RICARDO LUIS DOS REIS

ESTIMAÇÃO DE PARÂMETROS DE POPULAÇÕES COM BASE EM FREQÜÊNCIAS ALÉLICAS UTILIZANDO INFERÊNCIA BAYESIANA ATRAVÉS DO SOFTWARE LIVRE R

Monografia de graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do curso de Ciência da Computação para obtenção do título de Bacharel em Ciência da Computação.

LAVRAS
MINAS GERAIS – BRASIL
2006

RICARDO LUIS DOS REIS

ESTIMAÇÃO DE PARÂMETROS DE POPULAÇÕES COM BASE EM FREQÜÊNCIAS ALÉLICAS UTILIZANDO INFERÊNCIA BAYESIANA ATRAVÉS DO SOFTWARE LIVRE R

Monografia de graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do curso de Ciência da Computação para obtenção do título de Bacharel em Ciência da Computação.

Área de Concentração:
Inferência Bayesiana – Estatística Genética

Orientador:
Prof. Joel Augusto Muniz

LAVRAS
MINAS GERAIS – BRASIL
2006

Ficha Catalográfica

Reis, Ricardo Luis dos

Estimação de Parâmetros de Populações com Base em Freqüências Alélicas Utilizando Inferência Bayesiana através do Software Livre R/ Ricardo Luis dos Reis. Lavras – Minas Gerais, 2006. 94p: il.

Monografia de Graduação – Universidade Federal de Lavras. Departamento de Ciência da Computação.

1. Inferência Bayesiana. 2. Estatística Genética. 3. Simulação Computacional. I. REIS, R. L. dos. II. Universidade Federal de Lavras. III. Estimação de Parâmetros de Populações com Base em Freqüências Alélicas Utilizando Inferência Bayesiana através do Software Livre R

RICARDO LUIS DOS REIS

ESTIMAÇÃO DE PARÂMETROS DE POPULAÇÕES COM BASE EM FREQÜÊNCIAS ALÉLICAS UTILIZANDO INFERÊNCIA BAYESIANA ATRAVÉS DO SOFTWARE LIVRE R

Monografia de graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do curso de Ciência da Computação para obtenção do título de Bacharel em Ciência da Computação.

Aprovada em 20 de fevereiro de 2006.

_____

Profª. Olinda Nogueira Paes Cardoso

_____

Prof. Augusto Ramalho de Morais

_____

Prof. Joel Augusto Muniz
(Orientador)

LAVRAS
MINAS GERAIS – BRASIL

A Deus e a Santo Expedito
Aos meus pais Luzia e Pedro
À minha irmã Dinha
À Isabelle (in memorian)

"Acredite, que nenhum de nós,
Já nasceu com jeito pra super herói,
Nossos sonhos a gente é quem constrói.
É vencendo os limites,
Escalando as fortalezas,
Conquistando o impossível pela fé".
Jamily

AGRADECIMENTOS

Agradeço a DEUS e a Santo Expedito, por guiar sempre meus passos aonde quer que eu vá.

A meus pais, Luzia e Pedro, à minha irmã Dinha, pelo carinho, amor, alegria e incentivo em todos os momentos da minha vida. Que Deus abençoe vocês! Eu amo vocês!

A todos os meus amigos, em especial a Isabelle (*in memorian*) e ao Renato Agostini , que sempre estiveram comigo nos momentos bons e ruins da minha vida.

Ao professor Joel, por ter acreditado em mim e sempre me orientando para os caminhos certos.

Aos professores Fabyano, Thelma e Luiz Henrique, que tanto me ajudaram neste trabalho.

Ao professor Delly e a Luciana, por terem me indicado para uma bolsa se iniciação científica com o professor Joel.

Ao Dr. Carlos Alberto, por ter curado o meu pai de uma forte depressão.

Á minha tia Ana e ao meu tio Francisco, pela grande ajuda no começo dessa caminhada.

À minha madrinha Helena e ao meu padrinho José Valério, pelos momentos de alegria.

À minha avó Teresa (*in memorian*), que não teve a oportunidade de ver seu neto formado. Que Deus te ilumine aí do andar de cima.

À Adriana Ribeiro, pelos grandes conselhos e pelos momentos que com paciência me escutava desabafar.

A todas as pessoas que ajudaram a mim e a minha família a sermos o que somos hoje, uma família feliz, unida, com saúde e paz. Obrigado a todos vocês. Deus lhe pague.

