UNIVERSIDADE FEDERAL DE LAVRAS

MODELOS DE HERANÇA NO MELHORAMENTO VEGETAL: UMA ABORDAGEM BAYESIANA

MARIA IMACULADA DE SOUSA SILVA

2005

MARIA IMACULADA DE SOUSA SILVA

MODELOS DE HERANÇA NO MELHORAMENTO VEGETAL: UMA ABORDAGEM BAYESIANA

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

Orientador:

Prof. Dr. Eduardo Bearzoti

LAVRAS
MINAS GERAIS – BRASIL
2005

Ficha Catalográfica Preparada Pela Divisão de Processos Técnicos da Biblioteca Central da UFLA

Silva, Maria Imaculada de Sousa
Modelos de herança no melhoramento vegetal: uma abordagem bayesiana / Maria Imaculada de Sousa Silva. -- Lavras : UFLA, 2005.
77 p. : il.

Orientador: Eduardo Bearzoti.
Dissertação (Mestrado) – UFLA.
Bibliografia.

1. Herança genética. 2. Gene principal. 3. poligenes. 4. Amostrador de Gibbs. 5. Metropolis-Hastings. I. Universidade Federal de Lavras. II. título.

CDD-519.542
-575.1021

MARIA IMACULADA DE SOUSA SILVA

MODELOS DE HERANÇA NO MELHORAMENTO VEGETAL: UMA ABORDAGEM BAYESIANA

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

APROVADA em 28 de fevereiro de 2005

Prof. Dr. Heyder Diniz Silva	**UFU**
Profa. Dra. Thelma Sáfadi	**UFLA**
Prof. Dr. Tarcisio de Moraes Gonçalves	**UFLA**

Prof. Dr. Eduardo Bearzoti
UFLA
(Orientador)

LAVRAS
MINAS GERAIS – BRASIL

A Deus, pelo infinito amor
e pelo presente da vida,

Louvo e agradeço.

Ao meu amor, José Diniz, companheiro e amigo que
dedicou apoio incondicional aos meus sonhos.
Aos meus filhos, Wellington e Bruno, anjos de Deus,
orgulhos de minha vida,

Dedico.

Ao meu pai, Anicésio e à minha mãe, Juventina,
pelo orgulho que teriam se pudessem estar aqui.

Ofereço.

Se diante de mim não se abrir o mar, Deus
vai me fazer andar por sobre as águas

AGRADECIMENTOS

Ao professor Eduardo Bearzoti, por sua orientação e dedicação, de fundamental importância no desenvolvimento deste trabalho e, acima de tudo, pelo privilégio da convivência amiga, sempre transmitindo confiança, responsabilidade, determinação e exemplo de vida.

Ao professor da Universidade Federal de Uberlândia, Ednaldo Carvalho Guimarães, pelo incentivo, amizade e orientação na iniciação científica, fatores decisivos e responsáveis pelo início desse sonho realizado.

Ao meu marido, José Diniz e aos meus filhos, Wellington e Bruno, por serem a minha vida, o meu porto seguro, dividindo comigo as angústias e privações e compartilhando os momentos de alegria.

Aos meus pais, Anicésio e Juventina (in memorian), que mereciam muito dividir comigo o mérito dessa vitória.

A toda a minha família, especialmente aos meus sobrinhos e meus irmãos, Sebastiana, Pedro, José, Amantino, Aparecida, Lourdes, Luiza, Anicésio, Ildo, Erildo e Hélio, por me incentivarem e depositarem tanta confiança em mim, fazendo-me acreditar que sempre posso mais, e simplesmente por serem a minha família, que eu amo para sempre.

Às minhas grandes amigas Gisele e Nádia, pelos inesquecíveis anos de convivência; à Vânia e Rafaela, pelo apoio e carinho no momento certo.

A todos os colegas do curso de mestrado e doutorado em Estatística, por cada palavra ou por cada sorriso, capazes de amenizar o peso dos momentos difíceis e tornar eternos os momentos felizes.

Aos meus amigos Regilson e Cristina, José Waldemar e Elisângela, pelo apoio e amizade nesses dois anos.

A todos os professores do Departamento de Ciências Exatas, por todo o apoio, e a todos os funcionários, pela atenção e dedicação de verdadeiros amigos.

Aos queridos amigos do Grupo de Partilha e Perseverança (GPP), pelos momentos de apoio, amizade e oração.

À Universidade Federal de Lavras e ao Departamento de Ciências Exatas, pela oportunidade da realização deste curso.

À CAPES, pela bolsa de estudos essencial para a realização deste trabalho.

A todos os meus amigos que estão ou que estiveram na minha vida e que nunca deixarão de estar no meu coração.

A todos aqueles, citados aqui ou não, que incentivaram, acreditaram ou colaboraram para a minha vitória.

Obrigada!!!

SUMÁRIO

RESUMO

Silva, Maria Imaculada de Sousa. **Modelos de herança no melhoramento vegetal: uma abordagem bayesiana.** 2005. 77 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, MG. [1]

Em programas de melhoramento genético vegetal, freqüentemente, os estudos desenvolvidos têm o objetivo de inferir sobre a herança genética de uma característica contínua, a qual pode ser atribuída à ação de um único gene (gene principal), ou de vários genes de pequeno efeito (poligenes). Recentemente têm sido considerados modelos que investigam a existência das duas categorias de genes simultaneamente, admitindo, para tal, distribuições de probabilidades normais a cada genótipo. Com este enfoque, em programas de melhoramento animal, têm sido empregados métodos de inferência bayesiana no processo de estimação dos parâmetros do modelo, em substituição aos métodos numéricos iterativos, por vezes exigido pela complexidade funções de verossimilhança. No entanto, tais estudos não se aplicam diretamente aos dados de populações de plantas devido ao fato de estes dados apresentarem particularidades que em geral, não são consideradas em populações animais, como a existência das várias gerações delineadas em experimentos com plantas, e de efeitos de dominância. Diante disso, apresentou-se este trabalho com o objetivo de estender os estudos com enfoque bayesiano aplicados a dados de populações animais para os dados oriundos de experimentos do melhoramento de plantas, de forma a contemplar as suas particularidades. Para a construção do modelo genético, admitiu-se distribuição de probabilidade normal a cada genótipo, considerando um modelo linear misto contendo os parâmetros de interesse, sendo as inferências feitas a partir dos algoritmos de Metropolis-Hastings e do Amostrador de Gibbs. A partir de tal modelo, ajustou-se também um submodelo contendo apenas os parâmetros referentes ao gene principal. A metodologia proposta foi ilustrada com um conjunto de dados de um estudo da herança da partenocarpia em abobrinha. O ajuste do modelo de herança monogênica permitiu explicitar claramente as probabilidades de cada indivíduo apresentar um ou outro genótipo com relação ao gene principal.

[1] **Comitê Orientador:** Dr. Eduardo Bearzoti – UFLA (Orientador) e Dr. Julio Sílvio de Sousa Bueno Filho – UFLA (co-orientador).

ABSTRACT

SILVA, Maria Imaculada de Sousa. **Inheritance models in plant breeding: a Bayesian approach.** 2005. 77 p. Dissertation (Master in Agronomy/Major in Statistics and Agricultural Experimentation) – Federal University of Lavras, Lavras, MG[1].

In plant breeding programs, it is common to carry out studies to investigate the inheritance of traits of interest, whether there is a single gene controlling the trait (major gene) and/or polygenes of minor effect. Recently, models have been proposed to investigate the existence of both kinds of genes simultaneously, considering normal distributions with different means according to the genotype of the major gene. Such models have been used in animal breeding by means of a Bayesian approach, as an alternative to former suggestions of numerical maximizing of the likelihood function. However, this approach, as originally proposed in animal breeding, cannot be directly applied in plant populations, because in those there are, in general, different generations arisen from the cross of two inbred lines, and the interest to account for dominance is also frequent. This study aimed at fitting genetic models of inheritance using Bayesian inference and taking into account such characteristics of data sets of plant breeding experiments. Normal densities with different means, according to the major gene genotype, were considered in a mixed linear model with random individual polygenic effects and fixed effects to discriminate the generations. The distributions a posteriori were obtained using both Metropolis-Hastings and the Gibbs sampler algorithms. A submodel containing only major gene effects was also fitted. The approach proposed here was illustrated with an actual data set from a study of the inheritance of partenocarpy in *Cucurbita pepo* L. The fitting of the latter model yielded a posteriori probabilities for all individuals having each of the genotypes of the major gene.

[1] **Guidance committee**: Dr. Eduardo Bearzoti – UFLA, **(Adviser)**, Dr. Julio Sílvio de Sousa Bueno - UFLA.

1 INTRODUÇÃO

No melhoramento de plantas, ao se estudar a herança genética de uma característica de interesse, geralmente, os pesquisadores buscam identificar o número de genes envolvidos e saber como são seus efeitos.

Quando a característica de interesse é quantitativa, como, por exemplo, peso de fruto ou altura de planta, é comum testar a hipótese de herança monogênica (um único gene principal) por meio de testes de aderência de qui-quadrado.

Supondo a ação de poligenes, em geral, a inferência é feita por meio do teste de escala conjunto. As metodologias mais indicadas para inferir sobre herança poligênica e monogênica de características contínuas estão apresentadas nas seções 2.1 e 2.2, respectivamente.

Ainda com relação a características contínuas, estudos recentes propõem investigar simultaneamente a existência de gene principal e poligenes, admitindo que as distribuições de probabilidades nas gerações segregantes (como, por exemplo, F_2 e retrocruzamentos) são misturas de densidades normais. Exemplos de tais estudos encontram-se descritos na seção 2.3.

Os métodos da inferência bayesiana também têm sido usados para inferir sobre parâmetros relativos à ação de gene principal e poligenes, porém, aplicados a dados de populações animais (Janss et al., 1995). No entanto, a aplicação de tais métodos a dados de populações vegetais não é imediata, sendo necessárias algumas adaptações de forma a contemplar suas particularidades, como a existência de várias gerações e a consideração de efeitos de dominância, que, na maioria das vezes, não é feita no melhoramento animal. Na seção 2.4.4 encontram-se descritos exemplos de análise bayesiana em experimentos de melhoramento animal.

A inferência bayesiana, criada antes mesmo da análise clássica utilizada atualmente, ficou esquecida durante anos pelo fato de que sua aplicação, a princípio, exigiria o cálculo analítico de integrais muito complicadas.

Com a utilização de métodos de simulação via Cadeias de Markov, nos casos em que as integrais não podem ser resolvidas analiticamente, passou a ser possível encontrar uma solução aproximada para o problema. Entre estes métodos, destacam-se os algoritmos do Amostrador de Gibbs e do Metropolis-Hastings, descritos nas seções 2.4.1 e 2.4.2.

Com isso, a partir de 1990, a inferência bayesiana tem sido bastante usada para resolver problemas em diversas áreas e os resultados coerentes têm atraído cada vez mais a atenção dos pesquisadores. No entanto, a facilidade de implementação dos algoritmos não deve substituir e, sim, complementar o pensamento crítico sobre o problema, por parte do pesquisador. Além disso, é importante usar um dos critérios disponíveis para monitorar a convergência da cadeia gerada, evitando, assim, que se utilize esforço computacional além do necessário, ou que se pare o processo antes da convergência, o que, certamente, conduziria a conclusões incorretas. Alguns dos critérios de convergência mais comumente utilizados estão apresentados na seção 2.4.3.

Ao que parece, a inferência bayesiana não tem sido utilizada para estimar parâmetros referentes à ação de gene principal e poligenes em dados de populações vegetais, considerando suas particularidades.

Assim, os objetivos deste trabalho foram:

1. estender os métodos da inferência bayesiana utilizada em dados de populações animais, para dados de melhoramento de plantas, apresentando, assim, estimadores bayesianos de parâmetros referentes a gene principal e poligenes, incluindo efeitos de dominância, sobre os quais há interesse relevante em tais estudos;

2. ilustrar a metodologia implementada com um conjunto de dados reais.

2 REFERENCIAL TEÓRICO

2.1 Herança poligênica

Em programas de melhoramento genético vegetal, freqüentemente os pesquisadores têm interesse em analisar a herança de uma característica contínua, como, por exemplo, resistência a doenças, peso de frutos ou sementes, entre outras. Em tais casos, é comum admitir a existência de genes de pequeno efeito (poligenes) e influência ambiental.

Os componentes de média relativos aos poligenes são freqüentemente estimados pelo teste de escala conjunto (Mather & Jinks, 1984; Ramalho et al., 1993). Neste procedimento, podem ser utilizados dados de gerações genitoras contrastantes e de diferentes tipos de gerações oriundas de seu cruzamento, conforme detalhadamente descrito em Mather & Jinks (1984). Os componentes de variância podem ser, por vezes, estimados pela análise de variância.

O teste de escala conjunto é baseado no método dos quadrados mínimos ponderados, pelo qual utiliza-se o modelo linear:

$$Y = X\theta + \varepsilon \qquad (2.1),$$

sendo Y o vetor de observações com matriz de variâncias e covariâncias diagonal, representada por V, X a matriz de incidência, θ o vetor de parâmetros e ε o vetor de resíduos.

Sendo V diagonal (covariâncias nulas mas variâncias possivelmente distintas devido à existência de dados de diferentes gerações, as quais apresentam variabilidades genéticas distintas), V^{-1} também o é, e sua decomposição de Cholesky resulta em:

$$V^{-1} = LL' ,$$

sendo L uma matriz diagonal cujos elementos são as raízes quadradas dos elementos de V^{-1}. Pré-multiplicando (2.1) por L, tem-se:

$$LY = LX\theta + L\varepsilon \qquad (2.2),$$

o qual também é um modelo linear, porém, homocedástico e de variância igual a 1. De fato, e observando que neste caso L = L', tem-se:

$V(LY) = LV(Y)L' = L(LL')^{-1}L' = LL^{-1}L^{-1}L = I$

Utilizando o sistema de equações normais para estimar θ no modelo (2.2) e, admitindo que X tenha posto coluna completo, tem-se:

$$\hat{\theta} = [(LX)'(LX)]^{-1}(LX)'LY = [X'LLX]^{-1}X'LLY = (X'V^{-1}X)^{-1}X'V^{-1}Y \qquad (2.3).$$

Os resíduos de (2.2) são estimados por:

$Le = LY - LX\hat{\theta}$, sendo que a forma quadrática

$$(Le)'(Le) = e'L'Le = e'V^{-1}e \qquad (2.4),$$

sob normalidade, tem distribuição de qui-quadrado com número de graus de liberdade igual ao tamanho da amostra, menos o número de parâmetros estimados.

Os componentes de média dos poligenes podem ser definidos como sendo: μ uma constante de referência, [a] a soma dos efeitos aditivos dos poligenes e [d] a soma dos efeitos de dominância dos poligenes. Com estes componentes, e admitindo variâncias diferentes a cada geração, os valores genotípicos dos indivíduos de cada geração podem ser expressos por um modelo linear contendo a esperança da geração em questão mais um desvio aleatório. Assim, considerando as gerações genitoras contrastantes designadas por "P_1" e "P_2", o cruzamento entre elas (geração "F_1"), o cruzamento entre indivíduos P_1 e F_1, (RC_{11}), o cruzamento entre indivíduos P_2 e F_1, ("RC_{12}"), e o cruzamento entre indivíduos F_1, (geração "F_2"), de acordo com o modelo linear, tem-se (Mather & Jinks, 1984):

P_1: $Y_{1i} = \mu + [a] + \varepsilon_{1i}$, $\quad \varepsilon_{1i} \sim N(0, V_{P_1})$, $\quad i = 1, 2, \ldots, n_1$,

P_2: $Y_{2i} = \mu - [a] + \varepsilon_{2i}$, $\quad \varepsilon_{2i} \sim N(0, V_{P_2})$, $\quad i = 1, 2, \ldots, n_2$,

F_1: $Y_{3i} = \mu + [d] + \varepsilon_{3i}$, $\quad \varepsilon_{3i} \sim N(0, V_{F_1})$, $\quad i = 1, 2, \ldots, n_3$,

RC_{11}: $Y_{4i} = \mu + \frac{[a]}{2} + \frac{[d]}{2} + \varepsilon_{4i}$, $\quad \varepsilon_{4i} \sim N(0, V_{RC_{11}})$, $\quad i = 1, 2, \ldots, n_4$,

RC_{12}: $Y_{5i} = \mu - \frac{[a]}{2} + \frac{[d]}{2} + \varepsilon_{5i}$, $\quad \varepsilon_{5i} \sim N(0, V_{RC_{12}})$, $\quad i = 1, 2, \ldots, n_5$,

F_2: $Y_{6i} = \mu + \frac{[d]}{2} + \varepsilon_{6i}$, $\quad \varepsilon_{6i} \sim N(0, V_{F_2})$, $\quad i = 1, 2, \ldots, n_6$.

Se forem tomadas as médias de amostras aleatórias de cada geração, é possível representá-las de acordo com o modelo citado, por meio da relação matricial:

$$\begin{bmatrix} \bar{Y}_{1.} \\ \bar{Y}_{2.} \\ \bar{Y}_{3.} \\ \bar{Y}_{4.} \\ \bar{Y}_{5.} \\ \bar{Y}_{6.} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ 1 & -\frac{1}{2} & \frac{1}{2} \\ 1 & 0 & \frac{1}{2} \end{bmatrix} \cdot \begin{bmatrix} \mu \\ [a] \\ [d] \end{bmatrix} + \begin{bmatrix} \bar{\varepsilon}_{1.} \\ \bar{\varepsilon}_{2.} \\ \bar{\varepsilon}_{3.} \\ \bar{\varepsilon}_{4.} \\ \bar{\varepsilon}_{5.} \\ \bar{\varepsilon}_{6.} \end{bmatrix} \qquad (2.5).$$

$$Y = X \theta + \varepsilon$$

O método dos quadrados mínimos ponderados pode, então, ser utilizado, podendo os elementos de V (matriz diagonal cujos elementos são as variâncias de cada geração divididas pelos respectivos números de elementos) serem estimados calculando-se as variâncias amostrais dentro de cada geração. Assim, os parâmetros μ, [a] e [d] são estimados por (2.3).

Se o modelo for correto, a grandeza (2.4) terá distribuição aproximada (devido ao fato de se estimar V) de qui-quadrado. Assim, é conveniente expressar (2.4) como:

$$\chi_c^2 = \frac{n_1}{V_{P_1}}\left(\bar{Y}_1 - \hat{Y}_1\right)^2 + \frac{n_2}{V_{P_2}}\left(\bar{Y}_2 - \hat{Y}_2\right)^2 + \cdots + \frac{n_6}{V_{F_2}}\left(\bar{Y}_6 - \hat{Y}_6\right)^2 \qquad (2.6),$$

em que $\hat{Y}_j$ representa o estimador do valor genotípico esperado da geração j.

