

**UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS AGRÁRIAS E ENGENHARIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E MELHORAMENTO**

TIAGO DE SOUZA MARÇAL

**ASPECTOS COMPUTACIONAIS DA ESTIMAÇÃO E PREDIÇÃO EM MODELOS
LINEARES MISTOS PARA SELEÇÃO DE HÍBRIDOS DE MILHO EM ENSAIOS
PRELIMINARES**

ALEGRE-ES

2016

TIAGO DE SOUZA MARÇAL

**ASPECTOS COMPUTACIONAIS DA ESTIMAÇÃO E PREDIÇÃO EM MODELOS
LINEARES MISTOS PARA SELEÇÃO DE HÍBRIDOS DE MILHO EM ENSAIOS
PRELIMINARES**

Dissertação apresentada ao Programa de Pós-Graduação em
Genética e Melhoramento da Universidade Federal do Espírito
Santo, como requisito para obtenção do título de mestre em
Genética e Melhoramento.

Orientador: Prof. Dr. Adésio Ferreira.

ALEGRE-ES
2016

TIAGO DE SOUZA MARÇAL

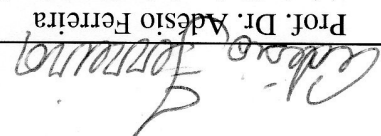
*

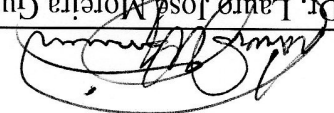
**ASPECTOS COMPUTACIONAIS DA ESTIMAÇÃO E PREDIÇÃO EM
MODELOS LINEARES MISTOS PARA SELEÇÃO DE HÍBRIDOS DE
MILHO EM ENSAIOS PRELIMINARES**

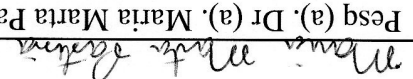
Dissertação apresentada ao Programa de Pós-Graduação em Genética e Melhoramento da
Universidade Federal do Espírito Santo, como requisito para obtenção do título de mestre em
Genética e Melhoramento.

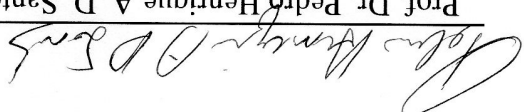
Aprovada: 30/06/2016.


COMISSÃO EXAMINADORA


Prof. Dr. Adesio Ferreira
Universidade Federal do Espírito Santo
Orientador


Pesq. Dr. Lauro José Moreira Guimarães
Embrapa Milho e Sorgo
Co-orientador


Pesq (a). Dr. (a). Maria Marta Pastina
Embrapa Milho e Sorgo
Co-orientadora


Prof. Dr. Pedro Henrique A. D. Santos
Pós-Doutorado UENF
Membro externo


Prof. Dr. Tércio da Silva de Souza
Instituto Federal do Espírito Santo
Membro externo

Dedico este trabalho á minha família e, em especial, aos meus pais Sebastião Marçal e Marcelina Maria Marçal, ao meu irmão Fábio de Souza Marçal, a minha cunhada Rogéria de Lourdes da Silva Marçal, aos meus sobrinhos Pedro Henrique da Silva Marçal e Gabriel da Silva Marçal e minha noiva Dinorah Moraes de Souza, pelo amor e confiança de todos.

AGRADECIMENTOS

A Deus!

À Universidade Federal do Espírito Santo e ao Programa de Pós-Graduação em Genética e Melhoramento, pela oportunidade de realizar o curso de mestrado.

Aos órgãos Fundação de Amparo a Pesquisa do Espírito Santo (FAPES), Conselho Nacional de Pesquisa e Desenvolvimento Científico e Tecnológico (CNPq) e Universidade Federal do Espírito Santo (UFES) pela concessão da bolsa e apoio financeiro e a EMBRAPA Milho e Sorgo pelo suporte técnico e intelectual.

Ao meu orientador Dr. Adésio Ferreira, pela orientação, ao apoio incondicional e principalmente pela amizade durante o curso.

Aos Coorientadores Dr. Lauro José Moreira Guimarães e Dr (a). Maria Marta Pastina, pelo apoio, auxílio no desenvolvimento dessa pesquisa e disponibilidade sempre prestadas.

A professora Dr (a). Marcia Flores Ferreira, pelo apoio incondicional e principalmente pela amizade durante o curso.

Aos amigos do Laboratório de Genética e Melhoramento Vegetal: Marina, Paula Henrique, Luina, Paula Mauri, Drielli, Coralina, Matheus e Iana, pelo companheirismo e os momentos de descontração.

Aos amigos do Laboratório de Biometria: José Henrique, Guilherme, Ramon, Cintia, Lidiane, Ana, Clemilton, Sephora e Sabrina, pela amizade e o alegre convívio.

Aos professores Dr. Rafael Fonseca Zanotti e Dr(a) Ludymila Brandão Motta pelo apoio, amizade e ensinamentos.

A todos os colegas de mestrado pela amizade e o alegre convívio.

Aos meus pais, Sebastião Marçal e Marcelina Maria Marçal, pelo apoio, dedicação, compreensão nos momentos difíceis, ao amor, aos esforços e sacrifícios que tiveram até eu atingir mais esta etapa de minha vida.

À meu irmão, Fabio de Souza Marçal, cunhada Rogéria Lourdes da Silva Marçal e sobrinhos Pedro Henrique da Silva Marçal e Gabriel da Silva Marçal, que ainda muito novos talvez não compreendam a importância deles para a minha vida, no entanto, agradeço a todos eles por sua dedicação, compreensão nos momentos difíceis, amor, esforços e sacrifícios que fizeram para que eu atingisse mais esta etapa de minha vida.

A minha noiva Dinorah Moraes de Souza, pela compreensão nos momentos de falta de tempo e atenção e por sempre me motivar.

A todos aqueles não citados aqui, que de alguma forma estiveram torcendo por mim durante o desenvolvimento deste trabalho.

BIOGRAFIA

TIAGO DE SOUZA MARÇAL, filho de Marcelina Maria Marçal e Sebastião Marçal, nasceu em 15 de abril de 1992, na Cidade de Santa Margarida, Estado de Minas Gerais. Em 2009, ingressou no curso de Bacharel em Agronomia do Centro de Ciências Agrárias da Universidade Federal do Espírito Santo, vindo a se graduar em julho de 2014, recebendo o título de Bacharel em Agronomia. Em agosto de 2014, ingressou no curso de mestrado do Programa de Pós-Graduação em Genética e Melhoramento do Centro de Ciências Agrárias da Universidade Federal do Espírito Santo (CCA–UFES/Alegre), submetendo-se à defesa de dissertação em junho de 2016.

“No meio da dificuldade encontra-se a oportunidade”
Albert Einstein

RESUMO

MARÇAL, Tiago de Souza, M.Sc., Universidade Federal do Espírito Santo, junho de 2016.
ASPECTOS COMPUTACIONAIS DA ESTIMAÇÃO E PREDIÇÃO EM MODELOS

LINEARES MISTOS PARA SELEÇÃO DE HÍBRIDOS DE MILHO EM ENSAIOS PRELIMINARES. Orientador: Adésio Ferreira. Coorientadores: Lauro José Moreira Guimarães e Maria Marta Pastina.

O milho (*Zea mays* L.) é uma espécie da família Poaceae, diplóide e alógama. Para esta cultura verifica-se o aumento do vigor com o acúmulo de *loci* heterozigotos, justificando assim a produção dos híbridos. Com o advento das drásticas previsões de mudanças climáticas e aumento populacional para os próximos anos é necessária à adoção, desenvolvimento e aprimoração de métodos que permitam maior eficiência na seleção e alcance de maior progresso genético em programas de melhoramento de culturas de importância agrícola poderão auxiliar na mitigação dos desafios para sustentar a segurança alimentar ainda neste século. Diante do exposto, este trabalho teve por objetivo implementar os algoritmos de primeira e segunda derivadas para o método REML (máxima verossimilhança restrita) em R, generalizáveis para diferentes modelos lineares mistos e capazes de incorporar matrizes de parentesco. Além de avaliar o impacto de simplificações matemáticas, matrizes esparsas, e diferentes taxas de erro de convergência na eficiência computacional destes algoritmos, visando a minimização do custo computacional para viabilizar o REML, em estudos com grande número de híbridos de milho e modelos complexos, em computadores de configuração simples. Os dados experimentais utilizados neste trabalho foram obtidos na safra 2013/14 em ensaio conduzido no delineamento de blocos aumentados com cinco testemunhas e 3352 híbridos simples de milho na Embrapa (Empresa Brasileira de Pesquisa Agropecuária) Milho e Sorgo situada em Sete Lagoas - MG. A variável analisada foi o rendimento de grãos, sendo esta submetida a análise através de modelos mistos com e sem a incorporação do *pedigree* utilizando-se diferentes algoritmos REML, em R, e a resposta computacional foi avaliada quanto aos critérios de convergência, taxas de erro de convergência, matrizes esparsas, computadores com diferentes capacidades de processamento, diferentes estimativas iniciais dos componentes de variância e número crescente de passos EM (*Expectation Maximization*) nos algoritmos combinados. Os algoritmos propostos foram equivalentes aos *softwares* testados (ASReml, Selegen e lme4) quanto as estimativas dos componentes de variância, indicando a coerência dos mesmos. Além disso, o uso de matrizes esparsas em associação com as otimizações propostas diminuíram o custo computacional dos algoritmos utilizando os coeficientes de determinação como critério de convergência e taxa de erro de convergência igual a 10^{-5} . A combinação híbrida do algoritmo EM, em dez passos, com o NR (*Newton Raphson*) reduziu o custo computacional e aumentou o percentual de convergência médio. Ainda observou-se que pesos uniformes para as estimativas iniciais dos componentes de variância devem ser evitados.

Palavras-chave: Algoritmos; REML; BLUP; *Zea mays*; Topcross.

ABSTRACT

MARÇAL, Tiago de Souza, M.Sc., Federal Univeristy of Espírito Santo, June, 2016. **COMPUTER ESTIMATION ASPECTS AND PREDICTION IN LINEAR MODELS FOR MIXED CORN HYBRID SELECTION IN PRELIMINARY TEST.** Advisor: Adésio Ferreira. Co-Adivsors: Lauro José Moreira Guimarães e Maria Marta Pastina.

Maize (*Zea mays* L.), is a specie from the Poaceae family, diploid and allogamous. In this culture, there is an increase with the accumulation of heterozygous *loci*, thus justifying hybrids productions. Due to drastic predictions of climate change and population growth in the coming years, it is necessary to adopt, develop and enhance methods that allow a greater efficiency in the selection and achieve greater genetic progress in crop improvement programs of agriculture importance that can help mitigation of challenges to sustain the food security of this century. Therefore, the objective of this study was to implement the algorithms of first and second derivatives for the REML (restricted maximum likelihood) method in R, generalizable for different mixed linear models and enable incorporate arrays of relationship. Moreover, to evaluate the impact of mathematical simplifications, sparse matrices and different convergence error rates in computational efficiency of these algorithms aiming to minimize the computational cost to enable REML in studies with a great number of maize hybrids and complex models, in computers with simple setup. The experimental data used in this study was obtained from harvest 2013/14 conducted in a randomized block design with five controls and 3352 simple maize hybrids in Embrapa (*Empresa Brasileira de Pesquisa Agropecuária*) Maize and Sorghum in the city of *Sete Lagoas*- MG. The analyzed variable was grain yield, which is subjected to analysis using mixed models with and without *pedigree* of incorporation using different REML algorithms, in R. Computation response evaluated the convergence criteria, error rates convergence, sparse matrices, computers with different processing capabilities, different initial estimates of variance components and increasing number of EM (Expectation Maximization) steps in combined algorithms. The proposed algorithms were equivalent for the tested software (ASReml, Selegen and Ime4) and the estimates of variance components indicating their coherence. Furthermore, the use of sparse matrices in association with the proposed optimizations, reduced the computational cost of the algorithms using coefficients of determination as a convergence criteria and convergence error rate equal to 10^{-5} . The hybrid combination of EM algorithm, in ten steps, with NR (Newton Raphson) reduced the computational cost and increased the average convergence percentage. Although, it was observed that uniform weights for the initial estimates of the variance components should be avoided.

Keywords: Algorithms, REML, BLUP, *Zea mays*, Topcross.

LISTA DE FIGURAS

Capítulo 1

Figura 1: Diferença percentual entre os critérios de convergência (componentes de variância-CV versus coeficientes de determinação-CD) para as estimativas dos parâmetros dos modelos M1 (com *pedigree*) ($\Delta CV.CD-M1$) e M2 (sem *pedigree*) ($\Delta CV.CD-M2$) (Tabelas 1 e 2 - Anexo 2) obtidas através dos algoritmos *Average Information* (AI), *Newton Rhapson* (NR), *Expectation Maximization* (EM) e os algoritmos combinados EM-AI e EM-NR para diferentes taxas de erro (TE), com algoritmos desenvolvidos para o software R. SNMIP: superou o número máximo de iterações permitido.....47

Figura 2: Diferença percentual entre as taxas de erro 10^{-6} , 10^{-7} , 10^{-8} e 10^{-9} versus 10^{-5} para as estimativas dos parâmetros dos modelos M1 (com *pedigree*) ($\Delta TE.CD-M1$) e M2 (sem *pedigree*) ($\Delta TE.CD-M2$) (Tabelas 1 e 2 - Anexo 2) obtidas através dos algoritmos *Average Information* (AI), *Newton Rhapson* (NR), *Expectation Maximization* (EM) e os algoritmos combinados EM-AI e EM-NR, para o critério de convergência CD, com algoritmos desenvolvidos para o software R. SNMIP: superou o número máximo de iterações permitido.....48

Figura 3: Número de iterações para a convergência do modelo M1 (com *pedigree*) pelos algoritmos *Average Information* (AI), *Newton Rhapson* (NR), *Expectation Maximization* (EM) e os algoritmos combinados EM-AI e EM-NR para diferentes taxas de erro (TE), considerando as diferenças entre coeficientes de determinação (CD) (Figuras CD+EM e CD-EM) e componentes de variância (CV) (Figuras CV+EM e CV-EM) como critério de convergência, para os algoritmos desenvolvidos para o software R. SNMIP: superou o número máximo de iterações permitido.....50

Figura 4: Número de iterações para a convergência do modelo M2 (sem *pedigree*) pelos algoritmos *Average Information* (AI), *Newton Rhapson* (NR), *Expectation Maximization* (EM) e os algoritmos combinados EM-AI e EM-NR para diferentes taxas de erro (TE), considerando as diferenças entre coeficientes de determinação (CD) (Figuras CD+EM e CD-EM) e componentes de variância (CV) (Figuras CV+EM e CV-EM) como critério de

convergência, para os algoritmos desenvolvidos para o software R. SNMIP: superou o número máximo de iterações permitido.....51

Figura 5: Diagnóstico de convergência dos algoritmos *Average Information* (AI), *Newton Rhapsion* (NR), *Expectation Maximization* (EM) e as combinações EM-AI e EM-NR, desenvolvidos para o software R, para a função de máxima verossimilhança residual (REML), considerando os modelos M1 (com *pedigree*) (Figuras 5 A e B), M2 (sem *pedigree*) (Figuras 5 C e D) e coeficientes de determinação (CD) como critério de convergência.....56

Capítulo 2

Figura 1: Diagrama do esquema de análise sendo P1, P2, P3 e P4 os pesos dos valores iniciais dos componentes de variância; 1 it., 5 it. e 10 it. indicam o número de iterações e EM (*Expectation Maximization*), AI (*Average Information*), FS (*Fisher's Scoring*) e NR (*Newton-Rapson*) representam os algoritmos utilizados para ajustar os 8 modelos em análise.....75

Figura 2: Heatmap da convergência dos algoritmos *Average Information* (AI), *Newton-Rapson* (NR), *Expectation Maximization* (EM) e combinações entre EM e aos demais algoritmos considerando uma iteração EM (EM1-AI, EM1-FS e EM1-NR), cinco iterações EM (EM5-AI, EM5-FS e EM5-NR) e dez iterações EM (EM10-AI, EM10-FS e EM10-NR) em função dos pesos dos valores iniciais dos componentes de variância (P1, P2, P3 e P4) para oito modelos (M1 à M8). A região gráfica destacada em cor preta no gráfico representa à convergência e a região branca a ausência de convergência.....80

Figura 3: Número de iterações para a convergência dos algoritmos *Average Information* (AI), *Newton-Rapson* (NR), *Expectation Maximization* (EM) e as combinações entre o algoritmo EM e aos demais algoritmos considerando uma iteração EM (EM1-AI e EM1-NR), cinco iterações EM (EM5-AI e EM5-NR) e dez iterações EM (EM10-AI e EM10-NR) em função dos pesos dos valores iniciais dos componentes de variância para oito modelos (+EM). A figura (-EM) possui todos os algoritmos descritos anteriormente exceto o EM para melhorar a discriminação dos resultados dos demais algoritmos. NC: não convergiu.....82

LISTA DE TABELAS

Capítulo 1

Tabela 1: Estimativas dos componentes de variância e logaritmo de REML [$\log(L)$] dos modelos M1 (com *pedigree*) e M2 (sem *pedigree*) nos softwares ASReml, Selegen, lme4, média aritmética (MA) das estimativas obtidas através dos algoritmos propostos implementados em R e diferenças percentuais dos softwares ASReml, Selegen e lme4 em relação a MA [$\Delta 1(\%)$, $\Delta 2(\%)$ e $\Delta 3(\%)$].....46

Tabela 2: Estimativa dos parâmetros genéticos para os modelos M1 (com *pedigree*) e M2 (sem *pedigree*), obtidas no software R, pelos algoritmos *Average Information* (AI), *Newton Rhapsion* (NR), *Expectation Maximization* (EM) e os algoritmos combinados EM-AI e EM-NR para diferentes taxas de erro (TE), considerando as diferenças entre coeficientes de determinação (CD) e componentes de variância (CV) como critério de convergência.....49

Tabela 3: Tempo médio de execução dos algoritmos desenvolvidos para o software R, por meio de *Average Information* (AI), *Newton Rhapsion* (NR), *Expectation Maximization* (EM) e os algoritmos combinados EM-AI e EM-NR, para os modelos M1 (com *pedigree*) e M2 (sem *pedigree*), utilizando-se ou não matrizes esparsas (ME e MN). Considerando taxa de erro de 10^{-5} e os coeficientes de determinação como critério de convergência.....53

Tabela 4: Tempo médio por iteração (T/I) dos algoritmos desenvolvidos para o software R, por meio de *Average Information* (AI), *Newton Rhapsion* (NR), *Expectation Maximization* (EM) e os algoritmos combinados EM-AI e EM-NR para os modelos M1 (com *pedigree*) e M2 (sem *pedigree*), utilizando-se ou não matrizes esparsas (ME e MN). Considerando taxa de erro de 10^{-5} e os coeficientes de determinação como critério de convergência.....54

Capítulo 2

Tabela 1: Porcentagem de convergência dos algoritmos *Average Information* (AI), *Newton-Rapson* (NR), *Expectation Maximization* (EM) e as combinações entre EM e aos demais algoritmos considerando uma iteração EM (EM1-AI, EM1-FS e EM1-NR), cinco iterações EM (EM5-AI, EM5-FS e EM5-NR) e dez iterações EM (EM10-AI, EM10-FS e EM10-NR)

em função dos pesos dos valores iniciais dos componentes de variância para os oito modelos utilizados.....78

LISTAS DE SIGLAS

REML - Máxima Verossimilhança Restrita

EM	- <i>Expectation Maximization</i>
AI	- <i>Average Information</i>
FS	- <i>Fisher's Scoring</i>
NR	- <i>Newton Raphson</i>
EM-AI	- <i>Expectation Maximization - Average Information</i>
EM-FS	- <i>Expectation Maximization - Fisher's Scoring</i>
EM-NR	- <i>Expectation Maximization - Newton Raphson</i>
CD	- Coeficientes de Determinação
CV	- Componentes de Variância
it.	- Número de iterações
P1, P2, P3, P4	- Pesos iniciais para as estimativas dos componentes de variância
EM1-AI	- Uma iteração <i>Expectation Maximization - Average Information</i>
EM1-FS	- Uma iteração <i>Expectation Maximization - Fisher's Scoring</i>
EM1-NR	- Uma iteração <i>Expectation Maximization - Newton Raphson</i>
EM5-AI	- Cinco iterações <i>Expectation Maximization - Average Information</i>
EM5-FS	- Cinco iterações <i>Expectation Maximization - Fisher's Scoring</i>
EM5-NR	- Cinco iterações <i>Expectation Maximization - Newton Raphson</i>
EM10-AI	- Dez iterações <i>Expectation Maximization - Average Information</i>
EM10-FS	- Dez iterações <i>Expectation Maximization - Fisher's Scoring</i>
EM10-NR	- Dez iterações <i>Expectation Maximization - Newton Raphson</i>

SUMÁRIO

CONTEÚDO	
INTRODUÇÃO GERAL.....	14
OBJETIVOS.....	14

OBJETIVO GERAL.....	14
OBJETIVOS ESPECÍFICOS.....	14
REFERÊNCIAS.....	15
CAPITULO 1.....	16
1.Introdução.....	16
2.Material e Métodos.....	17
2.1Modelo.....	17
2.2A função de verossimilhança residual.....	17
2.3Algoritmos e estratégias computacionais.....	18
2.3.1Expectation Maximization (EM).....	18
2.3.2Average information (AI).....	18
2.3.3Fisher's Scoring (FS).....	19
2.3.4Newton Raphson (NR).....	19
2.3.5Algoritmos Combinados (EM-AI, EM-FS e EM-NR).....	20
2.3.6Valores iniciais dos componentes de variância.....	20
2.3.7Diagnóstico de convergência.....	20
2.4Material genético estudado	20
2.5Modelos estatísticos estudados.....	21
3.Resultados.....	22
3.1Estimação de parâmetros.....	22
3.2Eficiência computacional.....	23
4.Discussão.....	27
5.Conclusão.....	28
6.REFERÊNCIAS.....	29
CAPITULO 2.....	43
1Introdução	43
2Material e Métodos.....	43
2.1Modelo.....	43
2.2A função de verossimilhança residual.....	43
2.3Algoritmos.....	43
2.3.1Diagnóstico de convergência.....	43
2.4Material genético estudado	43
2.5Modelos estudados e análises realizadas.....	44
2.5.1Modelos com efeitos de bloco fixos somados à média geral.....	44

2.5.2 Modelos com efeito fixo somente para a média geral.....	44
2.5.3 Modelos com efeitos de testemunha fixos somados a média geral.....	44
3 Resultados.....	44
3.1 Convergência dos algoritmos.....	44
3.2 Número de iterações para atingir a convergência.....	45
3.3 Estimativas de parâmetros.....	45
4 Discussão.....	47
5 Conclusão.....	47
6 Referências	47

INTRODUÇÃO GERAL

O milho (*Zea mays* L.) é uma espécie da família Poaceae, diploide com número de cromossomos $2n = 2x = 20$ (WANG et al., 2014) e alógama. Devido à alogamia, na cultura do milho observa-se grande variabilidade nas progênies oriundas do intercruzamento de indivíduos dentro de uma população. Entretanto, na ocorrência de endogamia verifica-se progressiva perda de vigor das progênies (BOREM; MIRANDA, 2009). Em contrapartida, o acúmulo de *loci* heterozigotos (heterose) promove o aumento do vigor. Devido a estas particularidades citadas anteriormente a maior parte das cultivares atuais de milho são híbridos simples, que tem demonstrado boa resposta produtiva em campo.

No ano de 2013 a cultura ocupou área mundial de aproximadamente 185,12 milhões de hectares que foi responsável por produção de 1,02 bilhões de toneladas e produtividade média de 5.499,7 kg ha⁻¹. Neste mesmo ano o Brasil foi responsável por 8,25% da área cultivada, 7,88% da produção mundial e atingiu produtividade média de 5.253,6 kg ha⁻¹, o que conferiu ao mesmo, terceira posição no *rank* mundial de produção (FAO, 2015).

Pode ser considerada uma cultura multiuso, pois serve de alimento para os seres humanos e animais e pode ser utilizado na produção de combustível. Seus grãos têm alto valor nutricional e são utilizados como matéria prima para a fabricação de muitos produtos industriais (AFZAL et al., 2009).

O milho é uma das três culturas base da segurança alimentar mundial juntamente com o arroz e o trigo. No ano de 2013 a produção mundial de milho superou a de arroz e trigo em 37,42% e 42,21%, respectivamente (FAO, 2015), destacando sua importância no cenário mundial. Entretanto, apesar de termos superado as péssimas previsões da teoria de Malthus, que acreditava que o crescimento demográfico iria ultrapassar a capacidade produtiva do mundo novamente a agricultura enfrentará novos desafios frente as mudanças climáticas e ao crescimento populacional (MALTHUS, 1798).

Os relatórios da FAO preveem que na segunda metade do século a população humana atingirá 9,1 bilhões de habitantes e a produção deve sofrer acréscimo de 70% para suportar este incremento populacional (FAO, 2009). Além disso, o planeta terra pode estar sujeito a severas mudanças climáticas em escala global como: aumento da temperatura e da concentração atmosférica de CO₂, além de secas mais frequentes (LONG; ORT, 2010).

Esse conjunto de mudanças poderá alterar totalmente o sistema produtivo atual no futuro, face a necessidade de dobrar a produção das principais culturas responsáveis pela segurança alimentar (RAY et al., 2013). Portanto, novas tecnologias devem ser desenvolvidas e as atuais aprimoradas para enfrentarmos os desafios. O aumento na produção é o objetivo

para os próximos anos, visando garantir a segurança alimentar. Estima-se que a produção de cereais tenha que aumentar em cerca de um bilhão de toneladas, mas apenas o milho vem mantendo o incremento significativo na produção mundial diferentemente do arroz e do trigo (FAO, 2009; LONG; ORT, 2010).

Nas últimas décadas o incremento na produção do milho se deve principalmente ao melhoramento genético, realizado por grande número de pesquisadores de instituições privadas e públicas no mundo inteiro. Entretanto, devido ao nível avançado de melhoramento da espécie, é necessário adotar novas estratégias para aumentar significativamente a produção.

Neste contexto, diversos pesquisadores têm optado pelo *rank* dos valores genotípicos preditos a partir dos valores fenotípicos no momento da seleção para diversas culturas como: cana de açúcar (LUCIUS et al., 2014), trigo (PIMENTEL et al., 2014), algodão (RESENDE et al., 2014), mamona (VIVAS et al., 2014), maracujá (ASSUNÇÃO et al., 2015; SANTOS et al., 2015a), eucalipto (SANTOS et al., 2015b) e milho (BERNARDO, 1996).

Os valores genotípicos preditos são obtidos da abordagem de modelos mistos originalmente proposta por Henderson (1949), para aplicação no melhoramento de gado leiteiro. Sendo apresentada formalmente em 1959 (HENDERSON, 1959). Dentre as classes de modelos lineares o BLUP (melhor predição linear não viciada) maximiza a acurácia seletiva (RESENDE; DUARTE, 2007), sendo considerado um procedimento ótimo de seleção (RESENDE, 2002; BORGES et al., 2009).

Este procedimento pode ser aplicado a dados desbalanceados ou não, conduzindo a estimativas e predições mais precisas dos efeitos fixos e valores genéticos (aleatórios), respectivamente. Suas principais vantagens são: permitir comparar genótipos através do tempo (gerações e anos) e espaço (locais e blocos); permite a simultânea correção para os efeitos ambientais, predição de valores genéticos; manipulação de estruturas complexas de dados (medidas repetidas, diferentes anos, locais e delineamentos); permite utilizar simultaneamente grande número de informações, provenientes de diferentes gerações, locais e idades (RESENDE, 2002).

Além de permitir a análise de dados oriundos de inúmeras configurações experimentais utilizadas em programas de melhoramento. A abordagem de modelos mistos ainda permite a decomposição da variação genética em termos aditivos, dominantes e epistáticos (em alguns casos), quando se conhece o *pedigree* dos indivíduos em análise. Desta forma conhecendo-se as informações de parentesco podem-se construir as matrizes de parentesco genético aditivo

(A) (HENDERSON, 1975) e de dominância (D) (COCKERHAM, 1954; HENDERSON, 1985), que podem ser inclusas no sistema de equações de modelos mistos.

As matrizes de parentesco genético devem ser invertidas antes de serem incorporadas no sistema de equações de modelos mistos, o que pode exigir um esforço computacional demasiado dependendo do número de indivíduos em análise. Portanto, Henderson (1976) apresentou um algoritmo para a obtenção indireta da inversa da matriz A e Henderson (1985) mencionou um artifício para a obtenção indireta dos efeitos de dominância utilizando a matriz D, sem a necessidade de sua inversão.

Apesar das enormes vantagens dos modelos mistos, sua eficácia seletiva só será evidenciada quando as estimativas dos componentes de variância associados aos efeitos aleatórios forem as mais fieis possíveis (RESENDE; SILVA; AZEVEDO, 2014). No entanto, vários métodos estão disponíveis e os mais comuns são: análise de variância (ANOVA) (SEARLE; CASELLA; MCCULLOCH, 1992), máxima verossimilhança (ML) (HARTLEY; RAO, 1967) e máxima verossimilhança residual (REML) (PATTERSON; THOMPSON, 1971).