A Universidade Federal de Lavras, por ter me ajudado em meus estudos, tanto com moradia (Brejão) como com alimentação (RU).

Aos todos os colegas de curso, mesmo aqueles que se transferiram, pelos momentos de estudo, mas principalmente pelo *dream time* que montamos na sala. Tetra campeão do Jocomp.

Aos meus amigos que conquistei nesses 4 anos aqui em Lavras, em especial aos que começaram essa caminhada comigo e aos que me aturaram no apartamento 107.

RESUMO

Estimação de Parâmetros de Populações com Base em Freqüências Alélicas Utilizando Inferência Bayesiana através do Software Livre R

O objetivo deste trabalho foi utilizar o método Bayesiano no ajuste do modelo de COCKERHAM. Realizou-se um estudo de simulação de dados para avaliar o método Bayesiano sob diferentes estruturas. Amostras das distribuições marginais *a posteriori* dos parâmetros do modelo de COCKERHAM foram obtidas pelo amostrador de Gibbs, utilizando o software livre R. As inferências foram feitas em cada população e os resultados mostraram que o método é eficiente no estudo da genética .

Palavras-chave: Amostrador de Gibbs, genética , Estatística Bayesiana.

ABSTRACT

Estimation of Parameters of Population with Based in Allelic Frequency Using Bayesian Inference** through **of the Free Software R

The objective of this work was to use the Bayesian method in the fitting of the COCHERHAM. The data were simulated for evaluating Bayesian method under several structures. The posterior marginal distributions for each parameter were obtained in the fitting of the COCHERHAM via Gibbs Sampler algorithm, utilizity the free software R . The inference was done for each population and the results showed the efficiency of this method in the genetics.

keywords: *Gibbs Sampler algorithm, genetics, Bayesian Statitics.*

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

Neste primeiro capítulo apresenta-se uma breve introdução ao trabalho proposto e os objetivos para a realização do mesmo. Ao final, um breve escopo do trabalho.

1.1. Considerações Iniciais

Atualmente, grande parte dos estudos de genética de populações naturais se dedica à caracterização da estrutura genética existente nas mais diferentes espécies, já que as informações daí obtidas podem contribuir de modo decisivo para o estabelecimento de estratégias mais adequadas para a conservação, o manejo e o melhoramento genético das espécies de interesse.

Nos estudos de genética e melhoramento, é importante determinar o coeficiente de endogamia e a taxa de fecundação cruzada, o que possibilita entender a dinâmica evolutiva da população através dos conceitos de genética de populações.

A estimação do coeficiente de endogamia e da taxa de fecundação cruzada para a caracterização da estrutura genética de uma população diplóide, conforme descrito por WEIR (1996) e VENCOVSKY (1992), pode ser feita por meio da análise de variância com freqüências alélicas. Esta análise considera um modelo estatístico aleatório definido em relação a uma variável y, que no caso de alelos múltiplos vale 1 quando um determinado alelo, por exemplo A_1 de um loco, está no indivíduo, e vale zero quando este alelo está ausente e presentes os alelos A_2, A_3,..., A_u. Na estimação do coeficiente de endogamia, ocorrem algumas dificuldades, pois a variável y descrita não tem distribuição normal e o estimador envolve o quociente entre duas variáveis aleatórias não tendo distribuição conhecida.

Outros métodos de se estimar esses parâmetros são os clássicos de NEI e CHESSER (1983) e de ROBERTSON e HILL (1984) e o Bayesiano de AYRES e BALDING (1998). ARMBORST (2005) relata que dentre esses três métodos, o último apresenta maior qualidade, pois respeita o espaço paramétrico onde o coeficiente de endogamia está definido.

Segundo HOLSINGER (2005) a metodologia Bayesiana envolve uma estrutura geral de análise na área de Genética de Populações, pois considera o grau de conhecimento prévio acerca de um conjunto de parâmetros de interesse, como por exemplo, freqüências

alélicas, coeficiente de endogamia, taxa de fecundação cruzada, herdabilidade entre outros. Assim, a abordagem Bayesiana possibilita que sejam consideradas várias hipóteses de interesse, bem como níveis de credibilidade, fornecendo elementos úteis na tomada de decisões. Um aspecto interessante da técnica é a possibilidade de considerar distribuições de probabilidades associadas aos parâmetros. Estas distribuições de probabilidades podem ser atualizadas utilizando os conceitos do teorema de Bayes, de modo a considerar as informações obtidas das observações.