As estimativas dos componentes de variâncias utilizadas no teste de escala conjunto podem ser obtidas por meio da análise de variância. É muito freqüente, no melhoramento de plantas, a utilização de um delineamento experimental com controle local para avaliar plantas de diferentes gerações, considerando cada uma destas como um tratamento de efeito fixo. Igualando-se os quadrados médios às suas esperanças, obtêm-se as estimativas necessárias (método dos momentos) para a implementação do teste de escala conjunto.

2.2 Herança monogênica

No estudo da herança de uma característica quantitativa, o conhecimento do número de genes envolvidos e da magnitude de seus efeitos está, geralmente, entre os objetivos mais relevantes do pesquisador. Nestes casos, apesar de a característica ser contínua, é interessante verificar se ela é determinada pela ação de um gene principal (herança monogênica) ou se a herança é essencialmente poligênica. Se há um gene principal, a variabilidade dentro de cada classe genotípica se deve à ação de efeitos ambientais, somada ou não à ação de poligenes.

Quando se tem interesse em investigar a existência de um gene principal influenciando uma característica quantitativa, várias propostas estão disponíveis na literatura. Muitas delas propõem testar a hipótese de herança monogênica por meio de testes de aderência de qui-quadrado, mediante a construção de classes. Dentre as metodologias que apresentam este enfoque, merecem destaque a metodologia apresentada por Arias et al. (1994) e a apresentada por Souza

Sobrinho (1998), ambas baseadas no cálculo de freqüências esperadas para as gerações segregantes, utilizando informações das gerações P_1, P_2 e F_1, conforme a segregação mendeliana de um único gene. No entanto, cada metodologia possui particularidades que serão apresentadas a seguir.

2.2.1 Metodologia de Arias et al. (1994)

Arias et al. (1994) estudaram a herança da resistência à doença mancha olho-de-rã, na cultura de soja, provocada por *Cercospora sojina*. Utilizando diferentes cruzamentos entre cultivares de soja, foi possível identificar alelos de resistência às raças Cs-4 e C-15 do patógeno. No referido estudo, o limite entre as classes correspondentes à resistência e susceptibilidade não é claramente definido, devido à influência ambiental. Assim, os autores sugerem que classes de resistência sejam definidas (por exemplo, mediante o uso de notas), categorizando as observações de cada geração, como apresentado na Tabela 1.

TABELA 1 Distribuições de freqüências absolutas observadas em k classes C_j. N_{ij} é o número de indivíduos da geração i, observados na classe j.

Geração	Classe				Totais
	C_1	C_2	....	C_k	
P_1	N_{11}	N_{12}	...	N_{1k}	$N_{1.}$
P_2	N_{21}	N_{22}	...	N_{2k}	$N_{2.}$
F_1	N_{31}	N_{32}	...	N_{3k}	$N_{3.}$
F_2	N_{41}	N_{42}	...	N_{4k}	$N_{4.}$

Adaptado de Arias et al. (1994).

A metodologia pressupõe que, em presença de um gene principal, as freqüências esperadas em cada classe da geração F_2 deveriam ser funções das freqüências observadas das gerações P_1, F_1 e P_2, com proporções de 1:2:1,

respectivamente. Assim, as freqüências absolutas esperadas na classe j da geração F_2 podem ser calculadas utilizando a expressão:

$$f_{e_j} = \left[\frac{N_{1j}}{N_1} + 2\frac{N_{3j}}{N_3} + \frac{N_{2j}}{N_2}\right]\frac{N_4}{4} \qquad (2.7).$$

Se $fe_j < 5$ para algum j, então classes podem ser fundidas, conforme usualmente sugerido em testes de aderência. De posse das freqüências esperadas em F_2, a seguinte estatística pode ser calculada:

$$\chi_c^2 = \sum_{j=1}^{k} \frac{\left(fe_j - N_{4j}\right)^2}{fe_j} \qquad (2.8).$$

Sob a hipótese de nulidade (um único gene), a estatística calculada tem distribuição aproximada de qui-quadrado com k-1 graus de liberdade.

No contexto de resistência a doenças, os autores utilizaram seis classes, relativas à porcentagem de área foliar lesionada, as quais foram posteriormente reunidas em apenas duas, correspondentes a resistentes e suscetíveis, utilizando um ponto de truncamento que pode variar de acordo com cada estudo.

Embora não ressaltado pelos autores, esta metodologia pode ser adaptada quando se dispõe de outras gerações segregantes.

2.2.2 Metodologia apresentada por Souza Sobrinho (1998)

A metodologia para estudo de herança monogênica, originalmente apresentada por Souza Sobrinho (1998), considera diferentes valores de grau médio de dominância (GMD) e admite que a característica sob estudo tenha distribuição normal. Um ponto de truncamento (PT) deve ser estabelecido de maneira a discriminar os genitores contrastantes.

Uma adaptação desta metodologia foi apresentada por Freitas et al. (2002), de forma a contemplar as particularidades de um estudo sobre o teor de zingibereno em folhas de tomateiro, para o qual constatou-se a existência de um

gene principal. Assim, descreve-se aqui a metodologia, conforme apresentada por Freitas et al. (2002). Segundo estes autores, considerando dados das gerações P_1, P_2, F_1 e F_2, a metodologia é implementada pelas seguintes etapas:

1. estimam-se a média e a variância das gerações P_1 e P_2, por meio de estimadores usuais. Com tais estimativas, calculam-se as freqüências esperadas nestas gerações, conforme a densidade normal, abaixo e acima do PT;
2. com cada valor de GMD, estima-se a média da geração F_1 por:

$$\overline{F_1} = \frac{\left(\overline{P_1} + \overline{P_2}\right)}{2} + \text{GMD}\left(\overline{P_1} - \overline{P_2}\right) \qquad (2.9),$$

   sendo $\overline{P_1}$ e $\overline{P_2}$ as médias das gerações P_1 e P_2. A variância na geração F_1 é estimada pela variância amostral entre plantas F_1. A partir da densidade normal, com tais parâmetros, são calculadas freqüências esperadas abaixo e acima do PT;
3. as freqüências esperadas na geração F_2 são calculadas a partir das freqüências esperadas obtidas para as gerações P_1, F_1 e P_2, com pesos que mantêm as proporções de 1:2:1, respectivamente;
4. freqüências esperadas em todas as gerações são confrontadas com as freqüências observadas mediante uma estatística de qui-quadrado e um teste de aderência é feito;
5. a estatística de qui-quadrado é calculada com diferentes valores de GMD e, para os valores menores que o nível nominal, há indícios de que a hipótese de herança monogênica deva ser rejeitada.

Os autores ainda mencionam a possibilidade de fusão de gerações, para evitar a ocorrência de freqüências esperadas nulas. No teste de aderência descrito, os parâmetros estimados (médias, variâncias e GMD) não são funções diretas das freqüências absolutas observadas. Neste caso, segundo Mood et al. (1974), o número de graus de liberdade da distribuição de qui-quadrado, sob a

hipótese de nulidade, é algo entre o valor k-1 e k-1-p, sendo k o número de classes e p o número de parâmetros estimados. No presente contexto, o termo (k-1) deve ser multiplicado pelo número de gerações efetivamente envolvidas no teste de aderência.

Comparando-se as duas metodologias, pode-se observar que a proposta de Souza Sobrinho (1998) difere da de Arias et al. (1994), por admitir distribuição normal e fornecer explicitamente um estimador para o GMD, dado por aquele valor ao qual corresponde um mínimo qui-quadrado. No caso da primeira, por tratar-se de um teste de aderência, várias classes também poderiam ser definidas para confrontar freqüências observadas e esperadas, e não apenas duas por geração. Isso poderia contribuir, eventualmente, para aumentar o poder da metodologia, pelo aumento do número de graus de liberdade.

Um dos pontos comuns entre as metodologias é o fato de que a maior parte ou toda a informação necessária para o cálculo das freqüências esperadas é obtida apenas das gerações P_1 , F_1 e P_2. No entanto, seria interessante que a informação contida na geração F_2 , bem como em outras gerações segregantes, fosse aproveitada no processo de estimação. Para tanto, o enfoque estatístico natural seria reconhecer que as distribuições de probabilidades nestas gerações são misturas de densidades, em geral normais.

2.3 Modelos de misturas em estudos de herança genética

Das metodologias existentes na literatura para estudos de herança genética, grande parte considera a estimação dos parâmetros referentes ao gene principal e aos poligenes separadamente, sem o uso de um modelo único e geral que considere as duas categorias de genes simultaneamente. No entanto, alguns estudos mais recentes têm considerado modelos com este enfoque, reconhecendo

que as densidades nas gerações segregantes (F_2 e retrocruzamentos) são misturas de densidades normais.

Lou & Zhu (2002) apresentaram um estudo para investigar, simultaneamente, efeitos genéticos de poligenes e de gene principal em plantas diplóides e em animais, tendo sido usado um modelo linear misto para se proceder às análises estatísticas. O interesse dos autores foi estimar o efeito do gene principal e a variância dos poligenes e, como em Changjian (1994), concluíram que o método da máxima verossimilhança baseado em modelos de misturas é adequado para estimar os parâmetros.

Um modelo considerando misturas de densidades normais foi proposto por Changjian et al. (1994) para estimar parâmetros relativos a gene principal e poligenes, simultaneamente. Tal modelo supõe homogeneidade de variâncias devido ao ambiente e ou poligenes, segregação independente do gene principal, invariância do efeito do gene principal com diferentes tipos de população e efeitos aditivos e de dominância.

Os modelos e submodelos de interesse são testados pelo teste de razão de verossimilhança generalizada, podendo ser aceito um modelo simplificado que aumente a viabilidade da análise. Os autores consideram dados das seis gerações já citadas na seção 2.1 e, posteriormente, estendem o modelo para incluir as gerações F_3, B_1 e B_2, obtidas pelo cruzamento entre indivíduos das gerações F_2, RC_{11} e RC_{12} respectivamente. Eles assumem, ainda, dominância completa em relação ao gene principal.

Apresentando o mesmo princípio básico do modelo de Changjian et al. (1994), porém, considerando qualquer grau de dominância em relação ao gene principal, Silva (2003) apresentou uma adaptação deste modelo, a qual se encontra descrita a seguir.

Silva (2003) considera que os dados são oriundos das seis gerações já citadas, as quais são, em geral, utilizadas por geneticistas de plantas.

Os valores genotípicos referentes ao gene principal são representados por μ+A, μ–A e μ+D, correspondentes aos dois homozigotos e ao heterozigoto, respectivamente, sendo μ uma constante de referência, A o efeito aditivo e D o efeito de dominância do gene principal.

Havendo poligenes, existem ainda os componentes adicionais de média [a] e [d], correspondentes à soma dos efeitos poligênicos aditivos e de dominância, respectivamente. Os componentes poligênicos de variância, V_A, V_D e S_{AD} são, respectivamente, a variância aditiva, a variância atribuída aos desvios de dominância e a variância referente aos produtos dos efeitos poligênicos aditivos pelos efeitos poligênicos de dominância.

Segundo o autor, a cada uma das gerações P_1, P_2 e F_1 tem-se associada uma única distribuição normal, enquanto que às gerações RC_{11}, RC_{12} e F_2 correspondem misturas de duas ou três normais, conforme apresentado pelas equações a seguir.

$$P_1 : N(\mu + [a] + A; \sigma^2) \qquad (2.10),$$

$$P_2 : N(\mu - [a] - A; \sigma^2) \qquad (2.11),$$

$$F_1 : N(\mu + [d] + D; \sigma^2) \qquad (2.12),$$

$$RC_{11} : \frac{1}{2}N\left(\mu + \frac{[a]}{2} + \frac{[d]}{2} + A; \sigma^2 + \frac{V_A}{2} + V_D + S_{AD}\right) + \\ + \frac{1}{2}N\left(\mu + \frac{[a]}{2} + \frac{[d]}{2} + D; \sigma^2 + \frac{V_A}{2} + V_D + S_{AD}\right) \qquad (2.13),$$

$$RC_{12} : \frac{1}{2}N\left(\mu - \frac{[a]}{2} + \frac{[d]}{2} - A; \sigma^2 + \frac{V_A}{2} + V_D - S_{AD}\right) + \\ + \frac{1}{2}N\left(\mu - \frac{[a]}{2} + \frac{[d]}{2} + D; \sigma^2 + \frac{V_A}{2} + V_D - S_{AD}\right) \qquad (2.14),$$

$$F_2 : \frac{1}{4}N\left(\mu + \frac{[d]}{2} + A; \sigma^2 + V_A + V_D\right) + \frac{1}{2}N\left(\mu + \frac{[d]}{2} + D; \sigma^2 + V_A + V_D\right)$$

$$+ \frac{1}{4}N\left(\mu + \frac{[d]}{2} - A; \sigma^2 + V_A + V_D\right) \qquad (2.15).$$

O autor sugere que a estimação dos parâmetros seja feita pelo método da máxima verossimilhança. A função de verossimilhança (L) é definida por:

$$L = \prod_{j=1}^{6} \prod_{i=1}^{n_j} f(y_{ij})$$

sendo $f(y_{ij})$ a função densidade de probabilidade do indivíduo i da geração j.

Aplicando-se o logaritmo na função L, obtém-se a denominada função suporte (S),

$$S = \sum_{j=1}^{6} \sum_{i=1}^{n_j} \ln f(y_{ij}).$$

Derivando a função suporte em relação a cada um dos parâmetros do modelo e, em seguida, igualando estas expressões a zero, obtêm-se as funções de verossimilhança. Devido à complexidade das equações, torna-se necessário o uso de métodos numéricos iterativos para a obtenção das estimativas dos parâmetros. Assim, Silva (2003) apresentou um programa de análise estatística que alterna o uso dos métodos de quase-Newton e de Powell, obtendo assim as estimativas de máxima verossimilhança.

Diferentes submodelos devem ser julgados pelo teste da razão de verossimilhança generalizada. Estes submodelos surgem da atribuição do valor zero a diferentes parâmetros, resultando nos modelos genéticos listados na Tabela 2.

TABELA 2 Modelos genéticos e seus respectivos parâmetros

		Efeito		
Modelo	**Herança**	Principal	Poligenes	Parâmetros
1	Principal e Poligenes	Aditivo e dominância	Aditivo e dominância	$\mu, A, D, [a], [d], V_A, V_D, S_{AD}, \sigma^2$
2	Principal e Poligenes	Aditivo e dominância	Aditivo	$\mu, A, D, [a], V_A, \sigma^2$
3	Principal e Poligenes	Aditivo	Aditivo e dominância	$\mu, A, [a], [d], V_A, V_D, S_{AD}, \sigma^2$
4	Principal e Poligenes	Aditivo	Aditivo	$\mu, A, [a], V_A, \sigma^2$
5	Poligenes	—	Aditivo e dominância	$\mu, [a], [d], V_A, V_D, S_{AD}, \sigma^2$
6	Poligenes	—	Aditivo	μ, A, V_A, σ^2
7	Principal	Aditivo e dominância	—	μ, A, D, σ^2
8	Principal	Aditivo	—	μ, A, σ^2
9	Nenhum	—	—	μ, σ^2

O teste da razão de verossimilhança é feito por meio da estatística LR (Mood et al., 1974), dada por:

$$LR = -2 \ln \frac{L(M_i)}{L(M_j)},$$

sendo $L(M_i)$ e $L(M_j)$ as funções de verossimilhança dos modelos i e j, em que o modelo i deve estar hierarquizado ao modelo j, ou seja, o modelo i deve ser um caso particular do modelo j.

Esta estatística segue uma distribuição aproximada de qui-quadrado. Um teste com aproximação α, quando H_0 é verdadeira, é determinado pela regra de decisão de rejeitar H_0 se, e somente se, $LR > \chi^2_{(1-\alpha,\upsilon)}$. O número de graus de liberdade ν é dado pela diferença entre os números de parâmetros dos modelos M_j e M_i.

Por exemplo, confrontando-se os modelos 1 e 5 pela estatística LR, está se testando a hipótese da existência de gene principal mais poligenes contra a

existência de apenas poligenes. Se a estatística LR for superior ao valor tabelado de qui-quadrado, então, o modelo 5 é rejeitado, aceitando-se o modelo 1, ou seja, não se aceita a hipótese da existência de apenas poligenes e concluindo-se que há um gene principal.

A metodologia foi aplicada a um conjunto de dados de um estudo de herança da partenocarpia em abobrinha, permitindo ao autor concluir que o fenômeno tem herança monogênica com efeitos aditivos e de dominância, não havendo, portanto, ação de poligenes.

2.4 Inferência bayesiana

A inferência bayesiana, ao contrário da inferência clássica, leva em conta, também, o conceito de probabilidade subjetiva, que mede o grau de incerteza que se tem sobre a ocorrência de um determinado evento do espaço amostral. Assim, a análise bayesiana descreve toda quantidade desconhecida por meio de probabilidades.

Assim como na inferência freqüentista, a inferência bayesiana trabalha na presença de observações y, cujo valor é inicialmente incerto e descrito por uma densidade $f(y|\theta)$. A quantidade θ serve como indexador da família de distribuições das observações, representando uma característica de interesse (Gamerman, 1997).

A diferença formal entre a inferência bayesiana e a inferência freqüentista é que, para a inferência bayesiana, o parâmetro θ é uma variável aleatória, possuindo, então, uma distribuição de probabilidade, enquanto que, para a inferência freqüentista, os parâmetros são valores fixos ou constantes. Além disso, a inferência bayesiana usa toda a informação disponível a priori, enquanto a freqüentista ignora esta informação.

A densidade conjunta de um grupo de observações $y_1, \ldots, y_n$, como função do parâmetro θ, é denominada função de verossimilhança e é representada por $L(y_1, \ldots, y_n | \theta)$, em que n é o número de observações. A função de verossimilhança fornece as chances de cada valor de θ ter levado àquele valor observado para y.

Ao incorporar à sua análise um conhecimento prévio que se tenha sobre o parâmetro θ, o pesquisador está especificando a densidade a priori de θ, $p(\theta)$, a qual contém a distribuição de probabilidade de θ antes da observação do valor de y. O parâmetro θ pode ser um escalar ou um vetor de parâmetros.

As distribuições a priori podem ser informativas ou não informativas. Quando o pesquisador tem alguma informação prévia sobre o parâmetro em questão, ele pode usar uma priori informativa, descrevendo uma densidade $p(\theta)$, por sua vez, especificada com o auxílio de constantes chamadas de hiperparâmetros, pois são os parâmetros da distribuição dos parâmetros. Inicialmente, os hiperparâmetros são considerados conhecidos e traduzem a informação que se tem sobre o parâmetro, antes da realização da amostra.