Destes o que tem sido mais utilizado é o REML, pois corrige o viés proporcionado por ML e permite o ajuste de modelos complexos com tratamentos e erros correlacionados mesmos sob desbalanceamento, que são situações comuns no melhoramento genético (RESENDE; SILVA; AZEVEDO, 2014). Além de produzir estimativas sempre positivas para os componentes de variância (SEARLE; CASELLA; MCCULLOCH, 1992).

A obtenção de componentes de variância via procedimento REML ocorre em um processo iterativo, onde a convergência é obtida após a maximização da função de verossimilhança restrita via algoritmos numéricos. Partindo-se de um valor arbitrário para os coeficientes de determinação ou valores absolutos dos componentes de variância, associados aos efeitos aleatórios.

A natureza iterativa do método REML implica em custo computacional diretamente proporcional ao volume de informações em análise. Em vista desta limitação diversos pesquisadores têm desenvolvido algoritmos para melhorar a eficiência do procedimento (RESENDE; SILVA; AZEVEDO, 2014). Os mais comuns são: EM (*Expectation Maximization*) (DEMPSTER; LAIRD; RUBIN, 1977), AI (*Average Information*) (JOHNSON; THOMPSON, 1995), NR (*Newton Raphson*) (SEARLE; CASELLA; MCCULLOCH, 1992) e FS (*Fisher's Scoring*) (SEARLE; CASELLA; MCCULLOCH, 1992).

Dos algoritmos citados o EM é o mais estável, entretanto, é também o mais lento exigindo, na maioria das vezes, grande número de iterações para atingir a convergência. Ele é baseado na primeira derivada da função de verossimilhança restrita, possui convergência linear e suas estimativas estão sempre dentro do espaço paramétrico.

Os demais algoritmos são baseados na segunda derivada da função de verossimilhança restrita, convergem em poucas iterações (convergência quadrática), entretanto às vezes produzem estimativas que fogem do espaço paramétrico. Adicionalmente estes algoritmos envolvem inversões de matrizes que possuem mesma ordem do conjunto de dados, aumentando assim o esforço computacional para a obtenção dos componentes de variância (RESENDE, 2002; RESENDE; SILVA; AZEVEDO, 2014).

Embora todos os algoritmos possuam vantagens e desvantagens para viabilizar a análise via metodologia REML/BLUP, têm se utilizado algoritmos combinados e simplificações matemáticas e numéricas visando diminuir o custo computacional (SEARLE; CASELLA; MCCULLOCH, 1992; JOHNSON; THOMPSON, 1995; RESENDE, 2002; RESENDE; SILVA; AZEVEDO, 2014).

Desde a elucidação dos procedimentos BLUP, matrizes de parentesco e REML estes métodos vêm sendo utilizados intensivamente no melhoramento animal, sendo considerado atualmente um procedimento corriqueiro. Entretanto, apesar dos métodos consistirem em uma poderosa ferramenta para o melhoramento, somente recentemente eles vêm sendo aplicados em larga escala no melhoramento vegetal. Evidenciando ainda pouco uso de modelos que incorporam as informações de parentesco em programas de melhoramento genético vegetal.

A menor utilização dos modelos mistos no melhoramento vegetal pode estar associada a menor familiaridade dos melhoristas com esta abordagem que exige uso intenso de recursos computacionais. Neste contexto, o *software* R assume importância especial por permitir a implementação ou a utilização de pacotes para o ajuste de diversos modelos mistos de forma gratuita e eficiente. Além disso, devido a grande popularidade do *software* R as novas ferramentas geradas podem ser facilmente difundidas e melhoradas, garantindo acesso dos pesquisadores a procedimentos cada vez mais aplicáveis e eficientes para ajustes de modelos mistos.

OBJETIVOS

OBJETIVO GERAL

Diante do exposto, este trabalho teve por objetivo implementar algoritmos de primeira e segunda derivadas para o método REML em R, generalizáveis para diferentes modelos

lineares mistos e capazes de incorporar matrizes de parentesco. Além de avaliar o impacto de simplificações matemáticas, matrizes esparsas, diferentes taxas de erro de convergência na eficiência computacional destes algoritmos, visando a minimização do custo computacional para viabilizar o REML, em estudos com grande número de híbridos de milho e modelos complexos, em computadores de configuração simples.

OBJETIVOS ESPECÍFICOS

- I. Implementar os algoritmos de primeira e segunda derivadas e algoritmos combinados para o método REML em R;
- II. Propor simplificações matemáticas que levem a otimização dos algoritmos em R;
- III. Verificar o impacto das matrizes esparsas sobre a eficiência computacional dos algoritmos REML em R;
- IV. Apontar o critério e a taxa de erro mais adequados para monitorar a convergência para os modelos utilizados na seleção de híbridos de milho;
- V. Indicar os algoritmos mais eficientes para ajustar os modelos mistos, aplicados a seleção de híbridos de milho, via REML.
- VI. Verificar a resposta destes algoritmos em diferentes computadores.

REFERÊNCIAS

- AFZAL, M.; NAZIR, Z.; BASHIR, M.; KHAN, B. Analysis of host plant resistance in some genotypes of maize against chilo partellus (SWINHOE) (PYRALIDAE: LEPIDOPTERA). **Pakistan Journal of Botany**, n. 1, v. 41, p. 421-428, 2009.
- ASSUNÇÃO, M. P.; KRAUSE, W.; DALLACORT, R.; SANTOS, P. R. J.; NEVES, L. G. Seleção individual de plantas de maracujazeiro azedo quanto à qualidade de frutos via REML/BLUP. **Revista Caatinga**, v. 28, n. 2, p. 57-63, 2015.
- BERNARDO, R. Best linear unbiased prediction of maize single-cross performance. **Crop Science**, v. 36, n. 1, p. 50-56, 1996.
- BORÉM, A.; MIRANDA, G.V. **Melhoramento de plantas**. 5.ed. Viçosa: UFV, 2009.
- BORGES, V.; SOARES, A. A.; RESENDE, M. D. V.; REIS, M. S.; CORNÉLIO, V. M. O.; SOARES, P. C. Progresso genético do programa de melhoramento de arroz de terras altas de Minas gerais utilizando modelos mistos. **Revista Brasileira de Biometria**, v. 27, n. 3, p. 478-490, 2009.
- COCKERHAM, C. C. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. **Genetics**, v. 39, n. 6, p. 859, 1954.
- DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society**, v. 39, n.1, p. 1-38, 1977.
- FAO. **Global agriculture towards 2050**. 2009. Disponível em: < http://www.fao.org/fileadmin/templates/wsfs/docs/Issues_papers/HLEF2050_Global_Agriculture.pdf>. Acesso em: 12 de jun. de 2015.
- FAO. **Word production**. Disponível em: < <http://faostat3.fao.org/download/Q/QC/E>>. Acesso em: 12 de jun. de 2015.
- HARTLEY, H. O.; RAO, J. N. K. Maximum-likelihood estimation for the mixed analysis of variance model. **Biometrika**, v. 54, n. 1-2, p. 93-108, 1967.
- HENDERSON, C. R. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. **Biometrics**, p. 69-83, 1976.
- HENDERSON, C. R. Best linear unbiased prediction of nonadditive genetic merits. **Journal of Animal Science**, v. 60, p. 111-117, 1985.

HENDERSON, C. R. Estimation of changes in herd environment. **Journal of Dairy Science**, v. 32, n. 8, p. 706, 1949.

HENDERSON, C. R.; KEMPTHORNE, O.; SEARLE, S. R.; VON KROSIGK, C. M. The estimation of environmental and genetic trends from records subject to culling. **Biometrics**, v. 15, n. 2, p. 192-218, 1959.

HENDERSON, C. R. Use of relationships among sires to increase accuracy of sire evaluation. **Journal of Dairy Science**, v. 58, n. 11, p. 1731-1738, 1975.

JOHNSON, D. L.; THOMPSON, R. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. **Journal of Dairy Science**, v. 78, n. 2, p. 449-456, 1995.

LONG, S. P.; ORT, D. R. More than taking the heat: crops and global change. **Current Opinion in Plant Biology**, n. 3, v. 13, p. 241-248, 2010.

LUCIUS, A. S. F.; DE OLIVEIRA, R. A.; DAROS, E.; BESPALHOK FILHO, J. C.; VERISSIMO, M. A. A. Desempenho de famílias de cana-de-açúcar em diferentes fases no melhoramento genético via REML/BLUP. **Semina: Ciências Agrárias**, v. 35, n. 1, p. 101-112, 2014.

MALTHUS T. R. **Ensaio sobre a população**. Cambridge, 1798.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, v. 58 n. 3, p. 545-554, 1971.

PIMENTEL, A. J. B.; GUIMARÃES, J. F. R.; SOUZA, M. A.; RESENDE, M. D. V.; MOURA, L. M.; RIBEIRO, G. Estimação de parâmetros genéticos e predição de valor genético aditivo de trigo utilizando modelos mistos. **Pesquisa Agropecuária Brasileira**, v. 49, n. 11, p. 882-890, 2014.

RAY, D. K.; MUELLER, N. D.; WEST, P. C.; FOLEY, J. A. Yield trends are insufficient to double global crop production by 2050. **PloS one**, v. 8, n. 6, p. e66428, 2013.

RESENDE, M. D. V.; DUARTE, J. B. PRECISÃO E CONTROLE DE QUALIDADE EM EXPERIMENTOS DE AVALIAÇÃO DE CULTIVARES. **Pesquisa Agropecuária Tropical**, v. 37, n. 3, p. 182-194, 2007.

RESENDE, M. A. V.; FREITAS, J. A.; LANZA, M. A.; RESENDE, M. D. V.; AZEVEDO, C. F. Divergência genética e índice de seleção via BLUP em acessos de algodoeiro para

características tecnológicas da fibra. **Pesquisa Agropecuária Tropical**, v. 44, n. 3, p. 334-340, 2014.

RESENDE, M. D. V. **Genética biométrica e estatística no melhoramento de plantas perenes**. Embrapa Informação Tecnológica, Colombo: Embrapa Florestas, 2002.

RESENDE, M. D. V.; SILVA, F. F. E. ; AZEVEDO, C. F. **ESTATÍSTICA MATEMÁTICA, BIOMÉTRICA E COMPUTACIONAL: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência**. 1ª ed. Visconde do Rio Branco: Suprema, 2014, v. 1, p. 448-502.

SANTOS, E. A.; VIANA, A. P.; OLIVEIRA FREITAS, J. C.; RODRIGUES, D. L.; TAVARES, R. F.; PAIVA, C. L.; SOUZA, M. M. Genotype selection by REML/BLUP methodology in a segregating population from an interspecific *Passiflora* spp. crossing. **Euphytica**, v. 204, n. 1, p. 1-11, 2015a.

SANTOS, G. A.; RESENDE, M. D. V.; SILVA, L. D.; HIGA, A.; ASSIS, T. F. Interação genótipos x ambientes para produtividade de clones de *Eucalyptus* L'Hér. no estado do rio grande do sul. **Revista Árvore**, v. 39, n. 1, p. 81-91, 2015b.

SEARLE, S. R.; CASELLA, G.; MCCULLOCH, C. E. **Variance Components**. New York: Wiley, 1992.

VIVAS, M.; SILVEIRA, S. F.; VIVAS, J. M. S.; VIANA, A. P.; AMARAL JUNIOR, A. T.; PEREIRA, M. G. Seleção de progênies femininas de mamoeiro para resistência a mancha-de-phoma via modelos mistos. **Bragantia**, v. 73, n. 4, p. 446-450, 2014.

WANG, K.; WU, Y.; ZHANG, W.; DAWE, R. K.; JIANG, J. Maize centromeres expand and adopt a uniform size in the genetic background of oat. **Genome research**, v. 24, n. 1, p. 107-116, 2014.

CAPITULO 1

Algoritmos eficientes em R para estimação de parâmetros genéticos em milho com modelos lineares mistos que incorporam informação de parentesco via pedigree

RESUMO: Os modelos mistos são uma poderosa ferramenta para a estimação de parâmetros e predição de valores genéticos, no entanto o custo computacional destes estudos é demasiadamente elevado para grandes conjuntos de dados e modelos complexos. Diante do exposto, este trabalho teve por objetivo implementar algoritmos utilizando o método REML (máxima verossimilhança residual), em R, para estimação de parâmetros genéticos de milho, considerando grande volume de dados de um programa de melhoramento e com incorporação de informações de parentesco via pedigree. Também foi avaliado o impacto de simplificações matemáticas, matrizes esparsas, critérios de convergência e taxas de erro de convergência na eficiência computacional destes algoritmos. Neste estudo, foram apresentadas simplificações matemáticas que levaram a melhorias no desempenho dos algoritmos estudados, diminuindo o tempo de execução dos mesmos e foi ratificada a relevância do uso de matrizes esparsas. Além disso, verificou-se que a convergência pode ser monitorada eficientemente por coeficientes de determinação, como a herdabilidade, e foi constatado que a taxa de erro 10^{-5} promoveu precisão satisfatória para a estimação de parâmetros genéticos.

Palavras-chave: Eficiência computacional; REML; BLUP; *Zea mays*.

1. INTRODUÇÃO

Em um futuro próximo o incremento populacional, as mudanças climáticas e a demanda por biocombustíveis serão responsáveis pelo aumento da pressão sobre a produção agrícola (FOLEY et al., 2011; RAY et al., 2013), e o avanço tecnológico no melhoramento genético será crucial para a superação dos desafios (TESTER; LANGRIDGE, 2010; ARAUS; CAIRNS, 2014). Neste contexto, novas tecnologias têm sido aplicadas nos programas de melhoramento visando melhor caracterização fenotípica e genética para identificação de cultivares superiores de forma mais acurada e rápida, preconizando o aumento da eficiência de seleção (COBB et al., 2013; ARAUS; CAIRNS, 2014; CASALE et al., 2015; LOH et al., 2015).

A estimação de parâmetros é um procedimento muito utilizado e fundamental em estudos de melhoramento genético animal e vegetal (THOMPSON; BROTHERSTONE; WHITE, 2005; MISZTAL, 2008; ROYCHOWDHURY; TAH, 2011). Durante muito tempo a análise de variância (ANOVA), desenvolvida por Ronald Aylmer Fisher, tem sido a principal

ferramenta utilizada no melhoramento vegetal, sendo reportada em vários trabalhos científicos (BOMMERT; NAGASAWA; JACKSON, 2013; JESSKE, T. et al., 2013; BAXTER, et. al., 2014; KÖNIG et al., 2014; LIU; YANG; HU, 2015). Entretanto, tal procedimento é limitado quando há necessidade de se modelar tratamentos e erros correlacionados e conjuntos de dados complexos e desbalanceados que são situações comuns em programas de melhoramento (RESENDE; SILVA; AZEVEDO, 2014). Adicionalmente, a técnica de ANOVA pode conduzir a estimativas negativas dos componentes de variância (SEARLE; CASELLA; MCCULLOCH, 1992), fato que pode ser embaraçoso para os melhoristas.

Com o intuito de produzir estimativas mais fiéis dos componentes de variância Patterson; Thompson (1971) conceberam o método de máxima verossimilhança residual (REML) que contorna o vício produzido pelo procedimento de máxima verossimilhança (ML), devido à remoção dos graus de liberdade utilizados na estimação dos efeitos fixos (THOMPSON, 2008; TENESA; HALEY, 2013). Além disso, REML mantém as propriedades vantajosas dos estimadores ML como o ajuste de modelos complexos com erros e tratamentos correlacionados e a não negatividade dos componentes de variância (SEARLE; CASELLA; MCCULLOCH, 1992; RESENDE; SILVA; AZEVEDO, 2014).

Devido a tais características o método REML em associação com os modelos mistos revolucionaram a prática de seleção, em especial em programas de melhoramento animal, (THOMPSON, 2008), e o método REML se tornou preferido na estimação de parâmetros genéticos em melhoramento animal (JOHNSON; THOMPSON, 1995; MEYER, 2008) e vem sendo difundido no melhoramento vegetal (VIRK et al., 2009; MÖHRING; MELCHINGER; PIEPHO, 2011; RESENDE; SILVA; AZEVEDO, 2014).

Apesar da robustez e aplicabilidade do método REML, tal procedimento demanda grandes recursos computacionais à medida que a complexidade dos modelos e o número de informações experimentais aumentam, devido à natureza iterativa do método (MEYER, 1991; HARVILLE, 2004; MISZTAL, 2008). Neste contexto, diversos pesquisadores vêm desenvolvendo e incorporando tecnologias que viabilizem computacionalmente o método REML para modelos complexos e grandes conjuntos de dados. Dentre estas, podemos citar o uso de matrizes esparsas (MISZTAL; PEREZ-ENCISO, 1993; JOHNSON; THOMPSON, 1995; THOMPSON et al., 2003), simplificações matemáticas (MEYER, 1983; MEYER, 1987; SEARLE; CASELLA; MCCULLOCH, 1992; GILMOUR; THOMPSON; CULLIS, 1995; JOHNSON; THOMPSON, 1995; MEYER, 2008), e o desenvolvimento de novos

algoritmos (JOHNSON; THOMPSON, 1995; FOULLEY; VAN DYK, 2000; DIFFEY; WELSH; CULLIS, 2013; LI; POURAHMADI, 2013).

Como já citado, a obtenção da estimativa dos parâmetros de um modelo misto via REML ocorre em um processo iterativo, onde a função de verossimilhança residual é maximizada. Portanto, algoritmos numéricos devem ser utilizados, sendo os mais comuns o EM (*Expectation Maximization*) (DEMPSTER; LAIRD; RUBIN, 1977), AI (*Average Information*) (JOHNSON; THOMPSON, 1995; GILMOUR; THOMPSON; CULLIS, 1995), FS (*Fisher's Scoring*) (PATTERSON; THOMPSON, 1971) e NR (*Newton Raphson*) (SEARLE; CASELLA; MCCULLOCH, 1992), sendo que o EM é baseado na primeira derivada e os demais na segunda derivada da função de verossimilhança residual.

O algoritmo de primeira derivada EM caracteriza-se por ser numericamente estável, sensível a diferentes conjuntos de valores iniciais atribuídos aos parâmetros do modelo e apresentar lentidão para atingir a convergência, diferentemente dos algoritmos de segunda derivada (AI, FS e NR) (MEYER; SMITH, 1996; RESENDE; SILVA; AZEVEDO, 2014). Devido às diferentes características de ambas as classes de algoritmos alguns autores têm sugerido a utilização de combinações entre as classes, visando explorar a estabilidade numérica e rápida convergência (MEYER, 2006; MEYER, 2007a).

O uso de procedimentos numéricos em REML implica na necessidade do uso de softwares capazes de processar este tipo de análise, dos quais podemos citar o ASReml (GILMOUR et al., 2015), o Genstat (PAYNE, 2009), o Wombat (MEYER, 2007b), o Selegen (RESENDE, 2007), o SAS (SAS Institute Inc, 2012) e o R (R CORE TEAM, 2016). Dentre estes, o R assume importância especial por ser livre e possibilitar a modelagem e a implementação dos mais vastos tipos de modelos mistos. Entretanto, apesar do software R apresentar diversas bibliotecas eficientes para o ajuste de modelos mistos via REML, basicamente as bibliotecas nlme (PINHEIRO et al., 2016), lme4 (BATES et al., 2015) e ASReml (BUTLER et al., 2009) permitem uma modelagem mais flexível. Entretanto, nlme e lme4 não admitem o uso de matrizes de parentesco e ASReml é um “pacote estatístico”, desenvolvido para rodar em ambiente R pelo grupo detentor do software ASReml, que não é gratuito.

Diante do exposto, este trabalho teve por objetivo implementar algoritmos para o método REML, em R, para estimação de parâmetros genéticos de milho, considerando grande volume de dados de híbridos topcrosses avaliados em ensaios em blocos aumentados e com incorporação de informações de parentesco via pedigree. Além disso, avaliou-se o impacto de

simplificações matemáticas, matrizes esparsas, critérios de convergência e taxas de erro de convergência na eficiência computacional destes algoritmos.

2. MATERIAL E MÉTODOS

2.1 MODELO

O modelo linear misto utilizado na implementação dos algoritmos em R foi uma generalização do originalmente proposto por Henderson et al. (1959) e pode ser escrito na forma matricial usando a notação

$$y = X\beta + \sum_{i=1}^I Z_i u_i + \varepsilon, \quad (1a)$$

ou, em sua forma compacta como é rotineiramente representado

$$y = X\beta + Zu + \varepsilon \quad \text{com } N_u = \sum_{i=1}^I N_{u_i}, \quad (1b)$$

onde y : é o vetor de $N \times 1$ de observações, β : é o vetor de $N_\beta \times 1$ de efeitos fixos e covariáveis fixas, X : é a matriz de incidência para efeitos fixos e covariáveis fixas de dimensão $N \times N_\beta$, N_{u_i} : número de níveis do vetor de efeitos aleatórios, u : é o vetor que concatena i vetores de efeitos aleatórios $u = (u_1, u_2, u_3 \dots u_i)$ de dimensão $N_u \times 1$, Z : é a matriz de incidência que agrupa i matrizes de efeitos aleatórios $Z = (Z_1, Z_2, Z_3 \dots Z_i)$ de dimensão $N \times N_u$ e ε : é o vetor de $N \times 1$ erros aleatórios.

Ainda foram assumidas as seguintes estruturas de (co)variância

$$\begin{aligned} \begin{bmatrix} u \\ \varepsilon \end{bmatrix} &\sim N\left(0, \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}\right), \quad G = \bigoplus_{i=1}^I G_i, \\ \text{var}(u_i) &= H_i \sigma_{u_i}^2 = G_i \quad \text{ou} \quad I \sigma_{u_i}^2 = G_i \quad \forall H_i = I, \\ \text{var}(\varepsilon) &= I \sigma_\varepsilon^2 = R, \\ \text{cov}(u_i, \varepsilon) &= 0, \\ \text{var}(y) &= V = ZGZ' + R, \end{aligned}$$

onde G : é a matriz de (co)variância dos efeitos aleatórios, \bigoplus : é a operação de soma direta, R : é a matriz de (co)variância residual, V : é a matriz de (co)variância de y , H_i : é a matriz de correlação entre os efeitos aleatórios de u_i , I : é uma matriz identidade, $\sigma_{u_i}^2$: é a variância de u_i e σ_ε^2 : é a variância do erro.

As soluções de máxima verossimilhança de β e $u = (u_1, u_2, u_3 \dots u_i)$ do modelo (1a) podem ser obtidas através da primeira derivada da função de densidade de probabilidade

conjunta $f(y,u)$, proposta por Henderson et al. (1959), em relação aos parâmetros β e u e então obtemos o sistema de equações de modelos mistos (SEMM)

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z_1 & X'R^{-1}Z_2 & \cdots & X'R^{-1}Z_j \\ Z_1'R^{-1}X & Z_1'R^{-1}Z_1 + G_1^{-1} & Z_1'R^{-1}Z_2 & \cdots & Z_1'R^{-1}Z_j \\ Z_2'R^{-1}X & Z_2'R^{-1}Z_1 & Z_2'R^{-1}Z_2 + G_2^{-1} & \cdots & Z_2'R^{-1}Z_j \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_i'R^{-1}X & Z_i'R^{-1}Z_1 & Z_i'R^{-1}Z_2 & \cdots & Z_i'R^{-1}Z_j + G_i^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \tilde{u}_1 \\ \tilde{u}_2 \\ \vdots \\ \tilde{u}_i \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z_1'R^{-1}y \\ Z_2'R^{-1}y \\ \vdots \\ Z_i'R^{-1}y \end{bmatrix}, \quad (2)$$

entretanto, assumindo que a estrutura de R^{-1} é uma matriz identidade (MRODE, 2014) podemos reescrever o SEMM como

$$\begin{bmatrix} X'X & X'Z_1 & X'Z_2 & \cdots & X'Z_j \\ Z_1'X & Z_1'Z_1 + H_1^{-1}\lambda_1 & Z_1'Z_2 & \cdots & Z_1'Z_j \\ Z_2'X & Z_2'Z_1 & Z_2'Z_2 + H_2^{-1}\lambda_2 & \cdots & Z_2'Z_j \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_i'X & Z_i'Z_1 & Z_i'Z_2 & \cdots & Z_i'Z_j + H_i^{-1}\lambda_i \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \tilde{u}_1 \\ \tilde{u}_2 \\ \vdots \\ \tilde{u}_i \end{bmatrix} = \begin{bmatrix} X'y \\ Z_1'y \\ Z_2'y \\ \vdots \\ Z_i'y \end{bmatrix}, \quad (3)$$

onde os lambdas são expressos pela seguinte expressão

$$\lambda_1, \lambda_2 \cdots \lambda_i = \frac{\sigma_\varepsilon^2}{\sigma_{u_1}^2}, \frac{\sigma_\varepsilon^2}{\sigma_{u_2}^2} \cdots \frac{\sigma_\varepsilon^2}{\sigma_{u_i}^2}.$$

A solução dos SEMM apresentadas em (2) e (3) assumem que $(\sigma_{u_1}^2, \sigma_{u_2}^2 \cdots \sigma_{u_i}^2)$ e σ_ε^2 são conhecidos. Porém, na prática tais parâmetros são desconhecidos e podem ser substituídos por suas estimativas, $(\hat{\sigma}_{u_1}^2, \hat{\sigma}_{u_2}^2 \cdots \hat{\sigma}_{u_i}^2)$ e $\hat{\sigma}_\varepsilon^2$, obtidas através de REML.

2.2 A FUNÇÃO DE VEROSSIMILHANÇA RESIDUAL

A função de verossimilhança residual assume que $y \sim N(X\beta, V)$ e pode ser escrita como (MEYER, 1989; JOHNSON; THOMPSON, 1995)

$$\begin{aligned} -2\log(L) &= \text{const.} + \log|V| + \log|X'V^{-1}X| + y'Py \\ &= \text{const.} + \log|R| + \log|G| + \log|C^*| + y'Py, \end{aligned} \quad (4)$$

onde $\log(L)$: é o logaritmo da função de verossimilhança residual (PATTERSON;

THOMPSON, 1971), const.: é uma constante da função de verossimilhança residual, P: é o projetor ortogonal de y no espaço coluna de X (GILMOUR; THOMPSON; CULLIS, 1995), C*: é a matriz de coeficientes do SEMM (2) ou o produto da matriz de coeficientes do SEMM (3) pelo inverso da variância residual.

As matrizes P e C* podem ser escritas como

$$P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1} = R^{-1} - R^{-1}W(C^*)^{-1}W'R^{-1} \quad (5)$$

$$W = [X:Z]$$

$$Z = [Z_1:Z_2:Z_3 \cdots Z_i]$$

$$C = \begin{bmatrix} X'X & X'Z_1 & X'Z_2 & \cdots & X'Z_j \\ Z_1'X & Z_1'Z_1 + H_1^{-1}\lambda_1 & Z_1'Z_2 & \cdots & Z_1'Z_j \\ Z_2'X & Z_2'Z_1 & Z_2'Z_2 + H_2^{-1}\lambda_2 & \cdots & Z_2'Z_j \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_i'X & Z_i'Z_1 & Z_i'Z_2 & \cdots & Z_i'Z_j + H_i^{-1}\lambda_i \end{bmatrix}$$

$$C^* = \frac{1}{\sigma_\varepsilon^2} C \text{ para } R = I\sigma_\varepsilon^2$$

2.3 ALGORITMOS E ESTRATÉGIAS COMPUTACIONAIS

Para este trabalho, os algoritmos REML implementados em R, foram o EM (*Expectation Maximization*) (DEMPSTER; LAIRD; RUBIN, 1977), AI (*Average Information*) (JOHNSON; THOMPSON, 1995; GILMOUR; THOMPSON; CULLIS, 1995), FS (*Fisher's Scoring*) (PATTERSON; THOMPSON, 1971), NR (*Newton Raphson*) (SEARLE; CASELLA; MCCULLOCH, 1992) e combinações entre o EM e os demais algoritmos (EM-AI, EM-FS e EM-NR).

As principais estratégias utilizadas para a melhoria da eficiência computacional dos algoritmos em R foram a soma do produto de Hadamard para a obtenção de traços matriciais, uso de matrizes esparsas através do pacote Matrix (BATES; MAECHLER, 2016) e cálculo indireto dos traços das matrizes de informação.

A utilização do produto de Hadamard se fundamentou na possibilidade de reduzir esforço computacional ao evitar a operação algébrica convencional de grandes matrizes, onde as linhas são multiplicadas pelas colunas, para a obtenção dos traços matriciais. Esta

estratégia foi relatada por Misztal; Perez-Enciso (1993) para obtenção de traços matriciais do algoritmo EM.

A técnica de matrizes esparsas foi utilizada visando a economia de memória e maior agilidade nas operações matriciais envolvidas no ajuste dos modelos mistos (MISZTAL; PEREZ-ENCISO, 1993; JOHNSON; THOMPSON, 1995; THOMPSON et al., 2003). Esta técnica foi utilizada pois, geralmente a maioria das posições das matrizes de incidência e do SEMM são preenchidas por zero e somente os elementos diferentes de zero são armazenados.

A primeira e segunda derivadas da função de verossimilhança residual (4) envolvem operações matriciais computacionalmente complexas (GILMOUR; THOMPSON; CULLIS, 1995; JOHNSON; THOMPSON, 1995) e exigem a obtenção da matriz P (5) que possui dimensão N x N. Neste contexto, optou-se pela simplificação dos termos de primeira e segunda derivadas dos algoritmos visando a minimização do esforço computacional para estimar os componentes de variância e parâmetros genéticos.