Softwares de uso mais geral em estatística também têm sido utilizados para implementação de técnicas Bayesianas. Dentre esses se destaca o R, o qual caracteriza-se como uma linguagem e ambiente para computação estatística e geração gráfica. È um software livre, pertencente ao projeto GNU, rede internacional de softwares livres que visa o desenvolvimento científico por meio da interação com os usuários. Este software disponibiliza uma grande variedade de métodos estatísticos (modelagem linear e não linear testes estatísticos clássicos, séries temporais,...) e técnicas gráficas.

1.2. Objetivos

O presente trabalho visa utilizar a metodologia Bayesiana para estimar o coeficiente de endogamia e a taxa de fecundação cruzada de uma população diplóide, considerando o modelo aleatório de COCHERHAM (1969) na análise de variância de freqüências alélicas utilizando o R.

Mais especificamente, os objetivos do trabalho são os seguintes:

a) Estabelecer a função de verossimilhança de acordo com a distribuição assumida para os dados;
b) Estabelecer distribuições *a priori* de cada componente do modelo aleatório de COCHERHAM;
c) Aplicar a técnica de Monte Carlo via Cadeias de Markov (MCMC) por meio dos algoritmos Amostrador de Gibbs e Metropolis-Hastings para a obtenção das distribuições marginais *a posteriori* para os componentes do modelo;
d) Obter a distribuição marginal *a posteriori* do coeficiente de endogamia e da taxa de fecundação cruzada;
e) Estimar o coeficiente de endogamia e a taxa de fecundação cruzada com base na sua distribuição marginal *a posteriori*.

f) Avaliar o efeito do número de indivíduos na estimação do coeficiente de endogamia e da taxa de fecundação cruzada.

1.3. Estrutura do Documento

Esta monografia está organizada como mostrado a seguir.

No capítulo 2, apresenta-se todo o procedimento de inferência bayesiana assim como todas as suas definições básicas.

No capítulo 3, descrevem-se os métodos computacionais bayesianos com ênfase aos métodos de amostragem de Monte Carlo via Cadeias de Markov (MCMC). Dá-se uma ênfase maior nos algoritmos iterativos e nas análises de convergência.

No capítulo 4, apresentam-se vários conceitos genéticos necessários para o entendimento desta monografia e também apresentamos as principais metodologias empregadas na estimação do coeficiente de endogamia e da taxa de fecundação cruzada.

No capítulo 5, demonstra-se a grande importância da utilização de software livre, em particular o R, que realiza várias análises estatísticas além de ser uma ambiente de programação funcional, e assim, facilitar o uso para pesquisadores que ainda não tem nenhum conhecimento sobre sua utilização.

No capítulo 6, descreve-se toda a metodologia bayesiana desenvolvida assim como a simulação dos dados para a aplicação desta. Ao final, dá-se uma explicação sobre as funções implementadas para as análises estatísticas.

No capítulo 7, são apresentados os resultados pela metodologia bayesiana assim como sua análise de convergência.

No capítulo 8, são tiradas as conclusões da realização deste trabalho e assim demonstrar os avanços da inferência bayesiana nos últimos anos.

No capítulo 9, são apresentadas as futuras linhas de pesquisa, já que esta área permite grandes avanços. Dá-se uma ênfase para a futura utilização das redes bayesianas aplicadas à genética de populações.

Ao final são apresentadas as funções implementadas na linguagem R.

2. INFERÊNCIA BAYESIANA

A teoria Bayesiana foi desenvolvida por Thomas Bayes em meados do século XVIII, o qual propôs uma teoria subjetiva de probabilidade, baseada principalmente em um conhecimento *a priori* em relação às incertezas envolvidas no estudo. No século XX, mais precisamente na década de 30, a análise Bayesiana ressurgiu, após um longo tempo de domínio dos métodos estatísticos frequentistas, com base em alguns estudos teóricos como o de JEFFREYES (1939), que reagindo contra a predominante posição frequentista conseguiu ressuscitar o bayesianismo, apresentando-lhe aplicações lógicas para a resolução de problemas estatísticos em diversas áreas da ciência. Porém, estes estudos exigiam resoluções de integrais complexas, o que fez com que os métodos Bayesianos se sujeitassem a mais algumas décadas de adormecimento.