Quando se tem pouca ou nenhuma informação sobre o parâmetro, pode-se usar uma priori não informativa. A idéia é pensar em todos os valores de θ como igualmente prováveis, ou seja, com uma distribuição a priori uniforme. Neste caso, $p(\theta) \propto k$ ($p(\theta)$ proporcional a uma constante k) significa que nenhum valor de θ tem preferência.

A informação de que dispomos sobre θ, resumida probabilisticamente por $p(\theta)$, pode ser aumentada observando-se uma quantidade aleatória Y relacionada com θ.

O Teorema de Bayes é a regra de atualização da informação que se tem sobre o parâmetro e é utilizado para quantificar esse aumento de informação,

sendo um elemento essencial na análise bayesiana, pois toda inferência é feita a partir da distribuição a posteriori.

Se $p(\theta)$ é a densidade a priori de θ, então a densidade a posteriori de θ, $p(\theta \mid y)$, é dada pelo Teorema de Bayes:

$$p(\theta|y) = \frac{L(Y|\theta)p(\theta)}{\int L(Y|\theta)p(\theta)d\theta},$$

em que $Y = \{y_1, y_2, \ldots, y_n\}$. O denominador funciona como uma constante normalizadora, já que não depende de θ. Assim, o teorema pode ser reescrito como:

$$p(\theta \mid Y) \propto p(Y \mid \theta)\, p(\theta) \qquad (2.16),$$

sendo que $\propto$ representa proporcionalidade.

No caso de θ ser multivariado, ($\theta = \theta_1, \theta_2, \ldots, \theta_p$), as distribuições marginais das componentes θ_i, a partir das quais as inferências para cada parâmetro serão feitas, podem ser obtidas da densidade conjunta a posteriori $p(\theta_1, \theta_2, \ldots, \theta_p \mid y_1, \ldots, y_n)$.

A densidade marginal a posteriori de θ_i é dada por:

$$p(\theta_i|Y) = \int p(\theta_1, \theta_2, \ldots, \theta_p|Y)d\theta_{-i}$$

sendo $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p)$ o vetor θ com sua i-ésima componente removida.

A resolução desta integral, analiticamente, é, em muitos casos, impraticável. Uma das alternativas neste caso são os métodos aproximados de inferência, que se baseiam em preceitos analíticos e determinísticos. Alguns dos principais métodos são a linearização e a aproximação pela normal, aproximação de Laplace e aproximação via quadratura gaussiana, cujas descrições podem ser encontradas em Gamerman (1997).

Em oposição aos métodos analíticos, os métodos baseados em simulação estocástica não dependem do tamanho da amostra observada e fornecem

aproximações que serão tanto melhores quanto maior for o número de valores gerados.

Os métodos de simulação estocástica consideram as distribuições condicionais completas a posteriori de cada parâmetro para gerar amostras que convergem para a densidade marginal, com o aumento do tamanho dessa amostra.

A distribuição condicional completa do parâmetro θ_i (denotada por $p(\theta_i | \theta_{-i}, Y)$) é obtida considerando que, na densidade conjunta, os demais parâmetros θ_{-i} são conhecidos e, assim, a expressão se torna menos complexa, já que as constantes podem ser desconsideradas.

Quando a expressão da condicional completa tem a forma de uma densidade conhecida e, portanto, fácil de ser amostrada, um método de simulação indicado é o Amostrador de Gibbs, um processo iterativo que gera valores que convergem para a densidade marginal, sem que se conheça a sua expressão.

Se a distribuição condicional completa a posteriori não é uma densidade conhecida, outros métodos de simulação são indicados; entre eles estão a amostragem por importância (Paulino et al., 2003), a amostragem por aceitação e rejeição e o Metropolis-Hastings (Chib & Greenberg, 1995; Hastings, 1970). Os algoritmos do amostrador de Gibbs e do Metropolis-Hastings estão descritos a seguir.

2.4.1 Amostrador de Gibbs

O amostrador de Gibbs é, essencialmente, um esquema iterativo de amostragem de uma cadeia de Markov, cujo núcleo de transição é formado pelas distribuições condicionais completas (Gamerman, 1997). Para descrever o algoritmo, suponha-se que a distribuição de interesse é $\pi(\theta)$ em que $\theta = (\theta_1, \theta_2,$

... , θ_p). Cada uma das componentes θ_i pode ser um escalar ou um vetor. A distribuição π não precisa ser uma distribuição a posteriori e o método pode ser aplicado em outro contexto, sem qualquer referência à inferência bayesiana. No entanto, aqui a distribuição $\pi(\theta)$ corresponde à distribuição a posteriori, $p(\theta \mid Y)$. Considere-se, ainda, que as densidades completas a posteriori $\pi(\theta_i \mid \theta_{-i}, Y)$, i = 1, ..., p estão disponíveis.

O interesse aqui é gerar amostras da densidade conjunta $\pi(\theta)$, mas, sendo esta geração extremamente complicada, o algoritmo de Gibbs fornece uma forma alternativa de gerações por meio das distribuições condicionais completas. Ele é descrito da seguinte forma:

1. Iniciar o contador de iterações da cadeia t = 1 e estabelecer valores iniciais

$$\theta^0 = (\theta_1^0, \theta_2^0, \ldots, \theta_p^0).$$

2. Obter um novo valor $\theta' = (\theta_1', \theta_2', \ldots, \theta_p')$ a partir de θ^{t-1} por meio de sucessivas gerações de valores:

$$\theta_1' \sim \pi(\theta_1 \mid \theta_2'^{-1}, \theta_3'^{-1}, \ldots, \theta_p'^{-1})$$
$$\theta_2' \sim \pi(\theta_2 \mid \theta_1', \theta_3'^{-1}, \ldots, \theta_p'^{-1})$$
$$\vdots$$
$$\theta_p' \sim \pi(\theta_p \mid \theta_1', \theta_2', \ldots, \theta_p')$$

3. Mudar o contador de t para t + 1 e voltar ao passo 2 até a convergência.

À medida que o número t de iterações aumenta, a seqüência de valores gerados se aproxima da distribuição de equilíbrio, ou seja, da densidade marginal desejada para cada parâmetro, quando se assume que a convergência foi atingida.

2.4.2 Metropolis-Hastings

A partir dos trabalhos de Hastings (1970) e Metropolis et al. (1953), foi desenvolvido um método de amostragem denominado Metropolis-Hastings, que tem tido bastante atenção dos estatísticos nos últimos anos.

Suponha-se que o objetivo é gerar valores θ de uma distribuição condicional completa de interesse $\pi(\theta)$, sendo esta tarefa muito complicada pelo fato de $\pi(\theta)$ não ter a forma de uma densidade conhecida e fácil de ser amostrada.

O método consiste em gerar valores candidatos θ^* de uma densidade auxiliar $q(\theta, \theta^*)$ que possa ser amostrada e rejeitar ou aceitar esses valores com probabilidade $\alpha(\theta, \theta^*)$. A expressão mais comumente usada para a probabilidade de aceitação é:

$$\alpha(\theta,\theta^*) = \min\left\{1, \frac{\pi(\theta^*)q(\theta^*,\theta)}{\pi(\theta)q(\theta,\theta^*)}\right\}$$

Em termos práticos, a simulação de uma amostra de $\pi(\theta)$ pode ser obtida usando o algoritmo Metropolis-Hastings esquematizado a seguir:

1. Iniciar com um valor arbitrário θ^j e com o contador de iterações j= 0.
2. Gerar um valor θ^* de $q(\theta^j, .)$ e um valor u de uma uniforme (0,1).
3. Calcular $\alpha(\theta^j, \theta^*)$.

   Se $u \leq \alpha$, fazer $\theta^{j+1} = \theta^*$. Caso contrário, $\theta^{j+1} = \theta^j$.
4. Mudar o contador de j para j + 1 e voltar ao passo (2) até a convergência.

O método define uma cadeia de Markov, pois, as transições dependem apenas das posições no estágio anterior.

O núcleo de transição q define apenas uma proposta de movimento que pode ou não ser confirmado pela probabilidade de aceitação α.

A densidade de equilíbrio π só interfere no algoritmo por meio da razão do teste, na forma

$$\frac{\pi(\theta^*)}{\pi(\theta)};$$

portanto, não é necessário conhecer a constante de proporcionalidade (Hastings, 1970).

Para implementar o algoritmo de Metropolis-Hastings, uma densidade candidata $q(\theta, \theta^*)$ para gerar as amostras deve ser selecionada de uma família de distribuições, com especificação de um parâmetro como escala ou posição. Segundo Chib & Greenberg (1995), várias são as famílias de densidades que podem ser escolhidas, algumas das quais já foram citadas por Hastings (1970).

Uma escolha bastante eficiente, quando disponível, é explorar a forma de $\pi(\theta)$ para especificar a densidade candidata. Por exemplo, se $\pi(\theta)$ pode ser escrita como $\pi(\theta) = \psi(\theta)h(\theta)$, sendo $h(\theta)$ possível de ser amostrada e $\psi(\theta)$ é uniformemente limitada, pode-se fazer $q(\theta, \theta^*) = h(\theta)$ para gerar as amostras candidatas. Neste caso, a probabilidade de movimento se reduz à seguinte expressão:

$$\alpha(\theta,\theta^*) = \min\left\{1, \frac{\Psi(\theta^*)}{\Psi(\theta)}\right\} \qquad (2.17).$$

2.4.3 Diagnóstico de convergência

À medida que o número de iterações aumenta, a cadeia gerada se aproxima da distribuição de equilíbrio, ou seja, da densidade marginal desejada

de cada parâmetro. Assim, assume-se que a convergência é atingida em uma iteração cuja distribuição esteja arbitrariamente próxima da distribuição de equilíbrio. A grande dificuldade é determinar quantas iterações são necessárias para se atingir esta condição.

Quando se trata de modelos complicados, os processos iterativos das cadeias de Markov exigem um grande esforço computacional. Com o intuito de alcançar a convergência minimizando esse esforço, vários métodos estão disponíveis para monitorar a convergência das seqüências geradas pelo amostrador de Gibbs.

Nogueira (2004) apresentou um estudo no qual avaliou alguns dos critérios de convergência já existentes para os casos uni e multivariado, além de propor dois novos critérios multivariados, o critério do traço e o critério do determinante. Estes novos critérios podem ser usados nos casos em que a monitoração da convergência pelo critério multivariado de Brooks e Gelman pode falhar se existe uma baixa correlação entre os parâmetros.

Nogueira (2004) sugeriu os seguintes passos para monitorar a convergência em problemas univariados;

a. aplicar o critério de Raftery e Lewis em uma amostra piloto para determinar o tamanho ideal da seqüência;
b. determinar o tamanho do "burn-in" pelo critério de Heidelberger e Welch;
c. monitorar a convergência das seqüências por meio do critério de Gelman e Rubin e do critério de Geweke.

Alguns dos princípios dos critérios sugeridos por Nogueira (2004) estão brevemente descritos a seguir. Maiores detalhes podem ser encontrados em Gamerman (1997) e Nogueira (2004).

2.4.3.1 Critério de Raftery eLewis

Ao se analisar a convergência de uma seqüência gerada por meio do amostrador de Gibbs, é comum descartar as primeiras iterações, em geral, de 40% a 50% do total (Gamerman, 1997), considerando-se que essa primeira parte esteja sendo influenciada pelos valores iniciais. Este início da cadeia é chamado de período de "aquecimento" ou "burn-in".

Outro aspecto importante refere-se à dependência entre as observações subseqüentes da cadeia. Para se obter uma amostra independente, as observações devem ser espaçadas por um determinado número de iterações, ou seja, considerar saltos ("thin") de tamanho k, usando, para compor a amostra, os valores a cada k iterações.

O critério de Raftery e Lewis fornece estimativas do número de iterações necessárias para se obter a convergência, do número de iterações iniciais que devem ser descartadas (burn-in) e da distância mínima (k) de uma iteração à outra para se obter uma amostra independente. Esses valores são calculados mediante especificações para garantir que um quantil u de uma determinada função $f(\theta)$ seja estimado com uma precisão predefinida.

2.4.3.2 Critério de Heidelberger e Welch

O critério de Heidelberger e Welch propõe testar a hipótese nula de estacionariedade da seqüência gerada, por meio de testes estatísticos. Se a hipótese nula é rejeitada para um dado valor, o teste é repetido depois de descartados os 10% valores iniciais da seqüência. Se a hipótese é novamente rejeitada, mais 10% dos valores iniciais são descartados e assim sucessivamente até serem descartados os 50% valores iniciais. Se a hipótese for novamente rejeitada, isto indica que é necessário um número maior de iterações. Caso

contrário, o número inicial de iterações descartadas é indicado como o tamanho do "burn-in".

O critério utiliza também o teste de Half-Width para verificar se a média estimada está sendo calculada com uma acurácia preespecificada, sendo testada a porção da seqüência que passou no teste de estacionariedade para cada parâmetro. Se o resultado for positivo, a média está sendo estimada com um erro aceitável, portanto, julgada ser a média da distribuição de interesse.

2.4.3.3 Critério de Geweke

Usando técnicas de análise espectral, o critério de Geweke fornece um diagnóstico para a ausência de convergência.

Considerando uma função real $t(\theta)$, sua trajetória $t^{(1)}$, $t^{(2)}$, ... $t^{(n)}$, construída a partir de $t^{(j)} = t(\theta^{(j)})$, j = 1, 2, ... , n, define uma série temporal; portanto, a média desta série pode ser estimada por:

$$\bar{t}_j = \frac{1}{n}\sum_{j=1}^{n} t^{(j)}$$

que é um estimador não viesado de $E[t(\theta)]$, cuja variância assintótica é dada por $S_t(0)/n$, sendo $S_t(\omega)$ a densidade espectral da série t. Em geral, S_t é desconhecida e estimada por $\hat{S}_t$, utilizando análise espectral.

Após um número suficientemente grande de iterações, determinam-se as médias $\bar{t}_A$ das primeiras n_A iterações, e $\bar{t}_B$ das últimas n_B iterações. Calculam-se também os estimadores independentes $\hat{S}_t^A(0)$ e $\hat{S}_t^B(0)$, das variâncias assintóticas de $\{t^{(j)}: j=1, ..., n_A\}$ e $\{t^{(j)}: j=n^*, ..., n_B\}$, respectivamente, sendo $n^* = n - n_B + 1$. O critério garante que, se n_A/n e n_B/n são fixos, com $(n_A + n_B)/n < 1$, quando $n \rightarrow \infty$, tem-se que, se a seqüência for estacionária, a diferença

padronizada entre as médias tem distribuição normal com média zero e variância um, ou seja:

$$\frac{\bar{t}_A - \bar{t}_B}{\sqrt{\left(\hat{S}_t^A(0)/n_A\right)+\left(\hat{S}_t^B(0)/n_B\right)}} \sim N(0,1)$$

Assim, se a diferença padronizada entre as médias for grande, existe indicação de ausência de convergência. Os valores aconselháveis para se determinar as médias são $n_A = 0,1n$ e $n_B = 0,5n$.

2.4.3.4 Critério de Gelman e Rubin

O critério de convergência de Gelman e Rubin pressupõe que m cadeias tenham sido geradas em paralelo, partindo de diferentes valores iniciais, num total de 2n iterações, das quais n são descartadas ("burn-in"). As m seqüências rendem m possíveis inferências. Se estas inferências são similares tem-se um indicativo de que a convergência foi alcançada ou está próxima.

Considerando m cadeias em paralelo e uma função real t(θ), têm-se m trajetórias $\{t_j^1, t_j^1, \cdots, t_j^n\}$, j = 1, 2, ... , m; portanto, podem ser obtidas as variâncias entre cadeias (E) e dentro de cadeias (D), dadas por:

$$E = \frac{n}{m-1}\sum_{j=1}^{m}\left(\bar{t}_j - \bar{t}_j\right)^2 \quad \text{e}$$

$$D = \frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(t_j^{(i)} - \bar{t}_j\right)^2,$$

sendo $\bar{t}_j$ a média das observações da cadeia j, $\bar{t}$ a média dessas médias e j = 1, ..., m. A variância de t pode ser estimada, de forma não viesada, por:

$$\hat{V}\left[t(\theta)\right] = \left(1 - \frac{1}{n}\right)D + \left(\frac{1}{n}\right)E$$

genótipo AA, (0 1 0) associado ao genótipo Aa, ou (0 0 1) associado ao genótipo aa, em que n_q é o número de indivíduos. Assim, tem-se que $P[w_i = (1\ 0\ 0)] = (1 - q)^2$, $P[w_i = (0\ 1\ 0)] = 2q(1 - q)$, e $P[w_i = (0\ 0\ 1)] = q^2$, podendo ser, então, definida a probabilidade de cada indivíduo apresentar um determinado genótipo, com base no genótipo de ambos os pais.

O autor apresenta, então, o seguinte modelo linear:

$$y = Xb + Za + ZWm + \varepsilon,$$

em que "y" é o vetor dos dados de ordem n, "b" é o vetor de "efeitos fixos", "a" é o vetor de efeitos genéticos aditivos de ordem n_q, $W' = (w_1, w_2, \ldots, w_{n_q})$, "m" foi definido anteriormente e ε é o vetor de resíduos. Assume-se distribuição normal para os dados e para o vetor "a" ou seja:

$$y \mid b, a, W, m, \sigma_e^2 \sim N(Xb+Za+ZWm, I\sigma_e^2) \text{ e } a \mid \sigma_a^2 \sim N(0, A\sigma_a^2).$$

Aos parâmetros "b" e "m" foram atribuídas distribuições a priori constantes. A distribuição a priori para W foi especificada de acordo com as probabilidades que cada indivíduo apresenta de atribuir um dos três valores para w_i. Para a freqüência do gene principal "q", foi assumida como priori uma distribuição beta com parâmetros "e" e "f", da forma:

$$p(q) \alpha\ q^{e-1}(1 - q)^{f-1}.$$

Para os componentes de variância σ_a^2 e σ_e^2, foram assumidas distribuições a priori qui-quadrado invertida com parâmetro de escala, da forma:

$$p(\sigma_i^2 \mid V_i, S_i^2) \alpha (\sigma_i^2)^{-((V_i/2)-1)} \exp\left[-\frac{V_i S_i^2}{2\sigma_i^2}\right], \quad (i = a, e).$$

De posse de todas as prioris e da função de verossimilhança, foi obtida a densidade conjunta a posteriori, a partir do Teorema de Bayes. Em seguida, as distribuições condicionais completas a posteriori foram apresentadas, viabilizando a implementação do amostrador de Gibbs e a inferência sobre os parâmetros do modelo.

investigar as conseqüências de diferentes seleções na melhoria da população sob estudo.