O estudo da eficiência computacional dos algoritmos foi realizado no R versão 3.2.2 em quatro computadores listados na Tabela 1 (Anexo 1), para averiguar se as simplificações citadas anteriormente oportunizam o ajuste de modelos mistos complexos em computadores com configuração simples e qual a diferença em tempo para computadores com maior memória e capacidade de processamento de dados. O tempo médio de processamento dos algoritmos foi registrado pelo pacote microbenchmark (MERSMANN, 2016), sendo consideradas cinco repetições para a execução de cada algoritmo.

As funções implementadas em R para os algoritmos REML (EM, AI, FS, NR, EM-AI, EM-FS e EM-NR) serão disponibilizadas em breve na página do Programa de Pós-Graduação em Genética e Melhoramento da UFES (<https://sites.google.com/site/geneticaemelhorentoufes>) ou poderão ser solicitadas diretamente ao autor.

2.3.1 *Expectation Maximization (EM)*

A implementação do algoritmo EM considerou o SEMM apresentado em (3) e uma generalização das equações apresentadas por Mrode (2014), sendo estas obtidas através do rearranjo das expressões de primeira derivada de (4)

$$\frac{\partial \log(L)}{\partial \sigma_i^2} = \frac{1}{2} \left[\text{tr} \left(P \frac{\partial V}{\partial \sigma_i^2} \right) - y' P \frac{\partial V}{\partial \sigma_i^2} P y \right],$$

$$\frac{\partial \log(L)}{\partial \sigma_i^2} = \frac{1}{2} \begin{bmatrix} \text{tr}(P) - y'PPy \\ \text{tr}(PZ_1H_1Z_1') - y'PZ_1H_1Z_1'Py \\ \text{tr}(PZ_2H_2Z_2') - y'PZ_2H_2Z_2'Py \\ \vdots \\ \text{tr}(PZ_iH_iZ_i') - y'PZ_iH_iZ_i'Py \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \sum_{i=1, k=2}^{I, K} \frac{\text{tr}(H_i^{-1}C^{kk})}{\sigma_{u_i}^2} + \left(\frac{N - r(X) - \sum_{i=1}^I N_{u_i}}{\sigma_\varepsilon^2} \right) \frac{\tilde{\varepsilon}'\tilde{\varepsilon}}{(\sigma_\varepsilon^2)^2} \\ \frac{N_{u_1}}{\sigma_{u_1}^2} - \frac{\tilde{u}_1'H_1^{-1}\tilde{u}_1}{(\sigma_{u_1}^2)^2} - \frac{\text{tr}(H_1^{-1}C^{22})\sigma_\varepsilon^2}{(\sigma_{u_1}^2)^2} \\ \frac{N_{u_2}}{\sigma_{u_2}^2} - \frac{\tilde{u}_2'H_2^{-1}\tilde{u}_2}{(\sigma_{u_2}^2)^2} - \frac{\text{tr}(H_2^{-1}C^{33})\sigma_\varepsilon^2}{(\sigma_{u_2}^2)^2} \\ \vdots \\ \frac{N_{u_i}}{\sigma_{u_i}^2} - \frac{\tilde{u}_i'H_i^{-1}\tilde{u}_i}{(\sigma_{u_i}^2)^2} - \frac{\text{tr}(H_i^{-1}C^{kk})\sigma_\varepsilon^2}{(\sigma_{u_i}^2)^2} \end{bmatrix},$$

e podem ser escritas como

$$[N - r(X)] [\hat{\sigma}_\varepsilon^2] = \left[(y - X\hat{\beta} - \sum_{i=1}^I Z_i\tilde{u}_i)'y \right],$$

$$\begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_i \end{bmatrix} \begin{bmatrix} \hat{\sigma}_{u_1}^2 \\ \hat{\sigma}_{u_2}^2 \\ \vdots \\ \hat{\sigma}_{u_i}^2 \end{bmatrix} = \begin{bmatrix} \tilde{u}_1'H_1^{-1}\tilde{u}_1 + \text{tr}(H_1^{-1}C^{22})\hat{\sigma}_\varepsilon^2 \\ \tilde{u}_2'H_2^{-1}\tilde{u}_2 + \text{tr}(H_2^{-1}C^{33})\hat{\sigma}_\varepsilon^2 \\ \vdots \\ \tilde{u}_i'H_i^{-1}\tilde{u}_i + \text{tr}(H_i^{-1}C^{kk})\hat{\sigma}_\varepsilon^2 \end{bmatrix},$$

onde $r(X)$ é o *rank* da matriz X e $(C^{22}, C^{33} \dots C^{kk})$ são obtidas da inversa da matriz de coeficientes do SEMM (3)

$$C^{-1} = \begin{bmatrix} C^{11} & C^{12} & C^{13} & \dots & C^{1j} \\ C^{21} & C^{22} & C^{23} & \dots & C^{2j} \\ C^{31} & C^{32} & C^{33} & \dots & C^{3j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C^{i1} & C^{i2} & C^{i3} & \dots & C^{ij} \end{bmatrix}.$$

Com o objetivo de testar a eficiência computacional do algoritmo EM foram desenvolvidas duas versões deste algoritmo. A primeira versão realiza o cálculo dos traços matriciais por meio da soma do produto de Hadamard (MISZTAL; PEREZ-ENCISO, 1993)

$$\text{tr}(H_i^{-1}C^{kk}) = \sum H_i^{-1} \# C^{kk} = \sum_{i=1, j=1}^{I, J} h_{ij}c_{ij}.$$

Já a segunda versão realiza o cálculo dos traços matriciais da forma convencional. Entretanto, ambas as versões possibilitam o uso da técnica de matrizes esparsas durante o procedimento numérico.

2.3.2 Average information (AI)

O algoritmo AI em R, em sua forma convencional, foi fundamentado nas derivadas de primeira (6) e segunda ordem (7a e 7b) de (4) apresentadas por Johnson; Thompson (1995)

$$\frac{\partial \log(L)}{\partial \sigma_i^2} = \frac{1}{2} \left[\text{tr} \left(P \frac{\partial V}{\partial \sigma_i^2} \right) - y' P \frac{\partial V}{\partial \sigma_i^2} Py \right] = \frac{1}{2} \begin{bmatrix} \text{tr}(P) - y' PPy \\ \text{tr}(PZ_1H_1Z_1') - y' PZ_1H_1Z_1'Py \\ \text{tr}(PZ_2H_2Z_2') - y' PZ_2H_2Z_2'Py \\ \vdots \\ \text{tr}(PZ_iH_iZ_i') - y' PZ_iH_iZ_i'Py \end{bmatrix}, \quad (6)$$

$$AI = \frac{1}{2} \left[\frac{\partial^2 \log(L)}{\partial \sigma_i^2 \partial \sigma_j^2} + E \left(\frac{\partial^2 \log(L)}{\partial \sigma_i^2 \partial \sigma_j^2} \right) \right] = \frac{1}{2} y' P \frac{\partial V}{\partial \sigma_i^2} P \frac{\partial V}{\partial \sigma_j^2} Py \quad (7a)$$

$$AI = \frac{1}{2} \begin{bmatrix} y' PPy & y' PPZ_1H_1Z_1'Py & y' PPZ_2H_2Z_2'Py & \cdots & y' PPZ_jH_jZ_j'Py \\ y' PZ_1H_1Z_1'Py & y' PZ_1H_1Z_1'PZ_1H_1Z_1'Py & y' PZ_1H_1Z_1'PZ_2H_2Z_2'Py & \cdots & y' PZ_1H_1Z_1'PZ_jH_jZ_j'Py \\ y' PZ_2H_2Z_2'Py & y' PZ_2H_2Z_2'PZ_1H_1Z_1'Py & y' PZ_2H_2Z_2'PZ_2H_2Z_2'Py & \cdots & y' PZ_2H_2Z_2'PZ_jH_jZ_j'Py \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y' PZ_iH_iZ_i'Py & y' PZ_iH_iZ_i'PZ_1H_1Z_1'Py & y' PZ_iH_iZ_i'PZ_2H_2Z_2'Py & \cdots & y' PZ_iH_iZ_i'PZ_jH_jZ_j'Py \end{bmatrix}. \quad (7b)$$

Contudo, as operações matriciais apresentados em (6, 7a e 7b) exigem a obtenção da matriz P (5), que possui dimensão equivalente ao número de informações experimentais utilizadas (N x N), tal fato inflaciona o esforço computacional a medida que o número de observações experimentais aumenta.

Baseando-se no SEMM apresentado em (3), e utilizando as simplificações matemáticas das equações de primeira derivada (4) e variáveis de trabalho (8a e 8b) apresentadas por Johnson; Thompson (1995) podemos reescrever (6) na forma de (9)

$$f(\sigma_\epsilon^2) = Py = \frac{\tilde{\epsilon}}{\sigma_\epsilon^2}, \quad (8a)$$

$$f(\sigma_{u_i}^2) = Z_iH_iZ_i'Py = \frac{Z_i\tilde{u}_i}{\sigma_{u_i}^2}, \quad (8b)$$

$$\frac{\partial \log(L)}{\partial \sigma_i^2} = \frac{1}{2} \left[\begin{array}{c} \frac{N - r(X)}{\sigma_\varepsilon^2} - \frac{\sum_{i=1}^I N_{u_i} - \sum_{i=1, k=2}^{I, K} \text{tr}(H_i^{-1} C_\lambda^{kk})}{\sigma_\varepsilon^2} - \frac{\tilde{\varepsilon}' \tilde{\varepsilon}}{(\sigma_\varepsilon^2)^2} \\ \frac{N_{u_1}}{\sigma_{u_1}^2} - \frac{\text{tr}(H_1^{-1} C_\lambda^{22})}{\sigma_{u_1}^2} - \left(\frac{\tilde{\varepsilon}}{\sigma_\varepsilon^2} \right)' \left(\frac{Z_1 \tilde{u}_1}{\sigma_{u_1}^2} \right) \\ \frac{N_{u_2}}{\sigma_{u_2}^2} - \frac{\text{tr}(H_2^{-1} C_\lambda^{33})}{\sigma_{u_2}^2} - \left(\frac{\tilde{\varepsilon}}{\sigma_\varepsilon^2} \right)' \left(\frac{Z_2 \tilde{u}_2}{\sigma_{u_2}^2} \right) \\ \vdots \\ \frac{N_{u_i}}{\sigma_{u_i}^2} - \frac{\text{tr}(H_i^{-1} C_\lambda^{kk})}{\sigma_{u_i}^2} - \left(\frac{\tilde{\varepsilon}}{\sigma_\varepsilon^2} \right)' \left(\frac{Z_i \tilde{u}_i}{\sigma_{u_i}^2} \right) \end{array} \right], \quad (9)$$

onde $(C_\lambda^{22}, C_\lambda^{33}, \dots, C_\lambda^{kk})$ são submatrizes da matriz C_λ^{-1}

$$C_\lambda^{-1} = \begin{bmatrix} C_\lambda^{22} & C_\lambda^{23} & \dots & C_\lambda^{2j} \\ C_\lambda^{32} & C_\lambda^{33} & \dots & C_\lambda^{3j} \\ \vdots & \vdots & \ddots & \vdots \\ C_\lambda^{i2} & C_\lambda^{i3} & \dots & C_\lambda^{ij} \end{bmatrix} = \begin{bmatrix} \lambda_1 C^{22} & \lambda_1 C^{23} & \dots & \lambda_1 C^{2j} \\ \lambda_2 C^{32} & \lambda_2 C^{33} & \dots & \lambda_2 C^{3j} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_i C^{i1} & \lambda_i C^{i2} & \dots & \lambda_i C^{ij} \end{bmatrix} \quad (10)$$

$$= \begin{bmatrix} C^{22} & C^{23} & \dots & C^{2j} \\ C^{32} & C^{33} & \dots & C^{3j} \\ \vdots & \vdots & \ddots & \vdots \\ C^{i1} & C^{i2} & \dots & C^{ij} \end{bmatrix} \# \begin{bmatrix} \lambda_1 & \lambda_1 & \dots & \lambda_1 \\ \lambda_2 & \lambda_2 & \dots & \lambda_2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_i & \lambda_i & \dots & \lambda_i \end{bmatrix}.$$

Os traços matriciais apresentados em (9) foram obtidos por meio da soma do produto de Hadamard (11)

$$\text{tr}(H_i^{-1} C_\lambda^{kk}) = \sum H_i^{-1} \# C_\lambda^{kk} = \sum_{i=1, j=1}^{I, J} h_{ij} c_{ij}. \quad (11)$$

Os elementos da matriz de informação AI (7a e 7b) podem ser obtidos através das operações matriciais apresentadas em (8a, 8b, 12, 13, 14, 15, 16, 17, 18a, 18b, 19, 20, 21, 22 e 23), como se segue:

$$\begin{bmatrix} X'X & X'Z_1 & X'Z_2 & \cdots & X'Z_j \\ Z_1'X & Z_1'Z_1 + H_1^{-1}\lambda_1 & Z_1'Z_2 & \cdots & Z_1'Z_j \\ Z_2'X & Z_2'Z_1 & Z_2'Z_2 + H_2^{-1}\lambda_2 & \cdots & Z_2'Z_j \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_i'X & Z_i'Z_1 & Z_i'Z_2 & \cdots & Z_i'Z_j + H_i^{-1}\lambda_i \end{bmatrix} \begin{bmatrix} \hat{\delta} \\ \tilde{v}_1 \\ \tilde{v}_2 \\ \vdots \\ \tilde{v}_i \end{bmatrix} = \begin{bmatrix} X'f(\sigma_\varepsilon^2) \\ Z_1'f(\sigma_\varepsilon^2) \\ Z_2'f(\sigma_\varepsilon^2) \\ \vdots \\ Z_i'f(\sigma_\varepsilon^2) \end{bmatrix}, \quad (12)$$

$$\tilde{\varepsilon}_{(\delta,v)} = Py - X\hat{\delta} - \sum_{i=1}^I Z_i\tilde{v}_i, \quad (13)$$

$$PPy = \frac{\tilde{\varepsilon}_{(\delta,v)}}{\sigma_\varepsilon^2}, \quad (14)$$

$$y'PPPy = \left(\frac{\tilde{\varepsilon}}{\sigma_\varepsilon^2} \right)' \left(\frac{\tilde{\varepsilon}_{(\delta,v)}}{\sigma_\varepsilon^2} \right), \quad (15)$$

$$\begin{bmatrix} X'X & X'Z_1 & X'Z_2 & \cdots & X'Z_j \\ Z_1'X & Z_1'Z_1 + H_1^{-1}\lambda_1 & Z_1'Z_2 & \cdots & Z_1'Z_j \\ Z_2'X & Z_2'Z_1 & Z_2'Z_2 + H_2^{-1}\lambda_2 & \cdots & Z_2'Z_j \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_i'X & Z_i'Z_1 & Z_i'Z_2 & \cdots & Z_i'Z_j + H_i^{-1}\lambda_i \end{bmatrix} \begin{bmatrix} \hat{\phi} \\ \tilde{\tau}_1 \\ \tilde{\tau}_2 \\ \vdots \\ \tilde{\tau}_i \end{bmatrix} = \begin{bmatrix} \text{RHS}_1 \\ \text{RHS}_2 \end{bmatrix}, \quad (16)$$

$$\hat{\phi} = [\hat{\phi}_{11} \quad \hat{\phi}_{12} \quad \cdots \quad \hat{\phi}_{1j}], \quad \tilde{\tau} = \begin{bmatrix} \tilde{\tau}_1 \\ \tilde{\tau}_2 \\ \vdots \\ \tilde{\tau}_i \end{bmatrix} = \begin{bmatrix} \tilde{\tau}_{11} & \tilde{\tau}_{12} & \cdots & \tilde{\tau}_{1j} \\ \tilde{\tau}_{21} & \tilde{\tau}_{22} & \cdots & \tilde{\tau}_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\tau}_{i1} & \tilde{\tau}_{i2} & \cdots & \tilde{\tau}_{ij} \end{bmatrix}, \quad (17)$$

$$\begin{aligned} \text{RHS}_1 &= X'[Z_1H_1Z_1'Py \quad Z_2H_2Z_2'Py \quad \cdots \quad Z_iH_iZ_i'Py] \\ &= X' \left[\frac{1}{\sigma_{u_1}^2} Z_1\tilde{u}_1 \quad \frac{1}{\sigma_{u_2}^2} Z_2\tilde{u}_2 \quad \cdots \quad \frac{1}{\sigma_{u_i}^2} Z_i\tilde{u}_i \right], \end{aligned} \quad (18a)$$

$$\begin{aligned} \text{RHS}_2 &= Z' [Z_1 H_1 Z_1' P y \quad Z_2 H_2 Z_2' P y \quad \cdots \quad Z_i H_i Z_i' P y] \\ &= Z' \left[\frac{1}{\sigma_{u_1}^2} Z_1 \tilde{u}_1 \quad \frac{1}{\sigma_{u_2}^2} Z_2 \tilde{u}_2 \quad \cdots \quad \frac{1}{\sigma_{u_i}^2} Z_i \tilde{u}_i \right], \end{aligned} \quad (18b)$$

$$y(\varphi, \tau) = \left[\frac{1}{\sigma_{u_1}^2} Z_1 \tilde{u}_1 \quad \frac{1}{\sigma_{u_2}^2} Z_2 \tilde{u}_2 \quad \cdots \quad \frac{1}{\sigma_{u_i}^2} Z_i \tilde{u}_i \right], \quad (19)$$

$$\left[y' P P Z_1 H_1 Z_1' P y \quad y' P P Z_2 H_2 Z_2' P y \quad \cdots \quad y' P P Z_j H_j Z_j' P y \right] = y(\varphi, \tau)' \left(\frac{\tilde{\varepsilon}_{(\delta, y)}}{\sigma_\varepsilon^2} \right), \quad (20)$$

$$\tilde{\varepsilon}_{(\varphi, \tau)} = y(\varphi, \tau) - X\hat{\varphi} - Z\tilde{\tau}, \quad (21)$$

$$\left[P Z_1 H_1 Z_1' P y \quad P Z_2 H_2 Z_2' P y \quad \cdots \quad P Z_i H_i Z_i' P y \right] = \frac{\tilde{\varepsilon}_{(\varphi, \tau)}}{\sigma_\varepsilon^2}, \quad (22)$$

$$\begin{bmatrix} y' P Z_1 H_1 Z_1' P Z_1 H_1 Z_1' P y & y' P Z_1 H_1 Z_1' P Z_2 H_2 Z_2' P y & \cdots & y' P Z_1 H_1 Z_1' P Z_j H_j Z_j' P y \\ y' P Z_2 H_2 Z_2' P Z_1 H_1 Z_1' P y & y' P Z_2 H_2 Z_2' P Z_2 H_2 Z_2' P y & \cdots & y' P Z_2 H_2 Z_2' P Z_j H_j Z_j' P y \\ \vdots & \vdots & \ddots & \vdots \\ y' P Z_i H_i Z_i' P Z_1 H_1 Z_1' P y & y' P Z_i H_i Z_i' P Z_2 H_2 Z_2' P y & \cdots & y' P Z_i H_i Z_i' P Z_j H_j Z_j' P y \end{bmatrix} = y(\varphi, \tau)' \left(\frac{\tilde{\varepsilon}_{(\varphi, \tau)}}{\sigma_\varepsilon^2} \right). \quad (23)$$

Após a obtenção dos termos das derivadas primeira (9) e segunda (15, 20, 23) de (4) pode-se utilizar a equação iterativa mostrada em (24) para obter as estimativas dos componentes de variância (JOHNSON; THOMPSON, 1995)

$$\theta^{[i+1]} = \theta^{[i]} - (A I^{[i]})^{-1} \frac{\partial \log(L)}{\partial \theta} \Big|_{\theta^{[i]}}, \quad (24)$$

onde o vetor θ concatena os componentes de variância do modelo (1a).

Tanto a versão direta como a otimizada possibilitam a utilização da técnica de matrizes esparsas durante o procedimento numérico.

2.3.3 Fisher's Scoring (FS)

A versão do algoritmo FS, em R, onde a matriz de informação é obtida de forma direta, foi fundamentada nas primeiras derivadas de (4) mostradas em (6) e na esperança matemática das segundas derivadas (25a e 25b) de (4) apresentadas por Searle; Casella; McCulloch (1992) e Johnson; Thompson (1995)

$$FS = E \left(\frac{\partial^2 \log(L)}{\partial \sigma_i^2 \partial \sigma_j^2} \right) = \frac{1}{2} \text{tr} \left(P \frac{\partial V}{\partial \sigma_i^2} P \frac{\partial V}{\partial \sigma_j^2} \right), \quad (25a)$$

$$FS = \frac{1}{2} \begin{bmatrix} \text{tr}(PP) & \text{tr}(PPZ_1H_1Z_1') & \text{tr}(PPZ_2H_2Z_2') & \cdots & \text{tr}(PPZ_jH_jZ_j') \\ \text{tr}(PZ_1H_1Z_1'P) & \text{tr}(PZ_1H_1Z_1'PZ_1H_1Z_1') & \text{tr}(PZ_1H_1Z_1'PZ_2H_2Z_2') & \cdots & \text{tr}(PZ_1H_1Z_1'PZ_jH_jZ_j') \\ \text{tr}(PZ_2H_2Z_2'P) & \text{tr}(PZ_2H_2Z_2'PZ_1H_1Z_1') & \text{tr}(PZ_2H_2Z_2'PZ_2H_2Z_2') & \cdots & \text{tr}(PZ_2H_2Z_2'PZ_jH_jZ_j') \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{tr}(PZ_jH_jZ_j'P) & \text{tr}(PZ_jH_jZ_j'PZ_1H_1Z_1') & \text{tr}(PZ_jH_jZ_j'PZ_2H_2Z_2') & \cdots & \text{tr}(PZ_jH_jZ_j'PZ_jH_jZ_j') \end{bmatrix}. \quad (25b)$$

Contudo, os traços matriciais apresentados em (25a e 25b) são computacionalmente complexos (GILMOUR; THOMPSON; CULLIS, 1995; JOHNSON; THOMPSON, 1995). Já as primeiras derivadas de (4), apresentadas em (6), necessitam da matriz P (5), que possui dimensão equivalente ao número de informações experimentais utilizadas (N x N), tal fato inflaciona o esforço computacional à medida que o número de observações experimentais aumenta.

Devido à complexidade para a obtenção dos termos de primeira derivada de (4), as simplificações mostradas em (9 e 11) foram utilizadas no desenvolvimento do algoritmo FS em R. Para simplificar os termos da matriz de informação FS nos baseamos nas simplificações sugeridas por Searle; Casella; Mcculloch (1992), utilizando a expressão (10) obtida do SEMM (3).

Através das operações matriciais apresentadas em (26, 27, 28, 29, 30, 31 e 32) podemos obter a matriz de informação FS, como se segue:

$$\begin{bmatrix} \text{tr}(\text{PP}) \\ \text{tr}(\text{PZ}_1\text{H}_1\text{Z}_1' \text{PZ}_1\text{H}_1\text{Z}_1') \\ \text{tr}(\text{PZ}_2\text{H}_2\text{Z}_2' \text{PZ}_2\text{H}_2\text{Z}_2') \\ \vdots \\ \text{tr}(\text{PZ}_i\text{H}_i\text{Z}_i' \text{PZ}_j\text{H}_j\text{Z}_j') \end{bmatrix} = \begin{bmatrix} \frac{N - r(\mathbf{X})}{(\sigma_\varepsilon^2)^2} - \frac{\sum_{i=1}^I N_{u_i} - \text{tr}(\mathbf{H}^{-1}\mathbf{C}_\lambda^{-1}\mathbf{H}^{-1}\mathbf{C}_\lambda^{-1})}{(\sigma_\varepsilon^2)^2} \\ \frac{N_{u_1}}{(\sigma_{u_1}^2)^2} - \frac{2\text{tr}(\mathbf{H}_1^{-1}\mathbf{C}_\lambda^{22})}{(\sigma_{u_1}^2)^2} + \frac{\text{tr}(\mathbf{H}_1^{-1}\mathbf{C}_\lambda^{22}\mathbf{H}_1^{-1}\mathbf{C}_\lambda^{22})}{(\sigma_{u_1}^2)^2} \\ \frac{N_{u_2}}{(\sigma_{u_2}^2)^2} - \frac{2\text{tr}(\mathbf{H}_2^{-1}\mathbf{C}_\lambda^{33})}{(\sigma_{u_2}^2)^2} + \frac{\text{tr}(\mathbf{H}_2^{-1}\mathbf{C}_\lambda^{33}\mathbf{H}_2^{-1}\mathbf{C}_\lambda^{33})}{(\sigma_{u_2}^2)^2} \\ \vdots \\ \frac{N_{u_i}}{(\sigma_{u_i}^2)^2} - \frac{2\text{tr}(\mathbf{H}_i^{-1}\mathbf{C}_\lambda^{\text{kk}})}{(\sigma_{u_i}^2)^2} + \frac{\text{tr}(\mathbf{H}_i^{-1}\mathbf{C}_\lambda^{\text{kk}}\mathbf{H}_i^{-1}\mathbf{C}_\lambda^{\text{kk}})}{(\sigma_{u_i}^2)^2} \end{bmatrix} \text{ para } i = j, \quad (26)$$

$$\begin{bmatrix} \text{tr}(\text{PZ}_1\text{H}_1\text{Z}_1' \text{P}) \\ \text{tr}(\text{PZ}_2\text{H}_2\text{Z}_2' \text{P}) \\ \vdots \\ \text{tr}(\text{PZ}_i\text{H}_i\text{Z}_i' \text{P}) \end{bmatrix} = \begin{bmatrix} \frac{\text{tr}(\mathbf{H}_1^{-1}\mathbf{C}_\lambda^{22}) - \text{tr}[(\mathbf{H}^{-1}\mathbf{C}_\lambda^{-1}\mathbf{H}^{-1}\mathbf{C}_\lambda^{-1})^{22}]}{\sigma_\varepsilon^2 \sigma_{u_1}^2} \\ \frac{\text{tr}(\mathbf{H}_2^{-1}\mathbf{C}_\lambda^{33}) - \text{tr}[(\mathbf{H}^{-1}\mathbf{C}_\lambda^{-1}\mathbf{H}^{-1}\mathbf{C}_\lambda^{-1})^{33}]}{\sigma_\varepsilon^2 \sigma_{u_2}^2} \\ \vdots \\ \frac{\text{tr}(\mathbf{H}_i^{-1}\mathbf{C}_\lambda^{\text{kk}}) - \text{tr}[(\mathbf{H}^{-1}\mathbf{C}_\lambda^{-1}\mathbf{H}^{-1}\mathbf{C}_\lambda^{-1})^{\text{kk}}]}{\sigma_\varepsilon^2 \sigma_{u_i}^2} \end{bmatrix}, \quad (27)$$

$$\text{tr}(\text{PZ}_i\text{H}_i\text{Z}_i' \text{PZ}_j\text{H}_j\text{Z}_j') = \frac{\text{tr}(\mathbf{C}_\lambda^{\text{ij}}\mathbf{C}_\lambda^{\text{ji}})}{\sigma_{u_i}^2 \sigma_{u_j}^2} \text{ para } i \neq j, \quad (28)$$

$$\mathbf{H}^{-1} = \bigoplus_{i=1}^I \mathbf{H}_i^{-1}, \quad (29)$$

$$\text{tr}(\mathbf{H}_i^{-1}\mathbf{C}_\lambda^{\text{kk}}\mathbf{H}_i^{-1}\mathbf{C}_\lambda^{\text{kk}}) = \sum (\mathbf{H}_i^{-1}\mathbf{C}_\lambda^{\text{kk}})\#(\mathbf{H}_i^{-1}\mathbf{C}_\lambda^{\text{kk}}), \quad (30)$$

$$\text{tr}(\mathbf{H}^{-1}\mathbf{C}_\lambda^{-1}\mathbf{H}^{-1}\mathbf{C}_\lambda^{-1}) = \sum (\mathbf{H}^{-1}\mathbf{C}_\lambda^{-1})\#(\mathbf{H}^{-1}\mathbf{C}_\lambda^{-1}), \quad (31)$$

$$\text{tr}(\mathbf{H}_i^{-1}\mathbf{C}_\lambda^{\text{kk}}) = \sum \mathbf{H}_i^{-1}\#\mathbf{C}_\lambda^{\text{kk}}, \quad \text{tr}(\mathbf{H}^{-1}\mathbf{C}_\lambda^{-1}) = \sum \mathbf{H}^{-1}\#\mathbf{C}_\lambda^{-1} \quad \text{e} \quad \text{tr}(\mathbf{C}_\lambda^{\text{ij}}\mathbf{C}_\lambda^{\text{ji}}) = \sum \mathbf{C}_\lambda^{\text{ij}}\#\mathbf{C}_\lambda^{\text{ji}}, \quad (32)$$

Após a obtenção dos termos das derivadas primeira (9) e segunda (26, 27 e 28) de (4) podemos utilizar a equação iterativa mostrada em (33) (HAN; XU, 2008; THOMPSON, 2008) para obter as estimativas dos componentes de variância:

$$\theta^{[i+1]} = \theta^{[i]} - (FS^{[i]})^{-1} \frac{\partial \log(L)}{\partial \theta} \Big|_{\theta^{[i]}}, \quad (33)$$

onde o vetor θ concatena os componentes de variância do modelo (1a).

A duas versões do algoritmo FS utilizam a técnica de matrizes esparsas durante o procedimento numérico.