Foi somente na década de 90 que o problema dos cálculos integrais foi solucionado de maneira alternativa com o trabalho de GELFAND et al. (1990), que exploraram um recurso de simulação dinâmica, o algoritmo Gibbs Sampler, elaborado por GEMAN e GEMAN (1984), o qual faz uso da teoria das Cadeias de Markov. Os resultados obtidos foram excelentes, atraindo assim outros pesquisadores, e devido ao sucesso de suas aplicações este algoritmo se consolidou como uma das ferramentas mais utilizadas em análise Bayesiana.

A inferência Bayesiana é o processo de encontrar um modelo de probabilidade para um conjunto de dados e resumir o resultado por uma distribuição de probabilidade sobre os parâmetros do modelo e sobre quantidades não observadas tais como predição para novas observações (GELMAN ET AL., 2003).

A metodologia Bayesiana consiste de informações referentes aos dados amostrais (função de verossimilhança), de um conhecimento prévio a respeito dos parâmetros (distribuição *a priori*) e, a partir destas duas informações, do cálculo de uma distribuição *a posteriori* dos parâmetros, na qual todas as decisões e inferências são realizadas.

Em relação à teoria Bayesiana, BROEMELING (1985) e GELMAN et al. (1997) relatam que a característica essencial está no uso explícito de probabilidades para quantificar as incertezas que se tem sobre o parâmetro θ. Portanto, a inferência Bayesiana é baseada no conceito de probabilidade subjetiva, que mede o grau de confiança que alguém deposita no acontecimento de um determinado evento do espaço amostral.

A figura abaixo representa a evolução do pensamento humano.

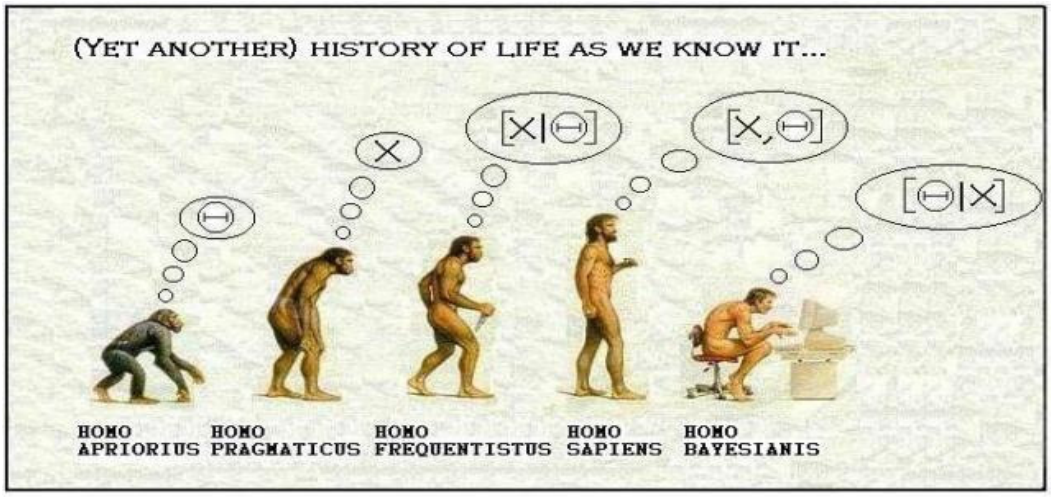

Figura 2.1: Evolução do pensamento humano

2.1. Teorema de Bayes

O Teorema de Bayes é um resultado simples de probabilidade condicional, porém é de importância fundamental na construção da inferência Bayesiana.

Portanto, a partir do momento que se opta por uma distribuição *a priori*, $P(\theta)$, seja ela informativa ou não, e obtém-se a função de verossimilhança, $L(\theta \mid Y)$, torna-se possível, através do Teorema de Bayes, obter a distribuição *a posteriori* de θ, $P(\theta \mid Y)$. A expressão matemática do Teorema de Bayes é:

$$P(\theta \mid Y) = \frac{L(\theta \mid Y)P(\theta)}{\int L(\theta \mid Y)P(\theta)d\theta}, \tag{1}$$

sendo $Y = \{y_1, y_2, ..., y_n\}$.