Janss et al. (1997) apresentaram um estudo com os objetivos de investigar se existe um gene principal influenciando características de qualidade da carne de suínos e mostrar a flexibilidade do uso do amostrador de Gibbs na estimação dos componentes de variância e dos componentes de média referentes aos poligenes. Foi utilizado um modelo linear misto e foram analisados dados de indivíduos da geração F_2 de um cruzamento entre linhagens contrastantes de suínos, considerando onze características que afetam a qualidade da carne. Em relação aos poligenes, foram considerados apenas efeitos aditivos. As análises estatísticas permitiram identificar a existência de um gene principal diferente dos genes já identificados para tais características em estudos anteriores. Amostras independentes da distribuição marginal dos parâmetros foram obtidas por meio do amostrador de Gibbs e a convergência foi encontrada para os componentes de variância em praticamente todos os casos. Na estimação dos componentes de média, segundo os autores, o amostrador de Gibbs mostrou-se mais flexível que os estimadores de máxima verossimilhança.

Uma metodologia para estudos de herança genética em populações animais foi apresentada por Sorensen (1996), considerando o modelo linear e algumas das pressuposições do modelo de Janss et al. (1997), com as devidas adaptações e particularidades. As inferências sobre os parâmetros foram feitas usando um enfoque bayesiano.

Sorensen (1996) assumiu que, para os alelos do gene principal, designados, por exemplo, por A e a, as respectivas freqüências são (1–q) e q. Tais freqüências, no caso de populações animais, são desconhecidas e estimadas. É definido um vetor $m' = (m_1, m_2, m_3)$, cujos elementos descrevem o efeito dos genótipos AA, Aa e aa. Associada a esses genótipos, define-se a variável aleatória w_i, $(i = 1, 2, ..., n_q)$, a qual pode assumir os valores (1 0 0) associado ao

genótipo AA, (0 1 0) associado ao genótipo Aa, ou (0 0 1) associado ao genótipo aa, em que n_q é o número de indivíduos. Assim, tem-se que $P[w_i = (1\ 0\ 0)] = (1 - q)^2$, $P[w_i = (0\ 1\ 0)] = 2q(1 - q)$, e $P[w_i = (0\ 0\ 1)] = q^2$, podendo ser, então, definida a probabilidade de cada indivíduo apresentar um determinado genótipo, com base no genótipo de ambos os pais.

O autor apresenta, então, o seguinte modelo linear:

$$y = Xb + Za + ZWm + \varepsilon,$$

em que "y" é o vetor dos dados de ordem n, "b" é o vetor de "efeitos fixos", "a" é o vetor de efeitos genéticos aditivos de ordem n_q, $W' = (w_1, w_2, \ldots, w_{n_q})$, "m" foi definido anteriormente e ε é o vetor de resíduos. Assume-se distribuição normal para os dados e para o vetor "a" ou seja:

$$y \mid b, a, W, m, \sigma_e^2 \sim N(Xb+Za+ZWm, I\sigma_e^2) \text{ e } a \mid \sigma_a^2 \sim N(0, A\sigma_a^2).$$

Aos parâmetros "b" e "m" foram atribuídas distribuições a priori constantes. A distribuição a priori para W foi especificada de acordo com as probabilidades que cada indivíduo apresenta de atribuir um dos três valores para w_i. Para a freqüência do gene principal "q", foi assumida como priori uma distribuição beta com parâmetros "e" e "f", da forma:

$$p(q) \alpha\ q^{e-1}(1 - q)^{f-1}.$$

Para os componentes de variância σ_a^2 e σ_e^2, foram assumidas distribuições a priori qui-quadrado invertida com parâmetro de escala, da forma:

$$p(\sigma_i^2 \mid V_i, S_i^2)\alpha(\sigma_i^2)^{-((V_i/2)-1)} \exp\left[-\frac{V_i S_i^2}{2\sigma_i^2}\right], \quad (i = a, e).$$

De posse de todas as prioris e da função de verossimilhança, foi obtida a densidade conjunta a posteriori, a partir do Teorema de Bayes. Em seguida, as distribuições condicionais completas a posteriori foram apresentadas, viabilizando a implementação do amostrador de Gibbs e a inferência sobre os parâmetros do modelo.

3 METODOLOGIA

3.1 Modelo genético

Considerando o modelo de misturas de normais, propõe-se, neste capítulo, uma metodologia para estimar parâmetros referentes à ação de gene principal e poligenes, aplicada a dados de populações vegetais.

As pressuposições do modelo são basicamente as mesmas já apresentadas por Silva (2003), porém, aqui, a metodologia que se segue baseia-se em métodos da inferência bayesiana para obterem-se as estimativas por ponto e por intervalo dos parâmetros de interesse. Além disso, procurou-se adaptar a metodologia aplicada a dados de populações animais, como já apresentado por Sorensen (1996) e Janss et al. (1997), para contemplar as particularidades dos dados de melhoramento de plantas. Entre essas particularidades, destaca-se a existência de várias gerações e a inferência acerca dos efeitos de dominância.

A modelagem leva em conta dados das gerações P_1, P_2, F_1, RC_{11}, RC_{12} e F_2, as quais foram definidas na seção 2.1, podendo, no entanto, ser adaptada para contemplar dados de outras gerações derivadas de linhagens contrastantes.

O modelo geral supõe a ação de gene principal mais poligenes com efeitos aditivos e de dominância, além de variâncias ambientais (σ^2) iguais em todas as gerações.

Os parâmetros do modelo referentes ao gene principal são A e D, sendo A o efeito aditivo e D o efeito de dominância do gene principal. Assim, havendo a ação de um gene principal, os valores genotípicos dos dois homozigotos e do heterozigoto são, respectivamente, $\mu + A$, $\mu - A$ e $\mu + D$, sendo μ uma constante de referência.

A ação de poligenes é representada no modelo pelos componentes de média [a] e [d], e pelos componentes de variância, V_A, V_D e S_{AD}, já definidos em Mather & Jinks (1984), sendo [a] o componente poligênico aditivo, [d] o

componente poligênico de dominância, V_A a variância atribuída aos efeitos poligênicos aditivos, V_D a variância atribuída aos desvios de dominância e S_{AD} a variância referente aos produtos dos efeitos poligênicos aditivos pelos efeitos poligênicos de dominância. Assim, têm-se atribuídos a cada geração, de acordo com o genótipo, os componentes poligênicos de média e de variância, conforme apresentado na Tabela 3 e cuja derivação pode ser encontrada em Mather & Jinks (1984) e em Silva (2003).

TABELA 3 Componentes poligênicos de média e de variância para as diferentes gerações.

Geração	Componente poligênico da média	Componente poligênico da variância
P_1	[a]	—
P_2	-[a]	—
F_1	[d]	—
F_2	$\frac{1}{2}[d]$	$V_A + V_D$
RC_{11}	$-\frac{1}{2}[a]+\frac{1}{2}[d]$	$\frac{1}{2}V_A + V_D - S_{AD}$
RC_{12}	$\frac{1}{2}[a]+\frac{1}{2}[d]$	$\frac{1}{2}V_A + V_D + S_{AD}$

Como se pode observar, o componente S_{AD} aparece na variância dos retrocruzamentos com sinais contrários, como apresentado em Mather & Jinks (1984). No entanto, não se pode afirmar, para cada caso, em qual das gerações este componente estaria sendo somado ou subtraído. Com o objetivo de inferir também sobre essa questão, está sendo considerado no modelo um parâmetro auxiliar t, que pode assumir os valores 1 e -1, multiplicando o componente S_{AD} que, por sua vez, foi mantido sempre positivo. Esta abordagem foi escolhida no sentido de facilitar a especificação de uma distribuição a priori para S_{AD}. Por se tratar de um parâmetro de segunda ordem, como os componentes de variância,

optou-se por mantê-lo positivo, de maneira a possibilitar o uso de uma gama invertida, como descrito mais adiante.

A probabilidade de o parâmetro t assumir cada valor é considerada a priori como sendo igual a 0,5 e deseja-se avaliar se um dos valores é mais provável.

É conveniente ressaltar que os componentes poligênicos de variância para as gerações P_1, P_2 e F_1 não existem, pois todos os indivíduos, em cada uma dessas gerações, possuem o mesmo genótipo e, portanto, toda a variação deve ser exclusivamente não herdável.

Com base nesses componentes de médias e admitindo distribuição de probabilidade normal aos dados, têm-se, associadas às gerações P_1, P_2 e F_1, densidades normais com médias e variâncias específicas, enquanto que às gerações segregantes RC_{11}, RC_{12} e F_2, por haver mais de um genótipo presente, correspondem misturas de densidades normais, conforme já apresentado na seção 2.3, nas equações de 2.10 a 2.16.

A natureza dos dados é representada de acordo com o seguinte modelo linear misto:

$$Y = X\beta + Wm + Zg + \varepsilon. \qquad (3.1).$$

Em tal modelo, "Y" é o vetor dos dados, β é um vetor de efeitos considerados fixos, de forma que $\beta' = (\mu \;\; [a] \;\; [d])$, sendo μ uma constante de referência, "[a]" e "[d]" os componentes poligênicos de média, aditivos e de dominância, "m" é um vetor de efeitos fixos contendo os componentes do gene principal, aditivo e de dominância, de forma que $m' = (A \;\; D)$. "X" e "Z" são as matrizes de delineamento de β e g, respectivamente e ε é o vetor de erros aleatórios.

A matriz W contém vetores aleatórios w_i para cada observação, referentes aos parâmetros A e D do gene principal, podendo, então, assumir os valores [-1 0], [1 0] ou [0 1]. O vetor w_i correspondente a cada elemento de Y é especificado com base na probabilidade que os indivíduos possuem de

apresentar um determinado genótipo, dependendo da geração a que pertencem. Assim, designando os alelos do gene principal, por exemplo, por B e b, sendo B o alelo que aumenta e b o alelo que diminui a expressão do caráter, o vetor w_i = [-1 0] está associado aos indivíduos homozigotos de genótipo bb, o vetor w_i = [1 0] está associado aos indivíduos homozigotos de genótipo BB e o vetor w_i = [0 1] está associado aos indivíduos heterozigotos. Portanto, para cada uma das gerações P_1, P_2 e F_1, por apresentarem apenas um genótipo, o vetor w_i pode assumir apenas um valor, diferente para cada uma delas, enquanto que, para as gerações RC_{11}, RC_{12} e F_2 existem duas ou três possibilidades. Assim, os vetores w_i para estas três últimas gerações constituem um parâmetro a ser estimado. Os possíveis valores para o vetor w_i, com as respectivas probabilidades a priori para cada geração, estão apresentados na Tabela 4 .

TABELA 4 Possíveis valores para o vetor w_i e suas probabilidades a priori.

	w_i		
	[-1 0]	[1 0]	[0 1]
Geração	Probabilidade		
P_1	1,0	0,0	0,0
P_2	0,0	1,0	0,0
F_1	0,0	0,0	1,0
RC_{11}	0,5	0,0	0,5
RC_{12}	0,0	0,5	0,5
F_2	0,25	0,25	0,5

O vetor "g" contém os efeitos genéticos poligênicos individuais de cada planta, "a" e "d", aditivos e de dominância, quando a geração em questão apresenta variação genética poligênica. Embora fosse possível fazer inferência sobre g, não há interesse, aqui, no efeito poligênico individual de cada planta. No entanto, é importante que conste do modelo, pois há interesse em se fazer

inferência sobre V_A, V_D e S_{AD}, que são os componentes da matriz (G) de variâncias e covariâncias de g.

Apenas a título de ilustração, suponha-se que se tenha uma amostra de 14 plantas, das quais, quatro são amostradas da geração F_2 e duas são amostradas de cada uma das demais gerações. O vetor g pode ser decomposto nos vetores "a" e "d", de forma que se tenha $Zg = Z_a a + Z_d d$. Assim, definem-se as matrizes de variâncias e covariâncias de a e d, e do produto de a por d como sendo $V(d) = V_D(I_8)$, e V(a) e V(ad) definidas abaixo, sendo que I_8 representa a matriz identidade de ordem 8.

$$V(a) = G_A = V_A \begin{bmatrix} 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$V(ad) = G_{AD} = t\frac{S_{AD}}{2} \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

A ordem das matrizes igual a 8 corresponde ao número de elementos das gerações que apresentam variação genética poligênica, sendo portanto as linhas das matrizes referentes aos 2 elementos do RC_{11}, 2 do RC_{12} e 4 da geração F_2. Com essas matrizes compõe-se a variância do vetor Y e do vetor g dadas por:

$V(Y) = V(Zg + \varepsilon) = V(Z_a a + Z_d d + \varepsilon) = V(Z_a a + Z_d d) + V(\varepsilon) =$

$= V(Z_a a) + V(Z_d d) + \text{cov}(Z_a a, Z_d d) + \text{cov}(Z_d d, Z_a a) + V(\varepsilon) =$

$= Z_a G_a Z_a' + Z_d G_d Z_d' + Z_a G_{ad} Z_d' + Z_d G_{ad}' Z_a' + I\sigma^2 =$

$= V_A(Z_a A_a Z_a') + V_D(Z_d A_d Z_d') + tS_{AD}/2(Z_a A_{ad} Z_d') + tS_{AD}/2(Z_d A_{ad} Z_a') + I\sigma^2 =$

$= ZGZ' + I\sigma^2$. Assim, $V(g) = ZGZ'$, sendo G a matriz apresentada a seguir.

$$\begin{bmatrix}
V_A/2 & 0 & -tS_{AD} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & V_A/2 & 0 & -tS_{AD} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-tS_{AD} & 0 & V_D & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -tS_{AD} & 0 & V_D & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & V_A/2 & 0 & tS_{AD} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & V_A/2 & 0 & tS_{AD} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & tS_{AD} & 0 & V_D & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & tS_{AD} & 0 & V_D & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & V_A & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & V_A & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & V_A & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & V_A & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & V_D & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & V_D & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & V_D & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & V_D
\end{bmatrix}$$

Da pressuposição de normalidade tem-se que:

$(Y \mid \beta, W, m) \sim N(X\beta + Wm, ZGZ' + I\sigma^2)$

$(g \mid V_A, V_D, S_{AD}) \sim N(0, G)$

A distribuição das observações, dados todos os parâmetros, é dada pela função de verossimilhança:

$$L(Y) = p(Y \mid \beta, W, m, V_A, V_D, S_{AD}, t, \sigma^2) = \left|ZGZ' + I\sigma^2\right|^{-\frac{1}{2}} x$$

$$x \exp\left\{-\frac{1}{2}(Y-(X\beta+Wm))'\left(ZGZ'+I\sigma^2\right)^{-1}(Y-(X\beta+Wm))\right\} \qquad (3.2).$$

3.2 Inferência bayesiana

Propõe-se aqui que a estimação dos parâmetros do modelo seja feita utilizando técnicas da inferência bayesiana. A derivação das expressões das densidades a priori e a posteriori para cada parâmetro encontra-se descrita nas seções seguintes.

3.2.1 Distribuições a priori

Alguns dos parâmetros apresentados na seção 3.1 já foram bastante investigados em estudos anteriores, principalmente em programas de melhoramento animal, também usando técnicas de inferência bayesiana. Portanto, existe uma informação prévia sobre esses parâmetros que não deve ser ignorada no processo de estimação. Esta informação para cada parâmetro foi traduzida por meio das distribuições a priori, apresentadas a seguir.

É comum na literatura encontrar, atribuída ao vetor de efeitos fixos β, uma priori não informativa, ou seja, uma distribuição a priori uniforme ou constante. Aqui também optou-se por atribuir prioris não informativas para os vetores de efeitos fixos β e m, ou seja, $p(\beta) \alpha$ constante e $p(m) \alpha$ constante.

Aos componentes de variância freqüentemente atribui-se, como priori, uma distribuição de qui-quadrado invertida, variando em cada estudo o valor dos hiperparâmetros, pois são eles que vão fornecer a informação prévia que se tem em cada caso. Foi considerado, então, que os componentes de variância V_A, V_D,

S_{AD} e σ^2 seguem, a priori, uma distribuição de qui-quadrado invertida com parâmetro de escala, conforme apresentado a seguir.

$$p(\sigma^2)\alpha\left(\frac{1}{\sigma^2}\right)^{\frac{v_e}{2}+1} e^{\frac{-T_e}{2\sigma^2}} \quad (3.3),$$

$$p(V_A)\alpha\left(\frac{1}{V_A}\right)^{\frac{v_a}{2}+1} e^{\frac{-T_a}{2V_A}} \quad (3.4),$$

$$p(S_{AD})\alpha\left(\frac{1}{S_{AD}}\right)^{\frac{v_{ad}}{2}+1} e^{\frac{-T_{ad}}{2V_{ad}}} \quad (3.5),$$

$$p(V_D)\alpha\left(\frac{1}{V_D}\right)^{\frac{v_d}{2}+1} e^{\frac{-T_d}{2V_d}} \quad (3.6).$$

A distribuição a priori para o parâmetro auxiliar t foi especificada mediante uma transformação de variáveis. Considere-se, inicialmente, uma variável aleatória x com distribuição de Bernoulli, ou seja:

$f(x) = p^x(1-p)^{1-x}$.

Associando-se os valores de x com os valores de t, de forma que quando x = 0, tem-se t = 1 e, quando x = 1, tem-se t = –1, obtém-se que t = –2x + 1, ou seja, x = (1 – t)/2, e (1 – x) = (t + 1)/2. Substituindo-se os valores de x em f(x), tem-se :

$$f(t) = p^{(1-t)/2}(1-p)^{(t+1)/2} \quad (3.7),$$

sendo que f(–1) = p e f(1) = 1 – p. Dessa forma, t também tem uma distribuição de Bernoulli, a qual foi usada como priori para parâmetro t do modelo, com p = 0,5.

A distribuição a priori para a matriz W é dada pela Tabela 4, apresentada na seção 3.1. Embora, nesse caso, a matriz W esteja sendo considerada como um

único parâmetro, ela poderia ser considerada como um conjunto de parâmetros representados por cada linha, ou seja, por cada vetor w_i, pois estes têm probabilidades a priori $p(w_i)$, dadas pela mesma Tabela 4, independentes dos demais vetores w_i ou parâmetros. Dessa forma, a distribuição a priori para a matriz W, lembrando que $W' = (w_1\ w_2\ ...\ w_n)$, é composta pelas prioris de cada vetor w_i.

O interesse na estimação da matriz W está em verificar se a informação a priori contida nesta Tabela é confirmada ou não pela informação a posteriori, ou seja, estas probabilidades, para cada indivíduo, poderiam ser atualizadas após a observação dos dados.