2.3.4 *Newton Raphson (NR)*

O algoritmo NR em R, em sua forma convencional foi fundamentado nas derivadas primeira e segunda (6, 34a e 34b) de (4), apresentadas por Johnson; Thompson (1995):

$$NR = \frac{\partial^2 \log(L)}{\partial \sigma_i^2 \partial \sigma_j^2} = \frac{1}{2} \left[-\text{tr} \left(P \frac{\partial V}{\partial \sigma_i^2} P \frac{\partial V}{\partial \sigma_j^2} \right) + 2y' P \frac{\partial V}{\partial \sigma_i^2} P \frac{\partial V}{\partial \sigma_j^2} P y \right], \quad (34a)$$

$$NR = \frac{\partial^2 \log(L)}{\partial \sigma_i^2 \partial \sigma_j^2} = \frac{1}{2} [-FS + 2AI]. \quad (34b)$$

Entretanto, os traços matriciais apresentados em (34a e 34b) são computacionalmente complexos (GILMOUR; THOMPSON; CULLIS, 1995; JOHNSON; THOMPSON, 1995). Já as primeiras derivadas de (4) apresentadas em (6) necessitam da matriz P (5), que devido a sua dimensão (N x N) aumenta o esforço computacional à medida que o número de observações experimentais aumenta.

Devido à complexidade para a obtenção dos termos de primeira derivada de (4) as simplificações apresentadas em (9 e 11) foram utilizadas no desenvolvimento do algoritmo NR em R. Para simplificar os termos da matriz de informação NR nos baseamos nas simplificações das matrizes AI e FS utilizando a expressão (10) obida do SEMM (3).

Após a obtenção dos termos das derivadas primeira (9) e segunda de (4) podemos utilizar a equação iterativa mostrada em (35) (MEYER; SMITH, 1996) para obter as estimativas dos componentes de variância:

$$\theta^{[i+1]} = \theta^{[i]} - (NR^{[i]})^{-1} \frac{\partial \log(L)}{\partial \theta} \Big|_{\theta^{[i]}}, \quad (35)$$

onde o vetor θ concatena os componentes de variância do modelo (1a).

A duas versões do algoritmo NR utilizam a técnica de matrizes esparsas durante o procedimento numérico.

2.3.5 Algoritmos Combinados (EM-AI, EM-FS e EM-NR)

Os algoritmos combinados foram construídos com a versão otimizada dos algoritmos EM, AI, FS e NR. A realização das análises considerou uma iteração EM e as demais realizadas pelos algoritmos AI, FS e NR respectivamente.

2.3.6 Valores iniciais dos componentes de variância

Os valores iniciais dos componentes de variância utilizados no algoritmo EM foram fixados para que os coeficientes de determinação dos efeitos aleatórios do modelo (1a) se igualassem a 0,1 seguindo o padrão do Selegen (RESENDE, 2007), caso o argumento *var* não seja declarado.

Para o algoritmo AI os valores iniciais dos componentes de variância foram 0,1⁰ enquanto a variação ambiental foi representada por ⁰, sendo ⁰ metade da variação do vetor *y* seguindo o padrão do ASReml (BUTLER et al., 2009), caso o argumento *var* não seja declarado.

Os valores iniciais dos componentes de variância dos algoritmos FS e NR foram 0,1⁰ enquanto a variação ambiental foi representada por 0,4⁰ caso o argumento *var* não seja declarado.

Os algoritmos combinados seguem o padrão do Selegen e após uma iteração EM as estimativas dos componentes de variância são utilizadas na inicialização dos algoritmos AI, FS e NR, caso o argumento *var* não seja declarado.

2.3.7 Diagnóstico de convergência

A convergência dos algoritmos foi monitorada através dos coeficientes de determinação (CD) (RESENDE, 2007) e componentes de variância (CV) (SEARLE; CASELLA; MCCULLOCH, 1992).

Os CD podem ser obtidos pela expressão (36)

$$cd_1 = \frac{\hat{\sigma}_{u_i}^2}{\hat{\sigma}_{u_1}^2 + \hat{\sigma}_{u_2}^2 \cdots + \hat{\sigma}_{u_i}^2 + \hat{\sigma}_{\varepsilon}^2}, \quad (36)$$

onde $(\hat{\sigma}_{u_1}^2, \hat{\sigma}_{u_2}^2, \dots, \hat{\sigma}_{u_i}^2, \hat{\sigma}_{\varepsilon}^2)$ podem ser obtidos dos algoritmos (EM, AI, FS, NR, EM-AI, EM-FS e EM-NR), descritos anteriormente.

Para detectar o momento da convergência via CD o vetor de estimativas dos CD da iteração $i + 1$ (ω^{i+1}) deve ser comparado com o vetor da iteração i (ω^i) e o maior valor da diferença absoluta ($\Delta \omega$) deve ser inferior a taxa de erro (TE) estipulada (ϕ):

$$\begin{cases} \Delta \omega = \|\omega^{i+1} - \omega^i\| \\ \max(\Delta \omega) < \phi \end{cases}, \omega = \begin{bmatrix} cd_1 \\ cd_2 \\ \vdots \\ cd_i \end{bmatrix}. \quad (37)$$

O procedimento utilizado para detectar o momento da convergência via CV (38) é idêntico ao descrito anteriormente para CD, diferindo apenas com relação ao vetor de estimativas (ω por κ):

$$\begin{cases} \Delta \kappa = \|\kappa^{i+1} - \kappa^i\| \\ \max(\Delta \kappa) < \phi \end{cases}, \kappa = \begin{bmatrix} \hat{\sigma}_{\varepsilon}^2 \\ \hat{\sigma}_{u_1}^2 \\ \hat{\sigma}_{u_2}^2 \\ \vdots \\ \hat{\sigma}_{u_i}^2 \end{bmatrix}. \quad (38)$$

As taxas de erro (ϕ) testadas neste trabalho foram 10^{-5} , 10^{-6} , 10^{-7} , 10^{-8} e 10^{-9} .

Para evitar *loops* infinitos estipulou-se que se o número de iterações para a taxa de erro (TE) $> 10^{-5}$ excedesse cinco vezes o número de iterações da TE imediatamente inferior o procedimento numérico seria interrompido.

No caso de modelos onde o único efeito aleatório é o erro experimental, por padrão os algoritmos realizam apenas uma iteração.

2.4 MATERIAL GENÉTICO ESTUDADO

Os dados experimentais utilizados foram obtidos na Embrapa Milho e Sorgo, situada em Sete Lagoas - MG. Os ensaios foram conduzidos no delineamento de blocos aumentados (grupo de experimentos), com cinco testemunhas comuns a todos os blocos (Experimentos) e 3352 híbridos simples, totalizando 3357 tratamentos genéticos. As parcelas foram compostas

de duas linhas de 4,2 m espaçadas em 0,7 m. Este conjunto tratamentos representou parte dos híbridos experimentais (topcrosses), de primeiro ciclo de seleção, do ano agrícola de 2013/14. Os híbridos topcrosses utilizados neste estudo foram obtidos a partir de cruzamentos entre progênies parcialmente endogâmicas, em S1, com testadores dos grupos heteróticos Flint e Dent, de forma complementar, de acordo com a classificação prévia dos grupos heteróticos das populações fonte. Assim, progênies S1 do grupo heterótico Dent foram cruzadas com o testador Flint e progênies Flint foram cruzadas com o testador Dent. A característica analisada foi o rendimento de grãos, em kg ha⁻¹, corrigido para a umidade de 13%, determinado com base na pesagem dos grãos de cada parcela experimental.

2.5 MODELOS ESTATÍSTICOS ESTUDADOS

Para o estudo da eficiência computacional foram utilizados dois modelos estatísticos. O primeiro se caracteriza pela presença da matriz de parentesco genético aditivo (M1 com *pedigree*):

$$y = Xu + Za + Tp.t + \varepsilon, \quad (39)$$

onde y : é o vetor de dados de produção, u : é a média geral (de efeito fixo), a : é o vetor de efeitos genéticos aditivos (aleatórios) com $a \sim N(0, A\sigma_a^2)$, $p.t$: é o vetor de efeitos da combinação população testador ou efeito de famílias (aleatórios) com $p.t \sim N(0, I\sigma_{p.t}^2)$ e ε : é o vetor de erros aleatórios com $\varepsilon \sim N(0, I\sigma^2)$. Sendo X , Z e T são as matrizes de incidência para os efeitos u , a e $p.t$ respectivamente.

A matriz de parentesco genético aditivo foi obtida através do pacote *nadiv* (WOLAK, 2012), a partir de anotações de *pedigree* das progênies extraídas das diversas populações fonte. Para o computo da inversa da matriz de parentesco aditivo necessário no sistema de equações de modelos mistos o pacote *nadiv* utiliza o método de Meuwissen e Luo (1992). A inversa da matriz de parentesco genético aditivo obtida através do pacote *nadiv* foi inserida nos algoritmos REML implementados em R associando-se a mesma a um termo do modelo.

O segundo modelo se caracteriza pela ausência de matrizes de parentesco (M2 sem *pedigree*) sendo o efeito genético predito somente com base nas informações experimentais:

$$y = Xu + Zg + Wb + \varepsilon, \quad (40)$$

onde y : é o vetor de dados de produção, u : é a média geral (de efeito fixo), g : é o vetor de efeitos genéticos (aleatórios) com $g \sim N(0, I^{\sigma_g^2})$, b : é o vetor de efeitos de blocos (aleatórios) com $b \sim N(0, I^{\sigma_b^2})$ e ε : é o vetor de erros aleatórios com $\varepsilon \sim N(0, I^{\sigma^2})$. Sendo X , Z e T são as matrizes de incidência para os efeitos u , g e b respectivamente.

O modelo M1 (com *pedigree*) apresentou um SEMM de 4237 equações, enquanto que o M2 (sem *pedigree*) foi representado por um SEMM com 3416 equações. Para testar a coerência dos algoritmos propostos na maximização da função de verossimilhança residual os modelos também foram ajustados nos pacotes ASReml (BUTLER et al., 2009), lme4 (BATES et al., 2015) e no programa Selegen (RESENDE, 2007).

3. RESULTADOS

3.1 ESTIMAÇÃO DE PARÂMETROS

Os algoritmos FS e sua combinação com o EM (EM-FS) não convergiram para os dois modelos utilizados. Não obstante, os demais algoritmos foram bem sucedidos no ajuste dos modelos M1 (com *pedigree*) e M2 (sem *pedigree*).

O modelo M1 (com *pedigree*) foi maximizado na mesma região de verossimilhança para todas as combinações entre os algoritmos, taxas de erro (TE) e critérios de convergência (CD e CV) com $\log(L)$ apresentando amplitude de -26192,41 a -26192,30 na Tabela 1 do Anexo 2, corroborando com $\log(L)$ estimado pelo *software* ASReml (-26192,30), apresentado na Tabela 1. Além disso, tomando-se as médias das estimativas dos componentes de variância $\hat{\sigma}_a^2$ (2541292,85), $\hat{\sigma}_{t,p}^2$ (266797,82) e $\hat{\sigma}^2$ (700678,82) obtidas dos algoritmos propostos para o *software* R, nos diferentes critérios de convergência e TE, em contraste às estimativas obtidas pelo ASReml, foram verificadas diferenças absolutas ínfimas, a ordem de 0,17%, 0,12% e 0,12%, respectivamente (Tabela 1).

As estimativas de $\log(L)$ do modelo M2 (sem *pedigree*) (Tabela 1) apresentaram divergência acentuada entre o *software* lme4 (-32349,14) com *softwares* ASReml (-29003,28) e Selegen (-29003,59), devido à consideração da constante $\{0,5[N-p(X)]\ln(2\pi) = 3345,86\}$ da função de verossimilhança residual neste dois últimos softwares. Somando esta constante a estimativa de $\log(L)$ do lme4 eliminamos a disparidade entre os *softwares* e a nova estimativa de $\log(L)$ do lme4 se torna -29003,28. Estes resultados corroboram com aqueles obtidos na Tabela 2 do Anexo 2 para todas as combinações entre os algoritmos, taxas de erro (TE) e critérios de convergência (CD e CV), onde os valores de $\log(L)$ apresentaram amplitude de -29003,59 a -29003,26, atestando que a maximização ocorreu em uma mesma região de verossimilhança.

As diferenças absolutas entre as médias dos componentes de variância $\hat{\sigma}_g^2$ (393585,09), $\hat{\sigma}_b^2$ (3500349,38) e $\hat{\sigma}^2$ (2475892,19), obtidas com algoritmos desenvolvidos para o *software* R, para o modelo M2 (sem *pedigree*), mostradas na Tabela 1, e os componentes de variância estimados pelos diferentes *softwares* (Tabela 1) foram da ordem de 0,12%, 0,03% e 0,02% para o ASReml, 0,33%, 0,01% e 0,05% para o Selegen e 0,01%, $2 \times 10^{-40}\%$ e $1 \times 10^{-30}\%$ para o lme4.

Tabela 1: Estimativas dos componentes de variância e logaritmo de REML [$\log(L)$] dos modelos M1 (com *pedigree*) e M2 (sem *pedigree*) nos *softwares* ASReml, Selegen, lme4, média aritmética (MA) das estimativas obtidas através dos algoritmos propostos implementados em R e diferenças percentuais dos *softwares* ASReml, Selegen e lme4 em relação a MA [$\Delta 1(\%)$, $\Delta 2(\%)$ e $\Delta 3(\%)$].

Mod.	Par.	ASReml	Selegen	lme4	MA [†]	$\Delta 1(\%)$	$\Delta 2(\%)$	$\Delta 3(\%)$
M1	$\log(L)$	-26192,30	---	---	-26192,31	4×10^{-5}	---	---
	$\hat{\sigma}_a^2$	2537065,09	---	---	2541292,85	0,17	---	---
	$\hat{\sigma}_{t,p}^2$	266469,49	---	---	266797,82	0,12	---	---
	$\hat{\sigma}^2$	699827,14	---	---	700678,82	0,12	---	---
M2	$\log(L)$	-29003,28	-29003,59	-32349,14	-29003,31	1×10^{-4}	1×10^{-3}	11,54
	$\hat{\sigma}_g^2$	393098,00	392283,30	393613,78	393585,09	0,12	0,33	0,01
	$\hat{\sigma}_b^2$	3499260,94	3500593,60	3500341,31	3500349,38	0,03	0,01	2×10^{-4}
	$\hat{\sigma}^2$	2475463,68	2477152,40	2475864,37	2475892,19	0,02	0,05	1×10^{-3}

$\hat{\sigma}_a^2$: variância genética aditiva, $\hat{\sigma}_{t,p}^2$: variância da interação testador x população, $\hat{\sigma}_g^2$: variância genética, $\hat{\sigma}_b^2$: variância dos blocos dentro de grupos de ensaios, $\hat{\sigma}^2$: variância ambiental, $\Delta 1(\%)$: diferença percentual entre o ASReml e MA, $\Delta 2(\%)$: diferença percentual entre o Selegen e MA e $\Delta 3(\%)$: diferença percentual entre o lme4 e MA.

[†]Média aritmética das estimativas obtidas para os parâmetros dos modelos M1 (com *pedigree*) e M2 (sem *pedigree*), com algoritmos desenvolvidos para o software R, através dos resultados mostrados Tabela 1 e 2 do Anexo 2, considerando os algoritmos (AI, NR, EM, EM-AI e EM-NR) nas diferentes taxas de erro de convergência (10^{-5} , 10^{-6} , 10^{-7} , 10^{-8} e 10^{-9}) e critérios de convergência (CD e CV).

As diferenças existentes entre as estimativas dos componentes de variância, obtidas com algoritmos desenvolvidos para o software R, para os critérios de convergência CD e CV, nas diferentes TE, foram sempre menores que 1%, podendo ser consideradas de baixa magnitude, em ambos os modelos. Desta forma, observou-se que as maiores diferenças absolutas encontradas nos modelos M1 (com *pedigree*) ($\hat{\sigma}_g^2$) e M2 (sem *pedigree*) ($\hat{\sigma}_g^2$) assumiram valores de 0,85% e 0,34%, respectivamente, e foram atribuídas ao algoritmo EM na TE de 10^{-5} (Figura 1).

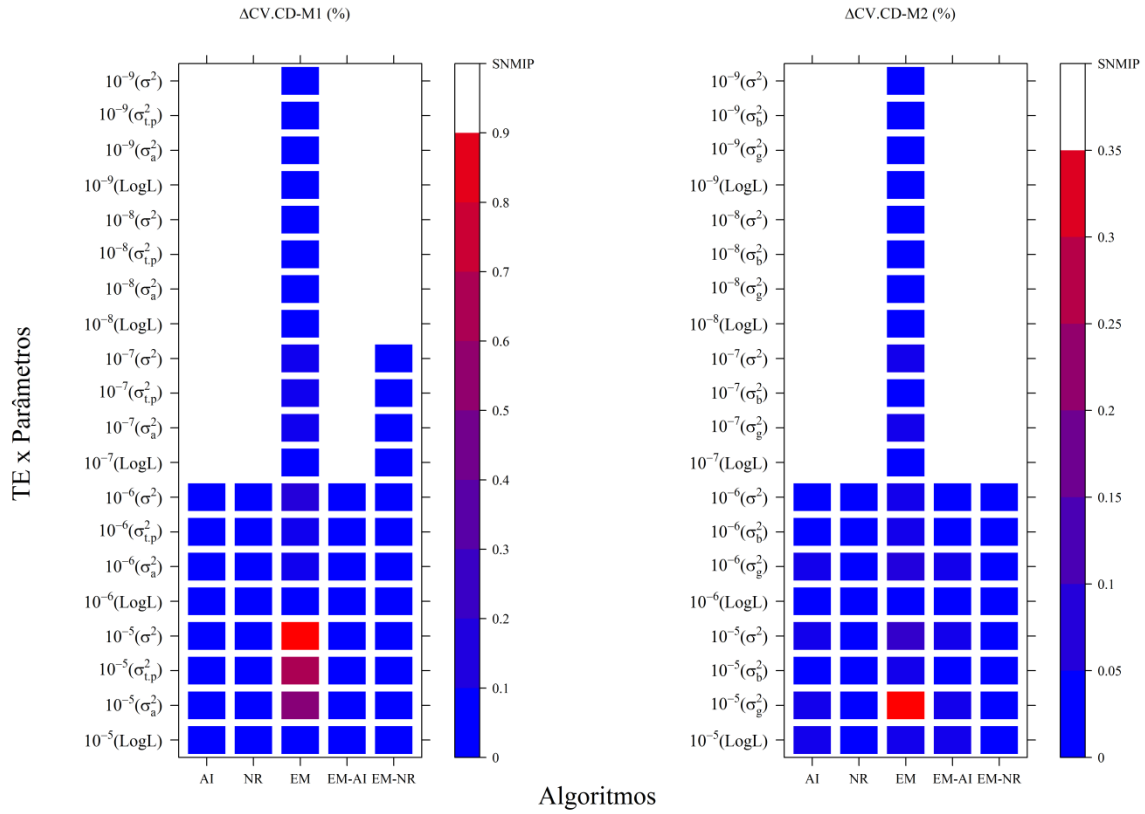


Figura 1: Diferença percentual entre os critérios de convergência (componentes de variância-CV versus coeficientes de determinação-CD) para as estimativas dos parâmetros dos modelos M1 (com *pedigree*) ($\Delta CV.CD-M1$) e M2 (sem *pedigree*) ($\Delta CV.CD-M2$) (Tabelas 1 e 2 - Anexo 2) obtidas através dos algoritmos *Average Information* (AI), *Newton Rhapson* (NR), *Expectation Maximization* (EM) e os algoritmos combinados EM-AI e EM-NR para diferentes taxas de erro (TE), com algoritmos desenvolvidos para o software R. SNMIP: superou o número máximo de iterações permitido.

Face ao exposto, comparando a TE de 10^{-5} com as TE 10^{-6} , 10^{-7} , 10^{-8} e 10^{-9} , quanto a diferença absoluta dos componentes de variância dentro dos critérios de convergência (CD e CV) e modelos (M1 com *pedigree* e M2 sem *pedigree*), foram observados valores ínfimos (menores que 1%) para CD (Figura 2) e nulo ou inexistente para CV (Tabelas 1 e 2 - Anexo 2) em ambos os modelos. As maiores diferenças observadas no critério de convergência CD foram para os parâmetros $\hat{\sigma}^2$ (M1 com *pedigree*) e $\hat{\sigma}_g^2$ (M2 sem *pedigree*), sendo ambas relativas ao algoritmo EM (Figura 2).

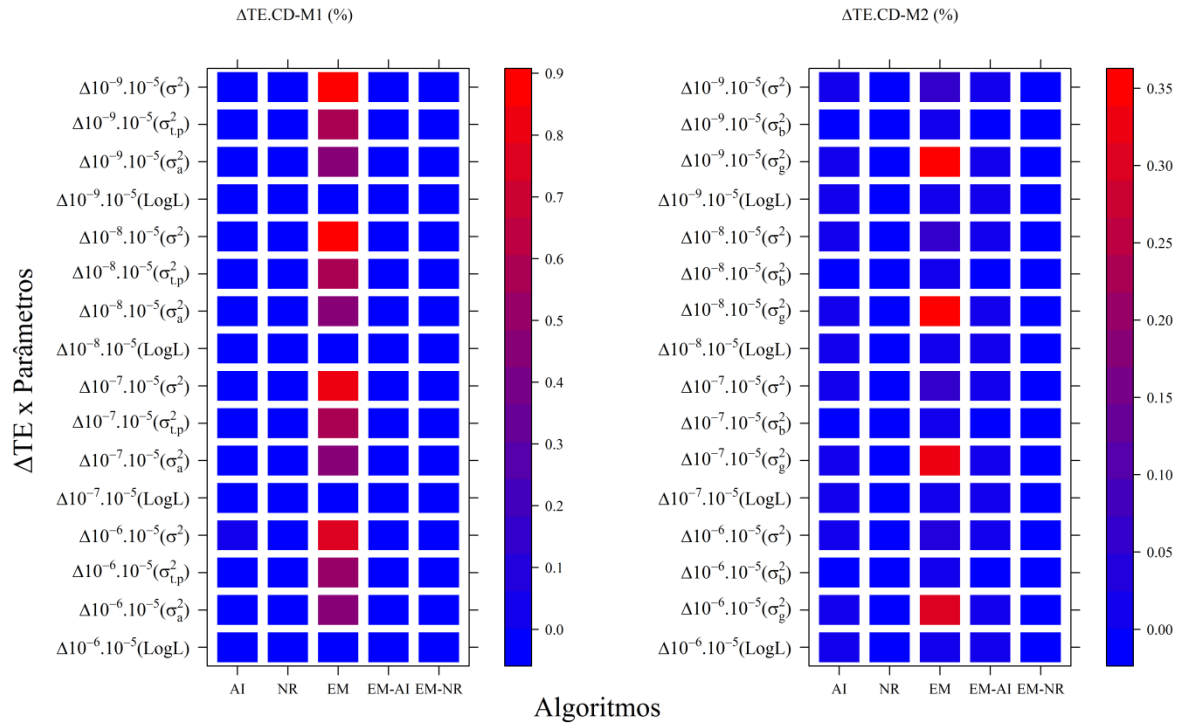


Figura 2: Diferença percentual entre as taxas de erro 10^{-6} , 10^{-7} , 10^{-8} e 10^{-9} versus 10^{-5} para as estimativas dos parâmetros dos modelos M1 (com *pedigree*) ($\Delta\text{TE.CD-M1}$) e M2 (sem *pedigree*) ($\Delta\text{TE.CD-M2}$) (Tabelas 1 e 2 - Anexo 2) obtidas através dos algoritmos *Average Information* (AI), *Newton Rhapsion* (NR), *Expectation Maximization* (EM) e os algoritmos combinados EM-AI e EM-NR, para o critério de convergência CD, com algoritmos desenvolvidos para o software R. SNMIP: superou o número máximo de iterações permitido.

Os dois modelos (M1 com *pedigree* e M2 sem *pedigree*) apresentaram problemas de convergência para $\text{TE} \leq 10^{-7}$ no critério de convergência CV, com exceção do algoritmo EM-NR no modelo M1 que demonstrou problema a partir de $\text{TE} < 10^{-7}$. Estes problemas foram atribuídos ao número de iterações que extrapolou o critério estipulado de cinco vezes o número de iterações da TE imediatamente anterior (Figura 1 e Tabelas 1 e 2 - Anexo 2).

Diferente dos componentes de variância (Tabelas 1 e 2 - Anexo 2) que apresentaram pequenas diferenças quanto ao critério de convergência (Figura 1) e os níveis de TE (Figura 2) as estimativas dos coeficientes de determinação, apresentados na Tabela 2, foram uniformes para as combinações entre algoritmos, critérios de convergência e TE para ambos os modelos.

Tabela 2: Estimativa dos parâmetros genéticos para os modelos M1 (com *pedigree*) e M2 (sem *pedigree*), obtidas no software R, pelos algoritmos *Average Information* (AI), *Newton Rhapson* (NR), *Expectation Maximization* (EM) e os algoritmos combinados EM-AI e EM-NR para diferentes taxas de erro (TE), considerando as diferenças entre coeficientes de determinação (CD) e componentes de variância (CV) como critério de convergência.

Mod	TE	Par.	CD					CV						
			AI	NR	EM	EM-AI	EM-NR	AI	NR	EM	EM-AI	EM-NR		
M1	10 ⁻⁵	h_a^2	0,7 2	0,7 2	0,72	0,72	0,72	0,72	0,72	0,72	0,72	0,72	0,72	
		$c_{t,p}^2$	0,0 8	0,0 8	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	
	10 ⁻⁶	h_a^2	0,7 2	0,7 2	0,72	0,72	0,72	0,72	0,72	0,72	0,72	0,72	0,72	
		$c_{t,p}^2$	0,0 8	0,0 8	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	
	10 ⁻⁷	h_a^2	0,7 2	0,7 2	0,72	0,72	0,72	SNMIP	SNMIP	0,72	SNMIP	0,72	SNMIP	
		$c_{t,p}^2$	0,0 8	0,0 8	0,08	0,08	0,08	SNMIP	SNMIP	0,08	SNMIP	0,08	SNMIP	
	10 ⁻⁸	h_a^2	0,7 2	0,7 2	0,72	0,72	0,72	SNMIP	SNMIP	0,72	SNMIP	SNMIP	SNMIP	
		$c_{t,p}^2$	0,0 8	0,0 8	0,08	0,08	0,08	SNMIP	SNMIP	0,08	SNMIP	SNMIP	SNMIP	
	10 ⁻⁹	h_a^2	0,7 2	0,7 2	0,72	0,72	0,72	SNMIP	SNMIP	0,72	SNMIP	SNMIP	SNMIP	
		$c_{t,p}^2$	0,0 8	0,0 8	0,08	0,08	0,08	SNMIP	SNMIP	0,08	SNMIP	SNMIP	SNMIP	
	M2	10 ⁻⁵	h^2	0,0 6	0,0 6	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0,06
			c_b^2	0,5 5	0,5 5	0,55	0,55	0,55	0,55	0,55	0,55	0,55	0,55	0,55
10 ⁻⁶		h^2	0,0 6	0,0 6	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0,06	
		c_b^2	0,5 5	0,5 5	0,55	0,55	0,55	0,55	0,55	0,55	0,55	0,55	0,55	
10 ⁻⁷		h^2	0,0 6	0,0 6	0,06	0,06	0,06	SNMIP	SNMIP	0,06	SNMIP	SNMIP	SNMIP	
		c_b^2	0,5 5	0,5 5	0,55	0,55	0,55	SNMIP	SNMIP	0,55	SNMIP	SNMIP	SNMIP	
10 ⁻⁸		h^2	0,0 6	0,0 6	0,06	0,06	0,06	SNMIP	SNMIP	0,06	SNMIP	SNMIP	SNMIP	
		c_b^2	0,5 5	0,5 5	0,55	0,55	0,55	SNMIP	SNMIP	0,55	SNMIP	SNMIP	SNMIP	
10 ⁻⁹		h^2	0,0 6	0,0 6	0,06	0,06	0,06	SNMIP	SNMIP	0,06	SNMIP	SNMIP	SNMIP	

c_b^2	0,5	0,5								
	5	5	0,55	0,55	0,55	SNMIP	SNMIP	0,55	SNMIP	SNMIP

h_a^2 : herdabilidade no sentido restrito (efeitos aditivos), $c_{i,p}^2$: coeficiente de determinação da interação testador versus população, h^2 : herdabilidade genética no sentido amplo, c_b^2 : coeficiente de determinação dos blocos e SNMIP: superou o máximo de iterações permitido (cinco vezes o número de iterações da taxa de erro anterior).

3.2 EFICIÊNCIA COMPUTACIONAL

Apesar da pequena variação observada nas estimativas dos componentes de variância considerando os níveis de TE e os critérios de convergência nos dois modelos, de maneira geral o número de iterações foi sempre superior para os algoritmos que utilizaram CV como critério de convergência em contraste àqueles que utilizaram o CD. Além disso, todos os algoritmos apresentaram aumento no número de iterações para $TE < 10^{-5}$, tanto em CD quanto em CV, contudo, os algoritmos NR e EM-NR atingiram a convergência com menor número de iterações (Figuras 1 e 2).