Uma forma equivalente de (1), visto que o denominador não depende de θ e, este então pode ser considerado constante é:

$$P(\theta \mid Y) \propto L(\theta \mid Y)P(\theta) \qquad (2)$$

A expressão (2), em que $\propto$ representa proporcionalidade, pode ser entendida como: **distribuição *a posteriori* $\propto$ função de verossimilhança x distribuição *a priori.***

Essa simples expressão contém o princípio técnico da inferência Bayesiana, em outras palavras, o Teorema de Bayes conta que a distribuição *a posteriori* é proporcional ao produto da distribuição a *priori* e a função de verossimilhança (BOX & TIAO, 1992). Pode-se justificar a presença do símbolo de proporcionalidade da seguinte maneira: quando se multiplica a função de verossimilhança por uma constante não se altera a inferência relativa ao parâmetro θ e, assim, a distribuição a *posteriori* não será alterada (LEANDRO, 2001).

2.2. Distribuição *a priori*

Nota-se que a fórmula de Bayes requer a especificação de $p(\theta)$ que é conhecida como distribuição *a priori* de θ. É através da especificação de uma distribuição *a priori*, para a quantidade de interesse θ, que se faz uso do conhecimento prévio em inferência Bayesiana. Tal distribuição deve representar probabilisticamente o conhecimento que se tem sobre θ antes de os dados serem obtidos (LEANDRO, 2001).

Esse conhecimento prévio sobre o parâmetro que se deseja estimar, antes que os dados tenham sido propriamente coletados, pode ser obtida através de análises anteriores, experiência do pesquisador na área em questão ou publicações sobre o assunto que se deseja tratar. A partir desta informação, é encontrada a distribuição *a priori* mais adequada para o caso.

A crença *a priori* sobre θ pode diferir de pesquisador para pesquisador (LEE, 1997). Há casos em que *a priori* dada é informativa, refletindo algum conhecimento prévio sobre o parâmetro desconhecido. Outras vezes, *a priori* é não-informativa, deixando que os dados mostrem todas as informações sobre θ.

2.3. Função de Verossimilhança

Os dados $y_1,...,y_n$, representados por uma amostra aleatória de uma população com densidade f, são utilizados na análise Bayesiana através de uma função de verossimilhança $L(\theta \mid y_1,...,y_n)$, que é a densidade conjunta destes dados.

É através da função de verossimilhança que os dados podem modificar o conhecimento que se tem *a priori* sobre θ. Dessa forma, esta pode ser considerada como a representação da informação de θ obtida dos dados, ou seja, esta função pode ser vista como a representação do que os dados têm a contar a respeito do parâmetro θ.

2.4. Distribuição *a posteriori*

A distribuição *a posteriori* de um parâmetro contém toda a informação probabilística a respeito deste parâmetro e um gráfico da sua função de densidade *a posteriori* é a melhor descrição do processo de inferência e todas as conclusões estão baseadas na distribuição condicional dos parâmetros sobre os dados observados. Toda a inferência com respeito ao parâmetro é feita através da distribuição *a posteriori* destes. A partir de tal distribuição, estima-se o parâmetro através de valores observados da distribuição *a posteriori* tais como média, mediana e moda *a posteriori*. A incerteza com respeito ao parâmetro é descrita através de intervalos de credibilidade obtidos através da distribuição *a posteriori*. .

Em resumo, a distribuição *a posteriori* incorpora, via Teorema de Bayes, toda a informação disponível sobre o parâmetro, isto é, a informação inicial mais a informação da experiência ou amostra (PAULINO, TURKMAN e MURTEIRA, 2003).

ROSA (1998) afirma que para se inferir com relação a qualquer elemento de θ, a distribuição *a posteriori* conjunta dos parâmetros, p(θ | Y), deve ser integrada em relação a todos os outros elementos que a constituem. Assim, se o interesse do pesquisador se concentra sobre determinado conjunto de θ, por exemplo, θ_1, tem-se a necessidade da obtenção da distribuição p(θ_1 | Y), dada por:

$$P(\boldsymbol{\theta}_1 \mid Y) = \int_{\boldsymbol{\theta} \neq \boldsymbol{\theta}_1} P(\boldsymbol{\theta} \mid Y) d\boldsymbol{\theta}_{\boldsymbol{\theta} \neq \boldsymbol{\theta}_1} \qquad (3)$$

A integração da distribuição conjunta *a posteriori* para a obtenção das marginais geralmente não é analítica, e, portanto, necessita de algoritmos iterativos especializados como o amostrador de Gibbs. No capítulo 3 será dada uma ênfase neste assunto.

