidimensional do vetor $\boldsymbol{\theta}$, a distribuição conjunta a posteriori dos parâmetros (multidimensional) deve ser integrada em relação a todos os outros parâmetros que a constituem, ou seja, deve-se procurar obter a distribuição marginal em relação a cada um dos parâmetros (PAULINO; TURKMAN; MURTEIRA, 2003) (BOX; TIAO, 1992).

De certa maneira existem dificuldades para a obtenção de uma forma analítica para a distribuição marginal. Essas dificuldades, em geral, devem-se à complexidade das distribuições conjuntas obtidas ou devido à dimensão do parâmetro θ em estudo. Assim, a obtenção dos parâmetros do modelo fica comprometida e formas alternativas são necessárias para seu cálculo.

Neste trabalho foi utilizado um algoritmo especial para obtenção de uma amostra da distribuição conjunta a posteriori, baseado no método de Monte Carlo via Cadeias de Markov (MCMC).

4.6 Método de Monte Carlo via Cadeias de Markov

Considera-se que uma cadeia de Markov é um processo estocástico no qual o próximo passo da cadeia, κ_{t+1}, depende somente do passo atual, κ_t e dos dados, (GAMERMAN; LOPES, 2006). Geralmente as primeiras interações são dominadas pelo passo inicial, κ_1 e por isso são descartadas. Esse processo é conhecido como aquecimento da cadeia (*burnin*). Considerando que as observações serão independentes entre si, para diminuir a alta correlação existente entre os valores amostrais, deve-se considerar um espaçamento entre as interações afim de garantir a independência das amostras, então digamos que k iterações, esse valor e conhecido como *thinning*.

Esperamos com o método MCMC obtenha uma amostra da distribuição conjunta dos parâmetros de interesse, por meio de um processo iterativo. Ao final de cada passo de atualização, os valores gerados são considerados amostras aleatórias da distribuição de probabilidade conjunta. Os algoritmos mais utilizados para obter amostras em inferência Bayesiana são o amostrador de Gibbs (GIBBS SAMPLER) e o Metropolis-Hastings.

4.7 Amostrador de Gibbs

Sob um enfoque Bayesiano para inferência, constantemente recaem em integrações de funções no espaço k-dimensional; na maioria das vezes essas integrais são analiticamente intratáveis. Existem diversas maneiras para contornar esse problema, porém, uma das mais conhecidas seja o método de monte Carlos via cadeias de markov (MCMC). Incorporado ao MCMC, o amostrador de Gibbs vem sendo o mais utilizado quando as distribuições condicionais completas permitem amostragem direta.

Segundo Paulino, Turkman e Murteira (2003), Geman e Geman (1984) introduziram o amostrador de Gibbs como um algoritmo de simulação de distribuições multivariadas complexas e de k-dimensão, onde é possível amostrar diretamente das distribuições a posteriori condicionais dos parâmetros, ao invés de marginalizar a distribuição integrando-se sobre uma distribuição conjunta (GELMAN *et al.*, 2014).

O algoritmo pode ser brevemente descrito, supondo-se que sua distribuição de interesse seja $\pi(\theta|y)$, em que $\pi = (\theta_1, \theta_2, ..., \theta_k)^T$.

Considere que o vetor de parâmetros θ seja um vetor de k-dimensional, $(\theta_1, \theta_2, ..., \theta_k)$ $(\theta_1, \theta_2, ..., \theta_k) \in \Theta \subseteq \mathbb{R}^k$. O amostrador de Gibbs consiste em selecionar um componente aleatório i de θ e para gerar uma amostra $(\theta_1, \theta_2, ..., \theta_k)$ de π procede-se do seguinte modo:
(1) Inicialize atribuindo valores aos parâmetros $\boldsymbol{\theta}^{(t)}=\left(\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_k^{(t)}\right)$ para t=0
(2) Para k=1, ..., k, os valores para o parâmetro θ_i são obtidos a partir da distribuição condicional completa, dados os valores atuais de todos os outros parâmetros do modelo e os dados observados. Ou seja, para cada parâmetro θ_i, obtêm-se o (t+1) -ésimo valor da cadeia partir de $p(\theta_i|\boldsymbol{y}, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k)$.

Um ciclo é dado sequencialmente, amostrando valores de

$$\begin{aligned}
\theta_1^{(t+1)} &\sim p\left(\theta_1|\boldsymbol{y}, \theta_2^{(t+1)}, \ldots, \theta_k^{(t+1)}\right) \\
\theta_2^{(t+1)} &\sim p\left(\theta_2|\boldsymbol{y}, \theta_1^{(t+1)}, \theta_3^{(t+1)} \ldots, \theta_k^{(t+1)}\right) \\
&\vdots \\
\theta_k^{(t+1)} &\sim p\left(\theta_k|\boldsymbol{y}, \theta_1^{(t+1)}, \theta_3^{(t+1)} \ldots, \theta_{k-1}^{(t+1)}\right)
\end{aligned}$$

(3) Repita a etapa (2) por um número grande de t iterações

A sequência $\boldsymbol{\theta}^{(t)}, t = 1, 2, \ldots t$, é uma efetuação de uma cadeia de Markov. Depois de muitas iterações quando $t \to \infty$, a amostra convergirá para aproximadamente a distribuição conjunta de todas as variáveis (BERNARDO; SMITH, 1994).

Sob condições bastante gerais, prova-se que a cadeia assim gerada é ergódica e que a distribuição de equilíbrio é precisamente a distribuição $\pi_\theta(\theta) = \pi(\theta_1, ..., \theta_p)$.

As amostras do início da cadeia podem não representar, com precisão, a distribuição desejada e são, geralmente, descartadas (o período de *burn-in*). Por outro lado, em cada iteração, o amostrador de Gibbs gera a amostragem de cada distribuição condicional e utiliza o novo valor para amostrar as outras distribuições condicionais. Neste caso, se houver uma autocorrelação levada entre os vários estados da cadeia, deve-se escolher amostras espaçadas de modo que garanta a não correlação (salto denominado *thinning*).

Frequentemente, o valor esperado de qualquer variável pode ser aproximado pela média de todas as amostras e passa a descrever o conhecimento sobre o parâmetro de interesse. A simulação da distribuição a posteriori pelo amostrador de Gibbs envolve a simulação das con-

dicionais completas. Porém, quando a simulação não é acessível, pode-se usar outras técnicas para simular as distribuições condicionais. Dentre as técnicas e detalhes sobre algoritmos, há vários trabalhos como (BERNARDO; SMITH, 1994), (GELMAN *et al.*, 2014) e (SORENSEN; GIANOLA, 2007).

Por meio das distribuições conjuntas a posteriori é possível obter os resumos a posteriori para todos os parâmetros de interesse, tais como: média, mediana, moda e intervalos de credibilidade.

4.8 Conceitos de genética

Para entender a metodologia da GWS é necessário conhecer alguns conceitos básicos de genética.

Genética é a parte da Biologia que estuda a hereditariedade, ou seja, é a ciência que investiga as razões de semelhanças que se manifestam nos organismos relacionados por descendência. *Quantitative trait loci* conhecido por QTLs são regiões do genoma responsáveis pela expressão de caracteres fenotípicos.

O indivíduo conhecido como diploide tem dois alelos em um determinado loco, quando isso acontece de dois alelos idênticos no mesmo loco, diz-se que ele é homozigoto: caso estes dois alelos sejam diferentes no mesmo loco em cada cromossomo, diz que ele é heterozigoto. Habitualmente emprega-se letras maiúsculas (A) para indicar alelos de genes dominantes e letra minúscula (a)para indicar alelos de genes recessivos, tais como: AA (homozigoto dominante), Aa(heterozigoto), aa (homozigoto recessivo) e ainda existe o indivíduo com alelos co-dominantes. Sendo identificados por (A- e aa).

4.9 Análise de marcadores

É considerado análise de marcador um estudo de associação em que cada marcador é considerado um candidato a QTL. Se o grupo de marcadores é extremamente denso, a maioria dos QTL serão potencialmente detectáveis por causa da ligação estreita com os marcadores. Como cada marcador será um candidato, então todos os marcadores serão incluídos na análise de QTL podendo deixar o modelo supersaturado. Espera-se que muitos dos parâmetros do modelo seja não significativo, levando a necessidade de um processo de seleção de variáveis

para incluir e excluir marcadores do modelo (YI; GEORGE; ALLISON, 2003); (YI *et al.*, 2005); (HÄKKINEN *et al.*, 1998)

4.10 Seleção genômica ampla (GWS)

A seleção genômica ampla (GWS) surgiu por volta do ano 2001, com o intuito de predizer o fenótipo de um indivíduo baseado em informações de marcadores moleculares utilizando métodos estatísticos associados com os dados genéticos (MEUWISSEN; HAYES; GODDARD, 2001).

A GWS é uma ferramenta importante no programa de melhoramento genético, levando a uma redução no tempo do ciclo do melhoramento de plantas e animais, podendo proporcionar aumentos nos ganhos de seleção por unidade de tempo (XU, 2013).

Os marcadores moleculares são definidos como um segmento do DNA por todo genoma o que permite fazer inferências diretas entre os indivíduos. O crescente avanço das técnicas de marcadores moleculares permitiu a geração de dados que estão disponíveis através de uma enorme literatura em forma de artigos científicos e revistas especializadas na área.

4.10.1 Polimorfismos de Nucleotídeo Único (SNPs)

No DNA onde encontramos as quatro bases nucleotídeos, as bases organizam-se aos pares em uma sequencia (Adenina-Timina, Citosina-Guanina). Single nucleotide polymorphisms (SNPs), ocorre entre trocas das bases purínicas (A/G) ou bases pirimídicas (G/T). Os SNPs são a classe mais abundante de variação genética encontrada em genoma, podendo representar 90 por cento do genoma humano (BROOKES, 1999).

Nos últimos anos, houve um avanço nas tecnologias de genotipagem de SNPs, o que levou o método de seleção genômica passar a ter grande interesse (DAETWYLER *et al.*, 2010). O método GWS proposta por Meuwissen, Hayes e Goddard (2001) consiste na análise de um grande número de marcadores amplamente distribuídos no genoma. Obtendo as informações genotípicas a partir dos marcadores moleculares, os efeitos podem ser estimados baseando-se em dados fenotípicos de uma população de estimação.

Estimando os efeitos genéticos aditivos do modelo, será possível obter a predição dos valores genéticos genômicos (VGG) que são os GEBV nos modelos. As acurácias dos modelos

são obtidas nas populações de validação, em seguida, eles são aplicados em populações de seleção (CROSSA *et al.*, 2011)

De acordo com Goddard e Hayes (2007) para utilizar a seleção genômica ampla são necessários três conjuntos de populações: população de estimação, população de validação e população de seleção. Na população de validação o conjunto de dados é menor que na população de estimação, portanto, essa amostra contem indivíduos genotipados e fenotipados para a características em análise.

Esta amostra é aplicada para testar e verificar a acurácia dos preditores usando os efeitos estimados dos valores genéticos genômicos (VGG). Os VGG são preditos usando os efeitos estimados da população de estimação e submetidos a análise de correlação com os valores fenotípicos observados (RESENDE, 2007). Vários métodos de predição de valores genéticos genômicos foram propostos, tais como: Bayes A, Bayes B e Blup/GWS (MEUWISSEN; HAYES; GODDARD, 2001) machine learning (Am de long et al 2007), regressão RKHS (reproducing kernel Hilbert Space) (GIANOLA; KAAM, 2008) , Lasso Bayesiano (CAMPOS *et al.*, 2009), regressão quantílica regularizada (NASCIMENTO *et al.*, 2017) entre outros.

4.11 Modelos de Regressão Empregados em Seleção Genômica

Com o surgimento dos marcadores moleculares (SNPs) que percorre todo genoma, sendo utilizados na seleção genômica para predição do mérito genético individual para características de interesse, mesmo assim sua aplicação ainda é um desafio, pois muitas vezes não é possível estimar livremente o efeito de cada SNPs sobre o fenótipo devido a problemas de multicolinearidade (números de parâmetros é maior que o número de observações). Segundo Gianola (2013) é necessário a utilização de métodos estatísticos que consideram a seleção de covariáveis (problemas de multicolinearidade) e a regulação do processo de estimação (problemas de dimensionalidade).

O modelo proposto por Meuwissen, Hayes e Goddard (2001) para seleção genômica é dada por:

$$y = 1\mu + \sum_{i=1}^{n} x_i g_i + e \tag{4.11}$$

Esse modelo ao assumir que $\mathrm{g_i} \sim \mathrm{N}\left(0, \sigma^2\right) \forall \mathrm{i} = 1, 2, \ldots, \mathrm{I}$, considerando que todos os marcadores dispõem de uma mesma variância, o modelo é denominado de Ridge Regression.

Quando ajustado por sob o ponto de vista frequentista, o mesmo é denominado de RR-BLUP e, quando ajustado sob o enfoque Bayesiano, de BRR (Bayesian Ridge Regression).

Conforme (MEUWISSEN; HAYES; GODDARD, 2001) que apresentaram também um modelo Bayesiano denominado Bayes A, na qual assume-se a priori que $g_i \sim N\left(0, \sigma_i^2\right)$, em que cada marcador apresenta uma variância especifica. Neste modelo na versão Bayesiana é considerada uma distribuição normal para o efeito de marcadores, uma distribuição uniforme para a média geral e uma distribuição inversa qui quadrada para a variância residual e para a variância do efeito de cada marcador.

Os autores acima citados perceberam um problema no método Bayes A: o fato de que as distribuições das variâncias dos efeitos dos marcadores não contemplavam uma massa de densidade no valor zero, o que não condiz com a realidade, pois alguns marcadores não apresentam variância genética. O método Bayes B já emprega densidade a priori com massa de densidade $\sigma_{gi}^2 = 0$ considera-se que $\sigma_{gi}^2 = 0$ com probabilidade π.

Do mesmo modo os autores já mencionados anteriormente dizem que regressão Bayesiana pode ser utilizada em casos em que se tem mais marcadores (covariáveis) do que observações; em que determinadas distribuições a priori impõem regularização no ajuste do modelo, sob forma de encolhimento dos coeficientes de regressão (Shrinkage), sendo uma forma de fazer esse encolhimento por meio da regressão Lasso (Least Absolute Shrinkage and Selection Operador) Tibshirani (1996) a qual combina seleção de variáveis via regularização dos coeficientes de regressão.

A versão Bayesiana da regressão Lasso (BL) para seleção genômica foi proposta por (CAMPOS *et al.*, 2009). Essa maneira que consiste na construção de estimadores de coeficientes de regressão do modelo original de (MEUWISSEN; HAYES; GODDARD, 2001) de forma que esse modelo (BL) tem um parâmetro λ que controla a força de regularização, de forma que $\lambda = 0$ não há regularização.

Devido ao fato que cada método leva uma letra do alfabeto, todos esses métodos foram formalmente conhecidos na literatura como alfabeto Bayesiano para seleção genômica (GIANOLA, 2013).

Na literatura existem os modelos estatísticos para seleção genômica, tal qual (VARONA *et al.*, 2008) propuseram a utilização de modelos lineares mistos com distribuições assimétricas nos resíduos para lidar com o problema no contexto da criação de animais, quando a informação pedigree está disponível. Nascimento *et al.* (2017) propuseram a Regressão Quantílica

Regularizada como uma maneira de superar a questão das distribuições não simétricas quando a informação do marcador está disponível.

4.12 Modelos Lineares Generalizados Mistos

Podemos entender que a classe dos Modelos Lineares Generalizados Mistos (GLMM) é uma extensão dos Modelos Lineares Mistos (MLM) e dos Modelos Lineares Generalizados (GLM) McCullagh e Nelder (1989). O GLMM pode ser utilizado em diversas áreas devido sua capacidade de modelar a super dispersão dos dados Williams (1982) e a dependência entre observações em estudos longitudinais Stiratelli, Laird e Ware (1984) ou em dados com medidas repetidas Breslow (1984), quando incorporamos efeitos aleatórios. Quando aplicado no contexto de genética quantitativa, a inclusão de tais efeitos é uma etapa fundamental na definição do modelo estatístico, pois é por meio do uso de componentes aleatórios que será modelada e inferida a dependência genética existente.

Além disto, a classe dos GLMM permite acomodar outras distribuições da família exponencial como as distribuições gama, inversa gaussiana, binomial e poisson, por exemplo, além de permitir o uso de funções de respostas não lineares.

4.12.1 Preditor Linear η

Através da teoria dos modelos lineares generalizados mistos, os efeitos fixos e aleatórios, que influenciam determinada característica, são combinados formando um preditor linear, descrito por:

$$\eta = X\beta + Za \tag{4.12}$$

em que,

η é o preditor linear

β é o vetor de efeitos fixos

a é o vetor dos efeitos genéticos aditivos

X *e* Z são respectivamente as matrizes de incidência dos efeitos de β *e* a.

Matricialmente podemos expressar o GLMM da seguinte maneira:

$$y = \eta + \varepsilon = X\beta + Za + \varepsilon \tag{4.13}$$

com $\varepsilon \sim \mathrm{N}(0, \mathrm{R})$. De acordo com a distribuição de probabilidade da característica analisada é definida a função de ligação apropriada. Quando a pressuposição de normalidade pode ser assumida, a função de ligação identidade é adotada.

Até esse ponto, tudo o que dissemos se aplica igualmente a modelos lineares mistos e a modelos lineares generalizados mistos. Agora vamos nos concentrar no que torna os GLMM únicos. De acordo com (BRUIN, 2011) o que é diferente entre "MLM"do GLMM é que as variáveis de resposta podem vir de diferentes distribuições além das gaussianas. Portanto, em vez de modelar as respostas diretamente, algumas funções de ligação são frequentemente aplicadas, como por exemplo a função de ligação log.

Seja o preditor linear η, como combinação dos efeitos fixos e aleatórios.

$$\eta = X\beta + Za$$

a função de ligação genérica é chamada $g(.)$. A função de ligação relaciona o resultado de y ao preditor linear η. Portanto:

$$\begin{aligned} &\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Za} \\ &g(\cdot) = \text{ função de ligação} \\ &h(\cdot) = g^{-1}(\cdot) = \text{ função de ligação inversa} \end{aligned} \tag{4.14}$$

Portanto, nosso modelo para a expectativa condicional de y (condicional porque é o valor esperado, dependendo do nível dos preditores) é:

$$g(E(y)) = \eta \tag{4.15}$$

também poderíamos modelar a expectativa de $\mathbf{y}$:

$$E(y) = h(\eta) = \mu \tag{4.16}$$

com $\mathbf{y}$ igual a:

$$\mathbf{y} = h(\boldsymbol{\eta}) + \boldsymbol{\varepsilon} \tag{4.17}$$

Para um resultado contínuo em que assumimos uma distribuição normal, a função de ligação mais comum é simplesmente a identidade. Nesse caso, existem algumas propriedades

especiais que simplificam as coisas:

$$\begin{aligned} g(\cdot) &= h(\cdot) \\ g(E(X)) &= E(X) = \mu \\ g(\text{Var}(X)) &= \text{Var}(X) = \Sigma^2 \\ PDF(X) &= \left(\frac{1}{\Sigma\sqrt{2}}\right) e^{\frac{-(x-\mu)^2}{2\Sigma^2}} \end{aligned} \tag{4.18}$$

Assim, podemos ver como, quando a função de ligação é a identidade; ela basicamente desaparece e voltamos à nossa especificação usual de médias e variações para a distribuição normal, que é o modelo usado para modelos mistos lineares típicos. Assim, os modelos mistos lineares generalizados podem acomodar facilmente o caso específico de modelos mistos lineares, mas generalizam ainda mais.