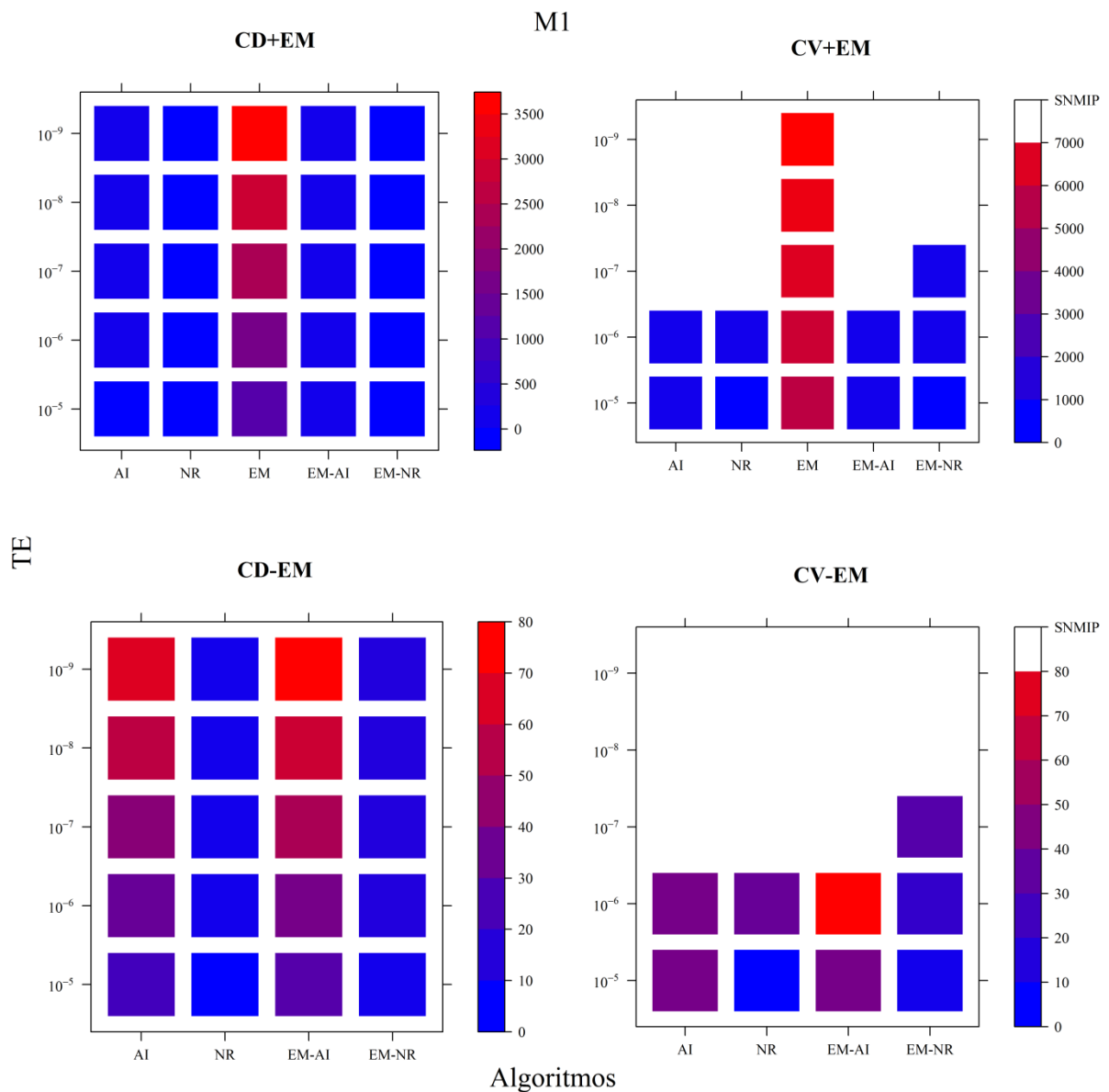


Figura 3: Número de iterações para a convergência do modelo M1 (com *pedigree*) pelos algoritmos *Average Information* (AI), *Newton Rhapson* (NR), *Expectation Maximization* (EM) e os algoritmos combinados EM-AI e EM-NR para diferentes taxas de erro (TE), considerando as diferenças entre coeficientes de determinação (CD) (Figuras CD+EM e CD-EM) e componentes de variância (CV) (Figuras CV+EM e CV-EM) como critério de convergência, para os algoritmos desenvolvidos para o software R. SNMIP: superou o número máximo de iterações permitido.

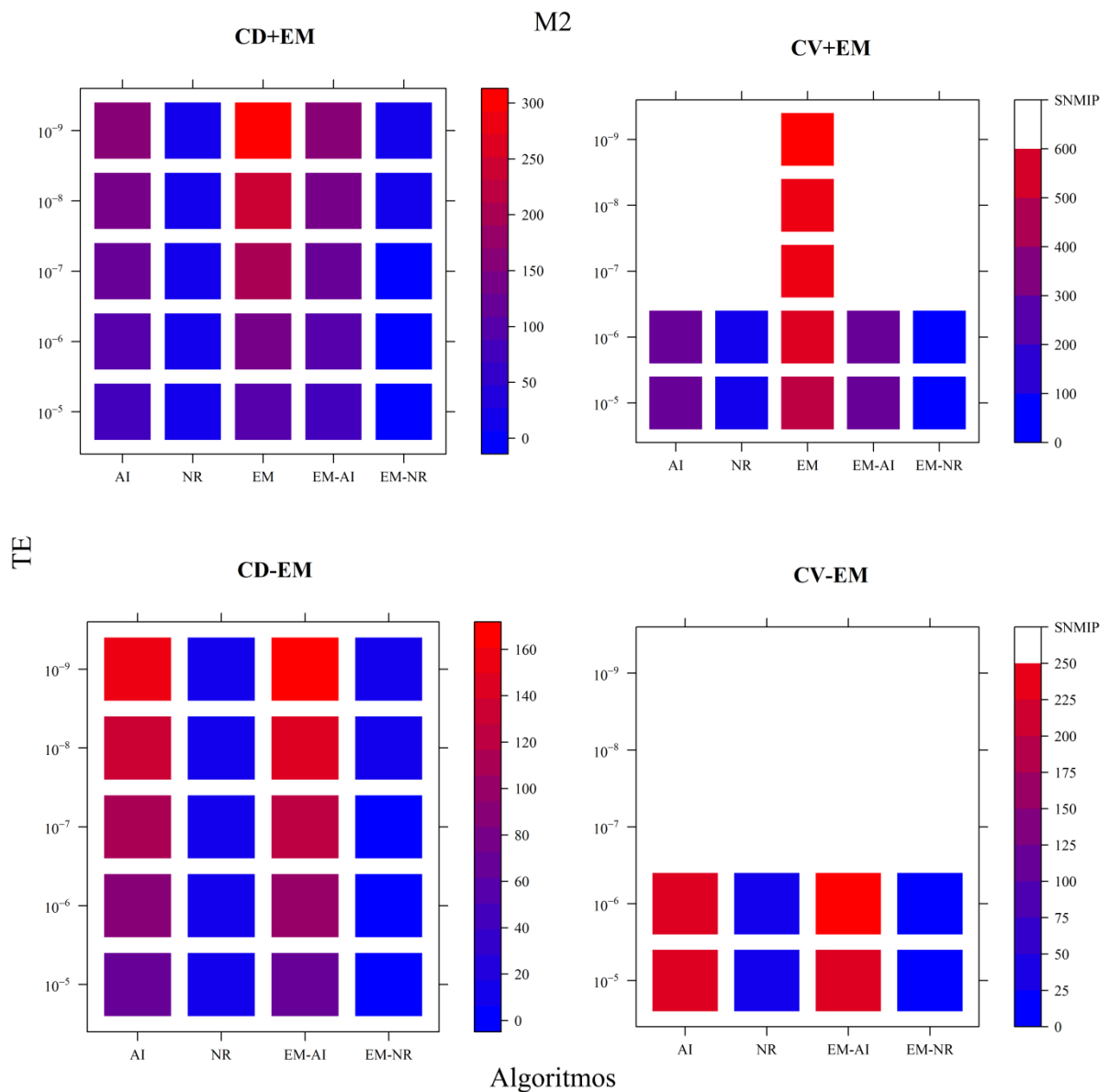


Figura 4: Número de iterações para a convergência do modelo M2 (sem *pedigree*) pelos algoritmos *Average Information* (AI), *Newton Rhapsion* (NR), *Expectation Maximization* (EM) e os algoritmos combinados EM-AI e EM-NR para diferentes taxas de erro (TE), considerando as diferenças entre coeficientes de determinação (CD) (Figuras CD+EM e CD-EM) e componentes de variância (CV) (Figuras CV+EM e CV-EM) como critério de convergência, para os algoritmos desenvolvidos para o software R. SNMIP: superou o número máximo de iterações permitido.

Dentre os algoritmos estudados o EM foi o que apresentou maior número de iterações em todas as combinações entre critérios de convergência (CD e CV) e níveis de TE, para ambos os modelos.

Devido à uniformidade das estimativas dos componentes de variância e coeficientes de determinação (Figura 1, Figura 2 e Tabela 2) entre os critérios de convergência e níveis de TE nos dois modelos, optou-se por apresentar o tempo médio de execução dos algoritmos somente para TE igual a 10^{-5} , utilizando CD como critério de convergência, pois nestas duas condições (10^{-5} e CD) pode-se obter maior eficiência computacional devido ao menor número de iterações para atingir a convergência (Figuras 3 e 4).

Analisando o tempo médio de execução dos algoritmos na Tabela 3 verifica-se que o uso da técnica de matrizes esparsas contribuiu significativamente para a diminuição do tempo de execução dos algoritmos otimizados, não otimizados e os algoritmos combinados, nos dois modelos (M1 com *pedigree* e M2 sem *pedigree*), viabilizando a maximização da função de verossimilhança residual em R. Não obstante, o uso de matrizes normais provocou falha de memória no computador C2 para o ajuste do modelo M1 (com *pedigree*) pelo algoritmo NR, em consequência da menor disponibilidade de memória RAM (2GB) (Tabela 1 - Anexo 1) e chegou a aumentar o tempo de processamento em aproximadamente 58 vezes (4293,70/74,35) para M1 (AI-Otimizado-C3) e 88 vezes (70261,60/797,69) para M2 (sem *pedigree*) (AI-Não otimizado-C2).

Além das matrizes esparsas as otimizações propostas foram efetivas na redução do tempo de execução dos algoritmos. A diminuição no tempo de execução (em segundos) chegou a aproximadamente 85 vezes (7100,74/83,62) para o M1 (com *pedigree*) (NR-C3-ME) e 104 vezes (6457,80/61,95) em M2 (sem) (NR-C3-ME) (Tabela 3).

Os algoritmos combinados EM-AI e EM-NR apresentaram resposta semelhante aos algoritmos AI e NR otimizados para ambos os modelos, respectivamente. Contudo, a utilização de matrizes esparsas para o algoritmo combinado EM-NR diminuiu, aproximadamente, pela metade o tempo de execução (15,78 segundos) e o número de iterações (6) em relação ao algoritmo NR (29,84 segundos e 11 iterações) no modelo M2 (sem *pedigree*), no computador C1 (Tabela 3 e Figuras 3 e 5).

Tabela 3: Tempo médio de execução dos algoritmos desenvolvidos para o software R, por meio de *Average Information* (AI), *Newton Rhapson* (NR), *Expectation Maximization* (EM) e os algoritmos combinados EM-AI e EM-NR, para os modelos M1 (com *pedigree*) e M2 (sem

pedigree), utilizando-se ou não matrizes esparsas (ME e MN). Considerando taxa de erro de 10^{-5} e os coeficientes de determinação como critério de convergência.

Algoritmos	Computador [†]	Tempo (segundos)					
		M1		M2			
		ME	MN	ME	MN		
Otimizados ¹	AI	C1	32,19	1459,52	95,05	4304,25	
	AI	C2	571,07	FM	320,30	13873,88	
	AI	C3	74,35	4293,70	201,58	12135,23	
	AI	C4	49,87	575,81	164,05	1759,11	
	NR	C1	36,48	1506,47	29,84	709,38	
	NR	C2	1406,42	FM	96,85	2341,01	
	NR	C3	83,62	FM	61,95	2046,01	
	NR	C4	55,77	535,27	47,38	293,48	
	EM	C1	2043,44	101426,60	65,69	4235,21	
	EM	C2	7266,53	311496,71	208,31	12563,42	
	EM	C3	5572,75	286981,80	172,45	11743,91	
	EM	C4	3343,79	39091,04	121,12	1620,79	
	Não Otmizados	AI	C1	839,16	3990,00	281,57	23038,06
		AI	C2	3547,80	FM	797,69	70261,60
		AI	C3	1915,22	8759,95	751,99	63057,59
		AI	C4	1144,78	1423,14	524,56	7557,71
NR		C1	2575,51	10268,84	2485,29	15127,75	
NR		C2	8328,55	FM	7716,22	46216,11	
NR		C3	7100,74	26877,28	6457,80	39294,11	
NR		C4	2147,42	3294,63	962,98	4752,31	
EM		C1	3159,74	170679,95	69,91	4226,76	
EM		C2	11941,78	FM	208,19	12556,18	
EM		C3	8828,96	451477,30	174,93	12036,28	
EM		C4	5227,25	58897,90	120,94	620,78	
Combinados ²		EM-AI	C1	37,60	1730,65	96,63	4446,29
		EM-AI	C2	1116,93	FM	324,00	13592,51
		EM-AI	C3	86,51	4350,75	213,72	12989,16
		EM-AI	C4	59,51	678,76	163,01	1813,84
	EM-NR	C1	39,26	1626,71	15,78	415,75	
	EM-NR	C2	1068,29	FM	49,50	1421,24	
	EM-NR	C3	87,87	3827,42	33,84	1207,88	
	EM-NR	C4	61,98	591,68	24,99	168,51	

¹Cálculo indireto dos termos da matriz de informação e/ou traço matricial obtido por meio do produto de Hadamard. ²Os algoritmos combinados EM-AI, EM-NR foram construídos com base nos algoritmos otimizados considerando uma iteração EM para gerar as estimativas iniciais dos componentes de variância utilizadas nos algoritmos AI e NR. O termo FM representa a falha de memória do computador. [†]A configuração dos computadores C1, C2, C3 e C4 estão listadas na tabela 1 do Anexo 1.

Tabela 4: Tempo médio por iteração (T/I) dos algoritmos desenvolvidos para o software R, por meio de *Average Information* (AI), *Newton Rhapsion* (NR), *Expectation Maximization* (EM) e os algoritmos combinados EM-AI e EM-NR para os modelos M1 (com *pedigree*) e

M2 (sem *pedigree*), utilizando-se ou não matrizes esparsas (ME e MN). Considerando taxa de erro de 10^{-5} e os coeficientes de determinação como critério de convergência.

Algoritmos	Computador [†]	T/I (s iteração ⁻¹)					
		M1		M2			
		ME	MN	ME	MN		
Otimizados ¹	AI	C1	2,68	121,63	1,36	61,49	
	AI	C2	47,59	FM	4,58	198,20	
	AI	C3	6,20	357,81	2,88	173,36	
	AI	C4	4,16	47,98	2,34	25,13	
	NR	C1	4,56	188,31	2,71	64,49	
	NR	C2	175,80	FM	8,80	212,82	
	NR	C3	10,45	FM	5,63	186,00	
	NR	C4	6,97	66,91	4,31	26,68	
	EM	C1	1,74	86,32	0,71	46,03	
	EM	C2	6,18	265,10	2,26	136,48	
	EM	C3	4,74	244,24	1,87	127,65	
	EM	C4	2,85	33,27	1,32	17,62	
	Não otimizados	AI	C1	69,93	332,50	4,02	329,12
		AI	C2	295,65	FM	11,40	1003,74
		AI	C3	159,60	730,00	10,74	900,82
		AI	C4	95,40	118,60	7,49	107,97
NR		C1	321,94	1283,61	225,94	1375,25	
NR		C2	1041,07	FM	701,47	4201,46	
NR		C3	887,59	3359,66	587,07	3572,19	
NR		C4	268,43	411,83	87,54	432,03	
EM		C1	2,69	145,26	0,76	45,94	
EM		C2	10,16	FM	2,26	136,48	
EM		C3	7,51	384,24	1,90	130,83	
EM		C4	4,45	50,13	1,31	6,75	
Combinados ²		EM-AI	C1	2,69	123,62	1,34	61,75
		EM-AI	C2	79,78	FM	4,50	188,78
		EM-AI	C3	6,18	310,77	2,97	180,41
		EM-AI	C4	4,25	48,48	2,26	25,19
	EM-NR	C1	4,36	180,75	2,63	69,29	
	EM-NR	C2	118,70	FM	8,25	236,87	
	EM-NR	C3	9,76	425,27	5,64	201,31	
	EM-NR	C4	6,89	65,74	4,17	28,09	

¹Cálculo indireto dos termos da matriz de informação e/ou traço matricial obtido por meio do produto de Hadamard. ²Os algoritmos combinados EM-AI, EM-NR foram construídos com base nos algoritmos otimizados considerando uma iteração EM para gerar as estimativas iniciais dos componentes de variância utilizadas nos algoritmos AI e NR. O termo FM representa a falha de memória do computador. [†]A configuração dos computadores C1, C2, C3 e C4 estão listadas na tabela 1 do Anexo 1.

Os algoritmos AI, NR, EM-AI e EM-NR demonstraram boa eficiência computacional no computador C1, que apresenta uma configuração razoável, com tempo de processamento variando de 32,19 a 39,26 segundos para M1 (com *pedigree*) e de 15,78 a 96,63 segundos para M2 com matrizes esparsas (Tabela 3). Quanto ao número de iterações, os mesmos

algoritmos apresentaram amplitude de 8 a 14 para M1 (com *pedigree*) e de 6 a 72 para M2 (sem *pedigree*) sob matrizes esparsas (Figuras 3 e 5).

Por outro lado, relacionando o tempo total de processamento com o número de iterações, observa-se que os algoritmos otimizados apresentaram menor tempo por iteração que os não otimizados, com exceção do EM no modelo M2 (sem *pedigree*). Além disso, a técnica de matrizes esparsas contribuiu para a diminuição do tempo gasto por iteração em todos os algoritmos e computadores estudados. Outro ponto a ser destacado é que apesar do bom desempenho observado para os algoritmos AI, EM-AI, NR e EM-NR os dois primeiros apresentaram tempo médio por iteração sempre inferior aos dois últimos (Tabela 4).

Considerando o critério de convergência CD e TE igual a 10^{-5} , fixados anteriormente para a exibição das estimativas dos tempos de execução, observa-se na Figura 5 que todos os algoritmos promoveram o aumento de $\log(L)$ e conduziram a maximização para uma mesma região de verossimilhança em cada um dos modelos.

Os algoritmos AI e sua combinação com o algoritmo EM (EM-AI) apresentaram rápido aumento de $\log(L)$ nas primeiras iterações para M1 (com *pedigree*), entretanto, a maximização da função de verossimilhança residual desacelerou a partir da quinta iteração e a convergência só ocorreu nas iterações 12 e 14 respectivamente. De forma contrária o algoritmo NR e sua combinação com o algoritmo EM (EM-NR) proporcionaram aumento de $\log(L)$ inferior ao observado para o AI e EM-AI nas primeiras iterações, entretanto, atingiram a convergência mais cedo, com 8 e 9 iterações, respectivamente. Ainda no modelo M1 (com *pedigree*), demonstrou-se que o algoritmo EM foi demasiadamente lento para atingir a convergência, que ocorreu com cerca de 1175 iterações (Figura 5).

Em contrapartida, no modelo M2 (sem *pedigree*) os algoritmos NR e EM-NR promoveram rápido aumento de $\log(L)$ e apresentaram rápida convergência com 11 e 6 iterações respectivamente. Não obstante, os algoritmos AI e EM-AI apresentaram lentidão equiparável ao EM para maximizar a função de verossimilhança residual atingindo a convergência com 70 e 72 iterações, respectivamente. Apesar do algoritmo EM ter sido o mais lento para M2 (sem *pedigree*) ele convergiu com 92 iterações, mostrando-se mais eficiente que no modelo M1 (com *pedigree*) (Figura 5).

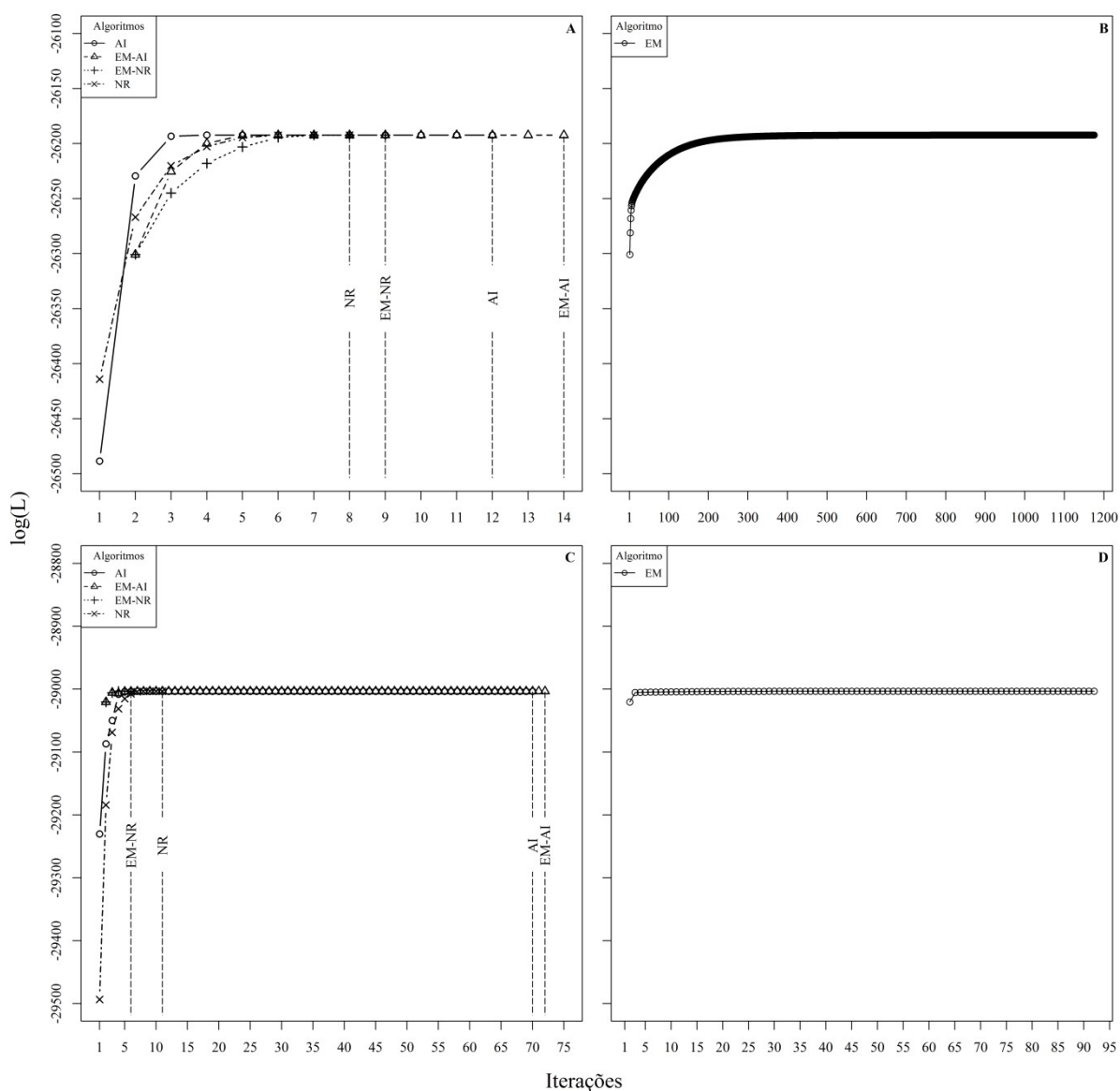


Figura 5: Diagnóstico de convergência dos algoritmos *Average Information* (AI), *Newton Rhapsion* (NR), *Expectation Maximization* (EM) e as combinações EM-AI e EM-NR, desenvolvidos para o software R, para a função de máxima verossimilhança residual (REML), considerando os modelos M1 (com *pedigree*) (Figuras 5 A e B), M2 (sem *pedigree*) (Figuras 5 C e D) e coeficientes de determinação (CD) como critério de convergência.

Apesar da rápida convergência dos algoritmos NR e EM-NR em ambos os modelos, é oportuno ressaltar que o tempo gasto por iteração foi sempre superior para estes algoritmos quando contrastados com AI e EM-AI, entretanto as otimizações propostas reduziram a disparidade entre os resultados obtidos por estes algoritmos.

4. DISCUSSÃO

A ausência de convergência do algoritmo FS para ambos os modelos estudados não é um resultado atípico, pois o gradiente imposto pela esperança matemática da segunda derivada da função de verossimilhança residual normalmente leva a um aumento rápido de $\log(L)$ e em alguns casos a rápida convergência (WANG, 2007). Entretanto, como a convergência desse algoritmo é quadrática (WANG, 2007) a fuga do espaço paramétrico torna-se inevitável em algumas situações. Não obstante, os demais algoritmos foram efetivos em maximizar a função de verossimilhança residual garantindo a convergência e estimativas positivas dos componentes de variância.

A combinação EM-FS também apresentou problemas de convergência e este resultado pode estar ligado ao número insuficiente de iterações EM (uma iteração) para fornecer uma boa estimativa inicial para os componentes de variância dos dois modelos (MEYER, 2006).

Apesar das estimativas dos parâmetros dos modelos M1 (com *pedigree*) e M2 (sem *pedigree*) terem sido obtidas de diferentes algoritmos combinados com diferentes critérios de convergência e TE observou-se que as soluções de REML foram equivalentes para todas as condições devido a pequena amplitude de $\log(L)$. Este fato demonstra a coerência entre os algoritmos, mostrando que independente do critério de convergência adotado ou taxa de erro estipulada os algoritmos conduziram a uma mesma solução de REML.

Adicionalmente, observou-se que as médias das estimativas de componentes de variância e de coeficientes de determinação obtidas dos diferentes algoritmos, combinados com diferentes critérios de convergência e TE, foram semelhantes às estimativas obtidas pelo *software* ASReml no modelo M1 (com *pedigree*), e aquelas alcançadas pelos *softwares* ASReml, Selegen e lme4 no modelo M2 (sem *pedigree*). Estes resultados respaldam os algoritmos propostos para o *software* R, atestando que os mesmos conduzem a estimativas coerentes para os componentes de variância.

As diferenças observadas entre as estimativas de $\log(L)$ do *software* lme4 e os *softwares* ASReml, Selegen é atribuída à constante da função de verossimilhança restrita, no entanto ela não afeta a estimação dos componentes de variância (RESENDE; SILVA; AZEVEDO, 2014) nem a comparação de modelos via razão de máxima verossimilhança.

Diante dos pequenos desvios observados para os componentes de variância e da uniformidade dos coeficientes de determinação em relação aos critérios de convergência e o gradiente de taxas de erro, verificou-se que os critérios de convergência CV e $TE < 10^{-5}$ contribuem apenas para onerar o esforço computacional devido ao aumento do número de iterações para atingir a convergência. Além disso, CV se mostrou ineficiente em atingir a

convergência para $TE \leq 10^{-7}$ inviabilizando REML. Portanto, para viabilizar computacionalmente o procedimento REML com grande volume de dados e garantir boa precisão nas estimativas dos componentes de variância pode-se adotar CD como critério de convergência e TE igual a 10^{-5} .

Dentre os algoritmos estudados o EM apresentou maior número de iterações para atingir a convergência em ambos os modelos. Este fato está ligado à propriedade de convergência linear do EM, garantindo aumento progressivo de $\log(L)$ (MEYER, 2006; RESENDE; SILVA; AZEVEDO, 2014), ao passo que os algoritmos de segunda derivada apresentam convergência quadrática, o que garante rápida convergência, entretanto nem sempre o aumento de $\log(L)$ é garantido (DEMPSTER; LAIRD; RUBIN, 1977; JOHNSON; THOMPSON, 1995; MEYER; SMITH, 1996).

Avaliando a eficiência computacional via tempo médio de execução dos algoritmos em R, observou-se que a utilização de matrizes esparsas promoveu impacto significativo na redução do tempo de execução dos algoritmos. Este resultado está associado à condição essencialmente esparsa do lado esquerdo do SEMM que resulta em menor demanda de memória e maior agilidade nas operações matriciais, pois somente os elementos diferentes de zero são armazenados (MISZTAL; PEREZ-ENCISO, 1993; LEE; VAN DER WERF, 2006; MASUDA et al., 2015).

Na ausência de matrizes esparsas o requerimento de memória aumenta devido ao armazenamento de todos os elementos do SEMM, o que pode limitar ou até mesmo inviabilizar o uso de REML para grandes bases de dados e modelos complexos. Este fato foi observado para o computador C2 que apresentou falha de memória e não foi capaz de maximizar a função de verossimilhança residual para o modelo M1 (com *pedigree*), em função da sua quantidade de memória limitada. Adicionalmente, Zhou e Stephens (2014) citam que a demanda por memória se mostrou como uma limitação para o ajuste de modelos com grande número de indivíduos, em virtude da operação matricial de decomposição de autovalores realizada na fase inicial do algoritmo proposto pelos autores.

Neste contexto, diversos autores vêm recomendando o uso de matrizes esparsas devido a melhoria da eficiência computacional (JOHNSON; THOMPSON, 1995; GILMOUR; THOMPSON; CULLIS, 1995; THOMPSON et al., 2003; LEE; VAN DER WERF, 2006) e esta técnica ganhou espaço em aplicações genéticas via modelos lineares mistos (ZHOU; CARBONETTO; STEPHENS, 2013).