3.2.2 Distribuição conjunta (Teorema de Bayes)

A distribuição conjunta de todos os parâmetros, dado que uma amostra aleatória foi realizada, é obtida a partir do Teorema de Bayes, multiplicando-se a verossimilhança e todas as prioris. Portanto, a distribuição conjunta a posteriori, supondo independência entre os parâmetros, é dada pela seguinte expressão:

$$p(\beta, W, m, V_A, V_D, S_{AD}, t, \sigma^2 \mid Y) \propto$$

$$L(Y)\, p(\beta)\, p(m)\, p(\sigma^2)\, p(V_A)\, p(S_{AD})\, p(V_D)\, p(t)\, p(W) \propto$$

$$\propto \left|ZGZ' + I\sigma^2\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(Y-(X\beta+Wm)\right)'\left(ZGZ'+I\sigma^2\right)^{-1}\left(Y-(X\beta+Wm)\right)\right\}$$

$$\times\ p(W) \left(\frac{1}{\sigma^2}\right)^{\frac{v_e}{2}+1} e^{\frac{-T_e}{2\sigma^2}} \left(\frac{1}{V_A}\right)^{\frac{v_a}{2}+1} e^{\frac{-T_a}{2V_A}} \left(\frac{1}{S_{AD}}\right)^{\frac{v_{ad}}{2}+1} e^{\frac{-T_{ad}}{2V_{ad}}} \times$$

$$\left(\frac{1}{V_D}\right)^{\frac{v_d}{2}+1} e^{\frac{-T_d}{2V_d}} p^{(1-t)/2}(1-p)^{(t+1)/2} \qquad (3.8).$$

É importante observar que, como não faz parte dos objetivos do presente trabalho inferir sobre os efeitos poligênicos individuais de cada planta contidos em g, a verossimilhança usada no Teorema de Bayes é marginal a este vetor .

As inferências devem ser feitas por meio de amostras da densidade marginal de cada parâmetro. Porém, devido ao fato, de neste caso, não se poder obter as suas expressões, foram obtidas distribuições condicionais completas a posteriori, a partir das quais podem-se obter amostras das desejadas marginais.

3.2.3 Distribuições condicionais completas

A distribuição condicional completa a posteriori para o vetor β é derivada como segue, considerando que, na equação (3.8), os demais parâmetros e os dados são constantes.

$p(\beta \mid W, m, V_A, V_D, S_{AD}, t, \sigma^2, Y) \propto L(Y)\, p(\beta) \propto$

$$\propto \left|ZGZ' + I\sigma^2\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(Y-(X\beta+Wm)\right)'\left(ZGZ'+I\sigma^2\right)^{-1}\left(Y-(X\beta+Wm)\right)\right\}$$

$$= \left|ZGZ' + I\sigma^2\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(X\beta-(Y-Wm)\right)'\left(ZGZ'+I\sigma^2\right)^{-1}\left(X\beta-(Y-Wm)\right)\right\}.$$

A partir desta última expressão, conclui-se que Xβ tem distribuição normal com média igual a (Y – Wm) e variância igual a ($ZGZ' + I\sigma^2$). Pré-multiplicando-se Xβ por $(X'X)^{-1}X'$, deduz-se que ($\beta \mid W, m, V_A, V_D, S_{AD}, t, \sigma^2, Y$) tem distribuição normal com matriz de variâncias igual a $(X'X)^{-1}X'(ZGZ'+I\sigma^2)X(X'X)^{-1}$ e vetor de médias igual a $(X'X)^{-1}X'(Y-Wm)$. Portanto, a distribuição condicional completa para β é da seguinte forma:

($\beta \mid W, m, V_A, V_D, S_{AD}, t, \sigma^2, Y$) ~

$\sim N((X'X)^{-1}X'(Y-Wm), (X'X)^{-1}X'(ZGZ'+I\sigma^2)X(X'X)^{-1})$ (3.9).

Utilizando o mesmo raciocínio, obtém-se a distribuição condicional completa para m.

$p(m \mid W, \beta, V_A, V_D, S_{AD}, t, \sigma^2, Y) \alpha L(Y)\, p(m) \alpha$

$$\alpha \left|ZGZ'+I\sigma^2\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(Y-(X\beta+Wm))'\left(ZGZ'+I\sigma^2\right)^{-1}(Y-(X\beta+Wm))\right\}$$

$$= \left|ZGZ'+I\sigma^2\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(Wm-(Y-X\beta))'\left(ZGZ'+I\sigma^2\right)^{-1}(Wm-(Y-X\beta))\right\}.$$

Da última expressão conclui-se que: $Wm \sim N((Y-X\beta), ZGZ' + I\sigma^2)$.

Pré-multiplicando-se Wm por $(W'W)^{-1}W'$, tem-se que:

$(m \mid W, \beta, V_A, V_D, S_{AD}, t, \sigma^2, Y)$ tem distribuição normal, isto é:

$(m \mid W, \beta, V_A, V_D, S_{AD}, t, \sigma^2, Y) \sim$

$\sim N((W'W)^{-1}W'(Y-X\beta), (W'W)^{-1}W'(ZGZ'+I\sigma^2)W(W'W)^{-1})$ (3.10).

As distribuições condicionais completas para os vetores β e m são fáceis de serem amostradas, portanto, pode-se utilizar, neste caso, o algoritmo do amostrador de Gibbs descrito na seção 2.3.1.

A distribuição condicional completa da matriz W, considerada como um único parâmetro, é dada por:

$p(W \mid \beta, m, S_{AD}, V_A, V_D, t, \sigma^2, Y) \alpha L(Y)\, p(W) \alpha$

$$\alpha \exp\left\{-\frac{1}{2}(Y-(X\beta+Wm))'\left(ZGZ'+I\sigma^2\right)^{-1}(Y-(X\beta+Wm))\right\}\ p(W) \quad (3.11).$$

Se cada vetor w_i é considerado como um parâmetro, os demais vetores e parâmetros são constantes em relação a w_i e, assim, tem-se a condicional completa: $p(w_i \mid \beta, m, w_{-i}, S_{AD}, V_A, V_D, t, \sigma^2, Y) \alpha L(Y)\, p(w_i)$, sendo que w_{-i} representa todos os vetores da matriz W, com exceção de w_i.

As expressões das condicionais completas para os demais parâmetros estão apresentadas a seguir. Deve-se ressaltar que a função de verossimilhança não é constante em relação aos parâmetros V_A, V_D, S_{AD} e t, pois estes estão contidos na matriz G.

$p(V_A \mid W, \beta, m, V_D, S_{AD}, t, \sigma^2, Y) \alpha L(Y)\, p(V_A)\, \alpha$

$$\alpha \left|ZGZ' + I\sigma^2\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(Y-(X\beta+Wm)\right)'\left(ZGZ'+I\sigma^2\right)^{-1}\left(Y-(X\beta+Wm)\right)\right\} x$$

$$x\left(\frac{1}{V_A}\right)^{\frac{v_a}{2}+1} e^{\frac{-T_a}{2V_A}} \qquad (3.12),$$

$p(S_{AD} \mid W, \beta, m, V_D, V_A, t, \sigma^2, Y) \alpha L(Y)\, p(S_{AD})\, \alpha$

$$\alpha \left|ZGZ' + I\sigma^2\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(Y-(X\beta+Wm)\right)'\left(ZGZ'+I\sigma^2\right)^{-1}\left(Y-(X\beta+Wm)\right)\right\} x$$

$$x\left(\frac{1}{S_{AD}}\right)^{\frac{v_{ad}}{2}+1} e^{\frac{-T_{ad}}{2S_{ad}}} \qquad (3.13),$$

$p(V_D \mid W, \beta, m, S_{AD}, V_A, t, \sigma^2, Y) \alpha L(Y)\, p(V_D)\, \alpha$

$$\alpha \left|ZGZ' + I\sigma^2\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(Y-(X\beta+Wm)\right)'\left(ZGZ'+I\sigma^2\right)^{-1}\left(Y-(X\beta+Wm)\right)\right\} x$$

$$x\left(\frac{1}{V_D}\right)^{\frac{v_d}{2}+1} e^{\frac{-T_d}{2V_D}} \qquad (3.14),$$

$p(\sigma^2 \mid W, \beta, m, S_{AD}, V_A, V_D, t, Y) \alpha L(Y)\, p(\sigma^2)\, \alpha$

$$\alpha \left|ZGZ' + I\sigma^2\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(Y-(X\beta+Wm)\right)'\left(ZGZ'+I\sigma^2\right)^{-1}\left(Y-(X\beta+Wm)\right)\right\} x$$

$$x\left(\frac{1}{\sigma^2}\right)^{\frac{\nu_e}{2}+1} e^{\frac{-T_e}{2\sigma^2}} \qquad (3.15),$$

$p(t \mid W, \beta, m, S_{AD}, V_A, V_D, \sigma^2, Y) \alpha L(Y)\, p(t)\, \alpha$

$$\alpha\left|ZGZ'+I\sigma^2\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(Y-(X\beta+Wm)\right)'\left(ZGZ'+I\sigma^2\right)^{-1}\left(Y-(X\beta+Wm)\right)\right\} x$$

$$x\ p^{(1-t)/2}(1-p)^{(t+1)/2} \qquad (3.16),$$

Como se pode observar pelas expressões de 3.11 a 3.16, as condicionais completas para os parâmetros W, V_A, V_D, S_{AD}, t e σ^2 não têm a forma de densidades conhecidas. Neste caso, os parâmetros não podem ser amostrados diretamente da condicional por meio do amostrador de Gibbs. Assim, é indicado, como alternativa, o algoritmo de Metropolis-Hastings. Tendo em vista que estas condicionais são todas resultantes do produto de duas densidades, sendo uma delas a priori para cada parâmetro e a outra a verossimilhança, esse fato pode ser explorado para especificar a densidade geradora para o Metropolis-Hastings. Em cada caso, a priori é uma densidade conhecida e possível de ser amostrada e a função de verossimilhança é uma densidade normal que se supõe uniformemente limitada, devido aos valores possíveis para os componentes de variância. Desse modo, o algoritmo de Metropolis-Hastings pode ser implementado utilizando, como densidade geradora, a priori de cada parâmetro. Assim, conforme descrito na seção 2.3.2, a probabilidade de aceitação se reduz à expressão 2.17.

Dessa forma, ficam definidas todas as densidades necessárias para se proceder à análise bayesiana dos parâmetros considerados no modelo.

3.3 Modelo genético para herança monogênica

A estimação dos parâmetros referentes a gene principal e poligenes simultaneamente, como apresentado na seção 3.1, pode ser uma opção vantajosa em relação às metodologias que consideram apenas um tipo de herança. No entanto, ainda assim pode haver interesse em estimar apenas os parâmetros relativos ao gene principal ou aos poligenes.

Para esta situação, o modelo da seção 3.1 pode ser facilmente adaptado para estimar apenas os parâmetros relativos ao gene principal, bastando para isso atribuir o valor zero para os parâmetros relativos aos poligenes. Com isso, o modelo linear misto dado pela expressão 3.1 tem sua forma reduzida à nova expressão dada por:

$$Y = X\beta + Wm + \varepsilon.$$

Neste modelo, o vetor β é composto apenas pela constante de referência μ, já que, no modelo anterior, os outros componentes de β eram referentes aos poligenes, que aqui não existem. Os vetores Y, ε e m, as matrizes X, W e Z são definidos da mesma maneira que em 3.1.

Assumindo distribuição normal, tem-se que:

$$(Y \mid \beta, W, m) \sim N(X\beta + Wm, I\sigma^2).$$

A função de verossimilhança passa a ser definida pela seguinte expressão:

$$L(Y) = p(Y \mid \beta, W, m, \sigma^2) \propto$$

$$\alpha\left(\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\left(Y-(X\beta+Wm)\right)'\left(I\sigma^2\right)^{-1}\left(Y-(X\beta+Wm)\right)\right\} \qquad (3.17).$$

3.3.1 Análise bayesiana do modelo

Os procedimentos para a análise bayesiana do modelo são similares aos adotados na seção 3.1, sendo as densidades a priori para cada parâmetro deste modelo as mesmas já consideradas naquela seção.

A distribuição conjunta de todos os parâmetros é dada pela expressão a seguir.

$$p(\beta, W, m, \sigma^2 \mid Y) = L(Y)\, p(\beta)\, p(m)\, p(\sigma^2)\, p(W) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{\nu}{2}+1} e^{\frac{-N_e}{2\sigma^2}} \times$$

$$\times \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\left(Y-(X\beta+Wm)\right)'\left(I\sigma^2\right)^{-1}\left(Y-(X\beta+Wm)\right)\right\} p(W) \qquad (3.18).$$

As expressões das densidades condicionais completas a posteriori, obtidas por meio da densidade conjunta, são derivadas a seguir.

$p(\beta \mid W, m, \sigma^2, Y) \propto L(Y)\, p(\beta) \propto$

$$\propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\left(Y-(X\beta+Wm)\right)'\left(I\sigma^2\right)^{-1}\left(Y-(X\beta+Wm)\right)\right\} \propto$$

$$\propto \exp\left\{-\frac{1}{2}\left(X\beta-(Y-Wm)\right)'\left(I\sigma^2\right)^{-1}\left(X\beta-(Y-Wm)\right)\right\}.$$

A partir desta última expressão, conclui-se que $X\beta$ tem distribuição normal com média igual a $(Y - Wm)$ e variância igual a $(I\sigma^2)$. Pré-multiplicando-se $X\beta$ por $(X'X)^{-1}X'$, deduz-se que β tem distribuição normal com vetor de médias igual a $(X'X)^{-1}X'(Y - Wm)$ e matriz de variâncias igual a $(X'X)^{-1}X'(I\sigma^2)X(X'X)^{-1}$. Portanto, a distribuição condicional completa para β é da seguinte forma:

$$(\beta \mid W, m, \sigma^2, Y) \sim N((X'X)^{-1}X'(Y - Wm), (X'X)^{-1}\sigma^2) \qquad (3.19).$$

Utilizando-se o mesmo raciocínio, obtém-se a distribuição condicional completa para m.

$p(m \mid W, \beta, \sigma^2, Y) \alpha L(Y)\, p(m)\, \alpha$

$$\alpha\left(\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\left(Y-(X\beta+Wm)\right)'\left(I\sigma^2\right)^{-1}\left(Y-(X\beta+Wm)\right)\right\}\alpha$$

$$\alpha \exp\left\{-\frac{1}{2}\left(Wm-(Y-X\beta)\right)'\left(I\sigma^2\right)^{-1}\left(Wm-(Y-X\beta)\right)\right\}.$$

Assim, $Wm \sim N((Y-X\beta), I\sigma^2)$.

Pré-multiplicando-se Wm por $(W'W)^{-1}W'$, tem-se que:

$(m \mid W, \beta, \sigma^2, Y) \sim N((W'W)^{-1}W'(Y-X\beta), (W'W)^{-1}\sigma^2)$ (3.20).

A variância ambiental (σ^2) tem sua condicional completa derivada como segue.

$$p(\sigma^2 \mid W, \beta, m, Y) \alpha L(Y)\, p(\sigma^2)\, \alpha \left(\frac{1}{\sigma^2}\right)^{\frac{v}{2}+1} e^{\frac{-S_e}{2\sigma^2}} \text{ x}$$

$$\text{x } \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\left(Y-(X\beta+Wm)\right)'\left(I\sigma^2\right)^{-1}\left(Y-(X\beta+Wm)\right)\right\}\alpha$$

$$\alpha\left(\frac{1}{\sigma^2}\right)^{\frac{n+v}{2}+1} \exp\left\{-\frac{1}{\sigma^2}\left[\frac{\left(Y-(X\beta+Wm)\right)'\left(Y-(X\beta+Wm)\right)+S_e}{2}\right]\right\} \quad (3.21).$$

Observando a expressão (3.21), conclui-se que a distribuição condicional completa para a variância ambiental é uma gama invertida com parâmetros:

$$\left(\frac{n+v}{2}\right) \text{ e } \frac{\left(Y-(X\beta+Wm)\right)'\left(Y-(X\beta+Wm)\right)+S_e}{2}.$$

A condicional completa para a matriz W, como no exemplo anterior, é dada pela seguinte expressão:

$p(W \mid m, \beta, \sigma^2, Y) \propto L(Y)\, p(W) \propto$

$$\propto \exp\left\{-\frac{1}{2}\left(Y-(X\beta+Wm)\right)'\left(I\sigma^2\right)^{-1}\left(Y-(X\beta+Wm)\right)\right\} p(W) \qquad (3.22).$$

No entanto, para cada vetor w_i, tem-se: $p(w_i \mid \beta, m, w_{-i}, S_{AD}, V_A, V_D, t, \sigma^2, Y) \propto L(Y)\, p(w_i)$.

Observando-se as expressões de 3.19 a 3.22, nota-se que, com exceção desta última, todas as condicionais completas têm a forma de uma densidade conhecida e, portanto, podem ser facilmente amostradas utilizando-se o algoritmo do amostrador de Gibbs. O algoritmo do Metropolis-Hastings apresenta-se como uma alternativa para gerar os valores candidatos a compor a amostra da condicional completa da matriz W, podendo ser indicada como densidade candidata a priori para W, conforme já foi indicado na seção 3.1.

Definidas tais densidades e os algoritmos indicados em cada caso, as amostras das condicionais completas podem ser obtidas, permitindo assim inferir sobre os parâmetros considerados no modelo.

4 EXEMPLO DE APLICAÇÃO

A metodologia proposta em 3.1 foi ilustrada com um conjunto de dados oriundos de um experimento desenvolvido no setor de Olericultura da Universidade Federal de Lavras, com a finalidade de estudar a herança do fenômeno da partenocarpia em abobrinha (*Curcubita pepo*). As plantas amostradas foram avaliadas com notas de 1 a 5, conforme maior ou menor ocorrência de partenocarpia, tomando-se as médias das notas por planta atribuídas por três avaliadores. Foram consideradas as seis gerações mencionadas anteriormente e, os genitores contrastantes considerados foram a variedade Caserta (P_1) e a variedade Whitaker (P_2). O número de plantas amostradas para cada geração encontra-se na Tabela 5.

TABELA 5 Número de plantas amostradas por geração

Geração	Número de plantas
P_1	94
P_2	51
F_1	53
RC_{11}	86
RC_{12}	75
F_2	203
Total	562

4.1 Distribuição a priori

A distribuição a priori atribuída para os parâmetros V_A, V_D , S_{AD}, e σ^2, conforme mencionado na seção 3.2.1, é uma qui-quadrado invertida com

parâmetro de escala. No entanto, ainda resta definir os valores para os hiperparâmetros. As estimativas de máxima verossimilhança para esses parâmetros foram obtidas por Silva (2003), sendo que até mesmo a maior delas (σ^2 = 0,8) foi inferior a 1. Assim, no intuito de oferecer alguma informação prévia, porém, pouco informativa sobre os parâmetros, optou-se aqui por considerar que esta distribuição de qui-quadrado tenha média igual a 1 e variância também igual a 1, abrangendo, dessa forma, tanto os valores maiores quanto os menores que as estimativas de máxima verossimilhança. De posse da média e da variância da qui-quadrado, calculam-se os valores dos hiperparâmetros da maneira descrita a seguir.