Assim, uma forma de representar um modelo pode ser descrito da seguinte maneira:

$$y|\beta, a, \sigma_\varepsilon^2 \sim \pi(y|\beta, a, \sigma_\varepsilon^2) = N(X\beta + Za, R) \tag{4.19}$$

4.13 Distribuição Normal Assimétrica

A distribuição normal é aplicada em diversas áreas de estudo para modelar dados simétricos e assimétricos. No entanto, algumas ocasiões, esta distribuição não representa suficientemente bem uma situação real, em especial quando os dados originam-se de uma distribuição com grau significativo de assimetria. Modelos que fazem uso de generalizações da distribuição normal podem considerar a curtose ou a assimetria da distribuição. Neste trabalho nos deteremos na assimetria. Dentro da classe da distribuição normal generalizada, pode-se encaixar uma variedade de distribuições, como a de Laplace, a normal, log-normal e a uniforme. Esta é considerada uma ferramenta para tornar análises e estimativas mais robustas (BOX; TIAO, 1973), (WALKER; GUTIÉRREZ-PENA, 1999) e (LIANG *et al.*, 2007).

A distribuição normal assimétrica, apresentada em Arellano-Valle, Gómez e Quintana (2003) e Arellano-Valle e Del pino (2003), em sua versão Bayesiana, representa uma extensão da distribuição normal na qual um parâmetro modela o sentido e a intensidade da assimetria na distribuição.

Alerta: OLIVEIRA (2010)

Para o que se segue até o final desta seção, ou seja, a apresentação formal da distribuição normal assimétrica com suas definições, lemas e proposições, poderíamos simplesmente nos remeter a OLIVEIRA (2010) e em certa medida, Dávila, Bolfarine e Arellano-Valle (2004). No entanto, para fins didáticos, reproduzimos aqui este desenvolvimento.

Distribuição Normal Assimétrica conforme OLIVEIRA (2010)

Sejam:

$$\phi_n(X|\mu,\Sigma) \quad e \quad \Phi_n(X|\mu,\Sigma) \tag{4.20}$$

respectivamente a função de densidade de probabilidade (fdp) e a função de distribuição (fdc) normal $N_n(\mu,\Sigma)$. Quando $\mu = 0$, e $\Sigma = I_n$ (n x n matriz de identidade), podemos escrever estas funções como $\phi_n(x)$ e $\Phi_n(x)$. Também admita-se os seguintes apontamentos sejam feitos: na matriz de variância e covariância composta de uma diag$(a_1,...,a_n)$, para representar uma matriz diagonal com elementos $(a_1,...,a_n)$ na sua diagonal principal e I_n para representar uma matriz identidade de dimensão $(n\ x\ n)$.

Uma definição importante para a compreensão deste trabalho, que será mostrada a seguir pode ser encontrada em (DÁVILA; BOLFARINE; ARELLANO-VALLE, 2004) e (OLIVEIRA, 2010).

DEFINIÇÃO 01: Um vetor aleatório sendo n-dimensional Y segue uma distribuição n-variada normal assimétrica com vetor de locação $\mu \in \mathbb{R}^n$ e matriz de variância-covariância Σ (n x n matriz identidade positiva) e com uma matriz de assimetria Δ (n x k), se a sua função de densidade é dada pela expressão 4.21;

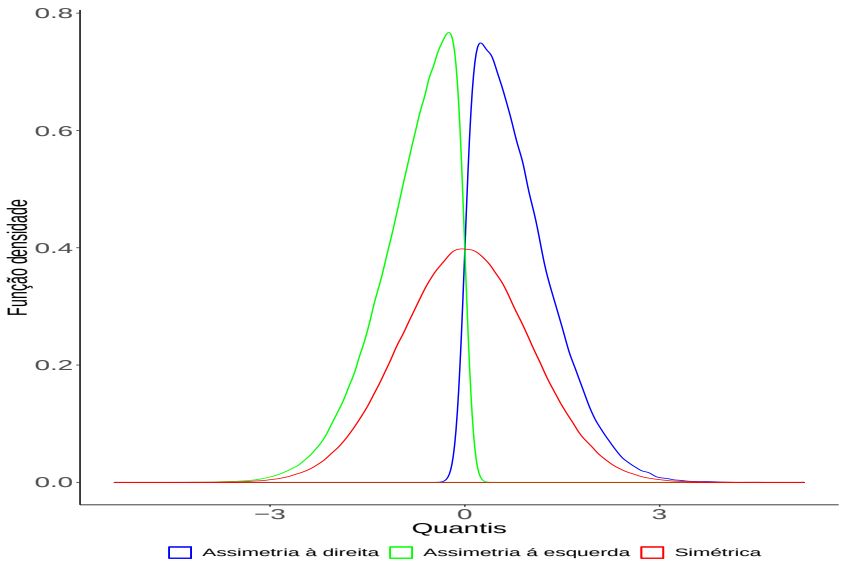
$$\begin{aligned} f(y|\mu,\Sigma,\Delta) =& 2^n \phi_n\left(y|\mu,\Sigma+\Delta\Delta^T\right)\Phi_k\left(\Delta^T\left(\Sigma+\Delta\Delta^T\right)^{-1}\right) \\ & (y-\mu)|0,\left(I_k+\Delta^T\Sigma^{-1}\Delta\right)^{-1} \end{aligned} \tag{4.21}$$

sendo assim, podemos empregar a seguinte notação: $Y \sim SN_{n,k}(\mu,\Sigma,\Delta)$ e quando n = k, obtemos $Y \sim SN_n(\mu,\Sigma,\Delta)$.

Podemos observar que para $\Delta = 0$ a equação acima, recuperamos a distribuição multivariada normal simétrica $N_n(\mu, \Sigma)$, considerando que $(I_k + \Delta^T \Sigma^{-1} \Delta)^{-1} = I_k - \Delta^T (\Sigma + \Delta\Delta^T)^{-1} \Delta$ em (ARELLANO-VALLE; BOLFARINE; LACHOS, 2007).

Neste caso podemos assumir que $\Delta\Delta^T = \Delta^2$, fazendo n=k, $\Delta = diag(\delta_1 ... \delta_n)$ e Σ é a diagonal que faz a equação 4.21 se tornar a equação proposta por (SAHU; DEY; BRANCO, 2003). Se δ for zero, temos uma distribuição normal simétrica, se δ for negativo, temos uma distribuição normal assimétrica a esquerda, se for positivo, temos uma distribuição normal assimétrica a direita. Conforme podemos observar na figura 4.1:

Figura 4.1 – Se δ for zero, tem uma distribuição normal simétrica, se δ for negativo, uma distribuição normal assimétrica a esquerda, se δ for positivo, uma distribuição normal assimétrica a direita.



Aqui será apresentado alguns contornos de densidades associadas a distribuição multivariada skew normal $SN(0, \Sigma, diag(\delta))$para diferentes valores de δ essa Figuras também podem ser encontradas em OLIVEIRA (2010) .

Note que especificando o modelo $Y \sim SN_n(\mu, \Sigma, \Delta)$ em que temos:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Delta = \begin{bmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{bmatrix} \tag{4.22}$$

Será usada a notação $SN((\mu_1, \mu_2)^T, diag(1,1), diag(0,0))$

Figura 4.2 – Distribuição $SN_2((0,0)^T, diag(1,1), diag(0,0))$

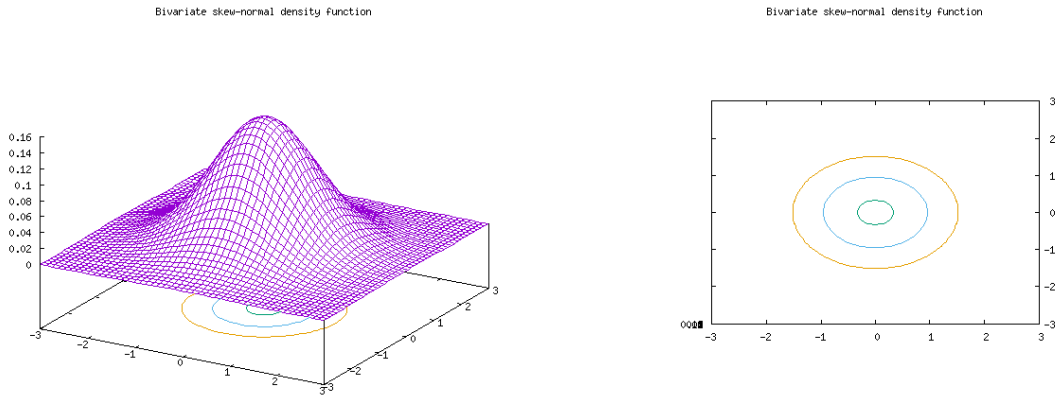


Figura 4.3 – Distribuição $SN_2((0,0)^T, diag(1,1), diag(0,-5))$

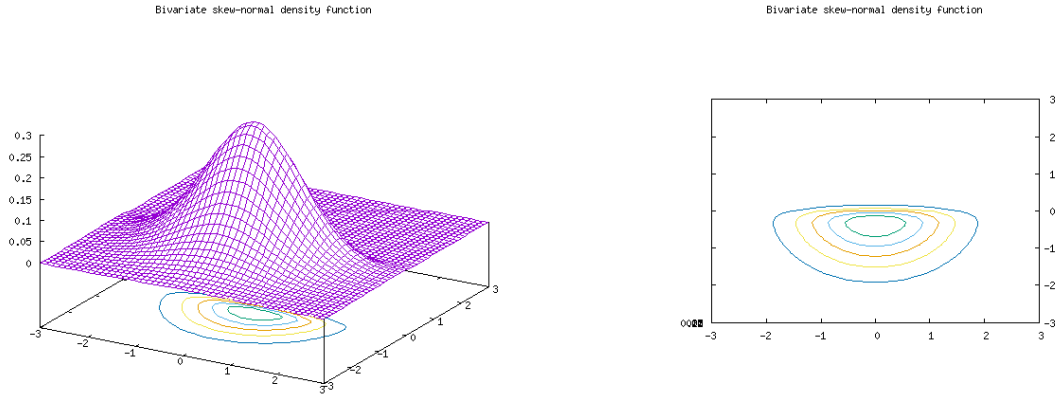


Figura 4.4 – Distribuição $SN_2((0,0)^T, diag(1,1), diag(5,0))$

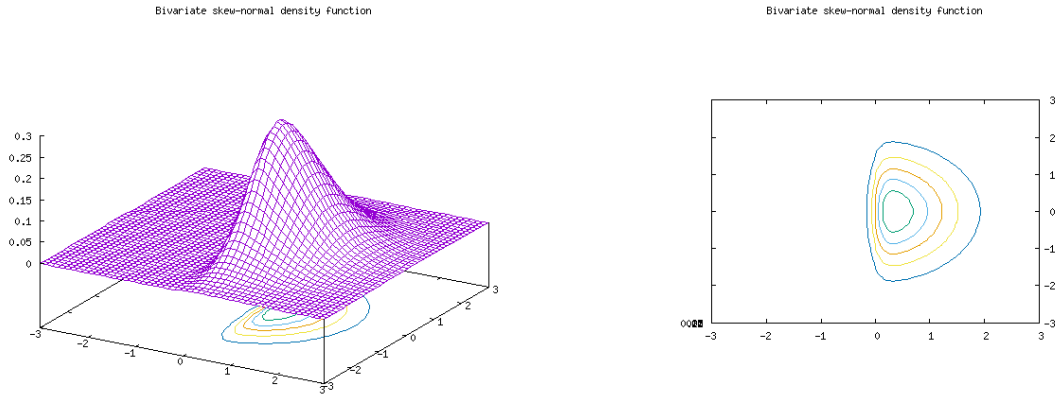


Figura 4.5 – Distribuição $SN_2((0,0)^T, diag(1,1), diag(5,5))$

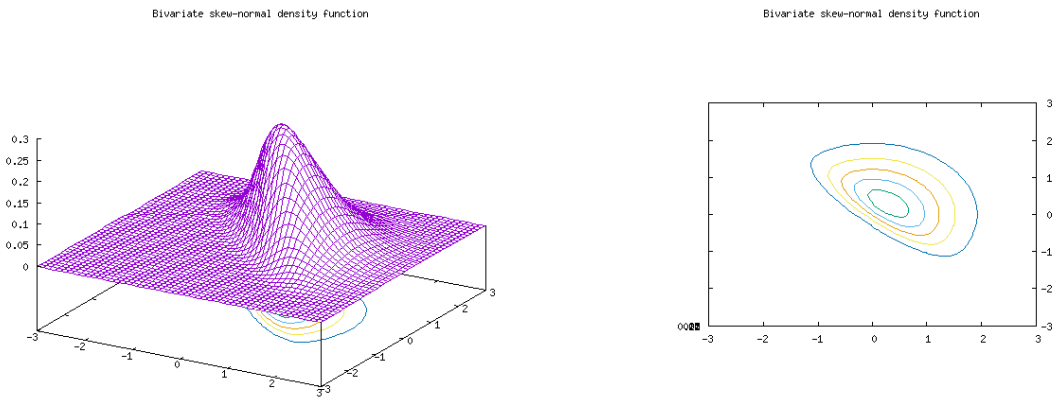
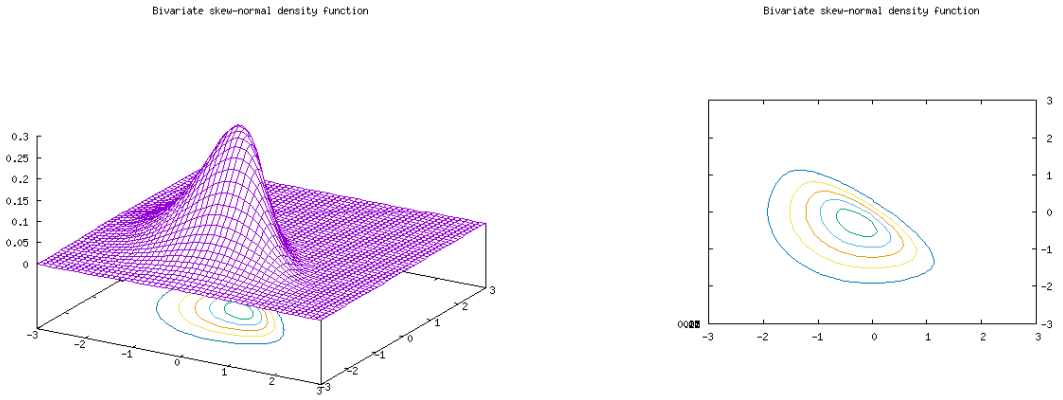


Figura 4.6 – Distribuição $SN_2((0,0)^T, diag(1,1), diag(-5,-5))$



Todas as figuras acima foram construídas no seguinte site: http://azzalini.stat.unipd.it/SN/plot-SN2.html

Vamos introduzir neste contexto, alguns conceitos, lema, corolário e proposições que auxiliará na compreensão das expressões matemáticas. Sendo que estas ajudaram na inferência da distribuição normal assimétrica multivariada.

O lema 1 que será apresentado a seguir pode ser encontrado no trabalho OLIVEIRA (2010)

LEMA 1:

Seja $Y|X = Y \sim N_p(\mu + Ax, \Sigma)$ e $X \sim N_q(\eta, \Omega)$.

podemos notar as seguintes expressões matemáticas que nos mostra que:

$$\phi_p(y|\mu + Ax, \Sigma)\phi_q(x|\eta, \Omega) = \phi_p(y|\mu + A\eta, \Sigma + A\Omega A^T)X\phi_q(x|\eta + \Lambda A^T\Sigma^{-1}(y - \mu - A\eta), \Lambda) \tag{4.23}$$

Reescrevendo o lado esquerdo da igualdade tem-se:

$$\frac{(2\pi)^{-p/2}}{\sqrt{|\Sigma|}}\exp\left[-\tfrac{1}{2}(y-\mu-Ax)^\top\Sigma^{-1}(y-\mu-Ax)\right]\times\frac{(2\pi)^{-q/2}}{\sqrt{|\Omega|}}\exp\left[-\tfrac{1}{2}(x-\eta)^\top\Omega^{-1}(x-\eta)\right]=$$

$$\frac{(2\pi)^{-p/2}(2\pi)^{-q/2}}{\sqrt{|\Sigma||\Omega|}}\exp\left\{-\tfrac{1}{2}\left[(y-\mu-Ax)^\top\Sigma^{-1}(y-\mu-Ax)+\ (x-\eta)^\top\Omega^{-1}(x-\eta)\right]\right\}=$$

Fazendo o mesmo para o lado direito da igualdade tem-se:

$$\begin{aligned}
&\frac{(2\pi)^{-p/2}}{\sqrt{|\Sigma+A\Omega A^\top}}\exp\left[-\frac{1}{2}(y-\mu-A\eta)^\top\left(\Sigma+A\Omega A^\top\right)^{-1}(y-\mu-A\eta)\right]\times\\
&\qquad\frac{(2\pi)^{-q/2}}{\sqrt{|\Lambda|}}\exp\left\{-\frac{1}{2}\left[x-\eta-\Lambda A^\top\Sigma^{-1}(y-\mu-A\eta)\right]^\top\Lambda^{-1}\times\right.\\
&\qquad\left.\left[x-\eta-\Lambda A^\top\Sigma^{-1}(y-\mu-A\eta)\right]\right\}=\\
&\frac{(2\pi)^{-p/2}(2\pi)^{-q/2}}{\sqrt{\left|\Sigma+A\Omega A^\top\right||\Lambda|}}\exp\left\{-\frac{1}{2}\left\{(y-\mu-A\eta)^\top\left(\Sigma+A\Omega A^\top\right)^{-1}(y-\mu-A\eta)+\right.\right.\\
&\quad\left.\left.\left[x-\eta-\Lambda A^\top\Sigma^{-1}(y-\mu-A\eta)\right]^\top\Lambda^{-1}\left[x-\eta-\Lambda A^\top\Sigma^{-1}(y-\mu-A\eta)\right]\right\}\right\}
\end{aligned}$$

em que $\Lambda=(\Omega^{-1}+A^T\Sigma^{-1}A)^{-1}$ e $z=y-\mu-A\eta$ e $w=x-\eta$