Além da utilização de matrizes esparsas as otimizações propostas foram efetivas na diminuição do esforço computacional para atingir a convergência em ambos os modelos resultando em menor tempo de processamento que no método direto. Tal fato demonstra que as estratégias adotadas contribuíram para minimizar o custo computacional para ajustar modelos mistos via REML. Ademais os futuros progressos na otimização de algoritmos juntamente com uso de matrizes esparsas devem impulsionar cada vez mais o uso de REML em aplicações cada vez mais arrojadas em estudos genéticos. Neste contexto, alguns autores vêm propondo otimizações em algoritmos para aumentar a eficiência computacional para viabilizar estudos genômicos com grande número de indivíduos com modelos uni ou multivariados que permitam a melhor compreensão da arquitetura genética de caracteres fenotípicos (LEE; VAN DER WERF, 2006; ZHOU; STEPHENS, 2014; HAN; XU, 2008).

Por outro lado, ainda foi observado que o tempo por iteração foi inferior nos algoritmos otimizados quando comparados com os não otimizados, atestando que essas conferiram maior agilidade para o procedimento numérico. Tais resultados corroboram com os obtidos por Lee; Van Der Werf (2006) e Zhou; Stephens (2014) que obtiveram êxito na redução do tempo de processamento em função de otimizações propostas pelos mesmos autores, evidenciando que as otimizações parecem ser uma tendência para o futuro da modelagem computacional dos modelos mistos, em um cenário de aumento progressivo no volume de dados genotípicos e fenotípicos gerados nos programas de melhoramento.

O algoritmo EM foi a única exceção entre algoritmos otimizados e não otimizados que apresentou resultado praticamente equivalente do tempo de processamento e tempo médio por iteração no modelo M2. Este resultado foi observado, pois nesta condição não houve necessidade de utilizar o produto de Hadamard, devido a inexistência de matrizes de parentesco neste modelo, portanto o custo computacional foi parecido em ambas as situações. Este procedimento utilizado para a simplificação dos traços matriciais foi descrito por Misztal; Perez-Enciso (1993) e trouxe impacto positivo na redução do tempo de processamento total e por iteração do EM para o modelo M1 otimizado em relação ao não otimizado.

Os algoritmos combinados apresentaram eficiência equivalente aos algoritmos otimizados AI e NR em relação ao tempo total de processamento. Entretanto, ao se comparar o algoritmo NR e a combinação EM-NR com os algoritmos AI e EM-AI verificou-se que a convergência dos dois primeiros ocorreu primeiro que os dois últimos para os dois modelos. Este resultado pode ser explicado pelo maior gasto de tempo por iteração dos algoritmos NR e

EM-NR em contraste aos algoritmos AI e EM-AI, devido à complexidade dos traços matriciais da esperança matemática da matriz de informação NR (GILMOUR; THOMPSON; CULLIS, 1995; JOHNSON; THOMPSON, 1995) denominada de informação de Fisher e utilizada no algoritmo FS. Esta complexidade ainda persistiu nos algoritmos otimizados, porém em menor grau, demonstrando a relevância das otimizações propostas.

Os algoritmos AI e NR convergiram com valores equivalentes a EM-AI e EM-NR, não apresentando problemas de convergência. Entretanto, é oportuno ressaltar que no presente trabalho foram testados apenas dois modelos para um conjunto de dados, relacionados ao rendimento de grãos de milho e, certamente para outros modelos e conjuntos de dados, podem ser observados problemas de convergência devido a problemas relacionados às estimativas iniciais dos componentes de variância. Nestes casos o esquema proposto por Meyer (2006, 2007a) deve ser adotado.

No modelo M1 (com *pedigree*) o algoritmo EM foi extremamente lento para atingir a convergência demonstrando sua menor eficiência computacional. Na preferência por este algoritmo uma solução é utilizar a versão PX (*Parameter Expanded*) do EM (FOULLEY; VAN DYK, 2000; DIFFEY; WELSH; CULLIS, 2013; LI; POURAHMADI, 2013) ou para casos mais específicos usar a reparametrização sugerida por Meyer (1987) e as absorções matriciais sugeridas Meyer (1987) e Resende; Silva; Azevedo (2014).

As otimizações propostas para os algoritmos aumentaram a eficiência computacional proporcionando a diminuição do tempo de execução para o ajuste dos modelos mistos em R, e podem servir de inspiração para serem adaptadas nos algoritmos AI e NR propostos por Lee; Van Der Werf (2006) e Zhou; Stephens (2014) para estudos genômicos. Ratificou-se, ainda, a importância da utilização de matrizes esparsas que, juntamente com as otimizações propostas, podem viabilizar o método REML em computadores com menor disponibilidade de memória. Além disso, verificou-se que uma boa estratégia é monitorar a convergência dos modelos por meio dos coeficientes de determinação, pois este procedimento contribuiu para redução do número de iterações, e conseqüentemente para a redução de tempo no processamento das análises. Ademais, foi constatado que a taxa de erro 10^{-5} promove a estimativa dos parâmetros com precisão satisfatória.

5. CONCLUSÃO

A utilização de matrizes esparsas associadas a otimizações para os algoritmos REML aumentaram a eficiência computacional proporcionando a diminuição do tempo de execução para o ajuste dos modelos mistos em R.

Estipular o critério de convergência de modelos mistos por meio dos coeficientes de determinação é uma estratégia eficiente para reduzir do número de iterações e o reduzir o tempo de processamento das análises, em relação à utilização de componentes de variância.

Adotar taxa de erro de 10^{-5} , para convergência de modelos mistos, promove estimativa de parâmetros genéticos de milho com precisão satisfatória e proporciona redução no tempo de análise em relação a taxas mais restritivas ($< 10^{-5}$).

6. REFERÊNCIAS

- ARAUS, J. L.; CAIRNS, J. E. Field high-throughput phenotyping: the new crop breeding frontier. **Trends in Plant Science**, v. 19, n. 1, p. 52-61, 2014.
- BATES, D.; MAECHLER, M. **Matrix: Sparse and Dense Matrix Classes and Methods**. R package version 1.2-3, 2016. Disponível em: < <https://cran.r-project.org/web/packages/Matrix/index.html> >. Acesso em: 10 de maio de 2016.
- BATES, D.; MAECHLER, M.; BOLKER, B.; WALKER, S. Fitting Linear Mixed-Effects Models Using lme4. **Journal of Statistical Software**, v. 67, n. 1, p. 1-48, 2015.
- BAXTER, I. R.; ZIEGLER, G.; LAHNER, B.; MICKELBART, M. V.; FOLEY, R.; DANKU, J.; ARMSTRONG, P.; SALT, D. E.; HOEKENGA, O. A. Single-kernel ionic profiles are highly heritable indicators of genetic and environmental influences on elemental accumulation in maize grain (*Zea mays*). **PLoS one**, v. 9, n. 1, p. e87628, 2014.
- BOMMERT, P.; NAGASAWA, N. S.; JACKSON, D. Quantitative variation in maize kernel row number is controlled by the FASCIATED EAR2 locus. **Nature genetics**, v. 45, n. 3, p. 334-337, 2013.
- BUTLER, D. G.; CULLIS, B. R.; GILMOUR, A. R.; GOGEL, B. J. **ASReml-R reference manual: mixed models for S language environments**. Queensland Department of Primary Industries, Queensland, Australia, 2009. Disponível em: < <http://discoveryfoundation.org.uk/downloads/asreml/release3/asreml-R.pdf> >. Acesso em: 10 maio de 2016.
- CASALE, F. P.; RAKITSCH, B.; LIPPERT, C.; STEGLE, O. Efficient set tests for the genetic analysis of correlated traits. **Nature methods**, v. 12, n. 8, p. 755-758, 2015.
- COBB, J. N.; DECLERCK, G.; GREENBERG, A.; CLARK, R.; MCCOUCH, S. Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. **Theoretical and Applied Genetics**, v. 126, n. 4, p. 867-887, 2013.
- DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the royal statistical society**, v. 39, n.1, p. 1-38, 1977.
- DIFFEY, S.; WELSH, A.; CULLIS, B. A faster and computationally more efficient REML (PX)EM algorithm for linear mixed models. **Centre for Statistical and Survey Methodology**, v. 8, p. 2-13, 2013.
- FOLEY, J. A.; RAMANKUTTY, N.; BRAUMAN, K. A.; CASSIDY, E. S.; GERBER, J. S.; JOHNSTON, M.; MUELLER, N. D.; O'CONNELL, C.; RAY, D. K.; WEST, D. K.; BALZER, C.; BENNETT, E. M.; CARPENTER, S. R.; HILL, J.; MONFREDA, C.;

- POLASKY, S.; ROCKSTRÖM, J.; SHEEHAN, J.; SIEBERT, S.; TILMAN, D.; ZAKS, D. P. M. Solutions for a cultivated planet. **Nature**, v. 478, n. 7369, p. 337-342, 2011.
- FOULLEY, J.-L.; VAN DYK, D. A. The PX-EM algorithm for fast stable fitting of Henderson's mixed model. **Genetics Selection Evolution**, v. 32, n. 2, p. 1-21, 2000.
- GILMOUR, A. R.; GOGEL, B. J.; CULLIS, B. R.; WELHAM, S. J.; THOMPSON, R. **ASReml user guide: Release 4.1 structural specification**. VSN International Ltd, Hemel Hempstead, HP1 1ES, 2015. Disponível em: <<http://www.vsn.co.uk/downloads/asreml/release4/UserGuideStructural.pdf>>. Acesso em: 10 de maio de 2016.
- GILMOUR, A. R.; THOMPSON, R.; CULLIS, B. R. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. **Biometrics**, v. 51, n. 4, p. 1440-1450, 1995.
- HAN, L.; XU, S. A Fisher scoring algorithm for the weighted regression method of QTL mapping. **Heredity**, v. 101, n. 5, p. 453-464, 2008.
- HARVILLE, D. A. Making REML computationally feasible for large data sets: use of the Gibbs sampler. **Journal of Statistical Computation & Simulation**, v. 74, n. 2, p. 135-153, 2004.
- HENDERSON, C. R.; KEMPTHORNE, O.; SEARLE, S. R.; VON KROSIGK, C. M. The estimation of environmental and genetic trends from records subject to culling. **Biometrics**, v. 15, n. 2, p. 192-218, 1959.
- JESSKE, T.; OLBERG, B.; SCHIERHOLT, A.; BECKER, H. C. Resynthesized lines from domesticated and wild Brassica taxa and their hybrids with *B. napus* L.: genetic diversity and hybrid yield. **Theoretical and Applied Genetics**, v. 126, n. 4, p. 1053-1065, 2013.
- JOHNSON, D. L.; THOMPSON, R. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. **Journal of dairy science**, v. 78, n. 2, p. 449-456, 1995.
- KÖNIG, J.; PEROVIC, D.; KOPAHNKE, D.; ORDON, F. Mapping seedling resistance to net form of net blotch (*Pyrenophora teres* f. *teres*) in barley using detached leaf assay. **Plant Breeding**, v. 133, n. 3, p. 356-365, 2014.

LEE, S. H.; VAN DER WERF, J. H. J. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. **Genetics Selection Evolution**, v. 38, n. 1, p. 1-19, 2006.

LI, E.; POURAHMADI, M. An alternative REML estimation of covariance matrices in linear mixed models. **Statistics & Probability Letters**, v. 83, n. 4, p. 1071-1077, 2013.

LIU, C.; YANG, Z.; HU, Y.-G. Drought resistance of wheat alien chromosome addition lines evaluated by membership function value based on multiple traits and drought resistance index of grain yield. **Field Crops Research**, v. 179, p. 103-112, 2015.

LOH, P. R.; BHATIA, G.; GUSEV, A.; FINUCANE, H. K.; BULIK-SULLIVAN, B. K.; POLLACK, S. J.; CONSORTIUM, S. W. G. P. G.; CANDIA, T. R.; LEE, S. H.; WRAY, N. R.; KENDLER, K. S.; O'DONOVAN, M. C.; NEALE, B. M.; PATTERSON, N.; PRICE, A. L. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. **Nature genetics**, v. 47, n. 12, p. 1385–1392, 2015.

MASUDA, Y.; AGUILAR, I.; TSURUTA, S.; MISZTAL, I. Technical note: Acceleration of sparse operations for average-information REML analyses with supernodal methods and sparse-storage refinements. **Journal of animal science**, v. 93, n. 10, p. 4670-4674, 2015.

MERSMANN, O. **microbenchmark: Accurate Timing Functions**. R package version 1.4-2.1, 2016. Disponível em: < <https://cran.r-project.org/web/packages/microbenchmark/index.html> >. Acesso em: 01 de junho de 2016.

MEUWISSEN, T. H. E.; LUO, Z. Computing inbreeding coefficients in large populations. **Genetics Selection Evolution**, v. 24, n. 4, p. 1, 1992.

MEYER K. PX×AI: Algorithmics for better convergence in restricted maximum likelihood estimation. **8th World Congress on Genetics Applied to Livestock Production**; Belo Horizonte, Brasil. 2006.

MEYER, K. Estimating variances and covariances for multivariate animal models by restricted maximum likelihood. **Genetics Selection Evolution**, v. 23, n. 1, p. 67-83, 1991.

MEYER, K. Maximum likelihood procedures for estimating genetic parameters for later lactations of dairy cattle. **Journal of Dairy Science**, v. 66, n. 9, p. 1988-1997, 1983.

MEYER, K. Parameter expansion for estimation of reduced rank covariance matrices. **Genetics Selection Evolution**, v. 40, p. 3-24, 2008.

MEYER, K. Performance of REML algorithms in multivariate analyses fitting reduced rank and factor-analytic models. **Proceedings of the Seventeenth Conference for the Advancement of Animal Breeding and Genetics**, p. 280-283, 2007a.

MEYER, K. Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. **Genetics Selection Evolution**, v. 21, n. 3, p. 1-24, 1989.

MEYER, K. Restricted maximum likelihood to estimate variance components for mixed models with two random factors. **Genetics Selection Evolution**, v. 19, n. 1, p. 49-68, 1987.

MEYER, K. WOMBAT—A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). **Journal of Zhejiang University Science B**, v. 8, n. 11, p. 815-821, 2007b.

MEYER, K.; SMITH, S. P. Restricted maximum likelihood estimation for animal models using derivatives of the likelihood. **Genetics Selection Evolution**, v. 28, n. 1, p. 23-49, 1996.

MISZTAL, I. Reliable computing in estimation of variance components. **Journal of animal breeding and genetics**, n. 6, v. 125, p. 363-370, 2008.

MISZTAL, I.; PEREZ-ENCISO, M. Sparse matrix inversion for restricted maximum likelihood estimation of variance components by expectation-maximization. **Journal of dairy science**, v. 76, n. 5, p. 1479-1483, 1993.

MÖHRING, J.; MELCHINGER, A. E.; PIEPHO, H. P. REML-based diallel analysis. **Crop science**, v. 51, n. 2, p. 470-478, 2011.

MRODE, R. A. **Linear models for the prediction of animal breeding values**. Cabi, 2014.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, v. 58, n. 3, p. 545-554, 1971.

PAYNE, R. W. GenStat. **Wiley Interdisciplinary Reviews: Computational Statistics**, v. 1, n. 2, p. 255-258, 2009.

PINHEIRO, J.; BATES, D.; DEBROY, S.; SARKAR, D.; R CORE TEAM. nlme: **Linear and Nonlinear Mixed Effects Models**. R package version 3.1-124, 2016. Disponível em: <<https://cran.r-project.org/web/packages/nlme/index.html>>. Acesso em: 10 de maio de 2016.

R CORE TEAM . R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2016. Disponível em: <<https://www.R-project.org/>>. Acesso em: 10 de maio de 2016.

RAY, D. K.; MUELLER, N. D.; WEST, P. C.; FOLEY, J. A. Yield trends are insufficient to double global crop production by 2050. **PloS one**, v. 8, n. 6, p. e66428, 2013.

RESENDE, M. D. V. **Selegen-Reml/Blup: Sistema Estatístico e Seleção Genética Computadorizada via Modelos Lineares Mistos**. Embrapa Florestas, Colombo, PR, Brazil, 2007.

RESENDE, M. D. V.; SILVA, F. F. E. ; AZEVEDO, C. F. **ESTATÍSTICA MATEMÁTICA, BIOMÉTRICA E COMPUTACIONAL: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência**. 1ª ed. Visconde do Rio Branco: Suprema, 2014, v. 1, p. 448-502.

ROYCHOWDHURY, R.; TAH, J. Evaluation of genetic parameters for agro-metrical characters in carnation genotypes. **African Crop Science Journal**, n. 3, v. 19, p. 183-188, 2011.

SARKAR, D. **Lattice: multivariate data visualization with R**. Springer Science & Business Media, 2008.

SAS Institute Inc. **SAS/STAT 12.1 User's Guide**. SAS Institute Inc., Cary, NC, 2012. Disponível em: <<https://support.sas.com/documentation/onlinedoc/stat/121/intro.pdf>>. Acesso em: 10 de maio de 2016.

SEARLE, S. R.; CASELLA, G.; MCCULLOCH, C. E. **Variance Components**. New York: Wiley, 1992.

TENESA, A.; HALEY, C. S. The heritability of human disease: estimation, uses and abuses. **Nature Reviews Genetics**, v. 14, n. 2, p. 139-149, 2013.

TESTER, M.; LANGRIDGE, P. Breeding Technologies to Increase Crop Production in a Changing World. **Science**, n. 5967, v. 327, p. 821-821, 2010.

THOMPSON, R. Estimation of quantitative genetic parameters. **Proceedings of the Royal Society of London B: Biological Sciences**, v. 275, n. 1635, p. 679-686, 2008.

THOMPSON, R.; BROTHERSTONE, S.; WHITE, I. M. S. Estimation of quantitative genetic parameters. **Philosophical Transactions of the Royal Society of London B: Biological Sciences**, v. 360, n. 1459, p. 1469-1477, 2005.

THOMPSON, R.; CULLIS, B.; SMITH, A.; GILMOUR, A. A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. **Australian & New Zealand Journal of Statistics**, v. 45, n. 4, p. 445-459, 2003.

VIRK, D. S.; PANDIT, D. B.; SUFIAN, M. A.; AHMED, F.; SIDDIQUE, M. A.; SAMAD, M. A.; RAHMAN, M. M.; ISLAM, M. M.; ORTIZ-FERRARA, G.; JOSHI, K. D.; WITCOMBE, J. R. REML is an effective analysis for mixed modelling of unbalanced on-farm varietal trials. **Experimental Agriculture**, 45, 77-91, 2009.

WANG, Y. Maximum likelihood computation based on the Fisher scoring and Gauss–Newton quadratic approximations. **Computational statistics & data analysis**, v. 51, n. 8, p. 3776-3787, 2007.

WOLAK, Matthew E. nadiv: an R package to create relatedness matrices for estimating non-

additive genetic variances in animal models. **Methods in Ecology and Evolution**, v. 3, n. 5, p. 792-796, 2012.

ZHOU, X.; CARBONETTO, P.; STEPHENS, M. Polygenic modeling with Bayesian sparse linear mixed models. **PLoS Genet**, v. 9, n. 2, p. e1003264, 2013.

ZHOU, X.; STEPHENS, M. Efficient algorithms for multivariate linear mixed models in genome-wide association studies. **Nature methods**, v. 11, n. 4, p. 407, 2014.

ZHOU, X.; STEPHENS, M. Genome-wide efficient mixed-model analysis for association studies. **Nature genetics**, v. 44, n. 7, p. 821-824, 2012.

Anexo 1: Configuração dos computadores utilizados nas análises

Tabela 1: Configuração dos computadores utilizados nas análises.

Computador	Processador	MR (GB)	VP (GHz)	SO (bit)	TC
C1	Intel Core I5	4	3,1	Windows 64	Desktop
C2	Intel Core 2 Duo	2	1,5	Windows 64	Notebook
C3	Intel Core I5	4	2,5	Windows 64	Notebook
C4	Intel Xeon	32	2,0	Linux	Servidor

MR - Memória RAM, VP - Velocidade de processamento, SO - Sistema Operacional e Tipo de computador.

Anexo 2: Estimativas de parâmetros para os modelos M1 (com *pedigree*) e M2 (sem *pedigree*)

Tabela 1: Estimativa de parâmetros (Par.) do modelo M1 (com *pedigree*) pelos algoritmos *Average Information* (AI), *Newton Rhapson* (NR), *Expectation Maximization* (EM) e os algoritmos combinados EM-AI e EM-NR para diferentes taxas de erro (TE), considerando as diferenças entre coeficientes de determinação (CD) e componentes de variância (CV) como critério de convergência.

TE	Par.	CD					CV				
		AI	NR	EM	EM-AI	EM-NR	AI	NR	EM	EM-AI	EM-NR
10 ⁻⁵	log(L)	-26192,35	-26192,32	-26192,41	-26192,33	-26192,33	-26192,30	-26192,30	-26192,30	-26192,30	-26192,30
	$\hat{\sigma}_a^2$	2541573,50	2541559,60	2529607,30	2541567,60	2541559,60	2541559,60	2541559,60	2541559,60	2541559,60	2541559,60
	$\hat{\sigma}_{t,p}^2$	266762,70	266764,80	268245,30	266763,60	266764,80	266764,80	266764,80	266764,80	266764,80	266764,80
	$\hat{\sigma}^2$	700538,20	700545,00	706540,50	700541,10	700545,00	700545,00	700545,00	700545,00	700545,00	700545,00
10 ⁻⁶	log(L)	-26192,30	-26192,30	-26192,32	-26192,30	-26192,30	-26192,30	-26192,30	-26192,30	-26192,30	-26192,30
	$\hat{\sigma}_a^2$	2541558,30	2541559,60	2540350,00	2541558,90	2541559,60	2541559,60	2541559,60	2541559,60	2541559,60	2541559,60
	$\hat{\sigma}_{t,p}^2$	266765,00	266764,80	266914,00	266765,00	266764,80	266764,80	266764,80	266764,80	266764,80	266764,80
	$\hat{\sigma}^2$	700545,70	700545,00	701151,80	700545,40	700545,00	700545,00	700545,00	700545,00	700545,00	700545,00
10 ⁻⁷	log(L)	-26192,31	-26192,30	-26192,31	-26192,30	-26192,30			-26192,30		-26192,30
	$\hat{\sigma}_a^2$	2541559,80	2541559,60	2541438,10	2541559,70	2541559,60			2541559,60		2541559,60
	$\hat{\sigma}_{t,p}^2$	266764,80	266764,80	266779,80	266764,80	266764,80	SNMIP	SNMIP	266764,80	SNMIP	266764,80
	$\hat{\sigma}^2$	700544,90	700545,00	700606,00	700545,00	700545,00			700545,00		700545,00
10 ⁻⁸	log(L)	-26192,30	-26192,30	-26192,30	-26192,30	-26192,30			-26192,30		
	$\hat{\sigma}_a^2$	2541559,60	2541559,60	2541547,50	2541559,60	2541559,60			2541559,60		
	$\hat{\sigma}_{t,p}^2$	266764,80	266764,80	266766,30	266764,80	266764,80	SNMIP	SNMIP	266764,80	SNMIP	SNMIP
	$\hat{\sigma}^2$	700545,00	700545,00	700551,10	700545,00	700545,00			700545,00		
10 ⁻⁹	log(L)	-26192,30	-26192,30	-26192,30	-26192,30	-26192,30			-26192,30		
	$\hat{\sigma}_a^2$	2541559,60	2541559,60	2541558,40	2541559,60	2541559,60			2541559,60		
	$\hat{\sigma}_{t,p}^2$	266764,80	266764,80	266765,00	266764,80	266764,80	SNMIP	SNMIP	266764,80	SNMIP	SNMIP
	$\hat{\sigma}^2$	700545,00	700545,00	700545,60	700545,00	700545,00			700545,00		

log(L): logaritmo de REML, $\hat{\sigma}_a^2$: variância genética aditiva, $\hat{\sigma}_{t,p}^2$: variância da interação testador versus população, $\hat{\sigma}^2$: variância ambiental e SNMIP: superou o número máximo de iterações permitido (cinco vezes o número de iterações da taxa de erro imediatamente anterior).

Tabela 2: Estimativa de parâmetros (Par.) do modelo M2 (sem *pedigree*) pelos algoritmos *Average Information* (AI), *Newton Rhapson* (NR), *Expectation Maximization* (EM) e os algoritmos combinados EM-AI e EM-NR para diferentes taxas de erro (TE), considerando as diferenças entre coeficientes de determinação (CD) e componentes de variância (CV) como critério de convergência.

TE	Par.	CD					CV				
		AI	NR	EM	EM-AI	EM-NR	AI	NR	EM	EM-AI	EM-NR
10 ⁻⁵	log(L)	-29003,54	-29003,29	-29003,59	-29003,55	-29003,29	-29003,29	-29003,29	-29003,29	-29003,29	-29003,29
	$\hat{\sigma}_g^2$	393642,50	393613,60	392283,30	393643,20	393613,60	393613,60	393613,60	393613,60	393613,60	393613,60
	$\hat{\sigma}_b^2$	3500344,10	3500343,80	3500593,60	3500344,10	3500343,80	3500343,80	3500343,80	3500343,80	3500343,80	3500343,80
	$\hat{\sigma}^2$	2475837,90	2475864,50	2477152,40	2475837,30	2475864,50	2475864,50	2475864,50	2475864,50	2475864,50	2475864,50
10 ⁻⁶	log(L)	-29003,31	-29003,29	-29003,32	-29003,26	-29003,29	-29003,29	-29003,29	-29003,29	-29003,29	-29003,29
	$\hat{\sigma}_g^2$	393616,60	393613,60	393484,80	393610,90	393613,60	393613,60	393613,60	393613,60	393613,60	393613,60
	$\hat{\sigma}_b^2$	3500343,80	3500343,80	3500368,00	3500343,80	3500343,80	3500343,80	3500343,80	3500343,80	3500343,80	3500343,80
	$\hat{\sigma}^2$	2475861,70	2475864,50	2475989,10	2475867,00	2475864,50	2475864,50	2475864,50	2475864,50	2475864,50	2475864,50
10 ⁻⁷	log(L)	-29003,28	-29003,29	-29003,29	-29003,28	-29003,29			-29003,29		
	$\hat{\sigma}_g^2$	393613,30	393613,60	393600,60	393613,30	393613,60			393613,60		
	$\hat{\sigma}_b^2$	3500343,80	3500343,80	3500346,20	3500343,80	3500343,80	SNMIP	SNMIP	3500343,80	SNMIP	SNMIP
	$\hat{\sigma}^2$	2475864,70	2475864,50	2475877,10	2475864,70	2475864,50			2475864,50		
10 ⁻⁸	log(L)	-29003,29	-29003,29	-29003,29	-29003,29	-29003,29			-29003,29		
	$\hat{\sigma}_g^2$	393613,60	393613,60	393612,30	393613,60	393613,60			393613,60		
	$\hat{\sigma}_b^2$	3500343,80	3500343,80	3500344,00	3500343,80	3500343,80	SNMIP	SNMIP	3500343,80	SNMIP	SNMIP
	$\hat{\sigma}^2$	2475864,50	2475864,50	2475865,80	2475864,50	2475864,50			2475864,50		
10 ⁻⁹	log(L)	-29003,29	-29003,29	-29003,29	-29003,29	-29003,29			-29003,29		
	$\hat{\sigma}_g^2$	393613,60	393613,60	393613,50	393613,60	393613,60			393613,60		
	$\hat{\sigma}_b^2$	3500343,80	3500343,80	3500343,80	3500343,80	3500343,80	SNMIP	SNMIP	3500343,80	SNMIP	SNMIP
	$\hat{\sigma}^2$	2475864,50	2475864,50	2475865,80	2475864,50	2475864,50			2475864,50		

log(L): logaritmo de REML, $\hat{\sigma}_g^2$: variância genética, $\hat{\sigma}_b^2$: variância dos blocos dentro de ensaio, $\hat{\sigma}^2$: variância ambiental e SNMIP: superou o número máximo de iterações permitido (cinco vezes o número de iterações da taxa de erro imediatamente anterior).

CAPITULO 2

Estabilidade de algoritmos REML para o ajuste de modelos mistos desbalanceados aplicados a seleção de híbridos de milho

RESUMO: Desde sua elucidação os modelos mistos vêm conquistando espaço em estudos de genética e de melhoramento genético em aplicações cada vez mais arrojadas. Entretanto, este procedimento pode ser limitado quando há grande número de informações experimentais e modelos complexos, que são situações comuns em programas de melhoramento de milho. Diante do exposto, este trabalho teve por objetivo avaliar a estabilidade de algoritmos REML (máxima verossimilhança residual) no ajuste de modelos desbalanceados aplicados à seleção de híbridos de milho, frente a diferentes estimativas iniciais para os componentes de variância. Os resultados mostraram que a estabilidade dos algoritmos de segunda derivada foi inferior à observada para o algoritmo EM (*Expectation Maximization*) que foi capaz de ajustar todos os modelos estudados. Entretanto, o algoritmo NR (*Newton Raphson*) quando combinado com dez passos EM apresentou percentual de convergência superior a 70% e rápida convergência. A utilização dos algoritmos AI em combinação com o EM deve ser restrita aos pesos uniforme (P1) ou peso que estabelece maior diferenciação entre a variação residual e os demais componentes de variância (P4), além disso, foi necessário grande número de iterações para atingir a convergência denotando baixa eficiência computacional. O peso uniforme (P1) deve ser evitado para os algoritmos de segunda derivada mesmo em combinações entre os algoritmos e na preferência pelo algoritmo NR o peso que denota maior magnitude de variação para o componente de variância ambiental (P4) deve ser evitado.