2.5. Intervalo de Credibilidade

Voltamos a enfatizar que a forma mais adequada de expressar a informação que se tem sobre um parâmetro é através de sua distribuição *a posteriori*. A principal restrição da estimação pontual é que quando estimamos um parâmetro através de um único valor numérico toda a informação presente na distribuição *a posteriori* é resumida através deste número. É importante também associar alguma informação sobre o quão precisa é a especificação deste número. Vamos introduzir o conceito de intervalo de credibilidade (ou intervalo de confiança Bayesiano) baseado na distribuição *a posteriori*.

O intervalo (*a,b*) é um intervalo de credibilidade de 100(1-α)% ou nível de credibilidade (1-α) para o parâmetro de interesse θ.

Note que a definição expressa de forma probabilística a pertinência ou não do intervalo. Assim, quanto menor for o tamanho do intervalo mais concentrada é a distribuição do parâmetro, ou seja, o tamanho do intervalo informa sobre a dispersão do parâmetro.

2.5.1. Intervalo de Máxima Densidade *a Posteriori* (HPD)

Um intervalo de credibilidade é um intervalo de máxima densidade *a posteriori* (HPD) quando a densidade para todo ponto pertencente ao intervalo é maior do que para qualquer ponto não pertencente ao intervalo. Além disso, deve-se considerar o menor intervalo possível (BOX & TIAO, 1992). Geralmente, esses intervalos são associados a 90, 95 ou 99% de probabilidade total.

3. MÉTODOS EM SIMULAÇÃO COMPUTACIONAL

O termo simulação refere-se ao tratamento de problemas reais a partir de reproduções em ambientes controlados pelo pesquisador, sendo que, muitas vezes esses ambientes são os equipamentos computacionais. Alguns problemas apresentam componentes aleatórios, os quais não podem ser descritos de forma exata e sim baseados em informações probabilísticas. Para estes problemas, o processo de simulação é estocástico, ou seja, baseado em distribuições de probabilidades.

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 é obtida considerando que, na densidade conjunta, os demais parâmetros são conhecidos e, assim, a expressão se torna menos complexa, já que as constantes podem ser desconsideradas.

Um dos tópicos mais ativos em Estatística Computacional é a inferência através de simulação iterativa, utilizando, especialmente o amostrador de Gibbs e o algoritmo Metropolis-Hastings.

Estes algoritmos, que demandam um uso intensivo de recursos computacionais, utilizam a simulação de Monte Carlo para gerar valores a partir de distribuições de probabilidades conhecidas e a teoria das Cadeias de Markov para representar a dependência entre os parâmetros, portanto o amostrador de Gibbs faz uso do método conhecido como MCMC, Monte Carlo – *Markov Chain*, (SORENSEN, 1996; GAMERMAN, 1996).

A idéia essencial da simulação iterativa é gerar valores de uma variável aleatória θ utilizando uma seqüência de distribuições que converge iterativamente para a distribuição de origem de θ. A simulação iterativa é menos eficiente do que as simulações diretas, que é simplesmente gerar amostras da distribuição fonte (original). No entanto, a simulação iterativa é aplicável em uma classe muito mais ampla de casos e desempenha um papel extremamente marcante em inferência Bayesiana no momento de calcular resumos de distribuições a *posteriori* extremamente complicadas.

3.1. Integração Monte Carlo

Métodos de simulação Monte Carlo são utilizados para calcular quantidades de interesse que são difíceis ou impossíveis de serem calculadas analiticamente. Um estudo de simulação é feito gerando milhares de observações de uma estrutura de interesse e estatísticas baseadas nesses valores simulados podem ser calculadas.

Suponha que X é uma variável aleatória com densidade $f(x)$. Sabe-se que:

$$\theta = E[g(X)] = \int g(x) f(x) dx. \tag{4}$$

A expressão (4) pode ser muito difícil, ou até impossível de ser calculada algebricamente, mas gerar elementos da distribuição com densidade $f(x)$ pode ser muito mais fácil.

3.2. Monte Carlo com Cadeias de Markov

Os métodos de Monte Carlo via cadeias de Markov (MCMC) são uma alternativa aos métodos não iterativos em problemas complexos. A idéia ainda é obter uma amostra da distribuição *a posteriori* e calcular estimativas amostrais de características desta distribuição. A diferença é que aqui usaremos técnicas de simulação iterativa, baseadas em cadeias de Markov, e assim os valores gerados não serão mais independentes.