Assim, o lado esquerdo da igualdade passa a ficar da seguinte forma

$$\frac{(2\pi)^{-p/2}(2\pi)^{-q/2}}{\sqrt{|\Sigma||\Omega|}}\exp\left\{-\frac{1}{2}\left[(z-Aw)^\top\Sigma^{-1}(z-Aw)+w^\top\Omega^{-1}w\right]\right\}$$

e o lado direito fica:

$$\begin{aligned}
&\frac{(2\pi)^{-p/2}(2\pi)^{-q/2}}{\sqrt{\left|\Sigma+A\Omega A^\top\right||\Lambda|}}\exp\left\{-\frac{1}{2}\left\{z^\top\left(\Sigma+A\Omega A^\top\right)^{-1}z+\right.\right.\\
&\qquad\left.\left.\left[w-\Lambda A^\top\Sigma^{-1}z\right]^\top\Lambda^{-1}\left[w-\Lambda A^\top\Sigma^{-1}z\right]\right\}\right\}
\end{aligned}$$

A prova é seguida por 2 partes. A primeira parte é demonstrar que as expressões apresentadas dentro da exponencial em ambos os lados esquerdo e direito são iguais, isto é,

$$\underbrace{(z-Aw)^\top \Sigma^{-1}(z-Aw)+w^\top \Omega^{-1}w}_{**} =$$

$$\underbrace{z^\top \left(\Sigma+A\Omega A^\top\right)^{-1} z+\left(w-\Lambda A^\top \Sigma^{-1}z\right)^\top \Lambda^{-1}\left(w-\Lambda A^\top \Sigma^{-1}z\right)}_{*}$$

A segunda parte é provar que os termos dentro da raíz em ambos os lados esquerdo e direito são iguais, ou seja, $\underbrace{|\Sigma||\Omega|}_{\bullet\bullet} = \underbrace{\left|\Sigma+A\Omega A^\top\|\Lambda\right|}_{\bullet}$, com $|A| = det(A)$

Parte I: Para facilitar a demonstração, inicia-se de $*$ para chegar em $**$:

$$z^\top \left(\Sigma+A\Omega A^\top\right)^{-1} z+\left(w-\Lambda A^\top \Sigma^{-1}z\right)^\top \Lambda^{-1}\left(w-\Lambda A^\top \Sigma^{-1}z\right) =$$

$$z^\top \left[\Sigma^{-1}-\Sigma^{-1}A\left(\Omega^{-1}+A^\top \Sigma^{-1}A\right)^{-1}A^\top \Sigma^{-1}\right] z+w^\top \Lambda^{-1}w$$

$$-w^\top \underbrace{\Lambda^{-1}\Lambda}_{I} A^\top \Sigma^{-1}z-\left(\Lambda A^\top \Sigma^{-1}z\right)^\top \Lambda^{-1}w+\left(\Lambda A^\top \Sigma^{-1}z\right)^\top \underbrace{\Lambda^{-1}\Lambda}_{I} A^\top \Sigma^{-1}z =$$

$$z^\top \Sigma^{-1}z-z^\top \Sigma^{-1}A\underbrace{\left(\Omega^{-1}+A^\top \Sigma^{-1}A\right)^{-1}}_{\Lambda}A^\top \Sigma^{-1}z+w^\top \Lambda^{-1}w$$

$$-(Aw)^\top \Sigma^{-1}z-z^\top \Sigma^{-1}A\underbrace{\Lambda\Lambda^{-1}}_{I}w+z^\top \Sigma^{-1}A\Lambda A^\top \Sigma^{-1}z =$$

$$z^\top \Sigma^{-1}z+w^\top \Lambda^{-1}w-(Aw)^\top \Sigma^{-1}z-z^\top \Sigma^{-1}Aw =$$

$$z^\top \Sigma^{-1}z+w^\top \left(\Omega^{-1}+A^\top \Sigma^{-1}A\right)w-(Aw)^\top \Sigma^{-1}z-z^\top \Sigma^{-1}Aw =$$

$$z^\top \Sigma^{-1}z+w^\top \Omega^{-1}w+(Aw)^\top \Sigma^{-1}Aw-(Aw)^\top \Sigma^{-1}z-z^\top \Sigma^{-1}Aw =$$

$$z^\top \Sigma^{-1}z-z^\top \Sigma^{-1}Aw-(Aw)^\top \Sigma^{-1}z+(Aw)^\top \Sigma^{-1}Aw+w^\top \Omega^{-1}w =$$

$$(z-Aw)^\top \Sigma^{-1}(z-Aw)+w^\top \Omega^{-1}w$$

Parte II: Inicia-se de • para chegar em ••:

$$\left|\Sigma+A\Omega A^\top \| \Lambda\right| = \left|\left(\Sigma+A\Omega A^\top\right)\Lambda\right| = \left|\left(\Sigma+A\Omega A^\top\right)\left(\Omega^{-1}+A^\top\Sigma^{-1}A\right)^{-1}\right| =$$

$$\left|\left(\Sigma+A\Omega A^\top\right)\left[\Omega-\Omega A^\top\left(\Sigma+A\Omega A^\top\right)^{-1}A\Omega\right]\right| =$$

$$|\Sigma\Omega-\{\underbrace{\Omega A^\top\left(\Sigma+A\Omega A^\top\right)^{-1}A\Omega}+A\Omega+A^\top\Omega-A\Omega A^\top\underbrace{\Omega A^\top\left(\Sigma+A\Omega A^\top\right)^{-1}A\Omega}|} =$$

$$\left|\Sigma\Omega-\left(\Sigma+A\Omega A^\top\right)\Omega A^\top\left(\Sigma+A\Omega A^\top\right)^{-1}A\Omega+A\Omega A^\top\Omega\right| =$$

$$\left|\Sigma\Omega-A\Omega\underbrace{\left(\Sigma+A\Omega A^\top\right)^{-1}\left(\Sigma+A\Omega A^\top\right)}_{I}A^\top\Omega+A\Omega A^\top\Omega\right| =$$

$$\left|\Sigma\Omega-A\Omega A^\top\Omega+A\Omega A^\top\Omega\right| = |\Sigma\Omega| = |\Sigma||\Omega|$$

Assim chegamos que $|\Sigma\Omega| = |\Sigma||\Omega|$

Conforme também é apresentado em Dávila, Bolfarine e Arellano-Valle (2004) e (OLIVEIRA, 2010), a seguinte proposição é dada.

Proposição 1:

Seja $Y \sim SN_n(\mu,\Sigma,\Delta)$. Então

$$Y \overset{d}{=} \Delta|X_0| + X_1, \qquad \text{em que } \overset{d}{=} \text{ significa converge em distribuição} \tag{4.24}$$

$$\delta = \frac{\lambda}{\sqrt{(1-\lambda^2)}}, \quad \lambda \text{ variando entre -1 a 1} \tag{4.25}$$

Onde $X_0 \sim N_n(0,I_n)$ e $X_1 \sim N_n(\mu,\Sigma)$, com X_0 e X_1 sendo independentes. Δ é considerada sendo uma matriz diagonal como em Sahu, Dey e Branco (2003) , a expressão 4.24 mostra que Y é assimétrica oriundo da soma de duas distribuições normais, sendo uma truncada positivamente, note que o parâmetro Δ foi incorporado ao modelo para absorver e controlar a assimetria gerada pelas distribuições associadas, sendo que quando Δ for zero recuperamos uma normal simétrica padrão.

Conforme Oliveira (2010), Suponha-se $U = \Delta|X_0| + X_1$ *e* $T = |X_0| \sim N_n(0,I_n)I_{t\geq 0}$, em que I é a função indicadora do domínio de T. Assumindo que a distribuição T é conhecida como sendo uma distribuição normal padrão truncada positiva (half normal).

Fazendo, $U|T = t \sim N_n(\Delta t+\mu,\Sigma)$, pois

$$\begin{aligned} E(U|T=t) &= E(\Delta|X_0|+X_1|T=t) \\ &= E(\Delta t+X_1|T=t) \\ &= \Delta t+E(X_1) \\ &= \Delta t+\mu \end{aligned}$$

$$\begin{aligned} Var(U|T=t) &= Var(\Delta t+X_1|T=t) \\ &= Var(X_1) \\ &= \Sigma \end{aligned}$$

Assim, tem-se que a distribuição marginal de U via a integral da distribuição conjunta de U e T em todo possível valor de T é dada por

$$f_U(w) = \int_{\mathbb{R}^n_+} f(w,t)dt$$
$$f_U(w) = \int_{\mathbb{R}^n_+} f(w|t)f(t)dt$$
$$f_U(w) = \int_{\mathbb{R}^n_+} \phi_n(w|\Delta t+\mu,\Sigma)2^n\phi_n(t)dt$$
$$f_U(w) = 2^n\int_{\mathbb{R}^n_+} \phi_n(w|\Delta t+\mu,\Sigma)\phi_n(t)dt \quad \text{Agora, utilizando o Lema 1, tem-se que}$$
$$f_U(w) = 2^n\int_{\mathbb{R}^n_+} \phi_n\left(w|\mu,\Sigma+\Delta^2\right)\times\phi_n\left(t|\left(I_n+\Delta\Sigma^{-1}\Delta\right)^{-1}\Delta\Sigma^{-1}(w-\mu),\left(I_n+\Delta\Sigma^{-1}\Delta\right)^{-1}\right)dt$$
$$f_U(w) = 2^n\phi_n\left(w|\mu,\Sigma+\Delta^2\right)\times\int_{\mathbb{R}^n_+} \phi_n\left(t|\Delta\left(\Sigma+\Delta^2\right)^{-1}(w-\mu),\left(I_n+\Delta\Sigma^{-1}\Delta\right)^{-1}\right)dt$$
$$f_U(w) = 2^n\phi_n\left(w|\mu,\Sigma+\Delta^2\right)\times\Phi_n\left(\Delta\left(\Sigma+\Delta^2\right)^{-1}(w-\mu)|0,\left(I_n+\Delta\Sigma^{-1}\Delta\right)^{-1}\right)$$

ou seja $U=Y\sim SN_n(\mu,\Sigma,\Delta)$, que conclui a prova

Corolário 1:

Seja $Y\sim N(\mu+\Delta X;\Sigma)$. logo a variância de Y são dadas por:

$$\begin{aligned} E[Y] &= E(\delta|X_0| + X_1) \\ &= \delta E(T) + E(X_1) \\ &= \delta\sqrt{\frac{2}{\pi}} + \mu \end{aligned} \tag{4.26}$$

em tal situação podemos representar equação 4.26 como no corolário 1 , como sendo

$$E[Y_j|a_j, \beta, \sigma_\varepsilon^2, \delta_\varepsilon] = X_j\beta + Z_j a_j + \delta_\varepsilon \left(\frac{2}{\pi}\right)^{1/2} 1_n, \tag{4.27}$$

1_n representa um vetor n-dimensional de uns.

$$\begin{aligned} Var[Y] &= Var(\Delta|X_0| + X_1) \\ &= \Delta^2 Var(T) + Var(X_1) \\ &= \Delta^2 I_n \left(1 - \frac{2}{\pi}\right) + \Sigma \\ &= \Sigma + \left(1 - \frac{2}{\pi}\right)\Delta^2 \end{aligned} \tag{4.28}$$

na mesma situação podemos representar a equação 4.28 de acordo com o corolário 1, sendo:

$$Var[Y_j|a_j, \beta, \sigma_\varepsilon^2, \delta_\varepsilon] = \left(\sigma_\varepsilon^2 + \delta_\varepsilon^2 \left(1 - \frac{2}{\pi}\right)\right) \tag{4.29}$$

onde $\delta = (\delta_1 ... \delta_n)^T$ é a diagonal da matriz Δ .

Proposição 2: Seja $Z \sim SN_n(0, I_n, \Delta)$ e considere a: transformação linear $Y = \mu + \Sigma^{1/2} Z$, onde Σ positiva definida. Então, $Y \sim SN_n(\mu, \Sigma, \Delta)$.

Pela Definição 1, quando um vetor aleatório Z possui distribuição $SN_n(0, I_n, \Delta)$, então sua densidade dada da seguinte maneira:

$$f_Z(z) = 2^n \phi_n\left(z|0, I_n + \Delta^2\right) \Phi_n\left(\Delta\left(I_n + \Delta^2\right)^{-1} z|0, \left(I_n + \Delta^2\right)^{-1}\right) \tag{4.30}$$

Como Σ é positiva definida, a prova segue do fato que $Z = (Y-\mu)\Sigma^{-1/2}$, isto é ,

$$f_Y(y) = |\Sigma|^{-1/2} f_Z\left(\Sigma^{-1/2}(y-\mu)\right)$$
$$f_Y(y) = |\Sigma|^{-1/2} 2^n \phi_n\left(\Sigma^{-1/2}(y-\mu)|0, I_n+\Delta^2\right) \times \Phi_n\left(\Delta\left(I_n+\Delta^2\right)^{-1}\Sigma^{-1/2}(y-\mu)|0, \left(I_n+\Delta^2\right)^{-1}\right)$$
$$f_Y(y) = 2^n \phi_n\left(y|\mu, \Sigma+\Delta^2\right) \times \Phi_n\left(\Delta\left(\Sigma+\Delta^2\right)^{-1}(y-\mu)|0, \left(I_n+\Delta\Sigma^{-1}\Delta\right)^{-1}\right)$$

5 METODOLOGIA

Neste trabalho foi utilizado o Modelo Hierárquico Generalizado Normal Assimétrico Bayesiano (MHGB). Foram atribuídas distribuições a priori para os parâmetros do modelo formulado. Foi especificada a função de verossimilhança para o vetor de observações tomando por base a distribuição normal assimétrica. Com estas especificações obteve-se a distribuição conjunta a posteriori para os parâmetros via teorema de Bayes e foram identificadas distribuições condicionais completas a posteriori para cada um dos parâmetros que serão objeto do processo de inferência. Este procedimento para um modelo GWS foi comparado a outros métodos padrão de análise GWS com dados normais (supondo simetria). Foram simulados vários cenários de comparação dos métodos.

5.1 Modelo Hierárquico Generalizado Normal Assimétrico Bayesiano

O MHGB pode proporcionar uma metodologia mais ampla e flexível para ajustar a complexidade de fatores genéticos e ambientais que afetam o desempenho de animais de produção em características biológicas complexas.

Nesta seção, com enfoque Bayesiano usando técnicas MCMC para o modelo normal assimétrico. O modelo Bayesiano 5.2 será representado e descrito da seguinte maneira o que vai de encontro com os trabalhos de (DÁVILA; BOLFARINE; ARELLANO-VALLE, 2004) e (OLIVEIRA, 2010).:

$$Y = \eta + \varepsilon$$

$$Y = X\beta + Za + \varepsilon \tag{5.1}$$

$$y|\beta, a, \sigma_\varepsilon^2 \sim \pi(y|\beta, a, \sigma_\varepsilon^2) = N(X\beta + Za, R) \tag{5.2}$$

em que temos:

Y tem dimensão $n \times 1$ é um vetor de observações fenotípicas;

X tem dimensão $n \times p$ é uma matriz de incidência dos efeitos fixos;

β tem dimensão $p \times 1$ é o vetor dos efeitos fixos;

Z tem dimensão $n \times q_a$ é uma matriz de incidência dos efeitos aditivos;

a tem dimensão $q_a \times 1$ é o vetor dos efeitos aditivos;

ε é o vetor de erros aleatórios ou resíduos de dimensão $n \times 1$.

Considerando que:

$$a \sim N_q(0, S_a), \qquad e \qquad \varepsilon \sim N_n(0, \Omega_\varepsilon) \tag{5.3}$$

Podemos supor que S_a e Ω_ε são matrizes de covariâncias entre os efeitos aleatórios $a_1 ... a_q$, e o resíduo $\varepsilon_1 ... \varepsilon_n$. aplicando as condições estabelecidas acima, podemos representar hierarquicamente o modelo como:

$$Y|\beta, a, \sigma_e^2 \sim N_n(X\beta + Za, I_n\sigma_\varepsilon^2) \tag{5.4}$$

$$a \sim N_q(0, S_a), \qquad \varepsilon \sim N_n(0, \Omega_\varepsilon) \tag{5.5}$$

A proposta é aplicar o Modelo Hierárquico Generalizado Normal Assimétrico Bayesiano, considerando que os efeitos genéticos a *e* ε tenham distribuição normal generalizada. Então o Modelo Hierárquico Generalizado Normal Assimétrico Bayesiano terá as seguintes suposições:

$$a \sim SN_q(\mu_a, S_a, \Delta_a), \qquad \varepsilon \sim N_n(\mu_\varepsilon, \Omega_\varepsilon, \Delta_\varepsilon) \tag{5.6}$$

em que $\Delta_a \in \Re^{q\times q}$ e $\Delta_\varepsilon \in \Re^{n\times n}$ são matrizes diagonais com os elementos $\delta_{a_1}, \ldots, \delta_{a_{qa}}$, $\delta_{d_1}, \ldots, \delta_{d_{q_d}}$ e $\delta_{\varepsilon_1}, \ldots, \delta_{\varepsilon_n}$, respectivamente, correspondentes aos parâmetros de assimetria. Portanto que: $\Delta_a = \delta_a I_{q_o}$, $\Delta_\varepsilon = \delta_\varepsilon I_n$, com $\delta_a \in \Re, \delta_\varepsilon \in \mathbb{R}$, salientando que $\delta = \dfrac{\lambda}{\sqrt{(1-\lambda^2)}}$. Sendo assim, obtemos a forma do modelo hierárquico da seguinte maneira:

$$Y = \eta + \Delta t + \varepsilon$$

$$Y = X\beta + Za + \Delta t + \varepsilon \tag{5.7}$$

$$y|\beta, a, \sigma_\varepsilon^2 \sim \pi(y|\beta, a, \sigma_\varepsilon^2, \delta_\varepsilon \sim SN_n\left(X\beta + Za, \sigma_\varepsilon^2 I_n, \delta_\varepsilon I_n\right) \tag{5.8}$$

$$a|\sigma_a^2, \delta_a \sim SN_{q_a}\left(\mu_a, S_a, \delta_a I_{q_a}\right) \tag{5.9}$$

Assim, chegamos a densidade condicional do vetor Y nos efeitos aleatórios, ou seja, pela verossimilhança dada analiticamente por:

$$\begin{aligned} f\left(y|\beta,a,\sigma_\varepsilon^2,\delta_\varepsilon\right) =& 2^n \phi_n\left(y|X\beta+Za,\left(\sigma_\varepsilon^2+\delta_\varepsilon^2\right)I_n\right)\times \\ & \Phi_n\left(\frac{\delta_\varepsilon}{\sigma_\varepsilon^2+\delta_\varepsilon^2}(y-X\beta-Za)|0,\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2+\delta_\varepsilon^2}I_n\right) \end{aligned} \tag{5.10}$$

Seja $\mathbf{y}=(y_1,\ldots y_n)^T$ e $\delta=(\delta_1,\ldots,\delta_n)^T$. Segue que a função de verossimilhança para $\boldsymbol{\theta}=\left(\mu,\sigma^2,\Delta,t\right)^T$ é dada por

$$L(\theta|y)=\frac{1}{\Sigma^{n/2}}\exp\left(-\frac{1}{2\sum}\sum_{i=1}^{n}\left(y_i-(X\beta_i+Za_i)\right)^2\right)\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\delta_i^2\right)\prod_{i=1}^{n}I_{i_i>0} \tag{5.11}$$

Uma parte fundamental neste ponto e especificar distribuições a prior para os parâmetros desconhecidos do modelo. De acordo com Arellano-Valle, Bolfarine e Lachos (2007) para garantir distribuições a posteriori próprias, adota-se o uso de prioris próprias.