Palavra chave: Convergência; Eficiência computacional; modelos mistos; *Zea mays*.

1 INTRODUÇÃO

Desde sua primeira descrição formal os modelos mistos (HENDERSON et al., 1959) vêm revolucionando os estudos genéticos e de melhoramento genético junto ao procedimento de máxima verossimilhança residual (PATTERSON; THOMPSON, 1971; THOMPSON, 2008). Atualmente no campo da genética encontramos estimativas de parâmetros genéticos de traços do sangue humano (WRIGHT et al., 2014) e doenças humanas (BULIK-SULLIVAN, 2015) utilizando estas técnicas. No melhoramento genético os modelos mistos são reportados em paralelo ao REML para ensaios em METs (*multi-environment trials*) no contexto fenotípico (KELLY et al., 2007) e genômico (MALOSETTI et al., 2008; OAKEY et al., 2016) visando o aumento da acurácia preditiva.

Visto a importância da estimação e predição em estudos genéticos, o método REML se tornou preferido (JOHNSON; THOMPSON, 1995; MEYER, 2008) em relação ao ML devida a remoção dos graus de liberdade utilizados na estimação dos efeitos fixos (THOMPSON, 2008; TENESA; HALEY, 2013).

Entretanto, apesar da robustez e aplicabilidade de REML, esta técnica pode ser limitada devido à complexidade dos modelos e quantidade de informações experimentais disponíveis, em função da natureza iterativa do método (MISZTAL, 2008; ZHOU; STEPHENS, 2012; ZHOU; STEPHENS, 2014).

Neste contexto, diversos autores vêm propondo algoritmos que apresentam rápida convergência e alta eficiência computacional (PATTERSON; THOMPSON, 1971; SEARLE; CASELLA; MCCULLOCH, 1992; GILMOUR; THOMPSON; CULLIS, 1995; JOHNSON; THOMPSON, 1995; LEE; VAN DER WERF, 2006; STEPHENS, 2012; ZHOU; STEPHENS, 2014). Dentre estes podemos destacar os algoritmos AI (*Average Information*) (JOHNSON; THOMPSON, 1995; GILMOUR; THOMPSON; CULLIS, 1995), FS (*Fisher's Scoring*) (PATTERSON; THOMPSON, 1971) e NR (*Newton Raphson*) (SEARLE; CASELLA; MCCULLOCH, 1992). Estes algoritmos são baseados na segunda derivada da função de verossimilhança residual e apresentam convergência quadrática (MEYER; SMITH, 1996), promovendo rápido aumento do logaritmo de verossimilhança residual [$\log(L)$] e rápida convergência quando comparados com o algoritmo de primeira derivada EM (*Expectation Maximization*). Contudo em alguns casos a convergência falha, devido as estimativas iniciais dos parâmetros do modelo serem inadequadas, gerando estimativas negativas para os componentes de variância (MEYER, 2006; KNIGHT, 2008).

Quando a convergência dos algoritmos citados anteriormente falha uma alternativa é utilizar o algoritmo EM (*Expectation Maximization*) (DEMPSTER; LAIRD; RUBIN, 1977) ou uma de suas variações denominadas de PX (*Parameter Expanded*) (FOULLEY; VAN DYK, 2000; DIFFEY; WELSH; CULLIS, 2013), devido à estabilidade numérica apresentada por estes algoritmos. Entretanto, apesar da estabilidade numérica estes algoritmos são extremamente lentos para atingir a convergência (MEYER, 2006; KNIGHT, 2008).

Para contornar os problemas de falha de convergência dos algoritmos de segunda derivada e aproveitar a estabilidade numérica dos algoritmos de primeira derivada alguns autores vêm utilizando esquemas híbridos entre as duas classes (MEYER, 2006; KNIGHT, 2008; ZHOU; STEPHENS, 2014) visando a melhoria da eficiência computacional.

Os múltiplos estudos visando a melhoria da eficiência computacional de REML se justificam pelas propriedades vantajosas de lidar com experimentos desbalanceados e permitir a modelagem de tratamentos e erros correlacionados (RESENDE; SILVA; AZEVEDO, 2014). O

desbalanceamento é uma situação comum em programas de melhoramento de milho (FIGUEIREDO et al., 2015), onde são desenvolvidos e testados milhares de híbridos em delineamentos não ortogonais capazes de acomodar todos os tratamentos genéticos. Além disso, nestes experimentos existe relacionamento genético entre indivíduos e gradientes ambientais atuando na expressão dos caracteres fenotípicos.

Nesta circunstância os modelos mistos em conjunto com REML se tornam uma poderosa ferramenta para os melhoristas de milho, pois maximizam a acurácia preditiva e por consequência a eficiência de seleção, contribuindo para maximizar o progresso genético da cultura (RESENDE; SILVA; AZEVEDO, 2014) frente a um cenário de incremento populacional e mudanças climáticas que ameaçam a segurança alimentar (TESTER; LANGRIDGE, 2010; FOLEY et al., 2011; RAY et al., 2013).

Diante do exposto, este trabalho teve por objetivo avaliar a estabilidade de algoritmos REML no ajuste de modelos desbalanceados, aplicados a seleção de híbridos de milho, frente a diferentes estimativas iniciais para os componentes de variância.

2 MATERIAL E MÉTODOS

2.1 MODELO

Os modelos mistos estudados foram uma extensão do originalmente proposto por Henderson et al. (1959) que foi descrito em detalhes no primeiro capítulo.

2.2 A FUNÇÃO DE VEROSSIMILHANÇA RESIDUAL

A função de verossimilhança residual utilizada na derivação dos algoritmos de primeira e segunda derivada para o ajuste dos modelos mistos no presente trabalho foi aquela descrita no capítulo um.

2.3 ALGORITMOS

Os algoritmos REML utilizados foram o EM (*Expectation Maximization*) (DEMPSTER; LAIRD; RUBIN, 1977), AI (*Average Information*) (JOHNSON; THOMPSON, 1995; GILMOUR; THOMPSON; CULLIS, 1995), FS (*Fisher's Scoring*) (PATTERSON; THOMPSON, 1971), NR (*Newton Raphson*) (SEARLE; CASELLA; MCCULLOCH, 1992) e combinações entre o EM e os demais algoritmos (EM-AI, EM-FS e EM-NR). Os algoritmos utilizados neste capítulo correspondem as versões otimizadas implementadas em R como foi descrito no capítulo um.

2.3.1 Diagnóstico de convergência

A convergência dos algoritmos foi monitorada através dos coeficientes de determinação (CD) (RESENDE, 2007) como foi descrito no primeiro capítulo. E a taxa de erro (ϕ) estipulada neste trabalho foi de 10^{-5} .

2.4 MATERIAL GENÉTICO ESTUDADO

Os dados experimentais foram obtidos a partir de ensaios montados em blocos aumentados, com um grupo de experimentos que agrupou híbridos topcrosses (TCs) derivados de progênies endogâmicas (S1) do grupo heterótico “Dent” em cruzamento com um testador “Flint”, e outro grupo de experimentos que reuniu híbridos TCs derivados de cruzamentos entre progênies (S1) “Flint” em cruzamento com um testador “Dent”. Os ensaios foram conduzidos no ano agrícola de 2013/14, na Embrapa Milho e Sorgo, localizada no município de Sete Lagoas no estado de Minas Gerais.

Na estrutura de blocos aumentados, utilizou-se 64 híbridos como número básico de tratamentos dentro de cada bloco, incluindo cinco testemunhas comuns a todos os blocos, mas houve experimentos com número diferente. Nessa estrutura, foram avaliados 3352 híbridos TCs (de primeiro ciclo) e mais cinco híbridos comerciais (utilizados como testemunhas comuns), totalizando 3357 tratamentos. As parcelas úteis foram compostas de duas linhas de 4,2 m espaçadas em 0,7 m. A característica analisada foi a produtividade de grãos (PG), em kg ha^{-1} , determinada com base no peso de grãos da parcela, e corrigida para a umidade de 13%.

2.5 MODELOS ESTUDADOS E ANÁLISES REALIZADAS

Para o estudo da estabilidade dos algoritmos foram utilizados oito modelos com diferentes estruturas de efeitos fixos. Portanto, pode-se subdividir estes modelos em três classes. Na primeira delas os blocos foram considerados como fixos, na segunda a média foi o único termo de efeito fixo e na terceira classe as testemunhas foram tratadas como efeito fixo.

As análises realizadas obedeceram ao esquema mostrado na Figura 1.

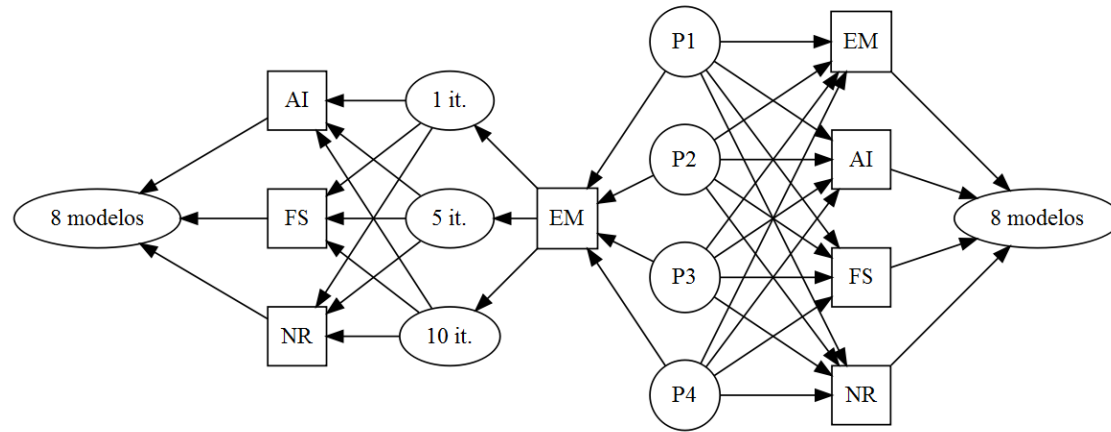


Figura 1: Diagrama do esquema de análise sendo P1, P2, P3 e P4 os pesos dos valores iniciais dos componentes de variância; 1 it., 5 it. e 10 it. indicam o número de iterações e EM (*Expectation Maximization*), AI (*Average Information*), FS (*Fisher's Scoring*) e NR (*Newton-Rapson*) representam os algoritmos utilizados para ajustar os 8 modelos em análise.

Os algoritmos mostrados no diagrama da Figura 1 foram avaliados em relação a quatro pesos para os valores iniciais dos componentes de variância. O primeiro deles foi o uniforme, onde todos os componentes tiveram mesmo peso P1 ($\hat{\sigma}_a^2 = v$, $\hat{\sigma}_d^2 = v$, $\hat{\sigma}_c^2 = v$, $\hat{\sigma}_g^2 = v$, $\hat{\sigma}_b^2 = v$ e $\hat{\sigma}^2 = v$) (KNIGHT, 2008). No segundo conjunto de pesos, o componente de variância associado aos efeitos de dominância assumiu menor importância e a variação ambiental apresentou peso 0,4, sendo P2 ($\hat{\sigma}_a^2 = 0,1v$, $\hat{\sigma}_d^2 = 0,01v$, $\hat{\sigma}_c^2 = 0,1v$, $\hat{\sigma}_g^2 = 0,1v$, $\hat{\sigma}_b^2 = 0,1v$ e $\hat{\sigma}^2 = 0,4v$) devido resultados prévios demonstrarem a menor magnitude do componente de variância associados aos efeitos de dominância. Para P3 ($\hat{\sigma}_a^2 = 0,1v$, $\hat{\sigma}_d^2 = 0,1v$, $\hat{\sigma}_c^2 = 0,1v$, $\hat{\sigma}_g^2 = 0,1v$, $\hat{\sigma}_b^2 = 0,1v$ e $\hat{\sigma}^2 = 0,4v$) apresentou peso de 0,1 para todos os componentes de variância com exceção do ambiental que teve peso de 0,4 em função dos resultados prévios mostrarem que os algoritmos NR e FS apresentam dificuldade de convergência quando o componente de variância residual inicial é muito superior aos demais componentes de variância. Por fim, P4 ($\hat{\sigma}_a^2 = 0,1v$,

$\hat{\sigma}_d^2 = 0,1v$, $\hat{\sigma}_c^2 = 0,1v$, $\hat{\sigma}_g^2 = 0,1v$, $\hat{\sigma}_b^2 = 0,1v$ e $\hat{\sigma}^2 = v$) contemplou o conjunto de pesos iniciais dos componentes de variância utilizados no ASReml (BUTLER et al., 2009). Sendo v metade da variação da variável resposta, $\hat{\sigma}_a^2$: é a variância aditiva, $\hat{\sigma}_d^2$: é a variância de dominância, $\hat{\sigma}_c^2$: é a variância entre testemunhas, $\hat{\sigma}_b^2$: é a variância dos blocos e $\hat{\sigma}^2$: é a variância ambiental.

Além disso, para os algoritmos combinados foram considerados um, cinco e dez passos EM para gerar as estimativas iniciais dos componentes de variância utilizados para dar início aos algoritmos AI, FS e NR. Ademais foram testados oito modelos para a seleção de híbridos de milho de primeiro ciclo de seleção.

Para mensurar a estabilidade dos algoritmos foi avaliado o percentual de convergência para as diversas condições impostas e realizado o agrupamento dos algoritmos e modelos versus pesos para os valores iniciais dos componentes de variância em um Heatmap pelo critério de UPGMA a partir da matriz de coincidência para a convergência entre pares de algoritmos ou pares de modelos versus pesos para os valores iniciais dos componentes de variância. Além disso, para aferir a eficiência computacional dos algoritmos avaliou-se o número de iterações para atingir a convergência.

Todas as análises foram realizadas no software R (R CORE TEAM, 2016) com o auxílio dos pacotes Matrix (BATES; MAECHLER, 2016), nadv (WOLAK, 2012) e lattice (SARKAR, 2008).

A inversa da matriz de parentesco genético aditivo foi obtida através do pacote nadv utilizando o método proposto por Meuwissen e Luo (1992). A matriz de parentesco de dominância também foi obtida através do pacote nadv utilizando o método citado por Wolak (2012).

2.5.1 Modelos com efeitos de bloco fixos somados à média geral

Nesta classe os modelos foram denominados de M1 (1), M2 (2) e M3 (3), onde o primeiro não apresenta informações de parentesco e o último trata as testemunhas como um fator aleatório adicional. Estes modelos podem ser descritos matricialmente como

$$y = Xb + Zg + \varepsilon, \quad (1)$$

$$y = Xb + Za + Zd + \varepsilon, \quad (2)$$

$$y = Xb + Za + Zd + Tc + \varepsilon, \quad (3) \text{ onde } y: \text{ é o vetor de dados de produção, } b: \text{ é o}$$

vetor de efeitos de bloco (fixos e somados a média geral) , g : é o vetor de efeitos genéticos (aleatórios) com $g \sim N(0, I^{\sigma_g^2})$, a : é o vetor de efeitos genéticos aditivos (aleatórios) com $a \sim N(0, A^{\sigma_a^2})$, d : é o vetor de efeitos genéticos de dominância (aleatórios) com $d \sim N(0, D^{\sigma_d^2})$, c : é o vetor de efeitos de testemunha (aleatórios) com $c \sim N(0, I^{\sigma_c^2})$, ε : é o vetor de erros (aleatórios) com $\varepsilon \sim N(0, I^{\sigma^2})$, X : é a matriz de incidência para b , Z : é a matriz de incidência para g , a e d e T : é a matriz de incidência para c .

2.5.2 Modelos com efeito fixo somente para a média geral

Para esta classe os modelos foram denominados de M4 (4), M5 (5) e M6 (6), onde o primeiro não apresenta informações de parentesco e o último trata as testemunhas como um fator aleatório adicional. Estes modelos podem ser descritos matricialmente como

$$y = Xu + Zg + Wb + \varepsilon, \quad (4)$$

$$y = Xu + Za + Zd + Wb + \varepsilon, \quad (5)$$

$$y = Xu + Za + Zd + Tc + Wb + \varepsilon, \quad (6) \text{ onde } y: \text{ é o vetor de dados de produção, } u: \text{ é a}$$

média geral (fixa), g : é o vetor de efeitos genéticos (aleatórios) com $g \sim N(0, I^{\sigma_g^2})$, a : é o vetor de efeitos genéticos aditivos (aleatórios) com $a \sim N(0, A^{\sigma_a^2})$, d : é o vetor de efeitos genéticos de dominância (aleatórios) com $d \sim N(0, D^{\sigma_d^2})$, c : é o vetor de efeitos de testemunha (aleatórios) com $c \sim N(0, I^{\sigma_c^2})$, b : é o vetor de efeitos de bloco (aleatórios) com $b \sim N(0, I^{\sigma_b^2})$, ε : é o vetor de erros (aleatórios) com $\varepsilon \sim N(0, I^{\sigma^2})$, X : é a matriz de incidência para u , Z : é a matriz de incidência para g , a e d , T : é a matriz de incidência para c e W : é a matriz de incidência para b .

2.5.3 Modelos com efeitos de testemunha fixos somados a média geral

Nesta classe os modelos foram denominados de M7 (7) e M8 (8) onde o primeiro não apresenta informações de parentesco e o último trata as testemunhas como um fator de efeito fixo. Estes modelos podem ser descritos matricialmente como

$$y = Xc + Zg + Wb + \varepsilon, \quad (7)$$

$$y = Xc + Za + Zd + Wb + \varepsilon, \quad (8) \text{ onde } y: \text{ é o vetor de dados de produção, } c: \text{ é o}$$

vetor de efeitos de testemunha (fixos e somados a média geral), g : é o vetor de efeitos genéticos (aleatórios) com $g \sim N(0, I \sigma_g^2)$, a : é o vetor de efeitos genéticos aditivos (aleatórios) com $a \sim N(0, A \sigma_a^2)$, d : é o vetor de efeitos genéticos de dominância (aleatórios) com $d \sim N(0, D \sigma_d^2)$, b : é o vetor de efeitos de bloco (aleatórios) com $b \sim N(0, I \sigma_b^2)$, ε : é o vetor de erros (aleatórios) com $\varepsilon \sim N(0, I \sigma^2)$, X : é a matriz de incidência para c , Z : é a matriz de incidência para g , a e d e W : é a matriz de incidência para b .

3 RESULTADOS

3.1 CONVERGÊNCIA DOS ALGORITMOS

O algoritmo FS e as combinações EM1-FS, EM5-FS e EM10-FS não convergiram para nenhum dos modelos e pesos iniciais para os componentes de variância. Não obstante, os demais algoritmos apresentaram boa taxa de convergência, somente para os pesos P2, P3 e P4 (Tabela 1). Adicionalmente, observou-se que o algoritmo NR apresentou valores positivos para os componentes de variância em P1, porém imprecisos.

Tabela 1: Porcentagem de convergência dos algoritmos *Average Information* (AI), *Newton-Rapson* (NR), *Expectation Maximization* (EM) e as combinações entre EM e aos demais algoritmos considerando uma iteração EM (EM1-AI, EM1-FS e EM1-NR), cinco iterações EM (EM5-AI, EM5-FS e EM5-NR) e dez iterações EM (EM10-AI, EM10-FS e EM10-NR) em função dos pesos dos valores iniciais dos componentes de variância para os oito modelos utilizados.

Algoritmos	Pesos [†]				Média (Algoritmos)
	P1	P2	P3	P4	
AI	0,0	87,5	87,5	87,5	65,63
NR	0,0	87,5	87,5	25,0	50,00
EM1-AI	0,0	87,5	50,0	87,5	56,25

EM1-NR	0,0	87,5	87,5	87,5	65,63
EM5-AI	0,0	87,5	50,0	87,5	56,25
EM5-NR	0,0	87,5	87,5	87,5	65,63
EM10-AI	0,0	87,5	50,0	87,5	56,25
EM10-NR	25,0	87,5	87,5	87,5	71,88
EM	100,0	100,0	100,0	100,0	100,00
Média (Pesos)	13,89	88,89	76,39	81,94	65,28*

†Pesos dos valores iniciais dos componentes de variância. Onde P1 ($\hat{\sigma}_a^2 = v$, $\hat{\sigma}_d^2 = v$, $\hat{\sigma}_c^2 = v$, $\hat{\sigma}_g^2 = v$, $\hat{\sigma}_b^2 = v$ e $\hat{\sigma}^2 = v$), P2 ($0,1 \hat{\sigma}_a^2 = 0,1 v$, $\hat{\sigma}_d^2 = 0,01 v$, $\hat{\sigma}_c^2 = 0,1 v$, $\hat{\sigma}_g^2 = 0,1 v$, $\hat{\sigma}_b^2 = 0,1 v$ e $\hat{\sigma}^2 = 0,4 v$), P3 ($0,1 \hat{\sigma}_a^2 = 0,1 v$, $\hat{\sigma}_d^2 = 0,1 v$, $\hat{\sigma}_c^2 = 0,1 v$, $\hat{\sigma}_g^2 = 0,1 v$, $\hat{\sigma}_b^2 = 0,1 v$ e $\hat{\sigma}^2 = 0,4 v$) e P4 ($0,1 \hat{\sigma}_a^2 = 0,1 v$, $\hat{\sigma}_d^2 = 0,1 v$, $\hat{\sigma}_c^2 = 0,1 v$, $\hat{\sigma}_g^2 = 0,1 v$, $\hat{\sigma}_b^2 = 0,1 v$ e $\hat{\sigma}^2 = v$). Sendo v metade da variação de y , $\hat{\sigma}_a^2$: variância aditiva, $\hat{\sigma}_d^2$: variância de dominância, $\hat{\sigma}_c^2$: variância entre testemunhas, $\hat{\sigma}_b^2$: variância dos blocos e $\hat{\sigma}^2$: variância ambiental. *Média geral.

O peso P1, ou uniforme para os valores iniciais dos componentes de variância, só apresentou convergência para os algoritmos EM e NR após 10 passos EM (EM10-NR). Dentro de P2 a convergência foi uniforme de 87,5% para todos os algoritmos, exceto, o EM que apresentou 100% de convergência. O peso P3 favoreceu a convergência do algoritmo NR e suas combinações com o algoritmo EM, que apresentaram média de 87,5% de convergência (NR = 87,5% versus EM1-NR = 87,5%, EM5-NR = 87,5%, EM10-NR = 87,5%), em contraste ao AI e suas combinações com EM (AI = 87,5% versus EM1-AI = 50%, EM5-AI = 50%, EM10-AI = 50%). No peso P4 o algoritmo NR foi desfavorecido e apresentou 25% de convergência (Tabela 1).

O algoritmo EM10-NR apresentou similaridade de convergência com o algoritmo EM alcançado a convergência em modelos onde somente o algoritmo EM obteve êxito (Figura 2). Além disso, o modelo M7 convergiu somente para o algoritmo EM.

Observando-se o algoritmo NR verifica-se sua ineficiência para o ajuste dos modelos propostos quando todos os componentes de variância apresentam mesma magnitude de valores para as estimativas iniciais (P1). Curiosamente o padrão de estimativas iniciais dos componentes de variância do *software* ASReml (P4) também foram ineficientes para o ajuste dos modelos propostos através do algoritmo NR, que não convergiu para nenhum dos modelos com informação de parentesco (M2, M3, M5, M6 e M8) (Figura 2).

O algoritmo AI teve sua eficiência reduzida quando combinado com o algoritmo EM utilizando o peso P3 para obter as estimativas iniciais dos componentes de variância, não atingindo convergência para os modelos com informação de parentesco.

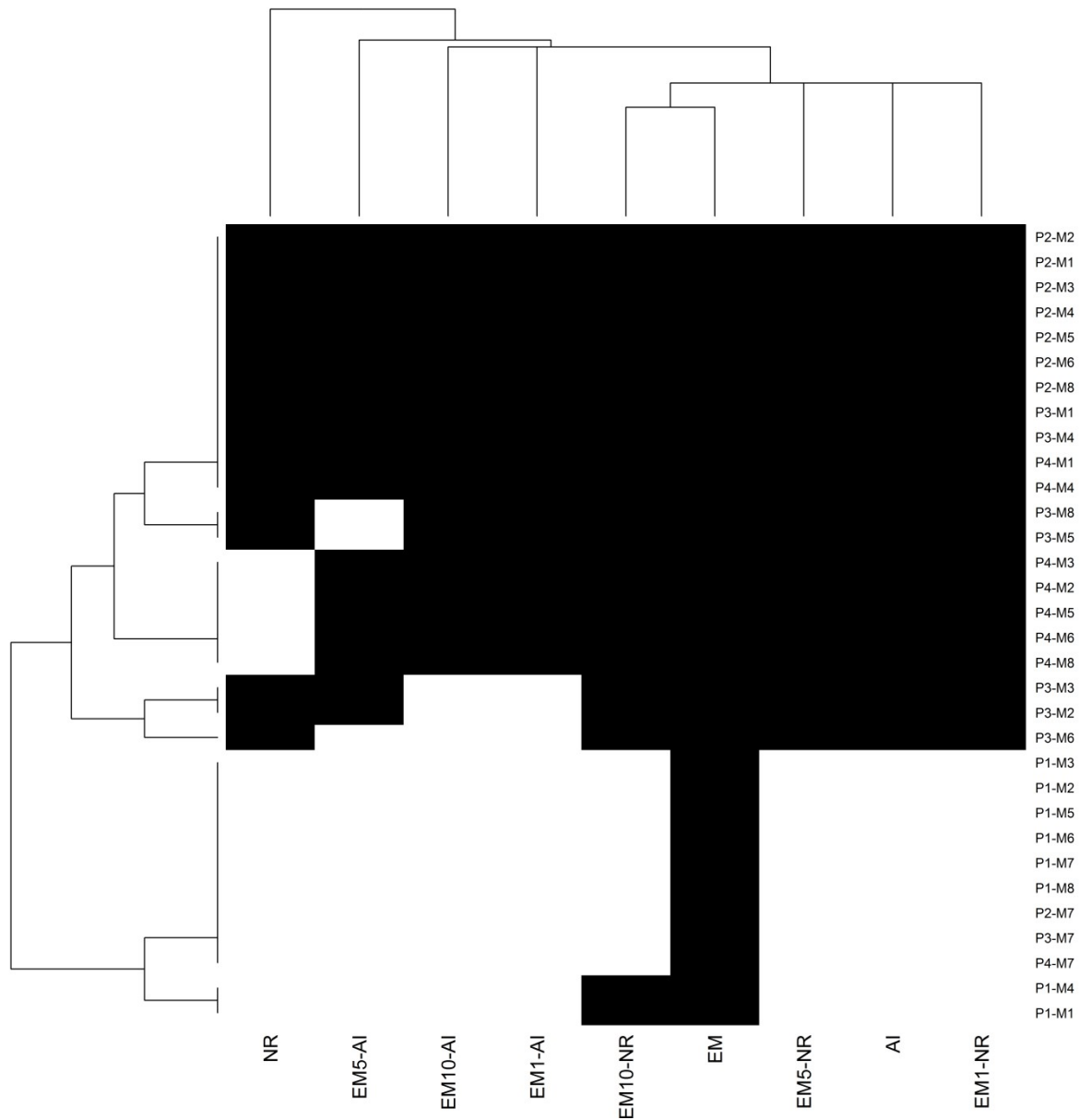


Figura 2: Heatmap da convergência dos algoritmos *Average Information* (AI), *Newton-Rapson* (NR), *Expectation Maximization* (EM) e combinações entre EM e aos demais algoritmos considerando uma iteração EM (EM1-AI, EM1-FS e EM1-NR), cinco iterações EM (EM5-AI,

EM5-FS e EM5-NR) e dez iterações EM (EM10-AI, EM10-FS e EM10-NR) em função dos pesos dos valores iniciais dos componentes de variância (P1, P2, P3 e P4) para oito modelos (M1 à M8). A região gráfica destacada em cor preta no gráfico representa a convergência e a região branca a ausência de convergência.

3.2 NÚMERO DE ITERAÇÕES PARA ATINGIR A CONVERGÊNCIA

O algoritmo EM convergiu para todos os modelos e pesos dos valores iniciais dos componentes de variância, entretanto, apresentou elevada lentidão com iterações variando de 102 à 1167. Os demais algoritmos apresentaram convergência mais rápida com número de iterações variando de 3 a 78 (Figuras 2 e 3 e Tabela 1 - Anexo 1).

Nas combinações entre pesos para os valores iniciais dos componentes de variância e modelos onde foi evidenciada a convergência, os algoritmos NR e suas combinações com o algoritmo EM (EM1-NR, EM5-NR, EM10-NR) convergiram mais rapidamente que os algoritmos AI e suas combinações (EM1-AI, EM5-AI, EM10-AI) (Figura 3). No primeiro caso a convergência ocorreu com número de iterações variando de 3 a 13 contra 20 à 78 no segundo caso (Tabela 1 - Anexo 1).

3.3 ESTIMATIVAS DE PARÂMETROS

Os parâmetros estimados pelos algoritmos estudados para todos os modelos foram extremamente uniformes e apresentaram baixo desvio padrão em relação aos pesos dos valores iniciais estipulados para os componentes de variância. Observa-se apenas uma leve diferença entre o algoritmo EM e os demais algoritmos para os parâmetros \hat{h}_a^2 nos modelos M2, M5 e M8, \hat{h}_d^2 nos modelos M3, M6 e M8, \hat{c}_c^2 no modelo M6 (Tabela 1 - Anexo 2).