Uma cadeia de Markov é um processo estocástico em que a probabilidade de estar em um certo estado em um tempo futuro pode depender do estado atual do sistema, mas não dos estados em tempos passados. Em outras palavras, em uma cadeia de Markov, dado o estado atual do sistema, o próximo estado é independente do passado (mas poderá depender do estado presente).

A idéia básica é simular um passeio aleatório no espaço do parâmetro que converge para uma distribuição estacionária, que é a distribuição de interesse no problema.

São listados abaixo os dois algoritmos estudados para a implementação das funções.

3.3. Algoritmo Metropolis-Hastings

O algoritmo de Metropolis-Hastings é utilizado para a obtenção da distribuição marginal *a posteriori* quando o amostrador de Gibbs não se mostra eficiente, ou seja, para parâmetros cuja distribuição condicional não se caracteriza como uma distribuição de probabilidade conhecida. Neste caso, geram-se valores do parâmetro a partir de uma distribuição proposta e esse é aceito ou não com uma certa probabilidade de aceitação.

Para descrever o algoritmo, suponha que a distribuição de interesse é a distribuição *a posteriori* $P(\theta \mid Y)$ com $\theta = (\theta_1, ..., \theta_S)$. Considere também que todas as condicionais completas *a posteriori* $P(\theta_i \mid \theta_{-i}, Y)$, com $i = 1, ..., n$ estejam disponíveis mas não se sabe gerar amostras diretamente de cada uma e que amostras de um novo valor de θ_i serão geradas a partir de uma distribuição proposta condicional ao valor atual de θ_i, $q(\theta_i^{(p)}, \theta_i^{(a)})$ onde $\theta_i^{(p)}$ é o valor proposto e $\theta_i^{(a)}$ é o valor atual, ou seja, o valor de θ exatamente antes da proposta ser gerada, ou seja, o valor atualizado da geração anterior, para $i = 1, ..., n$.

Os valores dos parâmetros gerados por esses algoritmos são utilizados para formar uma amostra aleatória $\{(\theta_1^{(i)}, \theta_2^{(i)}, ..., \theta_n^{(i)})^N\}$. À medida que o número de iterações aumenta, o conjunto de valores gerados aproxima de sua condição de equilíbrio. Assim assume-se que a convergência é atingida em uma iteração cuja distribuição esteja arbitrariamente próxima da distribuição de equilíbrio, ou seja, da distribuição marginal desejada.

O algoritmo de Metropois-Hastings é bastante geral, e pode, pelo menos em princípio, ser implementado com qualquer distribuição condicional completa *a posteriori* e para qualquer proposta. Entretanto sob o ponto de vista prático, a escolha da proposta é crucial para o bom desenvolvimento do algoritmo, ou seja, para sua convergência para a distribuição *a posteriori*:

A seguir é apresentado um esquema ilustrativo do algoritmo Metropolis-Hastings

Algoritmo Metropolis-Hastings

I *- Inicialize* $\theta^{(0)} = \theta_1^{(0)},...,\theta_S^{(0)}$ *e* $k = 1$*;*

II *- Obtenha um novo valor para* $\theta^{(k)}$ *a partir de* $\theta^{(k-1)}$ *através de sucessivas gerações de valores. Para* $i = 1$ *até* S*, faça:*

(i) Gere uma proposta para $\theta_i^{(k)}$ *de :*

$$\theta_i^{(p)} \sim q(\theta_i \mid \theta_i^{k-1})$$

(ii) Aceite a proposta com probabilidade de aceitação dada por:

$$\alpha = \min\left\{1, \frac{P(\theta_i^{(p)} \mid \theta_i^{(a)}, x) q(\theta_i^{(k-1)} \mid \theta_i^{(p)})}{P(\theta_i^{(k-1)} \mid \theta_i^{(a)}, x) q(\theta_i^{(p)} \mid \theta_i^{(k-1)})}\right\}$$

onde $\theta_{-i}^{(a)} = (\theta_1^{(k)},...,\theta_{i-1}^{(k)},\theta_{i+1}^{(k-1)},...,\theta_S^{(k-1)})$.

III *- Faça* $k = k+1$ *e volte para* ***II*** *e repita o procedimento até alcançar a convergência..*

3.4. Algoritmo Gibbs Sampler ou Amostrador de Gibbs

O algoritmo Gibbs Sampler ou 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. É uma técnica para gerar variáveis aleatórias de uma distribuição marginal sem que se conheça a sua densidade. Á 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.