5.1.1 Distribuições a priori para os parâmetros do modelo

Adotaremos uma distribuição normal multivariada para o vetor de parâmetros $\boldsymbol{\beta}$, com densidade

$$\pi\left(\beta|\beta_0,S_\beta\right)=\frac{1}{(2\pi)^{p/2}\sqrt{|S_\beta|}}\exp\left[-\frac{1}{2}\left(\beta-\beta_0\right)^\top S_\beta^{-1}\left(\beta-\beta_0\right)\right] \tag{5.12}$$

$$\beta\sim N_p(\beta_0,S_\beta) \tag{5.13}$$

Para o parâmetro t adotaremos uma distribuição normal multivariada com o vetor de parâmetros $\boldsymbol{t}$, com densidade

$$\pi\left(t|\mu_t,I\sigma_t^2\right)=\frac{1}{(2\pi)^{p/2}\sqrt{\left|I\sigma_t^2\right|}}\exp\left[-\frac{1}{2}\left(t-\mu_t\right)^\top\left(\sigma_t^2\right)^{-1}\left(t-\mu_t\right)\right] \tag{5.14}$$

$$t\sim N(\mu_t,I\sigma_t^2) \tag{5.15}$$

Para o parâmetro de assimetria Δ adotaremos uma distribuição normal multivariada com o vetor de parâmetros $\boldsymbol{\Delta}$ com densidade

$$\pi\left(\Delta|\mu_\delta, I\sigma_\delta^2\right) = \frac{1}{(2\pi)^{p/2}\sqrt{\left|I\sigma_\delta^2\right|}} \exp\left[-\frac{1}{2}\left(\Delta-\mu_\delta\right)^\top I(\sigma_\delta^2)^{-1}\left(\Delta-\mu_\delta\right)\right] \tag{5.16}$$

$$\Delta \sim N(\mu_\delta, I\sigma_\delta^2) \tag{5.17}$$

Admitimos uma distribuição a priori para cada efeito aleatório dos marcadores:

$$a_i \sim N\left(0, \sigma_{a_i}^2\right) \quad i = 1, \ldots, k \tag{5.18}$$

desta forma, a distribuição para os efeitos de marcadores condicionais as variâncias especificas de efeito de marca são independentes e, normalmente distribuídas com a função densidade:

$$\pi\left(\boldsymbol{a}|\boldsymbol{\sigma}_a^2\right) = \prod_{i=1}^{k}\left(2\pi\sigma_{a_i}^2\right)^{-\frac{1}{2}} \exp\left[-\frac{a_i^2}{2\sigma_{a_i}^2}\right] \quad i = 1, \ldots, k \tag{5.19}$$

no qual $\boldsymbol{\sigma}_a^2 = \left(\boldsymbol{\sigma}_{a_1}^2, \sigma_{a_2}^2, \ldots, \boldsymbol{\sigma}_{a_k}^2\right)$ é um vetor de $kx1$, com variâncias para cada marcador. As funções a priori para cada σ_{ai}^2, $i = 1, ..., k$ seguiram uma distribuição qui-quadrado invertida escalonada

$$\sigma_{a_i}^2|v_i, s_i^2 \sim \chi^{-2}\left(v_i, s_i^2\right) \quad \text{ou seja} \quad \pi\left(\sigma_{a_i}^2|v_i, s_i^2\right) \propto \left(\sigma_{a_i}^2\right)^{-\left(\frac{v_i+2}{2}\right)} \exp\left(\frac{-v_i s_i^2}{2\sigma_{a_i}^2}\right) \tag{5.20}$$

Os valores a priori dos hiperparâmetros assumiram para o parâmetro escala $v_0 = 0$ e $s_0^2 = 1$ graus de liberdade.

Para o parâmetro de escala σ^2 adotaremos uma distribuição normal multivariada. Onde $\sigma_g^2 \sim inv - \text{escalada} - \chi^2\left(v_g, S_g^2\right)$, considerando $v_g = 0$ e $S_g^2 = 0$ a distribuição priori reduz-se a $\frac{1}{\sigma_g^2}$

$\sigma_e^2 \sim inv - \text{escalada} - \chi^2\left(v_e, S_e^2\right)$ de forma análoga à considerada para variância genotípica, considerando os hiperparâmetros com valores iguais a zero tem-se $\frac{1}{\sigma_e^2}$. Para o vetor de efeito de genótipos $\mathbf{g}$ é atribuída também uma distribuição normal multivariada, com vetor σ_g^2. Decorrente da incerteza atribuída em relação à variância, essa distribuição a priori pode ser consi-

derada como informativa para genótipo. Para σ_g^2 e σ_e^2,como são componentes de variância, são atribuídas distribuições a priori qui-quadrado escaladas invertidas, que têm a seguinte forma:

$$\pi\left(\sigma^2|v,S^2\right) = \left(\frac{\frac{vS^2}{2}}{\Gamma\left(\frac{v}{2}\right)}\right)^{(v/2)} \cdot \frac{\exp\frac{-vS^2}{2\sigma^2}}{\left(\sigma^2\right)^{1+\frac{v}{2}}} \tag{5.21}$$

$$\sigma^2 \sim \chi^{-2} escalada(v_e, s_e^2) \tag{5.22}$$

Nessa distribuição v representa os graus de liberdade, S^2 é o parâmetro de escala e $\Gamma(\cdot)$ é a função gama. Os graus de liberdade v_g e v_e bem como os parâmetros de escala $\left(S_g^2 \text{ e } S_e^2\right)$ para σ_g^2 e σ_e^2, respectivamente, são considerados iguais a zero, nesta abordagem, do que resulta nas expressões $\frac{1}{\sigma_g^2}$ e $\frac{1}{\sigma_e^2}$ que também são consideradas distribuições a priori não informativas. Pelo fato da densidade a priori atribuída a σ_g^2 ser não informativa, a incerteza em relação a estimação de **g** é determinada basicamente pela função de verossimilhança, ou seja, a partir dos dados experimentais.

A densidade conjunta a posteriori de todas as quantidades envolvidas é dada por:

$$p(a,t,\Delta,\sigma^2|y) = \frac{1}{\Sigma^{n/2}}(\sigma^2)^{-(v_e+1)} \exp\left(-\frac{1}{2\Sigma}\sum_{i=1}^{n}(y_i-\mu-\Delta t_i)^2\right) \exp\left(-\frac{1}{2}\sum_{i=1}^{n} t_i^2\right) \prod_{i=1}^{n} I_{t_i>0}$$
$$\times \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu-\Delta t_i)^2 - \frac{(a-\mu_a)^2}{2\sigma_a^2} - \frac{(\Delta-\mu_\delta)^2}{2\sigma_\delta^2} - \frac{(t-\mu_t)^2}{2\sigma_t^2} - \frac{S_e^2}{\sigma^2}\right) \tag{5.23}$$

5.1.2 Distribuições a posteriori condicionais para os parâmetros

Assim, as densidades a posteriori marginais para os quatro parâmetros são dadas por:

$$p(\sigma^2|a,t,\Delta,,y) \propto \frac{1}{\Sigma^{n/2}}(\sigma^2)^{-(v_e+1)} \exp\left(-\frac{1}{2\Sigma}\sum_{i=1}^{n}(y_i-\mu-\Delta t_i)^2 - \frac{S_e^2}{\sigma^2}\right) \tag{5.24}$$

$$p(\Delta|a,t,\sigma^2,y) \propto \frac{1}{\Sigma^{n/2}} \exp\left(-\frac{1}{2\Sigma}\sum_{i=1}^{n}(y_i-\mu-\Delta t_i)^2 - \frac{(\Delta-\mu_\delta)^2}{2\sigma_\delta^2}\right) \tag{5.25}$$

$$p(t,|a,\Delta,\sigma^2,y) \propto \exp\left(-\frac{1}{2\Sigma}\sum_{i=1}^{n}(y_i-\mu-\Delta t_i)^2 - \frac{(t-\mu_t)^2}{2\sigma_t^2}\right) \tag{5.26}$$

$$p(a,|t,\Delta,\sigma^2,y) \propto \exp\left(-\frac{1}{2\Sigma}\sum_{i=1}^{n}(y_i-\mu-\Delta t_i)^2-\frac{(a-\mu_a)^2}{2\sigma_a^2}\right) \tag{5.27}$$

Onde todos os hiperparâmetros são especificados (conhecidos). Assim assumindo que $\mu_\delta = \mu_t = \mu_\mu$ são iguais a um vetor nulo e v_e sendo zero. Do mesmo modo assumindo que $\sigma_\delta^2, \sigma_\mu^2$ são variâncias muito alta, tornando as prioris não informativas. Para implementar amostrador de Gibbs necessitamos obter as distribuições a posteriori condicionais completas. Estas serão tratadas na seção seguinte.

5.1.3 Distribuições a posteriori condicionais completas para os parâmetros

Combinando as informações referentes aos dados (função de verossimilhança) com as densidades à priori, como estabelecidas em 5.10 e 5.11 por meio do teorema de Bayes, a densidade da distribuição conjunta a posteriori pode, então, ser estabelecida. Considera-se, inicialmente, a distribuição de verossimilhança. Aplicando o teorema de Bayes obtêm-se a distribuição conjunta a posteriori para os parâmetros. E, por meio de manipulações algébricas, e ainda completando quadrados em relação à distribuição multivariada normal, chega-se à seguinte expressão:

Distribuição condicional completa a posteriori para σ_ε^2

$$p(\sigma_\varepsilon^2|a,t,\Delta,y) \propto (\sigma_\varepsilon^2)^{-\frac{n\varepsilon}{2}} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(yi-\mu-\Delta t_i)^{'}(yi-\mu-\Delta t_i)\right\} \exp\frac{-\frac{1}{2}}{(\sigma_\varepsilon^2)^{1+\frac{v_\varepsilon}{2}}} \tag{5.28}$$

$$\vdots$$

$$p(\sigma_\varepsilon^2|a,t,\Delta,y) \propto (\sigma_\varepsilon^2)^{-\frac{n\varepsilon}{2}} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}e^{'}e\right\} \exp\frac{-\frac{1}{2}}{(\sigma_\varepsilon^2)^{1+\frac{v_\varepsilon}{2}}} \tag{5.29}$$

$$\vdots$$

$$p(\sigma_\varepsilon^2|a,t,\Delta,y) \propto (\sigma_\varepsilon^2)^{-(\frac{n_\varepsilon+v_\varepsilon}{2}+1)} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left[(n_\varepsilon+v_\varepsilon)\frac{n_\varepsilon\frac{\varepsilon^{'}\varepsilon}{n_\varepsilon}+v_\varepsilon S_\varepsilon^2}{(n_\varepsilon+v_\varepsilon)}\right]\right\} \tag{5.30}$$

$$\vdots$$

$$\sigma_\varepsilon^2 \sim \ X^{-2} escalada \left[(n_\varepsilon + v_\varepsilon); \frac{\varepsilon^{'}\varepsilon + v_\varepsilon S_\varepsilon^2}{(n_\varepsilon + v_\varepsilon)} \right] \tag{5.31}$$

Distribuição condicional completa a posteriori para σ_a^2

$$p(\sigma_a^2|a,t,\Delta,y) \propto \frac{1}{\Sigma^{n/2}} (\sigma_a^2)^{-(v_e+1)} \exp\left(-\frac{1}{2\sum} \sum_{i=1}^{n} (y_i - \mu - \Delta t_i)^2 - \frac{v_a S_a^2}{\sigma_a^2} \right)$$

$$\vdots$$

$$p(\sigma_a^2|...) \propto (\sigma_a^2)^{\frac{n_a}{2}} \exp\left\{ -\frac{1}{2\sigma_a^2} a^{'} G a \right\} \frac{\exp -\frac{v_a s_a^2}{2\sigma_a^2}}{(\sigma_a^2)^{1+\frac{v_a}{2}}} \tag{5.32}$$

$$\vdots$$

$$p(\sigma_a^2|...) \propto (\sigma_a^2)^{\frac{n_a}{2}} \exp\left\{ -\frac{1}{2\sigma_a^2} a^{'} G a \right\} (\sigma_a^2)^{-(1+\frac{v_a}{2})} \exp -\frac{v_a s_a^2}{2\sigma_a^2} \tag{5.33}$$

$$\vdots$$

$$p(\sigma_a^2|a,t,\Delta,y) \propto (\sigma_\varepsilon^2)^{-(\frac{n_a+v_a}{2}+1)} \exp\left\{ -\frac{1}{2\sigma_a^2} \left[(n_a + v_a) \frac{n_a \frac{a^{'} G a}{n_a} + v_a S^2}{n_a + v_a} \right] \right\} \tag{5.34}$$

$$\vdots$$

$$\sigma_a^2 \sim \ X^{-2} escalada \left[(n_a + v_a); \frac{a^{'} G a + v_a S_a^2}{(n_a + v_a)} \right] \tag{5.35}$$

Distribuição condicional completa a posteriori para Δ

$$p(\Delta|a,t,\sigma^2,y) \propto \frac{1}{\Sigma^{n/2}} \exp\left(-\frac{1}{2\sum} \sum_{i=1}^{n} (y_i - \mu - \Delta t_i)^2 - \frac{(\Delta - \mu_\delta)^2}{2\sigma_\delta^2} \right)$$

$$p(\Delta|a,t,\sigma^2,y) \propto \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} (yi - \mu - \Delta t_i)^{'} (yi - \mu - \Delta t_i) \right\} \exp\left(-\frac{1}{2\sigma_\delta^2} (\Delta - \mu_\delta)^{'} (\Delta - \mu_\delta) \right) \tag{5.36}$$

Fazendo com $\mu_\delta = 0$ e completando quadrado e fazendo algumas manipulações algébricas, chegamos ao seguinte;

$$p(\Delta|a,t,\sigma^2,y) \propto \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2}\left[\Delta - \left(t't+\frac{\sigma_\varepsilon^2}{\sigma_\delta^2}\right)^{-1} t'(y-X\beta-Za)\right]' \left(t't+\frac{\sigma_\varepsilon^2}{\sigma_\delta^2}\right) \left[\Delta - \left(t't+\frac{\sigma_\varepsilon^2}{\sigma_\delta^2}\right)^{-1} t'(y-X\beta-Za)\right]\right\} \tag{5.37}$$

Assim obtemos a condicional completa a posteriori sendo;

$$\Delta \sim N\left[\left(t't+\frac{\sigma_\varepsilon^2}{\sigma_\delta^2}\right)^{-1} t'(y-X\beta-Za); \left(t't+\frac{\sigma_\varepsilon^2}{\sigma_\delta^2}\right)(\sigma_\varepsilon^2)^{-1}\right] \tag{5.38}$$

Distribuição condicional completa a posteriori para t;

$$p(t|a,\Delta,\sigma^2,y) \propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(yi-\mu-\Delta t_i)'(yi-\mu-\Delta t_i)\right\} \exp\left(-\frac{1}{2\sigma_t^2}(t-\mu_t)'(\Delta-\mu_t)\right) \tag{5.39}$$

$$p(t|a,\Delta,\sigma^2,y) \propto \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2}\left[t - \left(\Delta'\Delta+\frac{\sigma_\varepsilon^2}{\sigma_t^2}\right)^{-1} \Delta'(y-X\beta-Za)\right]' \left(\Delta'\Delta+\frac{\sigma_\varepsilon^2}{\sigma_t^2}\right) \left[t - \left(\Delta'\Delta+\frac{\sigma_\varepsilon^2}{\sigma_t^2}\right)^{-1} \Delta'(y-X\beta-Za)\right]\right\} \tag{5.40}$$

$$\vdots$$

Completando quadrado e fazendo algumas manipulações algébricas chegamos a:

$$t \sim N\left[\left(\Delta'\Delta+\frac{\sigma_\varepsilon^2}{\sigma_t^2}\right)^{-1} \Delta'(y-X\beta-Za); \left(\Delta'\Delta+\frac{\sigma_\varepsilon^2}{\sigma_t^2}\right)(\sigma_\varepsilon^2)^{-1}\right] \tag{5.41}$$

Distribuição condicional completa a posteriori para a;

sendo θ igual $X\beta + \Delta t$

$$\begin{aligned} p(\mathbf{a}|\text{ outros }) &\propto \exp\left\{-\tfrac{1}{2\sigma_e^2}(\mathbf{y}-\mathbf{Za}-\boldsymbol{\theta})'(\mathbf{y}-\mathbf{Za}-\boldsymbol{\theta})\right\}\exp\left\{\tfrac{1}{2\sigma_a^2}\mathbf{a}'\mathbf{a}\right\} \\ p(\mathbf{a}|\text{ outros }) &\propto \exp\left\{-\tfrac{1}{2\sigma_e^2}\left[(\mathbf{y}-\mathbf{Za}-\boldsymbol{\theta})'(\mathbf{y}-\mathbf{Za}-\boldsymbol{\theta})+\mathbf{a}'I\tfrac{\sigma_e^2}{\sigma_a^2}\mathbf{a}\right]\right\} \end{aligned} \tag{5.42}$$

Por meio de manipulações algébricas, completando quadrados e ainda observando termos que são constantes em relação a **a** (esses termos podem ser absorvidos pela constante de normalização), obtém-se:

$$\begin{aligned} p(\mathbf{a}|\text{ outros }) \propto \exp\Bigg\{-\tfrac{1}{2\sigma_e^2}\Bigg[&\left(\mathbf{a}-\left(Z'Z+I\tfrac{\sigma_e^2}{\sigma_g^2}\right)^{-1}Z'(X\beta+Za)\right)'\left(Z'Z+I\tfrac{\sigma_e^2}{\sigma_g^2}\right) \\ &\times\left(\mathbf{a}-\left(Z'Z+I\tfrac{\sigma_e^2}{\sigma_g^2}\right)^{-1}Z'(X\beta+Za)\right)\Bigg]\Bigg\} \end{aligned} \tag{5.43}$$

$$\mathbf{a}|\cdots \sim N\left[\left(Z'Z+I\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1}Z'(\mathbf{y}-X\boldsymbol{\beta}),\left(Z'Z+I\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1}\sigma_e^2\right] \tag{5.44}$$

A herdabilidade no sentido restrito (h^2), será calculada para cada uma das amostras da distribuição posteriori conjunta, a cada passo da interação. A forma é a que segue na equação:

$$h^2_{restrito} = \frac{\sigma_a^2}{\sigma_a^2+\sigma_\varepsilon^2}\ . \tag{5.45}$$

Sendo assim, tem-se uma distribuição para representar a herdabilidade e estatísticas descritivas desta distribuição que permitirá inferir sobre a herdabilidade. Para aplicar esta metodologia será necessário atribuir valores iniciais para todas as variáveis do modelo e a cada passo das interações serão geradas amostras das distribuições condicionais apresentadas anteriormente até alcançar a convergência de acordo com os critérios que serão estabelecidos adiante.