Segundo Sorensen (1996), se uma variável aleatória x tem distribuição qui-quadrado da forma:

$$f(x)=\left[\Gamma\left(\frac{v}{2}\right)2^{\frac{v}{2}}\right]^{-1} S^{\left(\frac{v}{2}\right)}x^{-\left(\frac{v}{2}+1\right)}e\,\mathrm{xp}\left\{\frac{-S}{2x}\right\}\alpha\left(\frac{1}{x}\right)^{\frac{v}{2}+1}\exp\left\{\frac{-S}{2x}\right\},$$

então, a esperança da variável x é dada por $E[x]=\frac{S}{v-2}$.

A expressão para a esperança de x^2 foi obtida resolvendo a integral a seguir (usando um software matemático, MAPLE VR4).

$$E[x^2]=\frac{1}{C}\int_0^{\infty}x^2 f(x)dx=\frac{1}{C}2^{\left(\frac{v}{2}-2\right)}S^{\left(2-\frac{v}{2}\right)}\Gamma\left(\frac{v}{2}-2\right),$$

sendo a constante normalizadora $C=\int_0^{\infty}f(x)dx=\frac{\Gamma\left(\frac{v}{2}\right)}{\left(\frac{S}{2}\right)^{\frac{v}{2}}}$.

Assim, $E[x^2] = \frac{\left(\frac{S}{2}\right)^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}\right)} 2^{\left(\frac{v}{2}-2\right)} S^{\left(2-\frac{v}{2}\right)} \Gamma\left(\frac{v}{2}-2\right) = \frac{S^2}{(v-2)(v-4)}$.

Com esses valores, a variância de x pode ser calculada:

$$V(x) = E[x^2] - \left(E[x]\right)^2 = \frac{S^2}{(v-2)(v-4)} - \frac{S^2}{(v-2)^2} = \frac{2S^2}{(v-2)^2(v-4)}.$$

Finalmente, considerando cada um dos parâmetros de interesse como sendo a variável aleatória x, os valores dos hiperparâmetros são obtidos igualando-se essas expressões ao valor atribuído para a média e a variância, e resolvendo o sistema resultante.

$$\begin{cases} E[x] = 1 \\ V(x) = 1 \end{cases}$$

$$\begin{cases} \frac{S}{v-2} = 1 \\ \frac{2S^2}{(v-2)^2(v-4)} = 1 \end{cases}$$

$$\begin{cases} S = 4 \\ v = 6 \end{cases}$$

Portanto, a distribuição a priori para cada componente de variância V_A, V_D , S_{AD}, e σ^2, é uma qui-quadrado invertida com 6 graus de liberdade e parâmetro de escala S igual a 4. Para os demais parâmetros do modelo, as prioris foram apresentadas na seção 3.2.1.

4.2 Análise bayesiana

Para a realização das amostras das densidades condicionais completas, foi desenvolvido um programa utilizando o software estatístico SAS for Windows, SAS (2000), o qual está apresentado no Apêndice A.

No intuito de conferir maior agilidade ao processo iterativo, optou-se aqui por tomar aleatoriamente cinqüenta por cento das observações de cada geração, apresentadas na Tabela 5, obtendo-se assim um novo conjunto com 281 dados, os quais foram utilizados para a análise do modelo. Outro procedimento capaz de diminuir consideravelmente o esforço computacional, nesse caso, é considerar a matriz W como um único parâmetro, atualizando-a como um todo em cada iteração. Por esse procedimento, gera-se a matriz inteira, por meio das probabilidades a priori, apresentadas na Tabela 4 e toma-se a decisão de aceitar ou rejeitar essa nova matriz, de acordo com o algoritmo do Metropolis-Hastings. No entanto, se for considerado cada vetor w_i como um parâmetro distinto, o algoritmo de Metropolis-Hastings deve ser aplicado a cada um deles, ou seja, para cada uma das linhas de W correspondentes às gerações RC_{11}, RC_{12} e F_2. Assim, com o conjunto de dados adotado aqui, o algoritmo de Metropolis-Hastings seria aplicado a 183 parâmetros, apenas para a matriz W, ficando evidente que o esforço computacional envolvido seria consideravelmente maior. Dessa forma, justifica-se a atualização da matriz W como um todo, pois, caso seja comprovada a eficiência desse método, suas vantagens com relação à atualização da matriz a cada linha seriam incontestáveis. Para este exemplo, adotou-se, então, a atualização da matriz W como um todo, mas o outro procedimento também foi utilizado em uma aplicação descrita mais adiante.

No processo de verificação de convergência, utilizou-se o pacote BOA (Bayesian Output Analysis), executado através do software R, versão 1.9.0.

A convergência da matriz W e do parâmetro t foi tratada separadamente, pois, em tais casos, o objetivo foi determinar, entre apenas dois ou três valores possíveis, qual teve maior probabilidade de aceitação, ou seja, qual valor foi mais provável no modelo.

O critério de Raftery e Lewis indicou o tamanho ideal da seqüência (aproximadamente 50.000) e o tamanho do salto de uma iteração para a outra (nove para o parâmetro V_D, e menor que cinco para os demais parâmetros).

Para atender também ao critério de Gelman e Rubin, foram geradas duas cadeias com 50.000 iterações cada e as observações para compor a amostra da densidade a posteriori dos parâmetros foram tomadas a cada cinco iterações. As observações iniciais, dez por cento do total, como indicado pelo critério de Heidelberger e Welch, foram descartadas para evitar a influência dos valores iniciais. Os resultados da verificação de convergência estão apresentados na Tabela 6. Para os parâmetros gerados por meio do Metropolis-Hastings estão apresentadas também as freqüências relativas de aceitação.

TABELA 6 Critérios de convergência e freqüências relativas de aceitação para os parâmetros gerados por meio do algoritmo de Metropolis-Hastings.

Parâmetro	Geweke	Gelman e Rubin	Heidelberger e Welch Teste: aceita H_0		aceitação (%)
	p-valor	R	estacionária	Half-width	
V_A	0,8625	1,00124	sim	sim	48,22
V_D	0,8455	1,00001	sim	sim	41,13
S_{AD}	0,8646	0,99991	sim	sim	54,85
σ^2	0,8591	0,99989	sim	sim	16,27
μ	0,9672	1,00052	sim	sim	_
[a]	0,6610	1,00047	sim	sim	_
[d]	0,3981	1,00012	sim	sim	_
A	0,6751	1,00023	sim	não	_
D	0,6274	0,99996	sim	não	_

De acordo com os resultados apresentados na Tabela 6, o critério de Geweke não indicou falta de convergência para nenhum dos parâmetros, já que o p-valor foi maior que 0,05, em todos os casos.

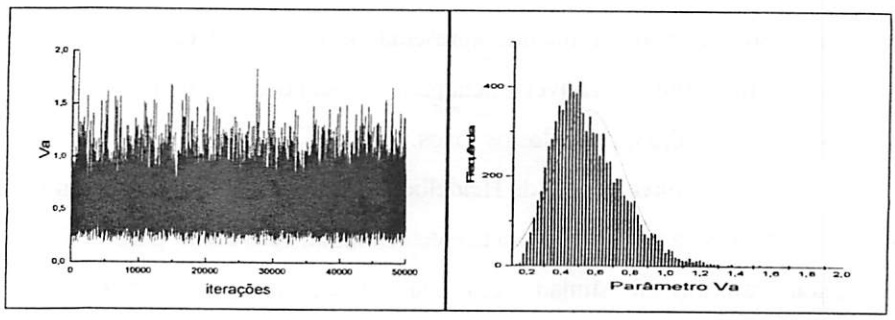
O critério de convergência de Heidelberger e Welch, que consiste em um teste para verificar se a seqüência gerada é estacionária, bem como para verificar se a média a posteriori foi estimada com uma acurácia preespecificada (teste de Half-Width), não levou à aceitação da hipótese de convergência para todos os parâmetros. A hipótese de estacionariedade da seqüência foi aceita para todos os parâmetros; porém, no teste de Half-Width, rejeitou-se a hipótese da acurácia da média para os parâmetros A e D, e aceitou-se a hipótese para os demais. Como a convergência só fica garantida quando se aceita a hipótese nula para todos os parâmetros nos dois testes, seria necessário um número maior de iterações, de acordo com o critério de Heidelberger e Welch.

Segundo o critério de Gelman e Rubin, a convergência foi alcançada para todos os parâmetros apresentados, pois, como se pode observar na Tabela 6, o fator de redução potencial de escala (R) ficou próximo de 1 em todos os casos.

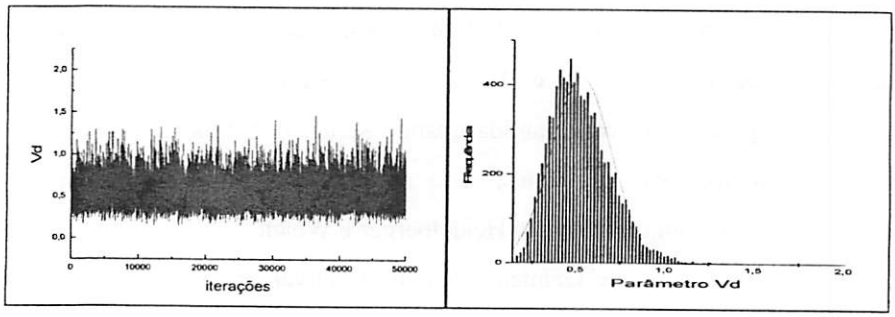
As freqüências relativas de aceitação (Tabela 6) apresentaram valores dentro ou próximos daqueles intervalos sugeridos como ideais na literatura, entre 40% e 50% (Chib & Greenberg, 1995), com exceção de σ^2, que teve probabilidades abaixo desses valores.

Embora o teste de Half-Width tenha indicado falta de convergência para dois parâmetros, optou-se por ainda assim analisar as estimativas a posteriori, levando em conta que os outros critérios sugeriram que houve convergência.

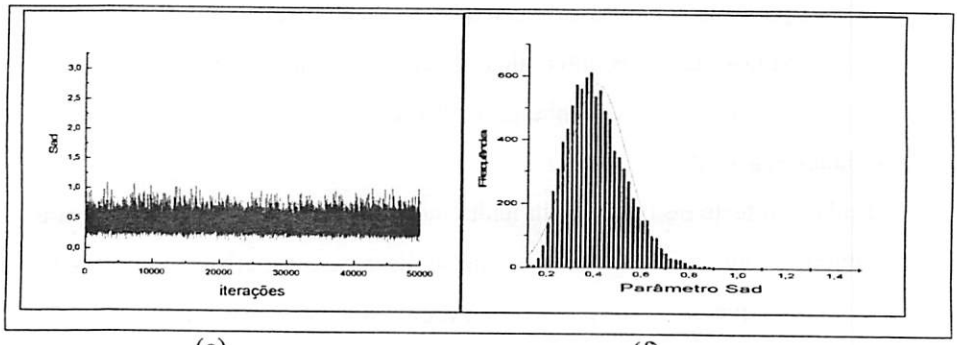
As distribuições a posteriori dos parâmetros estão ilustradas pelos histogramas a seguir (Figuras de 1 a 4), juntamente com as representações gráficas das cadeias completas geradas por meio do amostrador de Gibbs e do Metropolis-Hastings, com 50.000 iterações.



(a) (b)

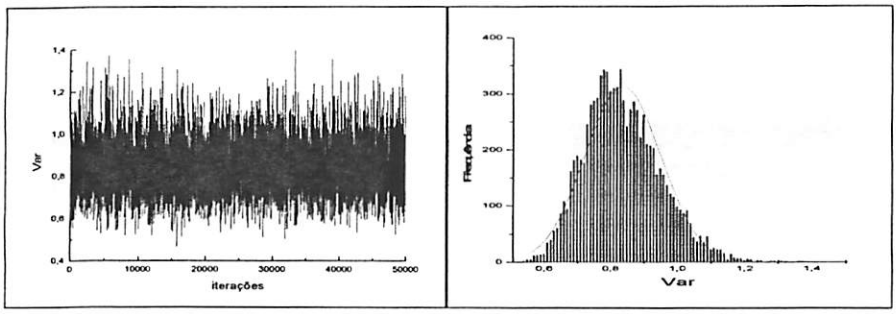


(c) (d)

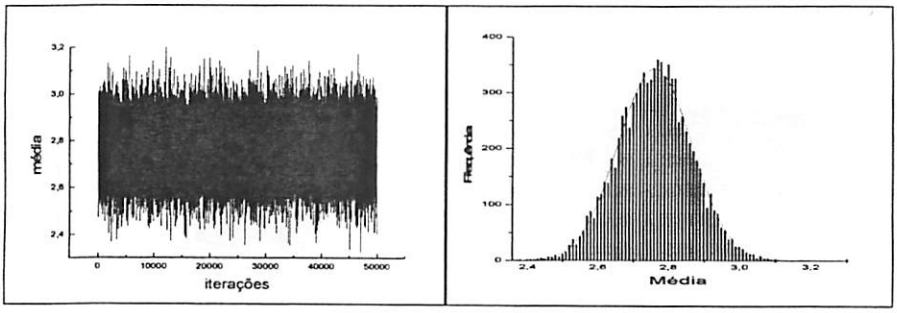


(e) (f)

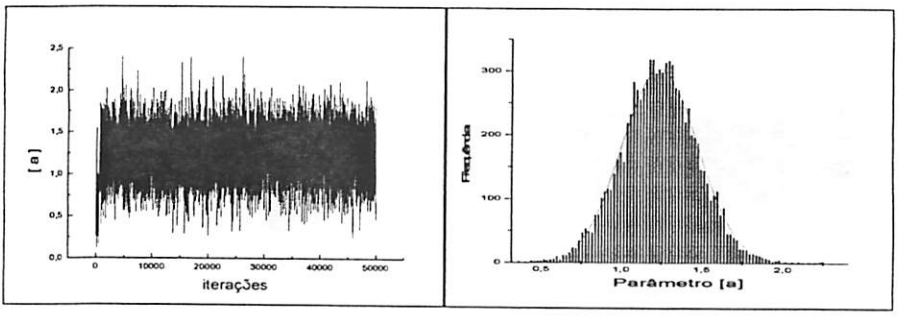
FIGURA 1 Representação gráfica das cadeias geradas pelo Metropolis-Hastings e da densidade a posteriori dos parâmetros V_A, V_D e S_{AD}.



(a) (b)

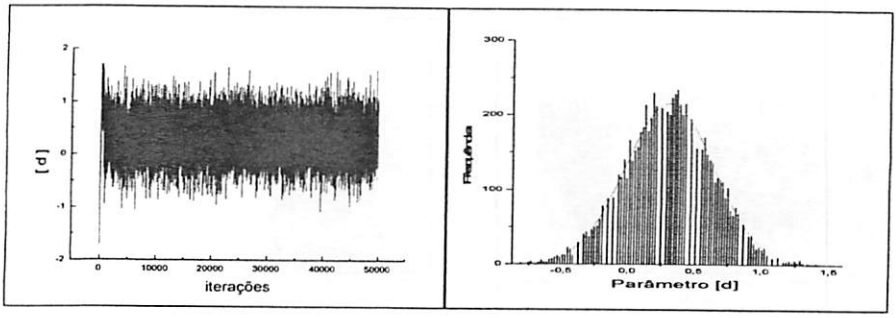


(c) (d)

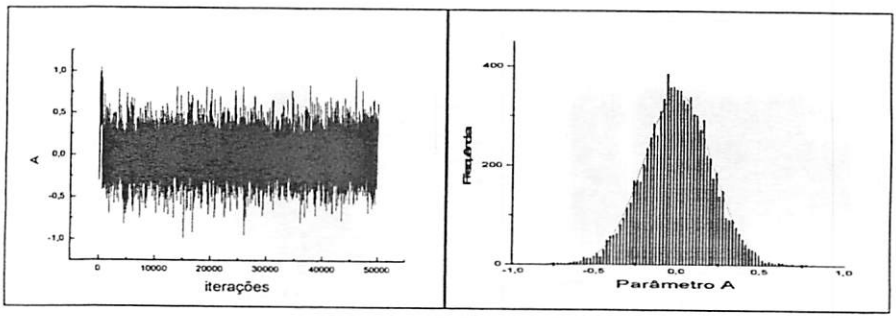


(e) (f)

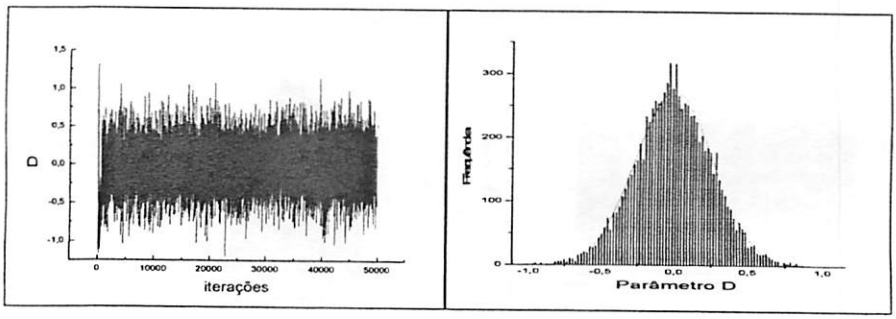
FIGURA 2 Representação gráfica das cadeias geradas e da densidade a posteriori dos parâmetros σ^2, μ e [a].



(a) (b)



(c) (d)



(e) (f)

FIGURA 3 Representação gráfica das cadeias geradas e da densidade a posteriori dos parâmetros [d], A e D.

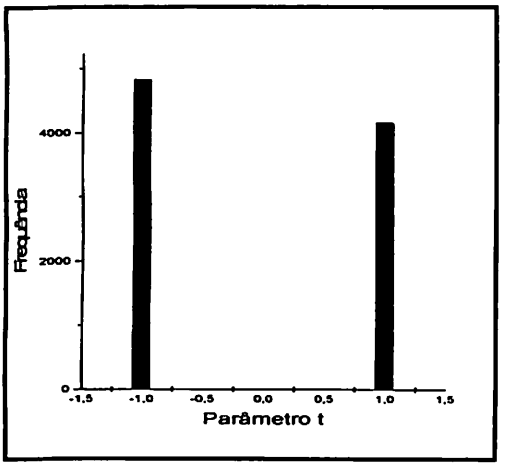


FIGURA 4 Representação gráfica da distribuição de probabilidade a posteriori do parâmetro t.

As amostras do parâmetro t apresentaram freqüências muito próximas para o valor negativo (4838) e para o valor positivo (4163), o que pode ser confirmado pelo gráfico da Figura 4. Portanto, o produto de t por S_{AD} considerado no modelo teve um sinal em 54% das iterações e um outro sinal em 46% dessas iterações, levando a crer que esse produto tanto pode ser somado como subtraído em qualquer um dos retrocruzamentos. Este fato sugere que a variância atribuída ao produto dos efeitos poligênicos aditivos pelos de dominância (S_{AD}) seja estatisticamente nula.