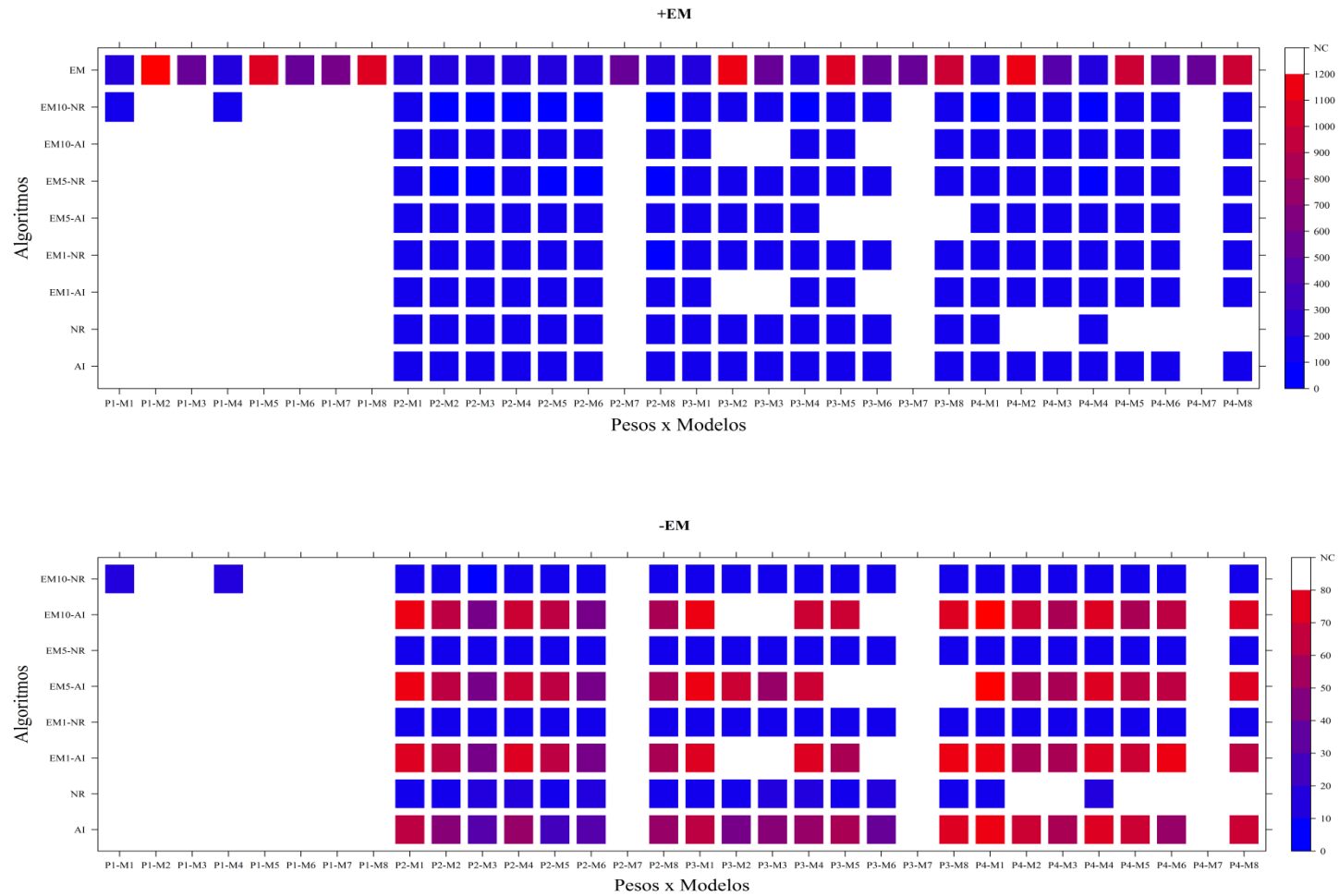


Figura 3: Número de iterações para a convergência dos algoritmos *Average Information* (AI), *Newton-Rapson* (NR), *Expectation Maximization* (EM) e as combinações entre o algoritmo EM e aos demais algoritmos considerando uma iteração EM (EM1-AI e EM1-NR), cinco iterações EM (EM5-AI e EM5-NR) e dez iterações EM (EM10-AI e EM10-NR) em função dos pesos dos valores iniciais dos componentes de variância para oito modelos (+EM). A figura (-EM) possui todos os algoritmos descritos anteriormente exceto o EM para melhorar a discriminação dos resultados dos demais algoritmos. NC: não convergiu.

4 DISCUSSÃO

O algoritmo FS e as combinações EM1-FS, EM5-FS e EM10-FS apresentaram falha de convergência para todos os modelos e pesos iniciais para os componentes de variância. Este fato está associado à natureza dos algoritmos de segunda derivada que apresentam convergência quadrática, mas nem sempre garantem o aumento de $\log(L)$ a cada iteração e, as vezes, geram estimativas fora do espaço paramétrico (MEYER; SMITH, 1996; MEYER, 2006; RESENDE; SILVA; AZEVEDO, 2014). Além disso, o algoritmo NR conduziu a estimativas restritas ao espaço paramétrico, porém, imprecisas considerando os pesos uniformes para os componentes de variância. Misztal (2008) comenta a respeito deste comportamento no algoritmo AI e atribuem este fato a imprecisão das estimativas iniciais dos parâmetros.

O peso uniforme para $\sigma_i^2 = v$ (P1) para os valores iniciais só apresentou convergências válidas para os algoritmos EM e EM10-NR. Estes resultados corroboram com os resultados de Meyer (2006) e Knight (2008) que trabalharam com o algoritmo AI. Estes autores afirmam que para evitar falhas de convergência ao se utilizar valores arbitrários para os componentes de variância, uma boa estratégia é aproveitar a estabilidade numérica do algoritmo EM nas primeiras iterações combinando o EM com algoritmos de segunda derivada.

Dentre os diferentes pesos associados às estimativas iniciais dos componentes de variância, P2 apresentou maior média de convergência. Este resultado está atrelado a menor importância atribuída a estimativa inicial do componente de variância associado aos efeitos de dominância ($\hat{\sigma}_d^2 = 0,01 v$), devido a sua menor contribuição na variação fenotípica total.

Curiosamente, as versões combinadas do algoritmo AI com o EM foram prejudicadas nos em P3 em contraste ao próprio AI, pois o número de passos EM testados foi insuficiente para fornecer uma estimativa mais acurada que as fornecidas por P3. Ainda observou-se para P4 que o algoritmo NR foi desfavorecido, pelo fato deste peso atribuir valor dez vezes superior para a variância residual em detrimento as demais.

Dos algoritmos citados o EM apresentou percentual máximo de convergência, evidenciando a estabilidade numérica do mesmo e sua eficiência em aumentar progressivamente $\log(L)$ até atingir a convergência (MEYER, 2006; RESENDE; SILVA; AZEVEDO, 2014). Não obstante, este algoritmo foi o que apresentou convergência mais lenta devido ao grande número de iterações. Este fato está ligado a propriedade de convergência linear do EM (MEYER, 2006) e limita a sua ampla utilização mesmo nas versões PX.

O algoritmo EM10-NR apresentou maior percentual de convergência favorecendo a convergência e favoreceu a convergência em grande número de modelos corroborando com Meyer (2006) e Knight (2008).

O número de iterações necessário para a convergência foi inferior para os algoritmos NR e suas variações combinadas em relação ao AI e suas combinações com o algoritmo EM demonstrando maior eficiência computacional dos primeiros. Lindstrom e Bates (1988) observaram comportamento similar para o algoritmo NR, que convergiu com menos de seis iterações. Além disso, Zhou e Stephens (2014) mencionam a utilização do algoritmo NR em combinação com o algoritmo PXEM no desenvolvimento de um *software* de alta performance para estudos de associação genômica.

Os parâmetros estimados pelos diferentes algoritmos e pesos para as estimativas iniciais dos componentes de variância se mostraram uniformes, com exceção de sensíveis variações observadas para o algoritmo EM. Estes resultados demonstram a acurácia dos algoritmos para a maximização da função de verossimilhança residual para os modelos testados.

5 CONCLUSÃO

A estabilidade dos algoritmos de segunda derivada foi inferior à observada para o algoritmo EM, que foi capaz de ajustar todos os modelos estudados. Entretanto, o algoritmo NR quando combinado com dez passos EM apresentou percentual de convergência superior a 70% e rápida convergência.

A utilização do algoritmo AI e suas combinações com o algoritmo EM devem ser restritas aos pesos iniciais dos componentes de variância P2, quando a variação devida aos efeitos de dominância apresentar pequena magnitude, ou P4 que já foi relatado para este algoritmo, além disso, foi necessário grande número de iterações para atingir a convergência.

O peso uniforme para os componentes de variância $\sigma_i^2 = v$ (metade da variação de $y = P1$) deve ser evitado para os algoritmos de segunda derivada mesmo em combinações com o EM e na preferência pelo algoritmo NR o peso inicial dos componentes de variância P4 ($\sigma^2 = v$ e $\sigma_i^2 = 0,1v$) deve ser evitado.

6 REFERÊNCIAS

BATES, D.; MAECHLER, M. **Matrix: Sparse and Dense Matrix Classes and Methods**. R package version 1.2-3, 2016. Disponível em: <<https://cran.r-project.org/web/packages/Matrix/index.html>>. Acesso em: 10 de maio de 2016.

BULIK-SULLIVAN, B.; FINUCANE, H. K.; ANTTILA, V.; GUSEV, A.; DAY, F. R.; PO-RU LOH, P.-R.; CONSORTIUM, R.; CONSORTIUM, P. G.; CONSORTIUM, G. C. A. N. W. T. C.; DUNCAN, L.; PERRY, J. R. B.; PATTERSON, N.; ROBINSON, E. B.; DALY, M. J.; PRICE, A. L.; NEALE, B. M. BULIK-SULLIVAN. An atlas of genetic correlations across human diseases and traits. **Nature genetics**, v. 47, p. 1236-1241, 2015.

BUTLER, D. G.; CULLIS, B. R.; GILMOUR, A. R.; GOGEL, B. J. **ASReml-R reference manual: mixed models for S language environments**. Queensland Department of Primary Industries, Queensland, Australia, 2009. Disponível em: <<http://discoveryfoundation.org.uk/downloads/asreml/release3/asreml-R.pdf>>. Acesso em: 10 maio de 2016.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the royal statistical society**, v. 39, n.1, p. 1-38, 1977.

DIFFEY, S.; WELSH, A.; CULLIS, B. A faster and computationally more efficient REML (PX)EM algorithm for linear mixed models. **Centre for Statistical and Survey Methodology**, v. 8, p. 2-13, 2013.

FIGUEIREDO, A. G.; VON PINHO, R. G.; SILVA, H. D.; BALESTRE, M. Application of mixed models for evaluating stability and adaptability of maize using unbalanced data. **Euphytica**, v. 202, n. 3, p. 393-409, 2015.

FOLEY, J. A.; RAMANKUTTY, N.; BRAUMAN, K. A.; CASSIDY, E. S.; GERBER, J. S.; JOHNSTON, M.; MUELLER, N. D.; O'CONNELL, C.; RAY, D. K.; WEST, D. K.; BALZER, C.; BENNETT, E. M.; CARPENTER, S. R.; HILL, J.; MONFREDA, C.; POLASKY, S.; ROCKSTRÖM, J.; SHEEHAN, J.; SIEBERT, S.; TILMAN, D.; ZAKS, D. P. M. Solutions for a cultivated planet. **Nature**, v. 478, n. 7369, p. 337-342, 2011.

FOULLEY, J.-L.; VAN DYK, D. A. The PX-EM algorithm for fast stable fitting of Henderson's mixed model. **Genetics Selection Evolution**, v. 32, n. 2, p. 1-21, 2000.

GILMOUR, A. R.; THOMPSON, R.; CULLIS, B. R. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. **Biometrics**, v. 51, n. 4, p. 1440-1450, 1995.

HENDERSON, C. R.; KEMPTHORNE, O.; SEARLE, S. R.; VON KROSIGK, C. M. The estimation of environmental and genetic trends from records subject to culling. **Biometrics**, v. 15, n. 2, p. 192-218, 1959.

JOHNSON, D. L.; THOMPSON, R. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. **Journal of dairy science**, v. 78, n. 2, p. 449-456, 1995.

KELLY, A. M.; SMITH, A. B.; ECCLESTON, J. A.; CULLIS, B. R. The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. **Crop Science**, v. 47, n. 3, p. 1063-1070, 2007.

KNIGHT, E. **Improved iterative schemes for REML estimation of variance parameters in linear mixed models**. 2008. 298p. Tese de Doutorado, University of Adelaide, Adelaide, 2008.

LEE, S. H.; VAN DER WERF, J. H. J. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. **Genetics Selection Evolution**, v. 38, n. 1, p. 1-19, 2006.

MALOSETTI, M.; RIBAUT, J. M.; VARGAS, M.; CROSSA, J.; VAN EEUWIJK, F. A. A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (*Zea mays* L.). **Euphytica**, v. 161, n. 1-2, p. 241-257, 2008.

MEUWISSEN, T. H. E.; LUO, Z. Computing inbreeding coefficients in large populations. **Genetics Selection Evolution**, v. 24, n. 4, p. 1, 1992.

MEYER K. PX×AI: Algorithmics for better convergence in restricted maximum likelihood estimation. **8th World Congress on Genetics Applied to Livestock Production**; Belo Horizonte, Brasil. 2006.

MEYER, K. Parameter expansion for estimation of reduced rank covariance matrices. **Genetics Selection Evolution**, v. 40, p. 3-24, 2008.

MEYER, K.; SMITH, S. P. Restricted maximum likelihood estimation for animal models using derivatives of the likelihood. **Genetics Selection Evolution**, v. 28, n. 1, p. 23-49, 1996.

MISZTAL, I. Reliable computing in estimation of variance components. **Journal of animal breeding and genetics**, n. 6, v. 125, p. 363-370, 2008.

OAKEY, H.; CULLIS, B.; THOMPSON, R.; COMADRAN, J.; HALPIN, C.; WAUGH, R. Genomic Selection in Multi-environment Crop Trials. **G3: Genes| Genomes| Genetics**, v. 6, n. 5, p. 1313-1326, 2016.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, v. 58, n. 3, p. 545-554, 1971.

- R CORE TEAM . R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2016. Disponível em: <<https://www.R-project.org/>>. Acesso em: 10 de maio de 2016.
- RAY, D. K.; MUELLER, N. D.; WEST, P. C.; FOLEY, J. A. Yield trends are insufficient to double global crop production by 2050. **PloS one**, v. 8, n. 6, p. e66428, 2013.
- RESENDE, M. D. V. **Selegen-Reml/Blup: Sistema Estatístico e Seleção Genética Computadorizada via Modelos Lineares Mistos**. Embrapa Florestas, Colombo, PR, Brazil, 2007.
- RESENDE, M. D. V.; SILVA, F. F. E. ; AZEVEDO, C. F. **ESTATÍSTICA MATEMÁTICA, BIOMÉTRICA E COMPUTACIONAL: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência**. 1ª ed. Visconde do Rio Branco: Suprema, 2014, v. 1, p. 448-502.
- SARKAR, D. **Lattice: multivariate data visualization with R**. Springer Science & Business Media, 2008.
- SEARLE, S. R.; CASELLA, G.; MCCULLOCH, C. E. **Variance Components**. New York: Wiley, 1992.
- TENESA, A.; HALEY, C. S. The heritability of human disease: estimation, uses and abuses. **Nature Reviews Genetics**, v. 14, n. 2, p. 139-149, 2013.
- TESTER, M.; LANGRIDGE, P. Breeding Technologies to Increase Crop Production in a Changing World. **Science**, n. 5967, v. 327, p. 821-821, 2010.
- THOMPSON, R. Estimation of quantitative genetic parameters. **Proceedings of the Royal Society of London B: Biological Sciences**, v. 275, n. 1635, p. 679-686, 2008.
- WOLAK, Matthew E. nadiv: an R package to create relatedness matrices for estimating non-additive genetic variances in animal models. **Methods in Ecology and Evolution**, v. 3, n. 5, p. 792-796, 2012.
- WRIGHT, F. A.; SULLIVAN, P. F.; BROOKS A. I.; ZOU, F.; SUN, W.; XIA, K.; MADAR, V.; JANSEN, R.; CHUNG, W.; ZHOU, Y.-H.; ABDELLAOUI, A.; BATISTA, S.; BUTLER, C.; CHEN, G.; CHEN, T.-H.; D'AMBROSIO, D.; GALLINS, P.; HA, M. J.; HOTTENGA, J. J.; HUANG, S.; KATTENBERG, M.; KOCHAR, J.; MIDDELDORP, C. M.; QU, A.; SHABALIN, A.; TISCHFIELD, J.; TODD, L.; TZENG, J.-Y.;

GROOTHEEST, G. V.; VINK, J. M.; WANG, Q.; WANG, W.; WANG, W.; WILLEMSSEN, G.; SMIT, J. H.; GEUS, E. J.; YIN, Z.; PENNINX, B. J. H.; BOOMSMA, D. I. Heritability and genomics of gene expression in peripheral blood. **Nature genetics**, v. 46, n. 5, p. 430-437, 2014.

ZHOU, X.; STEPHENS, M. Efficient algorithms for multivariate linear mixed models in genome-wide association studies. **Nature methods**, v. 11, n. 4, p. 407, 2014.

ZHOU, X.; STEPHENS, M. Genome-wide efficient mixed-model analysis for association studies. **Nature genetics**, v. 44, n. 7, p. 821-824, 2012.

Anexo 1

Tabela 1: Número de iterações para a convergência dos algoritmos *Average Information* (AI), *Newton-Rapson* (NR), *Expectation Maximization* (EM) e as combinações entre o algoritmo EM e os demais algoritmos considerando uma iteração EM (EM1-AI e EM1-NR), cinco iterações EM (EM5-AI e EM5-NR) e dez iterações EM (EM10-AI e EM10-NR) em função dos pesos dos valores iniciais dos componentes de variância para oito modelos.

P x M [†]	Algoritmos								
	AI	NR	EM1-AI	EM1-NR	EM5-AI	EM5-NR	EM10-AI	EM10-NR	EM
P1-M1	NC	NC	NC	NC	NC	NC	NC	12	142
P1-M2	NC	NC	NC	NC	NC	NC	NC	NC	1167
P1-M3	NC	NC	NC	NC	NC	NC	NC	NC	469
P1-M4	NC	NC	NC	NC	NC	NC	NC	12	125
P1-M5	NC	NC	NC	NC	NC	NC	NC	NC	1016
P1-M6	NC	NC	NC	NC	NC	NC	NC	NC	452
P1-M7	NC	NC	NC	NC	NC	NC	NC	NC	512
P1-M8	NC	NC	NC	NC	NC	NC	NC	NC	1012
P2-M1	59	6	72	5	76	5	74	5	119
P2-M2	41	6	61	5	60	4	60	4	131
P2-M3	28	13	39	6	37	4	37	3	108
P2-M4	51	11	71	6	64	5	67	4	102
P2-M5	20	7	60	5	59	4	57	4	127
P2-M6	29	13	39	6	39	4	40	4	110
P2-M7	NC	NC	NC	NC	NC	NC	NC	NC	494
P2-M8	48	7	52	4	53	4	54	4	150
P3-M1	59	6	72	5	76	5	74	5	119
P3-M2	36	5	NC	6	65	6	NC	6	1154

P3-M3	45	13	NC	8	48	6	NC	6	453
P3-M4	51	11	71	6	64	5	67	4	102
P3-M5	56	7	56	6	NC	6	64	5	1000
P3-M6	31	13	NC	8	NC	7	NC	6	433
P3-M7	NC	NC	NC	NC	NC	NC	NC	NC	494
P3-M8	68	7	76	8	NC	7	70	6	993
P4-M1	76	6	77	5	78	5	78	4	123
P4-M2	62	NC	54	6	55	6	63	6	1120
P4-M3	53	NC	55	6	55	5	53	5	408
P4-M4	70	11	70	5	69	4	69	4	107
P4-M5	62	NC	64	6	60	6	56	6	957
P4-M6	48	NC	73	6	57	5	57	5	379
P4-M7	NC	NC	NC	NC	NC	NC	NC	NC	456
P4-M8	62	NC	58	6	69	6	68	6	940

†Pesos x Modelos, NC: não convergiu.

Anexo 2

Tabela 1: Estimativas de parâmetros através dos algoritmos *Average Information* (AI), *Newton-Rapson* (NR), *Expectation Maximization* (EM) e as combinações entre o algoritmo EM e os demais algoritmos considerando uma iteração EM (EM1-AI e EM1-NR), cinco iterações EM (EM5-AI e EM5-NR) e dez iterações EM (EM10-AI e EM10-NR) e desvios padrões entre parênteses em função dos pesos dos valores iniciais dos componentes de variância para oito modelos.

Classe I						
	M1(\hat{h}_g^2)	M2(\hat{h}_a^2)	M2(\hat{h}_d^2)	M3(\hat{h}_a^2)	M3(\hat{h}_d^2)	M3(\hat{c}_c^2)
AI	0,14 (5,3e ⁻⁷)	0,27 (5,8e ⁻⁶)	0,03 (6,0e ⁻⁶)	0,05 (5,9e ⁻⁶)	0,00 (5,7e ⁻⁶)	0,79 (1,3e ⁻⁶)
EM	0,14 (2,0e ⁻⁴)	0,26 (8,4e ⁻³)	0,03 (8,9e ⁻³)	0,05 (2,8e ⁻³)	0,01 (3,6e ⁻³)	0,79 (8,7e ⁻⁴)
EM1-AI	0,14 (5,6e ⁻⁶)	0,27 (1,3e ⁻⁸)	0,03 (1,3e ⁻⁸)	0,05 (7,4e ⁻⁸)	0,00 (7,2e ⁻⁸)	0,79 (1,6e ⁻⁸)
EM1-NR	0,14 (1,7e ⁻⁶)	0,27 (7,7e ⁻⁸)	0,03 (7,4e ⁻⁸)	0,05 (8,1e ⁻⁷)	0,00 (5,9e ⁻⁷)	0,79 (4,2e ⁻⁶)
EM5-AI	0,14 (9,6e ⁻⁸)	0,27 (5,6e ⁻⁶)	0,03 (5,8e ⁻⁶)	0,05 (5,7e ⁻⁶)	0,00 (5,5e ⁻⁶)	0,79 (1,2e ⁻⁶)
EM5-NR	0,14 (5,1e ⁻⁸)	0,27 (2,8e ⁻⁶)	0,03 (6,2e ⁻⁷)	0,05 (1,2e ⁻⁶)	0,00 (3,5e ⁻⁷)	0,79 (1,9e ⁻⁶)
EM10-AI	0,14 (5,7e ⁻⁶)	0,27 (6,7e ⁻⁶)	0,03 (7,0e ⁻⁶)	0,05 (6,4e ⁻⁸)	0,00 (6,2e ⁻⁸)	0,79 (1,4e ⁻⁸)
EM10-NR	0,14 (1,5e ⁻⁶)	0,27 (6,3e ⁻⁷)	0,03 (1,4e ⁻⁷)	0,05 (4,6e ⁻⁶)	0,00 (1,9e ⁻⁶)	0,79 (1,8e ⁻⁶)
NR	0,14 (1,9e ⁻⁶)	0,27 (4,8e ⁻⁷)	0,03 (3,8e ⁻⁷)	0,05 (3,0e ⁻⁸)	0,00 (2,8e ⁻⁹)	0,79 (1,3e ⁻⁷)
Classe II						
	M4(\hat{h}_g^2)	M4(\hat{c}_b^2)	M5(\hat{h}_a^2)	M5(\hat{h}_d^2)	M5(\hat{c}_b^2)	----
AI	0,06 (5,7e ⁻⁶)	0,55 (2,3e ⁻⁷)	0,22 (5,6e ⁻⁶)	0,02 (4,5e ⁻⁶)	0,22 (3,5e ⁻⁷)	----
EM	0,06 (2,1e ⁻⁴)	0,55 (2,1e ⁻⁵)	0,21 (6,2e ⁻³)	0,02 (6,5e ⁻³)	0,22 (7,1e ⁻⁴)	----
EM1-AI	0,06 (3,3e ⁻⁸)	0,55 (1,2e ⁻⁹)	0,22 (4,4e ⁻⁴)	0,02 (7,5e ⁻⁴)	0,22 (7,4e ⁻⁶)	----
EM1-NR	0,06 (1,2e ⁻⁷)	0,55 (3,0e ⁻⁷)	0,22 (2,0e ⁻⁷)	0,02 (5,0e ⁻⁸)	0,22 (3,6e ⁻⁸)	----
EM5-AI	0,06 (5,9e ⁻⁶)	0,55 (2,3e ⁻⁷)	0,22 (5,7e ⁻⁶)	0,02 (5,8e ⁻⁶)	0,22 (3,0e ⁻⁷)	----
EM5-NR	0,06 (1,9e ⁻⁶)	0,55 (7,7e ⁻⁷)	0,22 (3,9e ⁻⁶)	0,02 (3,9e ⁻⁶)	0,22 (3,9e ⁻⁷)	----
EM10-AI	0,06 (5,9e ⁻⁶)	0,55 (2,3e ⁻⁷)	0,22 (3,4e ⁻⁷)	0,02 (3,5e ⁻⁷)	0,22 (1,8e ⁻⁸)	----
EM10-NR	0,06 (5,1e ⁻⁶)	0,55 (2,0e ⁻⁶)	0,22 (4,8e ⁻⁶)	0,02 (9,2e ⁻⁷)	0,22 (1,1e ⁻⁶)	----
NR	0,06 (1,7e ⁻⁹)	0,55 (1,4e ⁻⁸)	0,22 (1,5e ⁻⁸)	0,02 (7,1e ⁻¹⁰)	0,22 (5,0e ⁻⁸)	----
Classe II						
	M6(\hat{h}_a^2)	M6(\hat{h}_d^2)	M6(\hat{c}_c^2)	M6(\hat{c}_b^2)	----	----
AI	0,05 (6,0e ⁻⁶)	0 (5,7e ⁻⁶)	0,74 (1,3e ⁻⁶)	0,06 (1,8e ⁻⁸)	----	----
EM	0,05 (2,1e ⁻³)	0,01 (2,8e ⁻³)	0,73 (9,7e ⁻⁴)	0,06 (5,0e ⁻⁴)	----	----
EM1-AI	0,05 (7,0e ⁻⁶)	0 (6,7e ⁻⁶)	0,74 (1,5e ⁻⁶)	0,06 (2,1e ⁻⁸)	----	----
EM1-NR	0,05 (1,1e ⁻⁶)	0 (1,7e ⁻⁶)	0,74 (5,0e ⁻⁶)	0,06 (1,2e ⁻⁶)	----	----
EM5-AI	0,05 (6,7e ⁻⁶)	0 (6,4e ⁻⁶)	0,74 (1,4e ⁻⁶)	0,06 (2,0e ⁻⁸)	----	----
EM5-NR	0,05 (2,6e ⁻⁶)	0 (1,2e ⁻⁶)	0,74 (9,0e ⁻⁷)	0,06 (2,8e ⁻⁷)	----	----
EM10-AI	0,05 (3,8e ⁻⁷)	0 (3,6e ⁻⁷)	0,74 (7,9e ⁻⁸)	0,06 (7,1e ⁻¹⁰)	----	----
EM10-NR	0,05 (4,8e ⁻⁶)	0 (8,6e ⁻⁷)	0,74 (2,2e ⁻⁶)	0,06 (4,1e ⁻⁷)	----	----
NR	0,05 (2,5e ⁻⁸)	0 (1,4e ⁻⁹)	0,74 (1,4e ⁻⁷)	0,06 (3,2e ⁻⁸)	----	----
Classe III						
	M7(\hat{h}_g^2)	M7(\hat{c}_b^2)	M8(\hat{h}_a^2)	M8(\hat{h}_d^2)	M8(\hat{c}_b^2)	----
AI	NC	NC	0,19 (5,7e ⁻⁶)	0,01 (5,6e ⁻⁶)	0,23 (3,1e ⁻⁷)	----
EM	0,01 (2,0e ⁻⁴)	0,57 (3,8e ⁻⁵)	0,18 (4,2e ⁻³)	0,02 (4,9e ⁻³)	0,23 (4,9e ⁻⁴)	----
EM1-AI	NC	NC	0,19 (5,5e ⁻⁶)	0,01 (5,4e ⁻⁶)	0,23 (3,0e ⁻⁷)	----
EM1-NR	NC	NC	0,19 (5,9e ⁻⁸)	0,01 (3,2e ⁻⁸)	0,23 (2,6e ⁻⁷)	----

EM5-AI	NC	NC	0,19 (6,7e ⁻⁶)	0,01 (6,7e ⁻⁶)	0,23 (3,7e ⁻⁷)	-----
EM5-NR	NC	NC	0,19 (8,2e ⁻⁸)	0,01 (4,0e ⁻⁸)	0,23 (1,6e ⁻⁸)	-----
EM10-AI	NC	NC	0,19 (5,8e ⁻⁶)	0,01 (5,7e ⁻⁶)	0,23 (3,2e ⁻⁷)	-----
EM10-NR	NC	NC	0,19 (3,9e ⁻⁶)	0,01 (4,9e ⁻⁶)	0,23 (2,4e ⁻⁷)	-----
NR	NC	NC	0,19 (2,5e ⁻⁸)	0,01 (7,1e ⁻¹⁰)	0,23 (9,1e ⁻⁸)	-----

NC: Não convergiu.