5.2 MCMC

Para implementar o MCMC, usaremos a representação estocástica do modelo em 4.24, de modo que as distribuições normal-assimétricas definidas aqui podem ser representadas hie-

rarquicamente como:

$$\begin{aligned} Y|\beta, a, \sigma_\varepsilon^2, \delta_\varepsilon, t_\varepsilon &\sim N_n\left(X\beta + Za + \delta_\varepsilon t_\varepsilon, \sigma_\varepsilon^2 I_n\right) \\ t_\varepsilon &\sim N_n(0, I_n)\,\mathbb{I}_{w_\varepsilon > 0} \end{aligned} \tag{5.46}$$

$$\begin{aligned} a|\sigma_a^2, \delta_a, t_a &\sim N_{q_a}\left(\delta_a t_a, I\sigma_a^2\right) \\ t_a &\sim N_{q_a}\left(0, I_{q_a}\right)\mathrm{I}_{t_a > 0} \end{aligned} \tag{5.47}$$

Consideramos aqui que as variáveis t são as variáveis latentes com distribuição normal truncada positiva e I é uma função indicadora do domínio de variação t. Para implementar esta metodologia é necessário atribuir valores iniciais para todas as variáveis do modelo e as iterações geram amostras das distribuições condicionais apresentadas anteriormente até alcançar a convergência, que pode ser verificada e estudada através do pacote coda e também com o pacote Boa no software estatístico R.

5.3 Simulação dos Fenótipos

Com uso do *software R* (R Development Core Team, 2019), foram inicialmente simulados valores fenotípicos $y_{n\times 1}$, de uma distribuição normal truncada e uma distribuição normal, somada os valores das duas distribuições.

$$Y = vgg + \Delta|X_0| + X_1 \tag{5.48}$$

Y vetor (n $\times$ 1) de fenótipos, n= 300 é o número de observações;
ε vetor (n $\times$ 1) de resíduo calculado pela expressão ($\Delta|X_0| + X_1$);
W matriz (n $\times$ k) genotípica de SNPs, k=10.000 é o número de SNPs reais de uma espécie de milho.

As herdabilidades $h^2 \in (0,2; 0,5; 0,8)$ foram consideradas para compor dois cenários: oligogênico e poligênico. Desta forma, cada cenário relacionaram o valor dos efeitos genéticos de cada locus e criaram arquiteturas genéticas diferentes.

Gerou-se dois vetores; primeiro representou uma fração de 0,1% dos 10.000 marcadores com efeito, para o cenário oligogênico (apresentando herança controlada por poucos genes) e o segundo vetor com 1% dos 10.000 marcadores com efeito simulado representando o ce-

nário poligênico (refere-se as características fenotípicas determinados por muitos genes). Os QTLs (*Quantitative trait locus*) foram obtidos aleatoriamente, com efeitos simulados de uma distribuição normal padrão.

Em cada cenário oligogênico e poligênico foram considerados 3 graus de assimetria diferentes, sendo δ com os seguintes valores: (zero, duas vezes σ_ϵ^2 e dez vezes σ_ϵ^2). Desta maneira obtemos um total de 18 cenários diferentes.

As Figuras 5.1, 5.2 e 5.3 representam o cenário oligogênico, com as seguintes distribuições associadas aos dados: distribuição normal simétrica, distribuição normal assimétrica com médio grau de assimetria e distribuição normal assimétrica com alto grau de assimetria respectivamente.

Figura 5.1 – Fenótipo oligogênico simulado com herdabilidade 0,2 e $\lambda = 0$

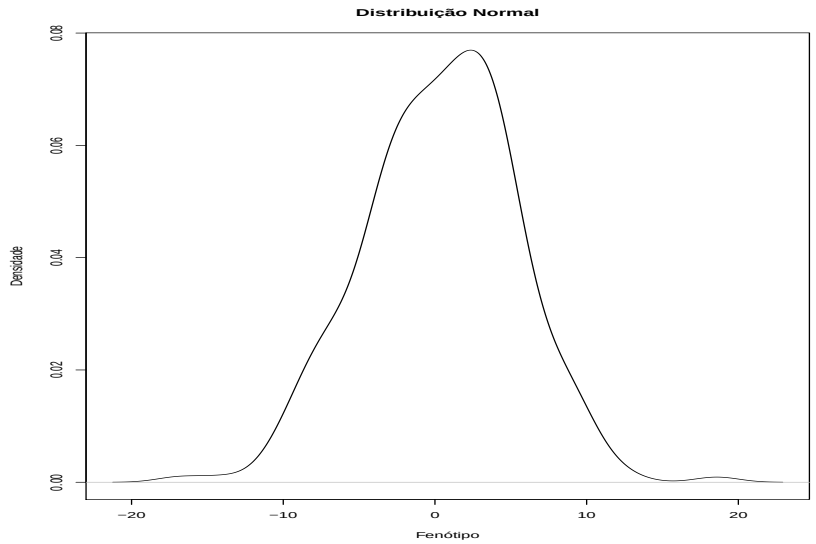


Figura 5.2 – Cenário Oligogênico com herdabilidade 0,2 e λ = 0,9996628, com validação cruzada

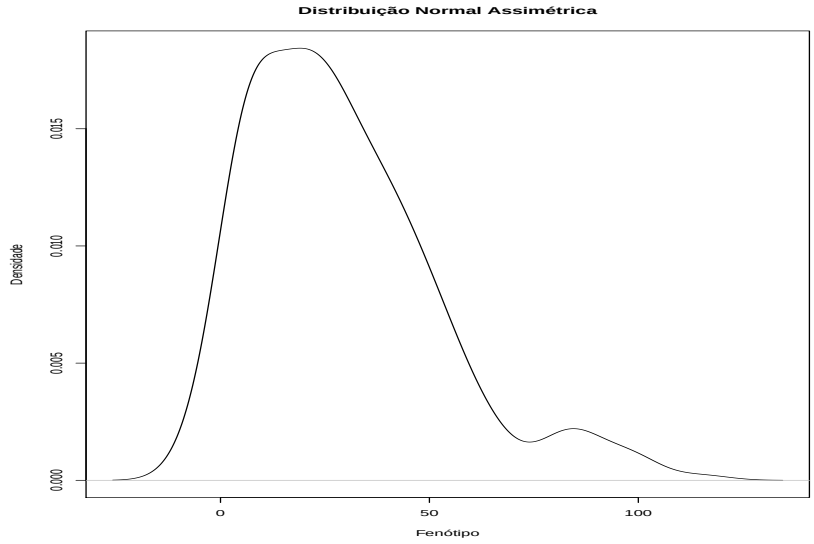
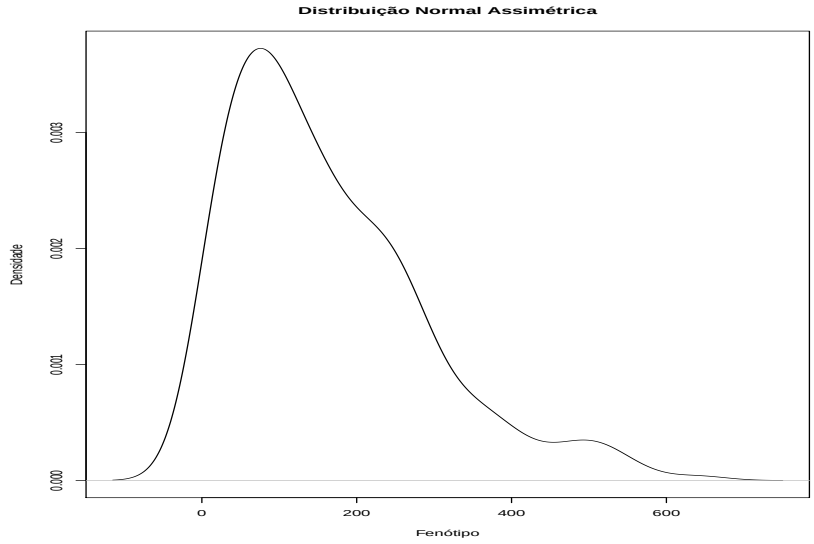


Figura 5.3 – Cenário Oligogênico com herdabilidade 0,2 e δ = 0,9999865, com validação cruzada



As Figuras 5.4, 5.5 e 5.6 representam o cenário poligênico, com as seguintes distribuições associadas aos dados: distribuição normal simétrica, distribuição normal assimétrica com médio grau de assimetria e distribuição normal assimétrica com alto grau de assimetria respectivamente.

Figura 5.4 – Cenário poligênico com herdabilidade 0,2 $\lambda = 0$, sem validação cruzada

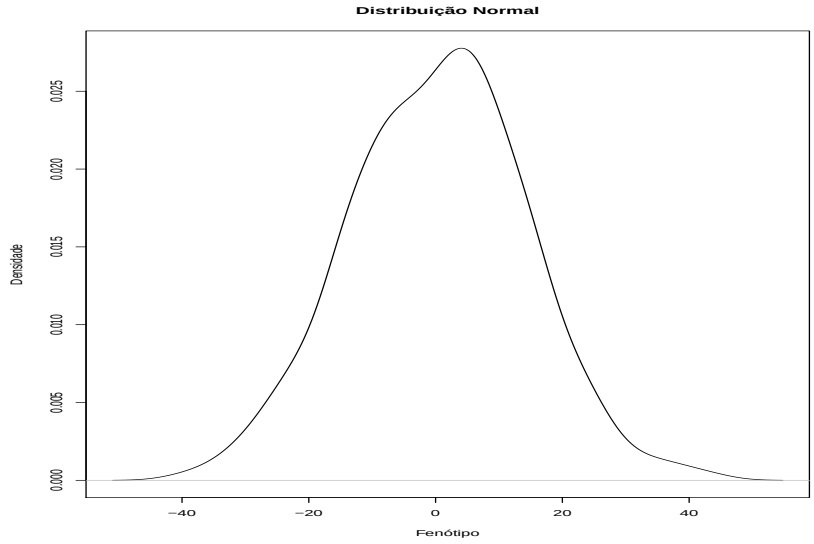


Figura 5.5 – Cenário poligênico com herdabilidade 0,2 e $\lambda = 0{,}9999955$, com validação cruzada

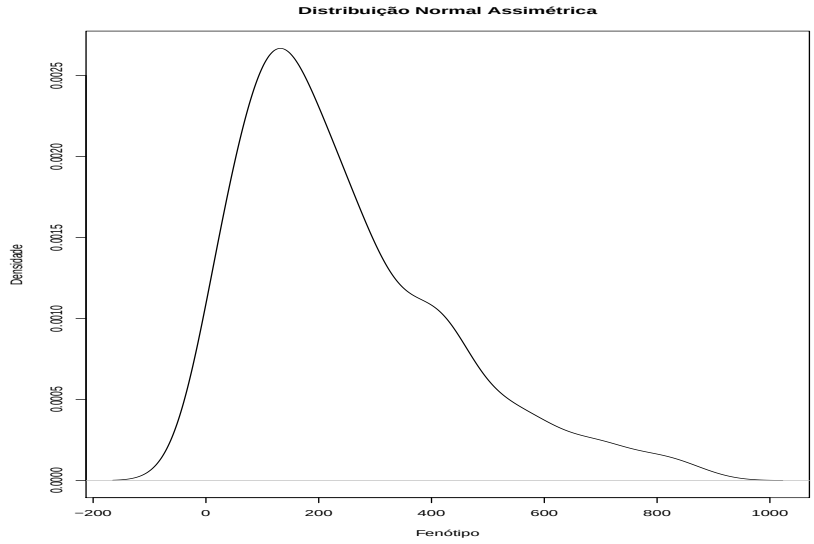
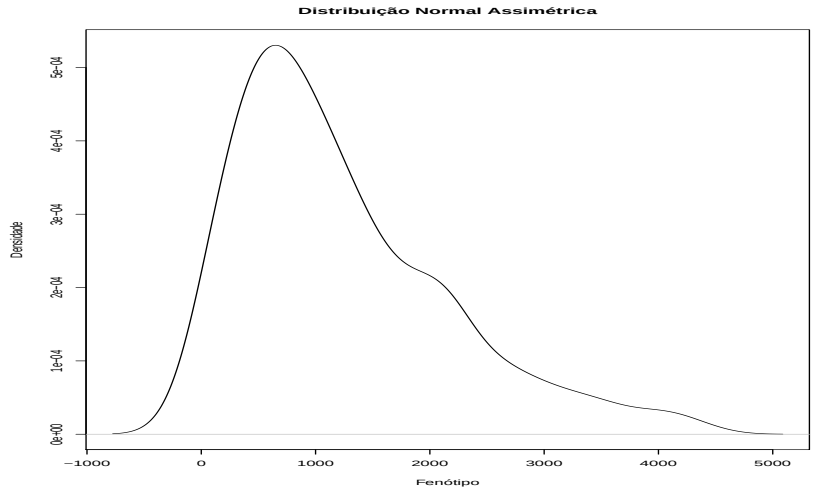


Figura 5.6 – Cenário poligênico com herdabilidade 0,2 e λ = 0,99999982 , com validação cruzada



Onde utilizou-se o amostrador de Gibbs implementado no *software R* (R Development Core Team, 2019). Utilizando para o modelo MHGB um total de 450.000 iterações com burn-in igual a 40.000 e um jump igual a 100 e um tamanho efetivo da cadeia de 4.100 observações, onde verificada a convergência do MCMC pelo critério de Raftery e Lewis (1992) e/ou Critério de Gelman e Rubin (1992a) .

5.4 Avaliação dos Modelos na Seleção Genômica

Para obtenção do efeito médio de cada marcador, levando em conta o modelo de predição utilizado para cada caráter sob avaliação. A soma desses efeitos é denominada de valor genético genômico (VGG), de forma que: o somatório em i dos efeitos de cada um dos marcadores, obtido pela multiplicação da matriz de incidência Z pelo vetor de efeitos de marcadores **a**, representa o VGG de cada indivíduo j, que se refere ao valor fenotípico predito pelo modelo de seleção genômica.

$$\widehat{VGG} = \hat{y}_i = \sum_i Z_i \widehat{a}_i \tag{5.49}$$

A acurácia é conceituada como a correlação entre o valor genético verdadeiro e aquele estimado a partir das informações genotípicas (marcadores) e/ou fenotípica dos indivíduos (JR *et al.*, 2012). Conforme Resende *et al.* (2010), como os conjuntos de dados utilizados na população de estimação e na validação do modelo são independentes, os resíduos associados aos VGG obtidos pelo modelo construído na população de estimação são independentes dos resíduos associados aos valores fenotípicos observados na população de validação. Assim, toda a correlação existente entre estes valores é de cunho genético, equivalendo à própria acurácia.

Alguns fatores afetam a acurácia da estimação dos valores genômicos como: herdabilidade da característica, tamanho da população de referência, abordagem estatística utilizada na estimação dos efeitos alélicos dos marcadores, desequilíbrio de ligação entres os marcadores e o lóco que participa do controle genético do caráter quantitativo (QTL) e a arquitetura genética do caráter em estudo (BASTIAANSEN *et al.*, 2012).

5.5 Validação do Modelo

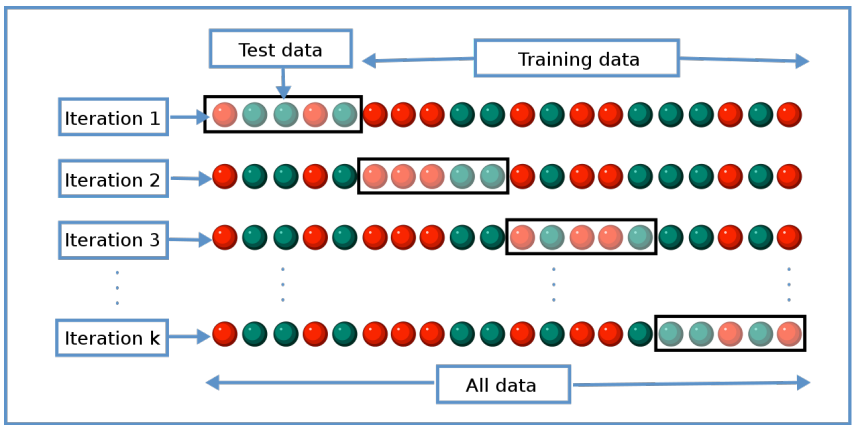
Já a validação cruzada é uma técnica para avaliar a capacidade preditiva de um modelo, a partir de um conjunto de dados (KOHAVI *et al.*, 1995). Esta técnica é amplamente utilizada em problemas onde o objetivo da modelagem é a predição. Busca-se então, estimar o quão preciso é este modelo na prática, ou seja, o seu desempenho para um novo conjunto de dados.

O conceito central das técnicas de validação cruzada é o particionamento do conjunto de dados em subconjuntos mutuamente exclusivos e, posteriormente, o uso de alguns destes subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento), sendo os subconjuntos restantes (dados de validação ou de teste) empregados na validação do modelo.

5.5.1 Método k-fold

Diversas formas de realizar o particionamento dos dados foram sugeridas, sendo as três mais utilizadas: o método holdout, o k-fold e o leave-one-out (KOHAVI *et al.*, 1995). Conforme podemos analisar na Figura 5.7.

Figura 5.7 – Diagrama de k-fold validação cruzada.



Fonte: Wikipédia (https://en.wikipedia.org/wiki/Cross-validation_(*statistics*))

O método de validação cruzada denominado k-fold consiste em dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir daí, um sub-

conjunto é utilizado para teste e os k-1 restantes são utilizados para estimação dos parâmetros, fazendo-se o cálculo da acurácia do modelo. Este processo é realizado k vezes alternando de forma circular o subconjunto de teste. Em seguida, os resultados são combinados obtendo a médias dos erros obtidos.