A partir das representações gráficas das cadeias de cada parâmetro, pode-se observar que o processo de geração das amostras não apresentou irregularidades, tais como valores discrepantes ou tendências de estabilização fora dos limites de convergência do processo iterativo.

Os histogramas apresentados nas Figuras 1, 2 e 3 sugerem que a distribuição a posteriori apresenta uma tendência à simetria, no caso dos

componentes de média e a uma ligeira assimetria no caso dos componentes de variância.

Para monitorar a convergência da matriz W optou-se por escolher aleatoriamente alguns elementos de cada geração. Assim, monitorou-se a convergência de 5 indivíduos das gerações RC_{11} e RC_{12} e de 10 indivíduos da geração F_2, no sentido de verificar a probabilidade que cada vetor w_i tem de corresponder a cada indivíduo. Na Tabela 7 estão apresentadas as probabilidades a posteriori de cada vetor w_i ser atribuído a cada um desses indivíduos.

TABELA 7 Probabilidades a posteriori de cada vetor w_i para cada geração.

Geração	Vetor w_i		
	[1 0]	[-1 0]	[0 1]
	Probabilidade		
RC_{11}	0,0	0,49	0,51
	0,0	0,49	0,51
	0,0	0,50	0,50
	0,0	0,49	0,51
	0,0	0,51	0,49
RC_{12}	0,51	0,0	0,49
	0,51	0,0	0,49
	0,48	0,0	0,52
	0,51	0,0	0,49
	0,49	0,0	0,51
F_2	0,25	0,49	0,25
	0,24	0,50	0,26
	0,25	0,50	0,25
	0,24	0,50	0,26
	0,24	0,50	0,26
	0,26	0,49	0,25
	0,25	0,5	0,25
	0,25	0,49	0,26
	0,26	0,49	0,25

Os resultados da Tabela 7 mostram que, ao considerar a matriz W como um único parâmetro (ou seja, alterando-a ou não como um todo em cada iteração), as probabilidades a priori de cada vetor w_i corresponder a um indivíduo das diferentes gerações são mantidas a posteriori. Este resultado não era o esperado, uma vez que se espera que os dados atualizem estas probabilidades. Por exemplo, um indivíduo da geração F_2 com o valor máximo (nota 5) para a característica em questão (partenocarpia) deve ter uma probabilidade maior de ser homozigoto dominante, em relação à probabilidade de ser homozigoto recessivo. Isto sugere que, talvez, a matriz W não tenha convergido (os critérios não foram usados com ela) e que assim, possivelmente, fosse melhor tentar alterar, em cada iteração, cada linha da matriz W individualmente. Esta estratégia foi tentada em uma outra análise descrita adiante.

As estimativas da média a posteriori de cada parâmetro com seus respectivos intervalos de credibilidade e erro de Monte Carlo estão apresentados na Tabela 8.

TABELA 8 Estimativas a posteriori por ponto e por intervalo (I. C. : intervalo de credibilidade, HPD: *highest probability density*), dos parâmetros e erro de Monte Carlo.

Parâmetro	média	I. C. - HPD		Erro de M. C.
		inferior	superior	
V_A	0,5535	0,2229	0,9762	0,0106
V_D	0,5268	0,2238	0,8563	0,0106
S_{AD}	0,4103	0,1972	0,6593	0,0106
σ^2	0,8358	0,6116	1,0525	0,0119
μ	2,7604	2,5559	2,9624	0,0107
[a]	1,2389	0,7482	1,6722	0,0107
[d]	0,3067	-0,3436	0,9508	0,0108
A	0,0009	-0,3954	0,4204	0,0107
D	-0,0038	-0,5213	0,5147	0,0107

De acordo com a Tabela 8, os intervalos de credibilidade para os parâmetros [d], A e D, os quais são referentes aos efeitos de dominância dos poligenes e aos efeitos aditivos e de dominância do gene principal, respectivamente, indicam que estes são estatisticamente iguais a zero. Portanto, considerando o conjunto com 281 observações, conclui-se que o fenômeno da partenocarpia em abobrinha não seria devido à ação de um gene principal. Aliás, as probabilidades da Tabela 7, que demonstram ausência de associação entre fenótipos e os genótipos do gene principal, foram coerentes com a não significância de A e D. Mas, como já realçado, é possível que a matriz W não tenha convergido. As estimativas dos componentes de variância e do componente aditivo dos poligenes foram todas diferentes de zero. No entanto, como o parâmetro t apresentou probabilidades semelhantes para o valor positivo e negativo, a soma de produtos dos efeitos poligênicos aditivos e de dominância pode ser considerada nula, conforme discutido anteriormente. O fato de [d] ter sido nulo, mas não V_D, não incorre em contradição, pois é sabido que os sinais envolvidos na soma de [d] podem eventualmente se cancelar. Nessas condições, considerando o conjunto com 281 observações, os resultados sugerem que a herança do fenômeno em questão deve-se à ação de poligenes sem efeitos de dominância.

Este resultado está aproximadamente semelhante às estimativas de máxima verossimilhança apresentadas por Silva (2003), apenas para os parâmetros μ e σ^2 ($\mu = 2,85$ e $\sigma^2 = 0,80$). Ao contrário do resultado obtido aqui, este autor concluiu que o fenômeno tem herança monogênica com efeito de dominância. Assim, poder-se-ia pensar que o tamanho reduzido da amostra e ou a eventual falta de convergência para a matriz W possam ter comprometido os resultados, uma vez que eram esperadas estimativas pontuais semelhantes às de máxima verossimilhança.

4.3 Exemplo de aplicação do modelo de herança monogênica

Atualizar a matriz W linha por linha em cada iteração, como comentado na seção anterior, demandaria um enorme esforço computacional. No entanto, os resultados do exemplo anterior dão indícios de que este talvez seja o procedimento mais adequado para a analise do modelo em questão. Assim, utilizando esse procedimento e o conjunto de dados completo, apresentado na Tabela 4, foi feito o ajuste do modelo contendo apenas os parâmetros μ, A, D e σ^2, além, é claro, dos vetores w_i. Este ajuste foi motivado pelo fato de que Silva (2003), analisando esse conjunto de dados concluiu pela ação de um único gene principal.

Foram geradas duas cadeias com 10.000 iterações cada, sendo descartadas as 2.000 iterações iniciais e tomadas as amostras a cada duas iterações.

Para este exemplo, foram analisadas as probabilidades de aceitação de cada vetor w_i, de forma independente. Assim, para cada nova linha gerada da matriz W, correspondente às gerações segregantes, tomava-se a decisão de aceitar ou rejeitar o novo valor, de acordo com o algoritmo de Metropolis-Hastings. Dessa forma, a matriz era quase sempre atualizada, pois, a cada iteração, cada uma das 364 linhas tinha uma oportunidade independente de ser substituída.

Da mesma forma que no exemplo anterior, foram escolhidos alguns elementos de cada geração segregante para verificar a probabilidade a posteriori de cada vetor w_i ser atribuído a estes indivíduos. Os resultados dessa verificação estão apresentados na Tabela 9.

TABELA 9 Probabilidades a posteriori de cada vetor w_i para indivíduos de cada geração.

Geração	Vetor w_i			Y
	[1 0]	[-1 0]	[0 1]	
	Probabilidade			
	0,0	0,9997	0,0003	1,000
	0,0	0,9995	0,0005	1,000
RC_{11}	0,0	0,9820	0,0180	1,667
	0,0	0,5326	0,4674	2,333
	0,0	0,0003	0,9997	3,667
	0,9105	0,0	0,0895	4,333
	0,1537	0,0	0,8463	3,333
RC_{12}	0,0	0,0	1,0000	1,000
	0,0003	0,0	0,9997	2,000
	0,0005	0,0	0,9995	2,000
	0,7133	0,0	0,2867	4,000
	0,0230	0,0100	0,9670	3,000
	0,0	0,9788	0,0212	1,667
	0,0	0,9730	0,0270	1,667
F_2	0,0	0,9635	0,0365	1,667
	0,9828	0,0	0,0172	5,000
	0,5404	0,0	0,4596	4,000
	0,0038	0,0812	0,9150	2,667
	0,9848	0,0	0,0152	5,000
	0,8155	0,0	0,1845	4,333

Pode-se ver na Tabela 9 que a estratégia de atualizar cada linha da matriz W levou a uma diferenciação quanto às probabilidades de cada genótipo. Na mesma Tabela 9 estão os fenótipos dos indivíduos em questão. Percebe-se uma clara coerência entre estes e aquelas. Por exemplo, na geração RC_{11}, os dois indivíduos com valor fenotípico 1,00 tiveram altas probabilidades de serem homozigotos, ou seja, de estarem associados ao vetor w_i igual a [-1 0]. Conseqüentemente, a probabilidade de que estes indivíduos sejam heterozigotos, ou seja, de estarem associados ao vetor w_i igual a [0 1], foi praticamente nula. Lembrando que as plantas foram avaliadas com notas de 1,00 a 5,00, de acordo com uma maior ou menor incidência de partenocarpia, fica evidente a coerência

do resultado, já que uma planta com a menor nota (1,00) é aquela em que a característica não se manifestou e, assim, seria mais lógico que ela apresentasse apenas os alelos que diminuem a expressão do caráter, como ocorre no indivíduo homozigoto desta geração. Na geração RC_{12}, o indivíduo com valor fenotípico igual a 1,00 apresentou probabilidade nula de ser homozigoto, associado ao vetor w_i igual a [0 1]. Isso também foi perfeitamente coerente, pois, com a menor nota, seria improvável que ele apresentasse dois alelos que aumentam a expressão do caráter, como ocorre com os indivíduos homozigotos desta geração. Ainda na geração RC_{12}, um indivíduo com nota igual a 4,33 apresentou uma alta probabilidade (0,9105) de ser homozigoto contra uma baixa probabilidade de ser heterozigoto, evidenciando que, com um valor fenotípico alto, é bem mais provável que ele apresente dois alelos que aumentam a expressão do caráter, do que apenas um, como ocorre com os indivíduos heterozigotos. Portanto, percebe-se claramente, pela Tabela 9, que, na geração RC_{11}, quanto maior o valor fenotípico, maior a probabilidade de o indivíduo ser heterozigoto e, na geração RC_{12}, maior a probabilidade de o indivíduo ser homozigoto dominante com o aumento do fenótipo. Comportamentos coerentes com a escala de valores fenotípicos também foram observados na geração F_2 (Tabela 9). Exemplo disso são as plantas avaliadas com notas mais altas (4,00 ou mais), que apresentaram probabilidades nulas de serem homozigotas recessivas (w_i = [-1 0]) e as plantas avaliadas com notas baixas (1,667), que apresentaram probabilidade nula de serem homozigotas dominantes (w_i = [1 0]).

A Tabela 9 ilustra uma grande vantagem da inferência bayesiana, no presente contexto, pois possibilita explicitar claramente probabilidades de cada indivíduo ser de um ou outro genótipo de um gene principal. Isto auxiliaria o geneticista de plantas a selecionar aqùeles indivíduos com maiores probabilidades de terem fixado o alelo favorável.

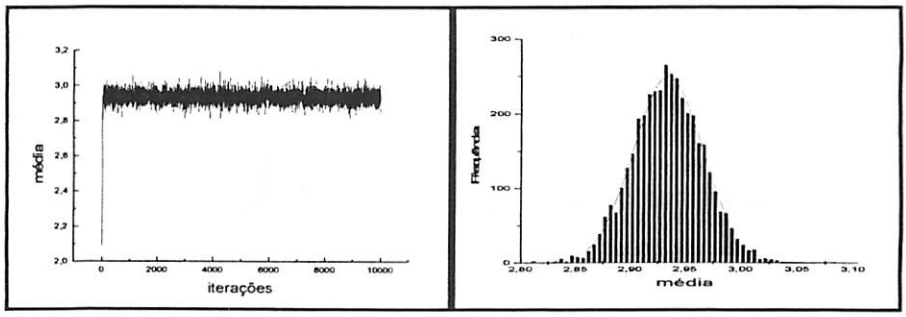
As cadeias geradas para os parâmetros do modelo foram submetidas aos critérios de convergência descritos anteriormente e os resultados estão dispostos na Tabela 10.

TABELA 10 Resultados dos critérios de convergência para os parâmetros do modelo.

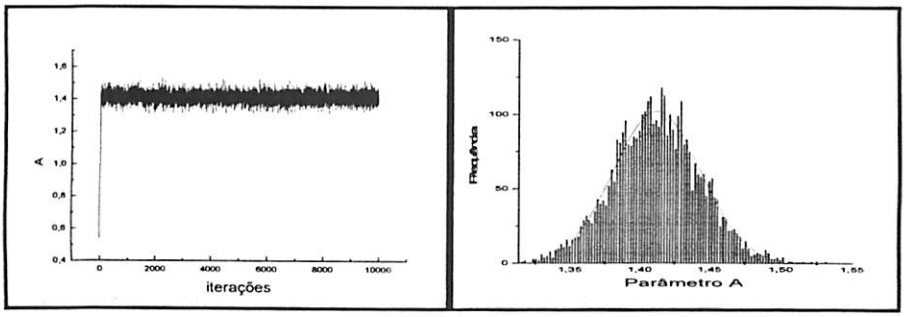
Parâmetro	Geweke	Gelman e Rubin	Heidelberger e Welch Teste: aceita H_0	
	p-valor	PSRF (R)	estacionária	Half-width
σ^2	0,9728	1,0004	sim	não
μ	0,9925	1,0006	sim	sim
A	0,9786	0,9997	sim	sim
D	0,9591	1,0008	sim	não

Os resultados da Tabela 10 apontam que dois parâmetros não passaram no teste de Half-Widht, ou seja, a média a posteriori não está sendo estimada com a acurácia predefinida pelo critério de Heidelberger e Welch. No entanto, pelos critérios de Geweke e de Gelman e Rubin, a convergência foi atingida para todos os parâmetros, sugerindo, assim, que se devam analisar as distribuições e as estimativas a posteriori de tais parâmetros.

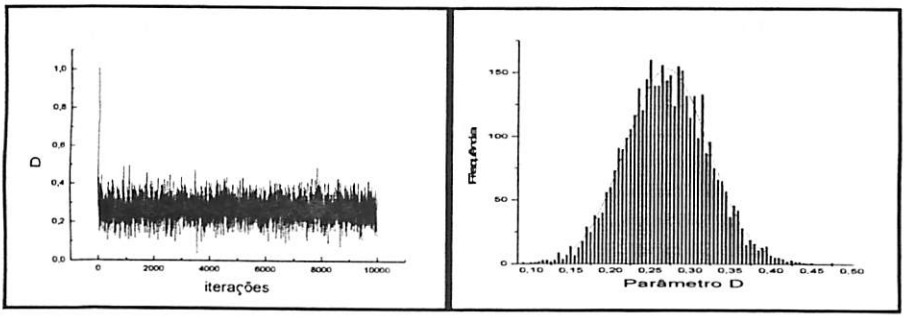
A distribuição a posteriori dos parâmetros e as seqüências completas geradas pelo amostrador de Gibbs estão representadas graficamente pelas Figuras 5 e 6.



(a) (b)

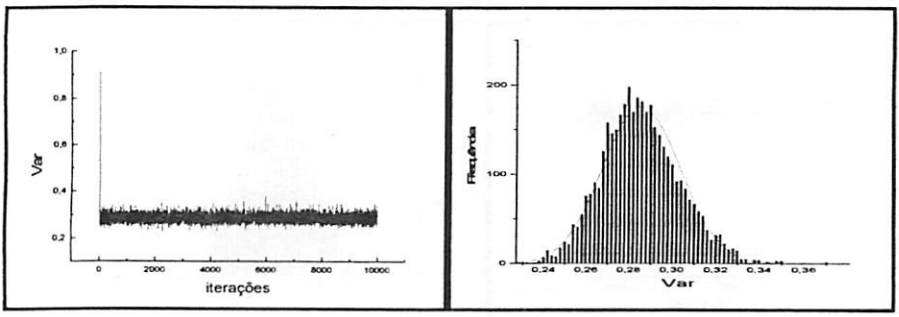


(c) (d)



(e) (f)

FIGURA 5 Representação gráfica da seqüência gerada e da distribuição a posteriori dos parâmetros A, D e μ (média).



(a) (b)

FIGURA 6 Representação gráfica da seqüência gerada e da distribuição a posteriori do parâmetro σ^2 (Var).

A partir das representações gráficas das cadeias de cada parâmetro geradas pelo amostrador de Gibbs (Figuras 5(a, c, e) e 6a), pôde-se observar que o processo de geração das amostras não apresentou irregularidades, tais como valores discrepantes ou tendências de estabilização fora dos limites de convergência do processo iterativo. As distribuições a posteriori para os parâmetros μ, A e D mostraram um comportamento simétrico (Figura 5(b, d, f)), o que não ocorreu no caso da variância ambiental (σ^2), cuja distribuição a posteriori mostrou uma ligeira assimetria (Figura 6b).

Na Tabela 11 estão dispostas as estimativas por ponto e por intervalo dos parâmetros do modelo, e do erro de Monte Carlo.

TABELA 11 Estimativas dos parâmetros do modelo (I. C.: intervalo de credibilidade, HPD: *highest probability density*) e erro de Monte Carlo.

Parâmetro	média	I. C. - HPD		Erro de Monte Carlo
		inferior	superior	
μ	2,93	2,87	2,99	0,016
A	1,41	1,34	1,47	0,016
D	0,27	0,17	0,37	0,016
σ^2	0,29	0,25	0,32	0,016

Com base nos resultados expressos na Tabela 11, verifica-se que, apesar de o modelo aqui sugerido levar em conta fatores antes não considerados, as estimativas a posteriori para os parâmetros μ, A e D tiveram valores aproximadamente iguais àqueles obtidos por Silva (2003), pelo método da máxima verossimilhança. Isto sugere que, mesmo alguns critérios de convergência não sendo satisfeitos, inferências válidas ainda são possíveis. Para a variância ambiental (σ^2), o valor obtido aqui (0,29) mostrou-se abaixo da estimativa de máxima verossimilhança (σ^2 = 0,8). Contudo, deve-se ressaltar que, rigorosamente, o modelo de Silva (2003) não é o mesmo usado aqui, em que a contribuição individual de cada planta, no tocante ao gene principal, é levado em conta. Assim, por exemplo, um indivíduo fenotipicamente extremo na geração RC_{11} tem probabilidades diferenciadas de ser homozigoto ou heterozigoto, no presente modelo, enquanto que, no modelo de Silva (2003), estas chances são iguais. Isto, provavelmente, leva a um aumento na estimativa de σ^2, a qual é, essencialmente, a variação dentro de cada classe genotípica, para a qual todos os indivíduos da geração contribuem com 50% de peso no modelo de Silva (2003). Aqui, essa contribuição é levada em conta com pesos mais adequados, levando a uma estimativa mais realista de σ^2.

Este último resultado evidencia o fato de que, no exemplo anterior, a não convergência da matriz W, provavelmente, foi a causa de estimativas

estatisticamente nulas para os parâmetros A e D, pois estes estão diretamente associados com os vetores w_i de W. Assim, existem indícios de que atualizar a matriz W a cada linha, naquele exemplo, teria levado à convergência da mesma e dos parâmetros A e D não nulos, encontrando-se, assim, as probabilidades do genótipo de cada planta com relação ao gene principal. Portanto, já que a viabilidade da aplicação desta alternativa foi comprovada por esta última análise, esta opção mostra-se, seguramente, ser a melhor e mais indicada para trabalhos futuros.

5 CONCLUSÕES

A metodologia proposta para estudo de herança monogênica e poligênica mostrou-se uma alternativa viável, com relação aos objetivos a que se propôs.

Por meio do modelo completo foi possível obter um submodelo considerando apenas gene principal, o qual mostrou-se adequado para obterem-se as estimativas dos parâmetros.

O modelo genético considerado foi adaptado com sucesso aos métodos bayesianos, possibilitando calcular as probabilidades do genótipo de cada indivíduo com relação ao gene principal.

A metodologia foi ilustrada com um conjunto de dados reais de um estudo de herança da partenocarpia em abobrinha, para o caso do modelo completo e do submodelo. A coerência deste último com resultados anteriores corrobora que a partenocarpia em abobrinha seja governada por um gene principal.

Os resultados também sugerem a conclusão de que a atualização da matriz W linha por linha conduz a uma maior convergência da mesma, ainda que demandando considerável esforço computacional a mais.

REFERÊNCIAS BIBLIOGRÁFICAS

ARIAS, C. A. A.; TOLEDO, J. F. F.; YORINORI, J. T. An improved procedure for testing theoretical segregation in qualitative genetic studies of soybeans. **Revista Brasileira de Genética**, Ribeirão Preto, v. 17, n. 3, p. 291-297, set. 1994.

CHANGJIAN, J.; XUEBIAO, P.; MINGHONG, G. The use of mixture models to detect effects of major genes on quantitative characters in a plant breeding experiment. **Genetics**, Baltimore, v. 136, n. 2, p. 383-394, Feb. 1994.

CHIB, S.; GREENBERG, E. Understanding the Metropolis-Hastings Algorithm. **The American Statistician**, Alexandria, v. 49, n. 4, p. 327-335, Nov. 1995.

FREITAS, J. A.; MALUF, W. R.; CARDOSO, M. G.; GOMES, L. A. A.; BEARZOTI, E. Inheritance of foliar zingiberene contents and their relationship to trichome densities and whitefly resistance in tomatoes. **Euphytica**, Wageningem, v. 127, n. 2, p. 275-287, 2002.

GAMERMAN, D. **Markov chain Monte Carlo:** Stochastic Simulation for Bayesian Inference. London: Chapman & Hall, 1997.

HASTINGS, W. K. Monte Carlo Sampling methods using Markov chains and their applications. **Biometrika**, London, v. 57, n. 1, p. 97-109, 1970.

JANSS, L. L. G.; THOMPSON, R.; VAN ARENDONK, J. A. M. Application of Gibbs sampling for inference in a major gene-polygenic inheritance model in animal populations. **Theoretical and Applied Genetics**, Berlin, v. 91, n. 6/7, p. 1137-1147, Nov. 1995.

JANSS, L. L. G.; VAN ARENDONK, J. A. M.; BRASCAMP, E. W. Bayesian statistical analyses for presence of single genes affecting meat quality traits in a crossed pig population. **Genetics**, Baltimore, v. 145, n. 2, p. 395-408, FEb. 1997.

LOU, X. Y.; ZHU, J. Analysis of genetics effects of major genes and polygenes on quantitative traits. **Theoretical and Applied Genetics**, Berlin, v. 104, n. 2/3, p. 414-421, Feb. 2002.

MATHER, K.; JINKS, J. L. **Introdução à genética biométrica**. Ribeirão Preto – SP: Sociedade Brasileira de Genética, 1984. 242 p.

METROPOLIS, N.; ROSEMBLUT, A. W.; ROSEMBLUT, M. N.; TELLER, A. H.; TELLER, E. Equations of state calculations by fast computing machines. **Journal of Chemical Physics**, New York, v. 21, p. 1087-1092, 1953.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of statistics**. 3. ed. Tokyo: McGraw-Hill Kogakusha, 1974. 564 p.

NOGUEIRA, D. A. **Proposta e avaliação de critérios de convergência para o método de Monte Carlo via Cadeias de Markov: casos uni e multivariados**. 2004. 121 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, MG.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. **Estatística bayesiana**. Lisboa: Portugal, 2003. 447 p.

R: Copyright 2004. **The R Foundation for Statistical Computing**, V. 1.9.0

RAMALHO, M. A. P.; SANTOS, J. B. dos; ZIMMERMANN, M. J. de O. **Genética quantitativa em plantas autógamas**: aplicação ao melhoramento do feijoeiro. Goiânia: UFG, 1993. 271 p.

SAS INSTITUTE. **SAS for Windows, Release 8.** Cary, N. C, 2000.

SILVA, W. P. **Estimadores de máxima verossimilhança em misturas de densidades normais**: uma aplicação em genética. 2003. 60 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, MG.

SORENSEN, D. **Gibbs sampling in quantitative genetics**. Denmark: Danish Institute of Animal Science, Department of Breeding and Genetics, 1996. n. 82, 186 p.

SOUZA SOBRINHO, F. **Herança da reação de resistência à raça 2 de *Meloidogyne incognita* na pimenta *Capsicum annuum* L. cv Carolina Cayenne.** 1998. 57 p. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Universidade Federal de Lavras, Lavras, MG.

WRIGHT, D. R.; STERN, H. S.; BERGER, P. J. Comparing traditional and Bayesian analyses of selection experiments in animal breeding. **Journal of Agricultural, Biological and Environmental Statistics,** Washington, v. 5, n. 2, p. 240-256, June 2000.

APÊNDICE A. PROGRAMA UTILIZADO NO PROCEDIMENTO DE OBTENÇÃO DAS AMOSTRAS DAS DENSIDADES CONDICIONAIS COMPLETAS.

Programa do modelo completo utilizando o conjunto de dados com 281 observações

```
data gen;
 infile 'C:\gene\dados281.txt';   /* importar os dados y */
 input x1 ;
 proc iml;
 use gen;
 read all  var _num_ into dados;
 y=j(281,1,0);
 y[1:281,1]=dados[1:281,1];
 close gen;

par = j(1,30,0);
 do p=1 to 30;
par[1,p]=p;
 end;
do cad = 1 to 2;       /*valores iniciais*/
if cad = 1 then
     do;
va=0.6; t=1; Sad=0.5; Vd=0.4;beta={4,9,1};m={1,2};Se= 1.2 ;
     end;
if cad = 2 then
     do;
va=1.5; t=1; Sad=1.2;Vd=0.9;beta={0.4,0.2,0.1};
m={0.2,0.5}; Se=0.5;
     end;

/*hiperparâmetros*/
Te=4; Ta=4; Tad=4;Td=4; ve=6; hva=6;hvd=6; vad=6;
/*Matriz X*/
X=j(281,3,0);
X[1:281,1] = 1;
X[1:47,2] = -1;
X[48:72,2] = 1;
X[99:141,2] = -0.5;
X[142:179,2] = 0.5;
X[73:98,3] = 1;
X[99:281,3] = 0.5;


/* Matriz W*/
```

```
Start matrizW;
Wteste=j(281,2,0);
do i=1 to 47;          /*valores para P1*/
  Wteste[i,1]= -1;
  Wteste[i,2]= 0;
end;
do i=48 to 72;         /*valores para P2*/
  Wteste[i,1]= 1;
  Wteste[i,2]= 0;
end;
do i= 73 to 98;        /*valores para F1*/
  Wteste[i,1]= 0;
  Wteste[i,2]= 1;
end;
do i=99 to 141;        /*valores para o RC11*/
  u = ranuni(0);
if u <= 0.5 then
  do;
    Wteste[i,1]= -1;
    Wteste[i,2]= 0;
  end;
  else
  do;
   Wteste[i,1]= 0;
   Wteste[i,2]= 1;
  end;
end;
do i=142 to 179;          /*valores para o RC12*/
 u= ranuni(0);
 if u <= 0.5 then
   do;
     Wteste[i,1]= 1;
     Wteste[i,2]= 0;
   end;
   else
   do;
    Wteste[i,1]= 0;
    Wteste[i,2]= 1;
   end;
end;
do i = 180 to 281;          /*valores para F2*/
 u= ranuni(0);
 if u <= 0.25 then
  do;
   Wteste[i,1]= -1;
   Wteste[i,2]= 0;
  end;
```

```
   else
    if  u<=0.5  then
     do;
      Wteste[i,1]= 1;
      Wteste[i,2]= 0;
     end;
     else
     do;
       Wteste[i,1]= 0;
       Wteste[i,2]= 1;
     end;
end;
finish matrizW;
run matrizW;
W = Wteste;

iter = 50000;

Abeta= j(iter,3,0); Am= j(iter,2,0); AVa=j(iter,1,0);
AVd=j(iter,1,0); ASad=j(iter,1,0); ASe=j(iter,1,0);
AW= j(iter,20,0); At= j(iter,1,0); cont_Va = 0;
cont_Vd = 0; cont_Sad = 0; cont_Se = 0;
cont_t = 0; cont_W = 0; cont = j(iter,1,0);
burnin = 0.1*iter; salto = 5; aux=2;
total = 1+(iter-burnin)/salto;  k=burnin;
thin1 =j(total+1,10,0); thin2 =j(total+1,10,0);
thin3 =j(total+1,10,0); thin1[1,] = par[1,1:10];
thin2[1,] = par[1,11:20]; thin3[1,] = par[1,21:30];

do n= 1 to iter;

  /*matriz ZGZ*/
Za=j(98,183,0)//i(183);
Aa=((0.5*i(81))||j(81,102,0))//(j(102,81,0)||i(102));
Aad = ((-1)*i(43)||j(43,140,0))//(j(38,43,0)||i(38)||
j(38,102,0))//j(102,183,0);
Ad = i(183);
ZGZ=
va*(Za*Aa*t(Za))+vd*(Za*Ad*t(Za))+t*Sad*(Za*Aad*t(Za));

/*gerando beta (Amostrador de Gibbs) */
 a= inv(x`*x)*x`;
 medbeta = a*(y-W*m);
 varbeta= a*(ZGZ + Se*I(281))*a`;
 B= j(3,1,0);
do j =1 to 3;
  B[j] = rannor(0);
```

```
 end;
 chol= root(varbeta);
 beta = chol`*B + medbeta;
 abeta[n,1:3] = t(beta);

/*gerando m (Amostrador de Gibbs) */
 c= inv(w`*w)*w`;
 medm = c*(y-x*beta);
 varm= c*(ZGZ + Se*I(281))*c`;
 D= j(2,1,0);
do j =1 to 2;
 D[j] = rannor(0);
end;
 chol= root(varm);
 m = chol`*D + medm;
 Am[n,1:2] = t(m);

/*Valores para Va (gama invertida) (Metropolis-Hastings)*/
  Vateste= 2*(Sad-Vd)-1;
do while (Vateste <= 2*(Sad-Vd) );
  alfag=(hva/2);
  betag=2/Ta;
  agama= rangam(10,alfag);
  Vateste= 1/(betag*agama);
end;
zgzatual= vateste*(Za*Aa*Za`)+ vd*(Za*Ad*Za`)+
t*Sad*(Za*Aad*Za`);
fatual= (det(zgzatual+I(281)*Se))**(-1/2)*exp(-0.5*
(Y-X*beta -W*m)`*inv(zgzatual+i(281)*Se)*(Y-X*beta -W*m));

fant= (det(ZGZ+I(281)*Se))**(-1/2)* exp(-0.5*(Y-X*beta -
W*m)`*inv(ZGZ+i(281)*Se)*(Y-X*beta -W*m));
if fant >0 then
 do;
  prob= min(1,fatual/fant);
 end;
 else
  prob=1;
un=ranuni(0);
if un <=prob then
 do;
  Va=vateste;
  cont_Va = cont_Va + 1;
 end;
AVa[n,1]=Va;

/*Valores para Vd (gama invertida) (Metropolis-Hastings)*/
```

```
W[179,1] ||W[190,1] ||W[201,1]||W[212,1]||W[223,1] ||
W[234,1] ||W[245,1] || W[256,1] || W[267,1] || W[278,1] ) ;
AW[n,] = pW ;

/*Valores para o parâmetro t*/
u = ranuni(0);
if u <= 0.5  then
  t2= -1;
else
  t2 = 1;
zgzat= va*(Za*Aa*Za`)+ vd*(Za*Ad*Za`)+t2*Sad*(Za*Aad*Za`);
fatual= (det(zgzat+I(281)*Se))**(-1/2)* exp(-0.5*(Y-X*beta-
        W*m)`*inv(zgzat+i(281)*Se)*(Y-X*beta -W*m));
fant= (det(ZGZ+I(281)*Se))**(-1/2)* exp(-0.5*(Y-X*beta -
W*m)`*inv(ZGZ+i(281)*Se)* (Y-X*beta -W*m));
if fant >0 then
 prob= min(1,fatual/fant);
else
 prob=1;
un=ranuni(0);
if un <=prob then
 do;
  t = t2;
  cont_t = cont_t + 1;
 end;
At[n,1]= t;

if k=n then
  do;
  cadeia = (AVa||AVd||ASad||ASe||At||abeta||am) ;
  thin1[aux,1:10] = cadeia[n,1:10];
  thin2[aux,1:10] = aW[n,1:10];
  thin3[aux,1:10] = aW[n,11:20];
  aux = aux + 1 ; k=k+salto ;
 end;
end;   /* do iter*/

if cad = 1 then
 do;
  cad1 = (par[1,1:10]//cadeia);
  thina1 = thin1;
  thina2 = thin2;
  thina3 = thin3;
  cont1 = j(6,1,0);
  cont1[1,1] = cont_Va;
  cont1[2,1] = cont_Vd;
  cont1[3,1] = cont_Sad;
```

```
 do;
  Sad=Sadteste;
  cont_Sad = cont_Sad + 1;
 end;
ASad[n,1]=Sad;

/*Valores para Se (gama invertida) (Metropolis-Hastings) */
 alfag=(ve/2);
 betag= 2/Te;
 agama= rangam(10,alfag);
 Seteste= 1/(betag*agama);

fatual= (det(zgzatual+I(281)*Seteste))**(-1/2)*exp(-0.5*(Y-
X*beta -W*m)`*inv(zgzatual+i(281)*Seteste)*(Y-X*beta-W*m));
fant= (det(ZGZ+I(281)*Se))**(-1/2)* exp(-0.5*(Y-X*beta -
W*m)`*inv(ZGZ+i(281)*Se)*(Y-X*beta -W*m));
if fant >0 then
  prob= min(1,fatual/fant);
else
  prob=1;
un=ranuni(0);
if un <=prob then
 do;
  Se=Seteste;
  cont_Se = cont_Se + 1;
 end;
ASe[n,1]=Se;

/* valores para W*/
run matrizW;
fatual= exp(-0.5*(Y-X*beta -Wteste*m)`*
        inv(zgzatual+i(281)*Se)*(Y-X*beta -Wteste*m));

fant= exp(-0.5*(Y-X*beta -W*m)`*inv(ZGZ+i(281)*Se)*
      (Y-X*beta -W*m));
if fant >0 then
 prob= min(1,fatual/fant);
else
 prob=1;
un=ranuni(0);
if un <=prob then
 do;
  W=Wteste;
  cont_W =  cont_W +1;
 end;
pW = ( W[99,1] ||W[108,1]||W[117,1] ||W[126,1] ||W[135,1]||
W[142,1]||W[150,1] ||W[159,1] ||W[168,1] ||W[177,1]||
```

```
W[179,1] ||W[190,1] ||W[201,1]||W[212,1]||W[223,1] ||
W[234,1] ||W[245,1] || W[256,1] || W[267,1] || W[278,1] ) ;
AW[n,] = pW ;

/*Valores para o parâmetro t*/
u = ranuni(0);
if u <= 0.5  then
  t2= -1;
else
  t2 = 1;
zgzat= va*(Za*Aa*Za`)+ vd*(Za*Ad*Za`)+t2*Sad*(Za*Aad*Za`);
fatual= (det(zgzat+I(281)*Se))**(-1/2)* exp(-0.5*(Y-X*beta-
        W*m)`*inv(zgzat+i(281)*Se)*(Y-X*beta -W*m));
fant= (det(ZGZ+I(281)*Se))**(-1/2)* exp(-0.5*(Y-X*beta -
W*m)`*inv(ZGZ+i(281)*Se)* (Y-X*beta -W*m));
if fant >0 then
 prob= min(1,fatual/fant);
else
 prob=1;
un=ranuni(0);
if un <=prob then
 do;
  t = t2;
  cont_t = cont_t + 1;
 end;
At[n,1]= t;

if k=n then
  do;
  cadeia = (AVa||AVd||ASad||ASe||At||abeta||am) ;
  thin1[aux,1:10] = cadeia[n,1:10];
  thin2[aux,1:10] = aW[n,1:10];
  thin3[aux,1:10] = aW[n,11:20];
  aux = aux + 1 ; k=k+salto ;
 end;
end;   /* do iter*/

if cad = 1 then
 do;
  cad1 = (par[1,1:10]//cadeia);
  thina1 = thin1;
  thina2 = thin2;
  thina3 = thin3;
  cont1 = j(6,1,0);
  cont1[1,1] = cont_Va;
  cont1[2,1] = cont_Vd;
  cont1[3,1] = cont_Sad;
```

```
   cont1[4,1] = cont_Se;
   cont1[5,1] = cont_t;
   cont1[6,1] = cont_W;
  end;

if cad = 2 then
 do;
  cad2 = (par[1,1:10]//cadeia);
  thinb1 = thin1;
  thinb2 = thin2;
  thinb3 = thin3;
  cont2 = j(6,1,0);
  cont2[1,1] = cont_Va;
  cont2[2,1] = cont_Vd;
  cont2[3,1] = cont_Sad;
  cont2[4,1] = cont_Se;
  cont2[5,1] = cont_t;
  cont2[6,1] = cont_W;
 end;
contador = cont1||cont2 ;

end; /* do cad */

filename cadeia 'c:\gene\cad1.txt';  /* Arquivo de saída */
 file cadeia;
 do i=1 to nrow(cad1);
  do j=1 to ncol(cad1);
   put (cad1[i,j]) +2 @;
  end;        put;        end;
closefile cadeia;
```