População de Estimação. Também denominada população de descoberta, de treinamento ou de referência.

População de Validação. Quando fisicamente disjunta da população de estimação, esse conjunto de dados é menor do que aquele da população de descoberta e contempla indivíduos avaliados para os marcadores SNPs e para os vários caracteres de interesse.

População de Seleção. Esse conjunto de dados contempla apenas os marcadores avaliados nos candidatos à seleção. Essa população não necessita ter os seus fenótipos avaliados.

Na sequência, a capacidade preditiva do modelo, onde os efeitos dos marcadores, pode ser verificada pelo ajuste em uma população de validação ou população de teste, para a qual existe há disponibilidade de dados genotípicos e fenotípicos. Nessa etapa, são estimados os valores genéticos com base apenas em dados genotípicos (GEBVs), utilizando o modelo estatístico ajustado na população de treinamento.

Para medir a capacidade preditiva dos modelos utilizados neste trabalho foi utilizado o coeficiente de correlação entre os valores fenotípicos observados e os valores fenotípicos estimados, dada pela expressão:

$$r_{y,\hat{y}} = \frac{cov(y,\hat{y})}{\sqrt{(var(y)var(\hat{y}))}} \tag{5.50}$$

Após isto, será calculado o R^2 que sempre assume um valor entre 0 e 1. Quanto mais próximo o R^2 é de um, melhor a adequação do modelo. Estas estatísticas foram assim calculadas e tornam-se úteis para validar a capacidade preditiva de seu modelo.

Depois foi calculado o Erro quadrático médio *(Mean squared error – MSE)*. dado pela expressão:

$$MSE = \frac{1}{n}\sum (y_i - \hat{y}_i)^2 \tag{5.51}$$

5.5.2 Pacotes estatísticos

Regressão linear generalizada Bayesiana (BGLR) implementa uma grande coleção de modelos de regressão bayesiana, incluindo seleção paramétrica de variáveis e métodos de con-

tração e procedimentos semiparamétricos (regressões de Hilbert no espaço de reprodução Bayesiana, RKHS). Extrai amostras da densidade posteriori usando um amostrador de Gibbs. Esta ferramenta suporta traços contínuos, bem como binários e ordinais.

O desenvolvimento dos modelos para a estimação dos VGG: Ridge Regression – Best Linear Unbiased Prediction (RR-BLUP) Meuwissen, Hayes e Goddard (2001); Whittaker, Thompson e Denham (2000) e BLUP genômico (GBLUP) VanRaden (2008); Clark e Werf (2013) implementados em contexto bayesiano, Bayes A e Bayes B (Meuwissen et al., 2001), Bayes C Habier *et al.* (2011) e Bayes LASSO de (PARK; CASELLA, 2008). Estes modelos foram implementados com o uso do pacote BGLR do R (CAMPOS; PÉREZ-RODRÍGUEZ, 2016) considerando *burn-in* de 30.000 iterações e *thinning* de 40. O tamanho resultante da cadeia suposta independente foi de 4.100 observações. Para o modelo GBLUP foi utilizado o pacote rrBlup do R. Já o modelo hierárquico normal assimétrico foi construído o código em R.

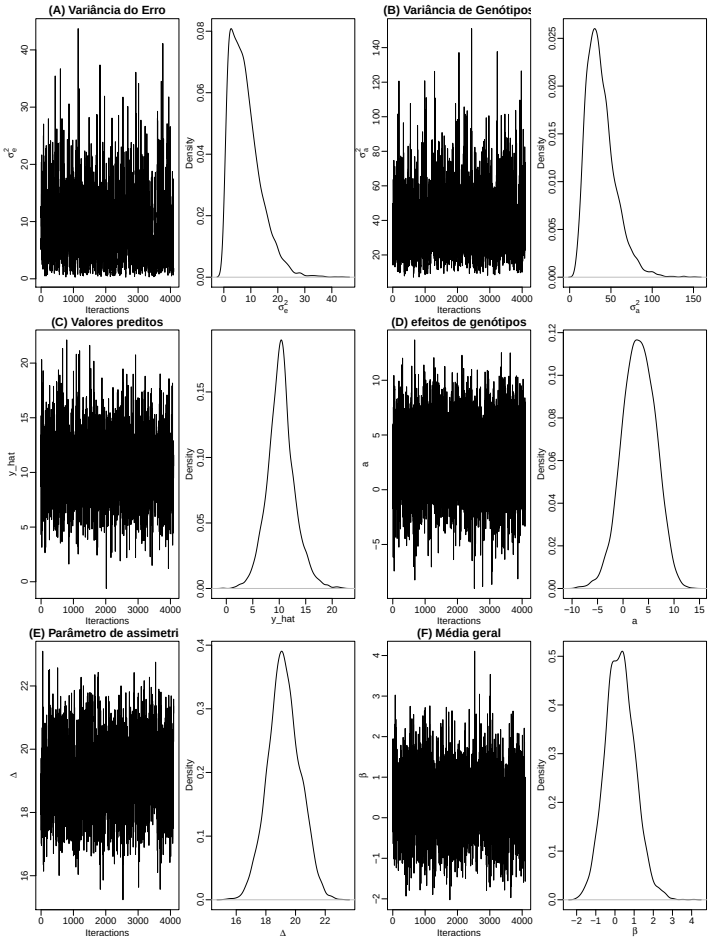
6 RESULTADOS E DISCUSSÃO

6.1 Propriedades das cadeias de Markov

Apresentamos abaixo a análise da convergência das cadeias amostradas para os parâmetros do modelo hierárquico generalizado normal assimétrico Bayesiano (que de agora em diante denominaremos MHGB). As cadeias amostradas para este modelo e também as das variâncias residuais dos demais modelos tiveram boas propriedades e podem ser consideradas independentes e utilizadas para inferência. A verificação foi feita com o pacote Boa (Bayesian Output Analysis) do R.

Os traço das cadeias dos parâmetros para o modelo MHGB no cenário com $\delta = 10$ e $h^2 = 0,2$ estão apresentados na Figura 6.1. De acordo com os critérios de Gelmam e Rubim (1992b) e Raftery e Lewis (1992), estas cadeias podem ser consideradas convergentes em probabilidade para os parâmetros em estudo.

Figura 6.1 – Amostras da distribuição a posteriori para os parâmetros do modelo hierárquico generalizado normal assimétrico Bayesiano. Os gráficos de traço e densidades representam, respectivamente: (A) Variância do Erro, (B) Variância de Genótipos, (C) Valores preditos, (D) efeitos de genótipos, (E) Parâmetro de assimetria e (F) Média Geral.



6.2 Apresentação dos resultados

Nas tabelas que se seguem nesta seção são apresentados os valores paramétricos simulados (VP), o parâmetro de assimetria (δ), capacidade preditiva (CP), a acurácia (ac), o ajuste do modelo (R^2), a média geral (β), a componente da variância genética (σ_a^2), a componente da variância residual (σ_ε^2), o parâmetro da herdabilidade (h^2) e quadrado médio do erro de predição (MSE) e as estimativas para eles obtidas nos diferentes cenários estudados pelos distintos modelos de análise considerados.

Na Tabela 6.1 apresenta-se os cenários oligogênicos e poligênicos, nas várias herdabilidades e sem assimetria (modelo normal). Para o parâmetro β, o modelo GBLUP de maneira geral foi superior ao demais modelos, mas no modelo proposto também se conseguiu estimativas satisfatórias. Para a variância genética (σ_a^2), os modelos GLUP e MHGB foram semelhantes em suas estimativas, sendo elas mais parecidas com os valores paramétricos do que as dos demais modelos (Bayes-X). Para a variância do erro (σ_ε^2) as estimativas pontuais mais distantes do valor paramétrico foram as do MHGB, embora estas estimativas serem razoáveis. Para a herdabilidade, no entanto, o MHGB ficou mais próximo do valor paramétrico, embora suas estimativas sejam ligeiramente superestimadas.

Pode-se concluir, portanto, que o MHGB obteve valores próximos aos demais modelos nos piores cenários para ele ser avaliado ($\Delta = 0$). Tanto para os cenário oligogênico ou poligênico o MHGB conseguiu produzir estimativas razoáveis, mostrando que a quantidade de genes e o valor de herdabilidade não o afeta de maneira expressiva ajuste do modelo. De acordo com Daetwyler *et al.* (2010), os modelos Bayes-X estimam melhor que o GBLUP quando o número de genes é pequeno; caso contrario, o GBLUP é melhor que os Bayes-X; no entanto, o MHGB é interessante em qualquer situação.

Na Tabela 6.2 são apresentadas as propriedades preditivas calculadas no método $k-fold$ para os diferentes modelos de ajuste. Os modelos tiveram estimativas semelhantes em suas acurácias seletivas nos distintos cenários avaliados, sendo difícil concluir qual seleciona melhor. Quanto à capacidade preditiva, nota-se que o MHGB foi superior aos demais modelos, mas percebe-se que o GBLUP e os modelos Bayes tiveram boas capacidades preditivas onde a herdabilidade foi alta. Verificamos também que o MHGB apresentou melhores ajustes nos distintos cenários, sendo superior aos demais modelos; este apresentou também melhores resultados quando analisamos o MSE. Isto é indício de que o MHGB pode ser uma alternativa viável, mesmo em cenários que não encontramos assimetria ou observações discrepantes.

As médias da distribuição marginal a posteriori para os parâmetros nas análises em que os dados foram simulados com o parâmetro de assimetria $\Delta = 2$ nos cenários oligogênicos e poligênicos estão na Tabela 6.3. Observa-se que no parâmetro β o modelo MHGB foi melhor que os demais modelos, fornecendo estimativas mais próximas do valor paramétrico. Para a componente da variância genética (σ_a^2) os modelos GLUP e Bayes não fizeram boas estimativas, e podem ser considerados inferiores ao MHGB, pois este apresentou os melhores resultados. Para a componente da variância residual (σ_ε^2) nota-se que todos os modelos geram estimativas distantes do valor paramétrico, exceto o MHGB. Observa-se que os modelos GBLUP e Bayes superestimaram este parâmetro, o que prejudica as estimativas de outros parâmetro dos modelos. No parâmetro (h^2) verifica-se que o MHGB foi também superior aos demais modelos, embora nesse caso tenda a superestimar o parâmetro. Aliás, as estimativas de (h^2) dos demais modelos podem ter sido comprometidas devido à piora na estimativa da variância residual, atribuindo erro excessivo ao que sabidamente era assimetria simulada.

Os resultados de predição nas simulações com assimetria estão resumidos na Tabela 6.4. Verifica-se que o MHGB foi superior a todos os modelos sob todos os aspectos e nos distintos cenários avaliados. Isto é mais um indício de que o MHGB é uma alternativa viável, em dados que apresentem assimetria de medida ou induzida por observações discrepantes.

Portanto, nas Tabelas 6.3, 6.4, 6.5 e 6.6 em que as observações assumem uma distribuição normal assimétrica, com diferentes graus de assimetria, nota-se que o MHGB foi de forma geral superior aos demais modelos. Observa-se que os modelos Bayes e GBLUP não fazem boas estimativas quando os dados são assimétricos ou possuem outliers.

As Tabelas 6.5 e 6.6 reforçam as ideias anteriores pois representam simulações com maior efeito de assimetria ($\Delta = 10$). De forma geral o MHGB foi superior aos demais modelos nos parâmetros e cenários avaliados, exceto quanto ao erro quadrático médio na predição para herdabilidade baixa. Isto confirma os resultados anteriores em que modelos gaussianos são inviáveis quando os dados apresentam forte assimetria ou muitos pontos discrepantes e que deve-se considerar a possibilidade de generalizar estas modelos empregando MHGB (assimétricos).

Uma ilustração da melhor capacidade preditiva dos MHGB em relação aos modelos concorrentes pode ser verificada nas Figuras 6.2 e 6.3, nas quais se percebe que os modelos GBLUP e Bayes são altamente afetados quando ao números de genes e herdabilidade, sendo

no cenário oligogênico os seus piores desempenhos. Mais uma vez o MHGB aparece como a alternativa viável quando estamos lidando com dados assimétricos.

Na Figura 6.2 podemos encontrar os gráficos do ajuste para os modelo Bayes L e MHGB no cenário oligogênico, com herdabilidade baixa ($h^2 = 0,2$) e erros com distribuição simétrica ($\Delta = 0$). Podemos notar que as predições do MHGB são ligeiramente mais correlacionadas aos fenótipos do que as do Bayes L. Na Figura 6.3 apresentamos gráfico análogo para o modelo poligênico e herdabilidade alta ($h^2 = 0,8$) e forte assimetria ($\Delta = 10$). Neste caso a correlação entre os valores preditos e o observado no testing set foi muito alta, indicando um ótimo ajuste do modelo assimétrico aos dados.

Tabela 6.1 – Médias da distribuição *a posteriori* marginal para cada um dos parâmetros comuns aos diversos modelos de ajuste. Cenários com diferentes herdabilidades e números de genes controladores da expressão do fenótipo e sem parâmetro de assimetria ($\Delta = 0$)

Parâmetros	Modelos	Cenários					
		Oligogênicos			Poligênicos		
		$h^2 = 0{,}2$	$h^2 = 0{,}5$	$h^2 = 0{,}8$	$h^2 = 0{,}2$	$h^2 = 0{,}5$	$h^2 = 0{,}8$
β	Valor Paramétrico	0	0	0	0	0	0
	MHGB	3,820	2,198	0,984	2,409	3,541	3,076
	GBLUP	0,375	0,187	0,094	0,412	0,206	0,103
	Bayes A	0,378	0,304	0,092	14,488	14,522	13,945
	Bayes B	0,728	0,297	-0,009	14,347	14,551	13,966
	Bayes C	0,708	0,317	0,029	15,174	14,497	13,980
	Bayes BL	0,685	0,357	0,057	14,267	14,530	13,940
	Bayes BRR	0,641	0,301	0,023	15,352	14,540	13,978
σ_a^2	Valor Paramétrico	4,82	4,82	4,82	41,68	41,68	41,68
	MHGB	3,230	4,062	4,642	28,54	35,05	48,75
	GBLUP	3,087	3,920	4,784	31,45	35,72	47,83
	Bayes A	0,690	0,288	0,009	14,71	14,55	13,95
	Bayes B	0,714	0,292	-0,005	14,45	14,56	13,96
	Bayes C	0,699	0,312	0,028	15,16	14,50	13,98
	Bayes BL	0,686	0,347	0,056	14,50	14,49	13,94
	Bayes BRR	0,632	0,299	0,030	15,36	14,53	13,96
σ_ε^2	Valor Paramétrico	19,269	4,817	1,204	166,743	41,683	10,420
	MHGB	15,013	3,314	0,994	133,544	32,048	5,510
	GBLUP	21,990	5,649	1,423	154,924	40,593	10,550
	Bayes A	21,893	5,628	1,502	154,787	40,658	11,177
	Bayes B	21,941	5,603	1,478	155,356	40,608	11,226
	Bayes C	21,746	5,683	1,519	153,356	40,933	11,383
	Bayes BL	21,445	5,720	1,568	154,336	39,050	11,903
	Bayes BRR	21,565	5,652	1,521	153,193	40,774	11,402
h^2	Valor Paramétrico	0,2	0,5	0,8	0,2	0,5	0,8
	MHGB	0,180	0,535	0,817	0,179	0,525	0,888
	GBLUP	0,123	0,409	0,771	0,169	0,468	0,819
	Bayes A	0,030	0,048	0,006	0,086	0,263	0,555
	Bayes B	0,031	0,049	-0,003	0,085	0,264	0,554
	Bayes C	0,031	0,052	0,018	0,090	0,261	0,551
	Bayes BL	0,031	0,057	0,035	0,086	0,270	0,540
	Bayes BRR	0,028	0,050	0,019	0,091	0,263	0,550

Tabela 6.2 – Propriedades preditivas dos diferentes modelos de ajuste. Cenários com diferentes herdabilidades e números de genes controladores da expressão do fenótipo e com parâmetro de assimetria ($\Delta = 0$) e dados apresentados como porcentagens (%) para coeficientes de herdabilidades, correlações e determinações em em notação científica indicada na linha para o MSE).

Parâmetros	Modelos	Cenários					
		Oligogênicos			Poligênicos		
		$h^2 = 0{,}2$	$h^2 = 0{,}5$	$h^2 = 0{,}8$	$h^2 = 0{,}2$	$h^2 = 0{,}5$	$h^2 = 0{,}8$
	MHGB	87,2	93,6	97,1	81,3	90,8	96,2
Acurácia	GBLUP	87,2	93,8	97,1	78,0	90,8	96,3
Seletiva	Bayes A	87,4	93,9	97,2	81,5	91,3	96,3
$r(vgg, \hat{a})$	Bayes B	87,5	94,1	97,4	81,5	91,0	96,4
	Bayes C	87,5	93,9	97,1	81,6	90,8	96,3
	Bayes BL	87,3	93,5	97,0	81,1	92,3	96,0
	Bayes BRR	87,6	93,9	97,1	82,0	91,1	96,2
	MHGB	90,2	92,0	96,0	77,4	88,1	96,9
Capacidade	GBLUP	44,7	70,9	90,7	51,2	59,8	92,2
Preditiva	Bayes A	45,9	72,0	90,6	52,0	75,8	92,1
$r(y, \hat{y})$	Bayes B	46,1	72,3	90,7	51,8	75,7	92,0
	Bayes C	47,3	71,4	90,4	53,5	75,3	91,9
	Bayes BL	47,3	70,9	90,1	51,6	75,6	91,5
	Bayes BRR	48,6	71,8	90,3	53,9	75,9	91,9
	MHGB	81,4	85,6	91,9	60,0	77,6	95,8
Coeficiente	GBLUP	20,0	50,2	82,2	26,0	56,0	85,1
de	Bayes A	21,0	51,6	82,1	26,8	57,3	85,0
Determinação	Bayes B	21,0	52,1	82,8	26,7	57,1	84,7
R^2	Bayes C	22,1	51,0	81,7	28,4	56,3	84,5
	Bayes BL	22,2	50,2	81,1	26,4	58,1	83,7
	Bayes BRR	22,4	51,5	81,6	28,8	57,1	84,5
	MHGB	2486,0	678,0	150,0	10827,2	3316,6	1186,0
Erro	GBLUP	2077,3	498,0	108,7	14461,9	3524,2	773,8
Quadrático	Bayes A	2097,3	493,5	109,6	34600,0	24500,0	20222,0
Médio na	Bayes B	2092,8	487,8	108,2	34956,0	24683,0	20299,0
Predição	Bayes C	2057,6	502,1	121,2	37397,0	24420,0	20351,0
	Bayes BL	2052,0	510,6	115,7	34741,2	24550,0	20272,0
	Bayes BRR	2011,9	493,9	112,9	37480,0	24589,0	20353,0

Tabela 6.3 – Médias da distribuição *a posteriori* marginal para cada um dos parâmetros comuns aos diversos modelos de ajuste. Cenários com diferentes herdabilidades e números de genes controladores da expressão do fenótipo e com parâmetro de assimetria ($\Delta = 2$)

Parâmetros	Modelos	Cenários					
		Oligogênicos			Poligênicos		
		$h^2 = 0{,}2$	$h^2 = 0{,}5$	$h^2 = 0{,}8$	$h^2 = 0{,}2$	$h^2 = 0{,}5$	$h^2 = 0{,}8$
β	Valor Paramétrico	0	0	0	0	0	0
	MHGB	-0,4	-0,285	0,918	-0,25	2,28	0,25
	GBLUP	29,2	7,380	1,880	250	62,50	15,70
	Bayes A	29,3	7,500	1,920	246	62,10	15,40
	Bayes B	29,4	7,460	1,890	243	61,10	15,10
	Bayes C	28,9	7,610	1,940	249	62,10	15,50
	Bayes BL	29,3	7,370	1,930	238	61,00	15,20
	Bayes BRR	29,1	7,180	1,870	243	62,60	15,40
σ_a^2	Valor Paramétrico	4,82	4,82	4,82	41,683	41,683	41,683
	MHGB	3,086	3,254	4,347	30,617	19,186	35,834
	GBLUP	5×10^{-7}	2,920	4,240	769,00	44,80	23,100
	Bayes A	29,349	7,492	1,922	247,11	62,120	15,426
	Bayes B	29,379	7,461	1,899	243,24	60,948	15,054
	Bayes C	28,993	7,609	1,940	248,50	62,099	15,174
	Bayes BL	29,279	7,367	1,926	238,14	60,520	15,168
	Bayes BRR	29,141	7,179	1,872	242,86	62,630	15,417
σ_ε^2	Valor Paramétrico	19,269	4,817	1,204	166,743	41,683	10,420
	MHGB	13,133	4,115	1,625	15,427	50,057	7,030
	GBLUP	537	38,5	3,560	34786,0	2141,00	145,00
	Bayes A	519,678	35,896	3,823	35054,0	1976,72	139,343
	Bayes B	530,930	35,322	3,599	33804,0	2202,92	136,804
	Bayes C	505,251	35,837	3,802	34607,9	2107,35	143,308
	Bayes BL	505,900	38,157	3,802	30154,3	2056,30	139,420
	Bayes BRR	515,461	36,149	3,584	33048,4	2171,16	140,990
h^2	Valor Paramétrico	0,2	0,5	0,8	0,2	0,5	0,8
	MHGB	0,203	0,429	0,705	0,443	0,262	0,785
	GBLUP	1×10^{-9}	0,070	0,544	0,022	0,021	0,138
	Bayes A	0,053	0,173	0,334	0,007	0,030	0,099
	Bayes B	0,052	0,174	-0,334	0,007	0,027	0,099
	Bayes C	0,054	0,175	0,336	0,008	0,020	0,097
	Bayes BL	0,055	0,161	0,336	0,008	0,028	0,098
	Bayes BRR	0,053	0,166	0,343	0,007	0,028	0,098

Tabela 6.4 – Propriedades preditivas dos diferentes modelos de ajuste. Cenários com diferentes herdabilidades e números de genes controladores da expressão do fenótipo e com parâmetro de assimetria ($\Delta = 2$) e dados apresentados como porcentagens (%) para coeficientes de herdabilidades, correlações e determinações em em notação científica indicada na linha para o MSE).

Parâmetros	Modelos	Cenários					
		Oligogênicos			Poligênicos		
		$h^2 = 0{,}2$	$h^2 = 0{,}5$	$h^2 = 0{,}8$	$h^2 = 0{,}2$	$h^2 = 0{,}5$	$h^2 = 0{,}8$
	MHGB	60,2	85,4	94,7	63,5	54,4	82,7
Acurácia	GBLUP	48,5	78,5	94,1	-15,0	-7,8	73,6
Seletiva	Bayes A	34,1	73,2	93,9	-12,1	-7,8	73,7
$r(vgg, \hat{a})$	Bayes B	28,5	74,1	94,0	-18,7	30,3	76,9
	Bayes C	25,6	73,9	93,5	-7,2	15,3	74,2
	Bayes BL	22,9	79,9	93,5	-6,7	16,6	73,0
	Bayes BRR	15,6	71,1	94,1	-17,4	16,6	77,0
	MHGB	90,2	95,9	96,0	90,1	90,9	91,0
Capacidade	GBLUP	20,5	37,9	79,6	25,4	25,4	47,3
Preditiva	Bayes A	36,2	39,6	79,8	32,1	2,7	49,7
$r(y, \hat{y})$	Bayes B	34,0	41,4	80,2	30,4	29,6	50,2
	Bayes C	34,1	42,8	79,6	35,5	34,6	50,9
	Bayes BL	48,4	39,2	78,6	27,4	39,0	63,6
	Bayes BRR	38,0	44,7	79,8	37,8	37,2	51,7
	MHGB	99,8	99,2	92,8	98,5	99,8	99,8
Coeficiente	GBLUP	0,4	14,2	63,3	6,2	6,1	22,1
de	Bayes A	16,2	15,4	63,5	10,0	0,2	24,5
Determinação	Bayes B	11,3	16,9	64,2	8,9	8,4	25,0
R^2	Bayes C	11,3	18,1	63,3	12,3	11,7	25,6
	Bayes BL	23,2	15,1	62,0	7,2	13,6	40,2
	Bayes BRR	14,2	19,7	63,6	14,0	13,6	26,2
	MHGB	129,6	87,0	75,4	67,0	1017,0	75,2
Erro	GBLUP	51880,6	3564,9	294,3	3365130,0	211510,0	13685,0
Quadrático	Bayes A	159212,0	10665,0	811,0	6371700,0	588140,0	62685,0
Médio na	Bayes B	155225,0	10750,0	808,0	6825650,0	493190,0	62347,0
Predição	Bayes C	155863,0	10862,0	830,0	6062630,0	467560,0	62756,0
	Bayes BL	166260,0	10627,0	841,0	7211440,0	441520,0	64658,0
	Bayes BRR	159232,0	10914,0	817,0	5828010,0	456200,0	63117,0

Tabela 6.5 – Médias da distribuição *a posteriori* marginal para cada um dos parâmetros comuns aos diversos modelos de ajuste. Cenários com diferentes herdabilidades e números de genes controladores da expressão do fenótipo e com parâmetro de assimetria ($\Delta = 10$)

Parâmetros	Modelos	Cenários					
		Oligogênicos			Poligênicos		
		$h^2 = 0{,}2$	$h^2 = 0{,}5$	$h^2 = 0{,}8$	$h^2 = 0{,}2$	$h^2 = 0{,}5$	$h^2 = 0{,}8$
β	Valor Paramétrico	0	0	0	0	0	0
	MHGB	-1	-0,2	-0,17	1228	0,4	2,859
	GBLUP	145	36,3	9,10	1245	311	77,9
	Bayes A	161	35,9	9,05	1268	310	77,3
	Bayes B	147	36,3	9,07	1213	304	74,9
	Bayes C	147	37,1	9,09	1240	309	76,9
	Bayes BL	144	36,9	9,19	1231	307	76,5
	Bayes BRR	145	35,5	9,11	1212	313	76,5
σ_a^2	Valor Paramétrico	4,82	4,82	4,82	41,683	41,683	41,683
	MHGB	4,512	3,370	3,090	17	33,769	23,707
	GBLUP	1×10^{-5}	9×10^{-7}	2,730	25441	1405,0	67,900
	Bayes A	161,240	35,932	9,047	1267	309,962	77,296
	Bayes B	146,528	36,324	9,067	1213	303,822	74,896
	Bayes C	146,670	37,102	9,095	1240	309,468	76,925
	Bayes BL	144,920	36,931	9,191	1230	307,142	76,545
	Bayes BRR	144,716	35,490	9,108	1212	312,857	76,472
σ_ε^2	Valor Paramétrico	19,269	4,817	1,204	166,743	41,683	10,420
	MHGB	2,252	3,38	1,083	867064,0	10,088	40,501
	GBLUP	12788,0	809,00	53,000	854150,0	53320,0	3332,0
	Bayes A	15416,6	747,93	49,797	881122,0	48516,0	3379,0
	Bayes B	12651,4	805,70	50,119	845407,0	54954,0	3032,0
	Bayes C	12207,1	763,13	50,126	868391,0	52298,0	3276,9
	Bayes BL	11418,9	763,13	52,133	783969,0	51754,0	3129,8
	Bayes BRR	12442,1	737,83	50,600	824829,0	54021,0	3180,2
h^2	Valor Paramétrico	0,2	0,5	0,8	0,2	0,5	0,8
	MHGB	0,508	0,447	0,641	2×10^{-5}	0,480	0,364
	GBLUP	1×10^{-9}	1×10^{-9}	0,049	0,029	0,006	0,020
	Bayes A	0,010	0,046	0,154	0,001	0,006	0,022
	Bayes B	0,012	0,043	0,153	0,001	0,005	0,024
	Bayes C	0,013	0,046	0,153	0,001	0,006	0,024
	Bayes BL	0,012	0,048	0,150	0,001	0,005	0,023
	Bayes BRR	0,011	0,046	0,152	0,001	0,006	0,023

Tabela 6.6 – Propriedades preditivas dos diferentes modelos de ajuste. Cenários com diferentes herdabilidades e números de genes controladores da expressão do fenótipo e com parâmetro de assimetria ($\Delta = 10$) e dados apresentados como porcentagens (%) para coeficientes de herdabilidades, correlações e determinações em em notação científica indicada na linha para o MSE).

Parâmetros	Modelos	Cenários					
		Oligogênicos			Poligênicos		
		$h^2 = 20$	$h^2 = 50$	$h^2 = 80$	$h^2 = 20$	$h^2 = 50$	$h^2 = 80$
	MHGB	15,9	48,6	87,1	-7,0	64,0	62,0
Acurácia	GBLUP	27,9	45,2	75,6	-21,4	-40,7	21,6
Seletiva	Bayes A	-27,0	42,1	70,2	-25,3	-40,7	21,6
$r(vgg, \hat{a})$	Bayes B	7,7	20,7	67,7	24,0	-7,2	-6,1
	Bayes C	-0,5	25,4	71,5	-26,3	-14,6	-2,2
	Bayes BL	-9,1	12,0	69,8	-16,5	-10,2	15,2
	Bayes BRR	-1,2	9,7	71,5	-26,0	-15,2	11,2
	MHGB	92,1	93,2	92,7	95,2	94,7	93,1
Capacidade	GBLUP	19,8	20,0	34,2	26,2	25,5	24,9
Preditiva	Bayes A	4,9	26,8	35,7	31,8	31,1	30,9
$r(y, \hat{y})$	Bayes B	5,2	29,7	38,2	31,1	30,0	29,8
	Bayes C	4,8	33,9	40,6	34,8	35,5	34,0
	Bayes BL	4,6	26,7	34,3	30,4	31,7	28,4
	Bayes BRR	4,6	38,5	42,8	37,9	37,6	37,2
	MHGB	90,0	91,0	60,0	58,9	90,0	99,2
Coeficiente	GBLUP	4,0	3,7	26,0	6,5	63,0	6,0
de	Bayes A	0,1	6,7	26,8	9,8	9,4	0,4
Determinação	Bayes B	0,1	8,5	26,7	9,4	8,7	0,4
R^2	Bayes C	0,1	11,2	28,4	11,8	12,3	0,3
	Bayes BL	0,1	6,8	26,4	9,0	9,7	0,1
	Bayes BRR	0,1	14,5	28,8	14,0	13,9	1,7
	MHGB	1,070	0,177	0,133	$236{,}6 \times 10^4$	0,612	10,29
Erro	GBLUP ($\times 10^3$)	12,38	0,781	0,049	841,72	52,59	3,286
Quadrático	Bayes A ($\times 10^4$)	3,431	2,231	0,015	152,821	10,282	0,756
Médio na	Bayes B ($\times 10^4$)	3,398	2,281	0,015	155,459	10,839	0,734
Predição	Bayes C ($\times 10^4$)	3,644	2,365	0,016	144,439	9,265	0,685
	Bayes BL ($\times 10^4$)	3,764	2,238	0,015	165,228	9,851	0,758
	Bayes BRR($\times 10^4$)	3,437	2,416	0,016	137,512	8,935	0,662

Tabela 6.7 – Propriedades assimétricas dos diferentes modelos de ajuste. Cenários com diferentes herdabilidades e números de genes controladores da expressão do fenótipo e com parâmetro de assimetria ($\Delta = 0$), ($\Delta = 2$) e ($\Delta = 10$)

Parâmetros	Modelos	Cenários					
		Oligogênicos			Poligênicos		
		$h^2 = 20$	$h^2 = 50$	$h^2 = 80$	$h^2 = 20$	$h^2 = 50$	$h^2 = 80$
Parâmetro de	Valor Paramétrico	0	0	0	0	0	0
Assimetria	MHGB	-4,340	-2,531	-1,117	-2,6	-4,2	-3,7
($\Delta = 0$)	MC	—	—	—	—	—	—
Parâmetro de	Valor Paramétrico	38,5	9,63	2,41	333,79	83,367	20,841
Assimetria	MHGB	37,4	9,627	2,278	305,09	75,350	19,250
($\Delta = 2$)	MC	—	—	—	—	—	—
Parâmetro de	Valor Paramétrico	192	48,172	12,043	1667	416,8	104,20
Assimetria	MHGB	183	46,022	11,679	4	377,5	94,09
($\Delta = 10$)	MC	—	—	—	—	—	—

Os modelos GBLUP, Bayes A, Bayes B, Bayes C, Bayes RR e Bayes L serão agrupados nesta tabela e denominados por Modelos concorrentes (MC).

Na Tabela 6.7 nota-se que o MHGB fez ótimas estimativas sobre o parâmetro de assimetria. Desta forma, foi capaz de captar a direção e a intensidade da assimetria presente nos dados, o que torna o modelo promissor sempre que os dados apresentarem assimetria ou pontos discrepantes, que é um diferencial no MHGB, pois os modelos MC não têm está capacidade de estimativa.

Figura 6.2 – Os gráficos ilustram os ajustes dos modelos MHGB (esquerda) e Bayes L (direita). Cenário oligogênico com herdabilidade $h^2 = 0,2$ e sem assimetria $\Delta = 0$, sem validação cruzada

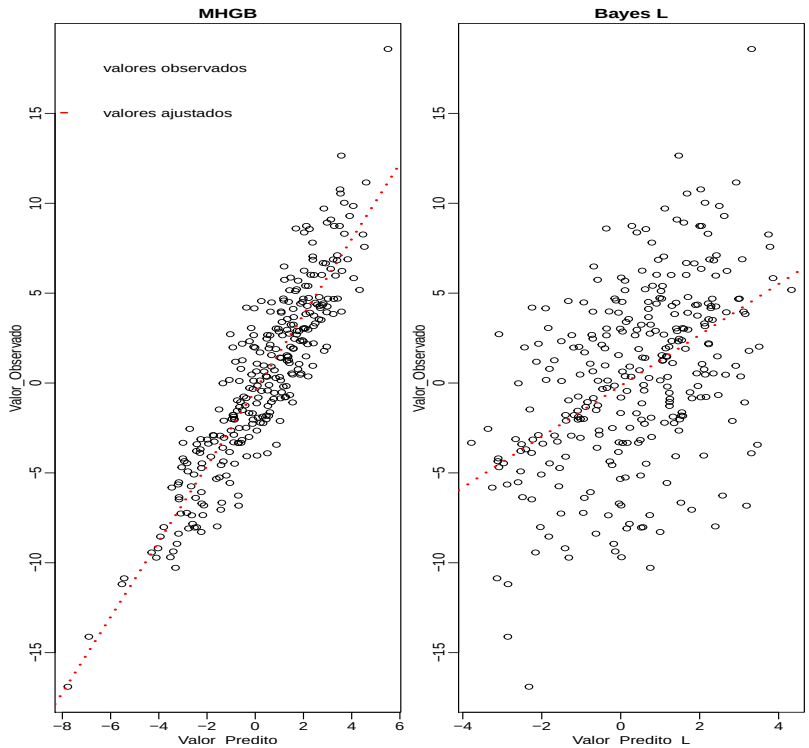
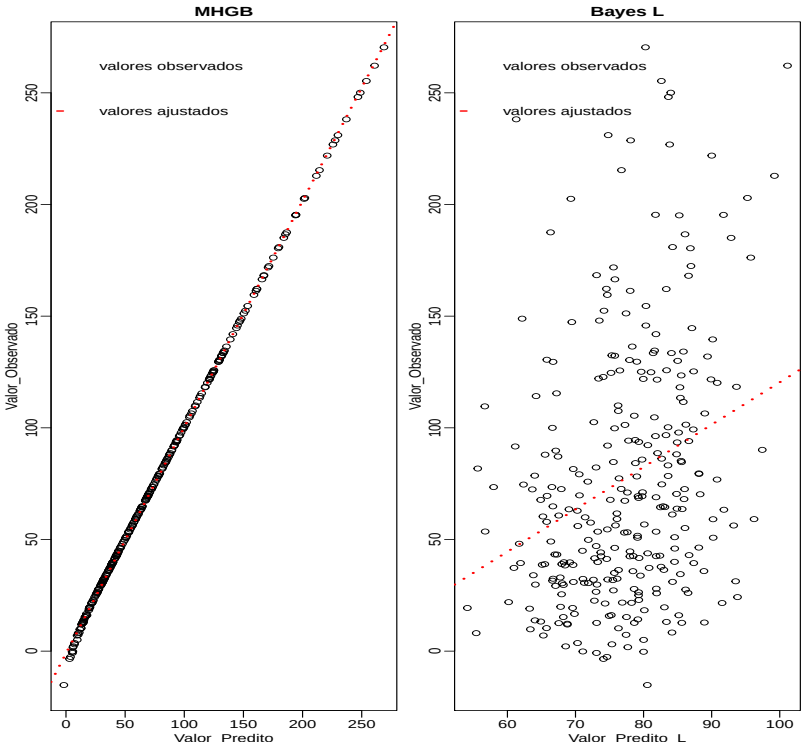


Figura 6.3 – Os gráficos ilustram os ajustes dos modelos MHGB (esquerda) e Bayes L (direita). Cenário poligênico com herdabilidade $h^2 = 0,8$ com parâmetro de assimetria $\Delta = 10$, com validação cruzada.



6.3 Discussão

De acordo com o que foi apresentado até aqui, além do número de genes, a herdabilidade é outra dimensão importante da arquitetura genética. Os modelos Bayes e GBLUP são sensíveis à herdabilidade e apresentam desempenho fraco para características com baixa herdabilidade, principalmente quando os fenótipos possuem alta dispersão. Portanto, o modelo MHGB se mostra uma alternativa viável para qualquer situação.

A obtenção de predições precisas de valores genéticos ou fenotípicos não observados para características complexas em populações de animais, plantas e humanos é possível através da seleção genômica ampla (GWS). É uma das suas preocupações é capacidade preditiva, que geralmente é medida com a correlação entre os valores genéticos preditos e os verdadeiros valores genéticos e/ou fenotípicos em uma população de validação Meuwissen, Hayes e Goddard (2001). Nas tabelas acima apresentadas pode-se notar que o MHGB obteve capacidade preditiva satisfatória em praticamente todos cenários, o que já não ocorreu com os demais modelos avaliados. Assim sendo, observa-se que principalmente quando os dados sejam assimétricos ou tenham *outliers* o MHGB é uma alternativa viável para seleção genômica.

VanRaden (2008), Clark e Werf (2013) observaram a superioridade da metodologia GBLUP em relação as outras metodologias utilizadas, Bayesian RR-BLUP, Bayes A, Bayes B, Bayes C e BLASSO. Nossos resultados corroboram esta conclusão para os cenários considerados. A superioridade da metodologia GBLUP também foi destacada por Azevedo *et al.* (2013), em Eucalyptus spp., em comparação ao método BLASSO. Em nosso trabalho de forma geral o GBLUP apresenta resultados ligeiramente melhores que os modelos Bayes. Há estudos de simulação em que foi verificada uma superioridade nas acurácias obtidas pela metodologia Bayes B em comparação ao GBLUP Meuwissen, Hayes e Goddard (2001) e Habier, Fernando e Dekkers (2007), de forma diferente do que observamos neste estudo. Isso pode ser reflexo da arquitetura genética ou devido a presença de cenários com forte assimetria dos dados. De modo geral o modelo MHGB foi igual ou superior aos modelos Bayes e ao GBLUP, mesmo nos cenários onde os modelos concorrentes são consagrados (referencia), ou seja, quando os fenótipos, efeitos aleatório e resíduos seguem uma distribuição normal, sem assimetria.

Em nosso estudo, os modelos Bayes e o GBLUP não apresentaram bons resultados. Isto reforça as hipóteses da justificativa deste trabalho. A herdabilidade é um parâmetro importante e é um dos fatores responsáveis por afetar a acurácia dos modelos de seleção genômica. Em geral, espera-se que quanto maior o coeficiente de herdabilidade do modelo, maior seria a sua capacidade preditiva. De modo geral, observamos isto neste trabalho, onde os modelos com maior estimativas de herdabilidade apresentaram melhores acurácias.

Da mesma forma, as maiores capacidade preditiva do MHGB nos cenários com assimetria, mas que também se mantém boa nos cenários sem assimetria foi um bom resultado em favor das hipóteses levantadas. É claro que modelos que resultam em maiores estimativas de herdabilidade tendem a ser mais acurados e apresentar melhor capacidade preditiva.

Outro fato importante é que os modelos em geral apresentaram bom ajuste no R^2, sendo MHGB o que apresentou melhores ajustes em praticamente todos os cenários avaliados. além disso, os modelos com melhores R^2 e herdabilidade, apresentaram menores MSE.

Conjecturamos que o MHGB pode ser uma ferramenta importante para analisar dados originalmente assimétricos (devido à escala de medida, por exemplo) ou com assimetria induzida por pontos discrepantes, uma vez que modelar a assimetria diminui estes problemas Gianola *et al.* (2018); Lange, Little e Taylor (1989); Seber e Wild (2003); Seber e Lee (2003). Sugere-se que o MHGB pode lidar com algumas das limitações dos modelos de regressão convencionais empregados na GWS. Os modelos GBLUP e Bayes têm boas propriedades caso

se verifiquem suas pressuposições para o estudo genético, conforme levantados na justificativa deste trabalho; mas podem fornecer resultados enganosos, caso contrário. Nas Tabelas 6.3, 6.4, 6.5 e 6.6 verificam-se resultados que corroboram esta afirmação. Segundo Gianola *et al.* (2018) a melhor previsão imparcial linear genômica (GBLUP) é muito popular no melhoramento de animais e plantas para prever características complexas, no entanto, o GBLUP é geralmente implementado sob uma suposição gaussiana para os resíduos, portanto pode ser sensível a valores extremos (genéticos ou ambientais) e, principalmente dados assimétricos conforme foi apresentado neste estudo.

Pode-se observar que nos cenários sem assimetria o modelo GBLUP é ligeiramente melhor que o MHGB e estes dois melhores que os demais. Por outro lado, à medida em que cresce o grau de assimetria o MHGB superou rapidamente o GBLUP e os demais modelos simétricos. Esse resultado está de acordo com outras pesquisas sobre regressão robusta Hampel *et al.* (2011), que embora não seja o objetivo deste trabalho, serve como suporte para validar nossos resultados.

A principal vantagem de usar o MHGB em vez do GBLUP e os modelos Bayes é que o primeiro funciona utilizando uma distribuição normal assimétrica na variável de resposta em vez da distribuição normal padrão da variável de resposta, como o GBLUP e os modelos Bayes. Conforme apontado por Koenker e Basset (1978), os modelos de regressão linear padrão (ou seja, aqueles que modelam a média condicional da variável resposta) podem ser imprecisos e viesados em contextos assimétricos ou na presença de *outliers*. Por outro lado, o modelo MHGB é menos sensível quando há assimetria ou outliers e, com isto, produz estimativas menos afetadas por este tipo de contaminação dos dados, pois vimos que na Tabela 6.7 o MHGB fez boas estimativas, tornando o modelo promissor nestes cenários assimétricos e podendo contribuir para a seleção genômica.

De forma geral, não há indícios ou evidências que demonstrem a existência de uma metodologia que seja melhor em todos os contextos, mas algumas que sobrepõem em situações específicas (PÉREZ; CAMPOS, 2014); é impossível prever como a metodologia proposta se comportaria se aplicada a outros conjuntos de dados. No entanto, nossos resultados são promissores e incentiva uma investigação mais aprofundada.

7 CONCLUSÕES

Investigamos a aplicação do Modelo Hierárquico Generalizado Normal Assimétrico Bayesiano (MHGB) ao problema da seleção genômica ampla (GWS). Em cenários sem assimetria, o modelo mostrou se quase tão acurado quanto o método consagrado GBLUP. Em cenários com assimetria, superou em muito esta referência principal e os demais modelos considerados.

A superestimação da herdabilidade com o uso do MHGB precisa ser melhor estudada. O modelo considerado é muito promissor para estudos de genômica, visto que o mesmo relaxa as suposições que outros modelos necessitam atender.

REFERÊNCIAS BIBLIOGRÁFICAS

ARELLANO-VALLE, R.; BOLFARINE, H.; LACHOS, V. Bayesian inference for skew-normal linear mixed models. *Journal of Applied Statistics*, Taylor & Francis, v. 34, n. 6, p. 663–682, 2007.

ARELLANO-VALLE, R. B.; GÓMEZ, H. W.; QUINTANA, F. A. A new class of skew-normal distributions. Citeseer, 2003.

AZEVEDO, C. F.; RESENDE, M. D. V. d.; SILVA, F. F.; LOPES, P. S.; GUIMARÃES, S. E. F. Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. *Pesquisa Agropecuária Brasileira*, SciELO Brasil, v. 48, n. 6, p. 619–626, 2013.

BALDONI, P. L. *et al.* Modelos lineares generalizados mistos multivariados para caracterização genética de doenças. [sn], 2014.

BASTIAANSEN, J. W.; COSTER, A.; CALUS, M. P.; ARENDONK, J. A. van; BOVENHUIS, H. Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. *Genetics Selection Evolution*, Springer, v. 44, n. 1, p. 3, 2012.

BERNARDO, J. M.; SMITH, A. F. Bayesian theory. Wiley, 1994.

BOLKER, B. M.; BROOKS, M. E.; CLARK, C. J.; GEANGE, S. W.; POULSEN, J. R.; STEVENS, M. H. H.; WHITE, J.-S. S. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, Elsevier, v. 24, n. 3, p. 127–135, 2009.

BORÉM, A. *Melhoramento de plantas. Viçosa: UFV*. 1997.

BOX, G.; TIAO, G. *Bayesian inference in statistical analysis, 113-122*. [S.l.]: Addison-Wesley, Reading, Massachusetts, 1992.

BOX, G. E.; TIAO, G. C. *Bayesian inference in statistical analysis*. [S.l.], 1973.

BRESLOW, N. E. Extra-poisson variation in log-linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 33, n. 1, p. 38–44, 1984.

BROOKES, A. J. The essence of snps. *Gene*, Elsevier, v. 234, n. 2, p. 177–186, 1999.

BRUIN, J. *newtest: comando para calcular o novo teste @ONLINE*. 2011. Disponível em: <https://stats.idre.ucla.edu/stata/ado/analysis/>.

CAMPOS, G. D. L.; NAYA, H.; GIANOLA, D.; CROSSA, J.; LEGARRA, A.; MANFREDI, E.; WEIGEL, K.; COTES, J. M. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, Genetics Soc America, v. 182, n. 1, p. 375–385, 2009.

CAMPOS, G. de los; PÉREZ-RODRÍGUEZ, P. Bglr: Bayesian generalized linear regression. *R package version*, v. 1, n. 5, 2016.

CLARK, S. A.; WERF, J. van der. Genomic best linear unbiased prediction (gblup) for the estimation of genomic breeding values. In: *Genome-Wide Association Studies and Genomic Prediction*. [S.l.]: Springer, 2013. p. 321–330.

CROSSA, J.; PÉREZ, P.; CAMPOS, G. de los; MAHUKU, G.; DREISIGACKER, S.; MAGOROKOSHO, C. Genomic selection and prediction in plant breeding. *Journal of Crop Improvement*, Taylor & Francis, v. 25, n. 3, p. 239–261, 2011.

DAETWYLER, H. D.; PONG-WONG, R.; VILLANUEVA, B.; WOOLLIAMS, J. A. The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, Genetics Soc America, v. 185, n. 3, p. 1021–1031, 2010.

DÁVILA, V. H. L.; BOLFARINE, H.; ARELLANO-VALLE, R. B. Modelos lineares mistos assimétricos. 2004.

GAMERMAN, D.; LOPES, H. F. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. [S.l.]: Chapman and Hall/CRC, 2006.

GAMERMAN, D.; MIGON, H. S. Dynamic hierarchical models. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 55, n. 3, p. 629–642, 1993.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; RUBIN, D. B. Bayesian data analysis. 2003.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; RUBIN, D. B. Bayesian data analysis (vol. 2). *Boca Raton, FL: Chapman*, 2014.

GEMAN, S.; GEMAN, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE transactions on pattern analysis and machine intelligence*, v. 6, n. 6, p. 721, 1984.

GIANOLA, D. Priors in whole-genome regression: the bayesian alphabet returns. *Genetics*, Genetics Soc America, v. 194, n. 3, p. 573–596, 2013.

GIANOLA, D.; CECCHINATO, A.; NAYA, H.; SCHÖN, C.-C. Prediction of complex traits: robust alternatives to best linear unbiased prediction. *Frontiers in genetics*, Frontiers, v. 9, p. 195, 2018.

GIANOLA, D.; KAAM, J. B. van. Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*, Genetics Soc America, v. 178, n. 4, p. 2289–2303, 2008.

GODDARD, M.; HAYES, B. Genomic selection. *Journal of Animal breeding and Genetics*, Wiley Online Library, v. 124, n. 6, p. 323–330, 2007.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, Genetics Soc America, v. 177, n. 4, p. 2389–2397, 2007.

HABIER, D.; FERNANDO, R. L.; KIZILKAYA, K.; GARRICK, D. J. Extension of the bayesian alphabet for genomic selection. *BMC bioinformatics*, Springer, v. 12, n. 1, p. 186, 2011.

HÄKKINEN, K.; NEWTON, R. U.; GORDON, S. E.; MCCORMICK, M.; VOLEK, J. S.; NINDL, B. C.; GOTSHALK, L. A.; CAMPBELL, W. W.; EVANS, W. J.; HÄKKINEN, A. *et al.* Changes in muscle morphology, electromyographic activity, and force production characteristics during progressive strength training in young and older men. *The Journals of*

Gerontology Series A: Biological Sciences and Medical Sciences, The Gerontological Society of America, v. 53, n. 6, p. B415–B423, 1998.

HAMPEL, F. R.; RONCHETTI, E. M.; ROUSSEEUW, P. J.; STAHEL, W. A. *Robust statistics: the approach based on influence functions*. [S.l.]: John Wiley & Sons, 2011. v. 196.

JR, M. R.; MUNOZ, P.; ACOSTA, J.; PETER, G.; DAVIS, J.; GRATTAPAGLIA, D.; RESENDE, M.; KIRST, M. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytologist*, Wiley Online Library, v. 193, n. 3, p. 617–624, 2012.

KOENKER, R.; BASSET, G. Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, v. 73, n. 363, p. 618–22, 1978.

KOHAVI, R. *et al.* A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA. *Ijcai*. [S.l.], 1995. v. 14, n. 2, p. 1137–1145.

LANGE, K. L.; LITTLE, R. J.; TAYLOR, J. M. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 84, n. 408, p. 881–896, 1989.

LEE, Y.; NELDER, J. A. Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 58, n. 4, p. 619–656, 1996.

LIANG, H.; SARAF, N.; HU, Q.; XUE, Y. Assimilation of enterprise systems: the effect of institutional pressures and the mediating role of top management. *MIS quarterly*, JSTOR, p. 59–87, 2007.

MCCULLAGH, P.; NELDER, J. A. *Generalized Linear Models*. [S.l.]: CRC Press, 1989. v. 37.

MEUWISSEN, T.; HAYES, B.; GODDARD, M. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, Genetics Society of America, v. 157, n. 4, p. 1819–1829, 2001.

NASCIMENTO, M.; SILVA, F. e; RESENDE, M. de; CD, C.; NASCIMENTO, A.; VIANA, J.; AZEVEDO, C.; BARROSO, L. Regularized quantile regression applied to genome-enabled prediction of quantitative traits. Genetics and Molecular Research, 2017.

OLIVEIRA, D. C. R. de. Modelos mistos normais assimétricos em dados de microarrays originados de pedigrees complexos. *Rev. Bras. Biom*, v. 28, n. 2, p. 137–160, 2010.

OLIVEIRA, G. F. *REGRESSÃO QUANTÍLICA APLICADA À SELEÇÃO GENÔMICA PARA CARACTERÍSTICAS OLIGOGÊNICAS EM MELHORAMENTO DE PLANTAS AUTÓGAMAS*. Tese (Doutorado) — Universidade Federal de Viçosa, 2019.

PARK, T.; CASELLA, G. The bayesian lasso. *Journal of the American Statistical Association*, Taylor & Francis, v. 103, n. 482, p. 681–686, 2008.

PAULINO, C.; TURKMAN, M.; MURTEIRA, B. estatística bayesiana fundação clouste gulbenkian lisboa. 2003.

PÉREZ, P.; CAMPOS, G. de los. Bglr: a statistical package for whole genome regression and prediction. *Genetics*, v. 198, n. 2, p. 483–495, 2014.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2019. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>.

RESENDE, M. d. Selegen-reml/blup: sistema estatístico e seleção genética computadorizada via modelos lineares mistos. *Colombo: Embrapa Florestas*, 2007.

RESENDE, M. D. V. de; JUNIOR, M. R.; AGUIAR, A. M.; ABAD, J. I. M.; MISSIAGGIA, A. A.; SANSALONI, C.; PETROLI, C.; GRATTAPAGLIA, D. Computação da seleção genômica ampla (gws). *Embrapa Florestas-Documentos (INFOTECA-E)*, Colombo: Embrapa Florestas, 2010., 2010.

RESENDE, M. D. V. de; LOPES, P. S.; SILVA, R. L. da; PIRES, I. E. Seleção genômica ampla (gws) e maximização da eficiência do melhoramento genético. *Pesquisa florestal brasileira*, n. 56, p. 63, 2008.

SAHU, S. K.; DEY, D. K.; BRANCO, M. D. A new class of multivariate skew distributions with applications to bayesian regression models. *Canadian Journal of Statistics*, Wiley Online Library, v. 31, n. 2, p. 129–150, 2003.

SAMPAIO, A. V. Estimação da equação de salário para o brasil, o paraná e o rio grande do sul em 2007—uma abordagem quantílica. *Indicadores Econômicos FEE*, v. 37, n. 2, 2009.

SEBER, G. A.; LEE, A. J. Confidence intervals and regions. *Hoboken, NJ: JohnWiley*, 2003.

SEBER, G. A.; WILD, C. *Nonlinear Regression*. [S.l.]: John Wiley & Sons, 2003. v. 503.

SILVA, E. N. d.; JÚNIOR, P.; SILVA, S. da. Sistema financeiro e crescimento econômico: uma aplicação de regressão quantílica. *Economia aplicada*, SciELO Brasil, v. 10, n. 3, p. 425–442, 2006.

SORENSEN, D.; GIANOLA, D. *Likelihood, Bayesian, and MCMC methods in quantitative genetics*. [S.l.]: Springer Science & Business Media, 2007.

STIRATELLI, R.; LAIRD, N.; WARE, J. H. Random-effects models for serial observations with binary response. *Biometrics*, JSTOR, p. 961–971, 1984.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996.

VANRADEN, P. M. Efficient methods to compute genomic predictions. *Journal of dairy science*, Elsevier, v. 91, n. 11, p. 4414–4423, 2008.

VARONA, L.; IBAÑEZ-ESCRICHE, N.; QUINTANILLA, R.; NOGUERA, J.; CASELLAS, J. Bayesian analysis of quantitative traits using skewed distributions. *Genetics research*, Cambridge University Press, v. 90, n. 2, p. 179–190, 2008.

WALKER, S. G.; GUTIÉRREZ-PENA, E. Robustifying bayesian procedures. *Bayesian statistics*, Oxford University Press, v. 6, p. 685–710, 1999.

WHITTAKER, J. C.; THOMPSON, R.; DENHAM, M. C. Marker-assisted selection using ridge regression. *Genetics Research*, Cambridge University Press, v. 75, n. 2, p. 249–252, 2000.

WILLIAMS, D. A. Extra-binomial variation in logistic linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 31, n. 2, p. 144–148, 1982.

XU, S. *Principles of statistical genomics*. [S.l.]: Springer, 2013.

YI, N.; GEORGE, V.; ALLISON, D. B. Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics*, Genetics Soc America, v. 164, n. 3, p. 1129–1138, 2003.

YI, N.; YANDELL, B. S.; CHURCHILL, G. A.; ALLISON, D. B.; EISEN, E. J.; POMP, D. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics*, Genetics Soc America, v. 170, n. 3, p. 1333–1344, 2005.