O algoritmo Gibbs Sampler tem sido extremamente útil na resolução de problemas que envolvem a estimação de mais de um parâmetro, como é o caso do modelo de COCHERHAM (1969), porém para utilização desse algoritmo exige-se que as distribuições condicionais *a posteriori* dos parâmetros tenham formas conhecidas, ou seja, que sejam dadas por uma distribuição de probabilidade conhecida.

Para descrever o algoritmo, suponha que a distribuição de interesse é a distribuição *a posteriori* $P(\theta \mid Y)$ com $\theta = (\theta_1,...,\theta_S)$ e considere também que todas as condicionais completas *a posteriori* $P(\theta_i \mid \theta_{-i}, Y)$ com $i = 1,...,n$ estejam disponíveis e que sabe-se gerar amostras de cada uma delas. A seguir é apresentado o esquema do Amostrador de Gibbs:

Amostrador de Gibbs

I *- Inicialize* $\theta^{(0)} = \theta_1^{(0)},...,\theta_S^{(0)}$ *e* $k = 1$*;*

II *- Obtenha um novo valor para* $\theta^{(k)}$ *a partir de* $\theta^{(k-1)}$ *através de sucessivas gerações de valores. Para* $i = 1$ *até* S*, faça:*

Gere um valor para $\theta_i^{(k)}$ de:

$$\theta_i^{(k)} \sim P(\theta_i \mid \theta_1^{(k)},\ldots,\theta_{i+1}^{(k-1)},\ldots,\theta_S^{(k-1)},Y)$$

III *- Faça* $k = k+1$ *e volte para* ***II*** *e repita o procedimento até alcançar a convergência*

3.5. Verificação de Convergência

Os métodos de MCMC são uma ótima ferramenta para resolução de muitos problemas práticos na análise Bayesiana. Porém, algumas questões relacionadas à convergência nestes métodos ainda merecem bastante pesquisa.

Uma questão que pode surgir é "Quantas iterações deve ter o processo de simulação para garantir que a cadeia convergiu para o estado de equilíbrio?".

A resposta definitiva para esta questão poderá nunca ser dada, visto que a distribuição estacionária será na prática desconhecida, mas pode-se sempre avaliar a convergência das cadeias detectando problemas fora do período de aquecimento. Uma análise de convergência em métodos de simulação pode ser feita preliminarmente analisando os gráficos ou medidas descritivas dos valores simulados da quantidade de interesse. Os gráficos mais freqüentes são o gráfico de θ ao longo das iterações e um gráfico da estimativa da distribuição *a posteriori* de θ, por exemplo, um histograma ou uma densidade kernel. As estatísticas usuais são a média, o desvio padrão e os quantis (2,5%; 50%; 97,5%).

Os algoritmos Gibbs Sampler e Metropolis-Hastings são iterativos, ou seja, necessitam da constatação de convergência, para que realmente se possa inferir sobre seus resultados como sendo valores das distribuições marginais dos parâmetros do modelo considerado. Para avaliação da convergência desses algoritmos optaram-se pelos testes de diagnóstico de GEWEKE (1992), GELMAN E RUBIN (1992), RAFTERY-LEWIS (1992B) e o de HEIDELBERG E WELCH (1983). Todos esses testes encontram-se disponíveis no pacote BOA (Bayesian Output Analisys) do software livre R.

3.5.1. Critério de Raftery e Lewis

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, 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 á 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 seja estimado com uma precisão predefinida.

3.5.2. 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 HalfWidth para verificar se a média estimada está sendo calculada com uma acurácia pré especificada, 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.

3.5.3. Critério de Geweke

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

Este propõe o diagnóstico de convergência para Cadeias de Markov baseados no teste de igualdade de médias da primeira e última parte da cadeia de Markov (geralmente dos primeiros 10% e dos últimos 50%).

3.5.4. Critério de Gelman e Rubin

O critério 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 esta próxima.

3.5.5. Descrição da Autocorrelação

Este teste calcula a função de autocorrelação da cadeia de Markov com defasagens ("lags") de 0, 1, 5, 10 e 50. Autocorrelações altas dentro das cadeias indicam cadeias lentas e, usualmente, a sua convergência também será lenta. Uma forma de fácil visualização é através do gráfico de autocorrelação, sendo possível tirar todas as conclusões a respeito da convergência.

3.5.6. Gráfico da Densidade de Kernel

Um procedimento que também é útil, neste sentido, é a construção de um gráfico da densidade *a posteriori* utilizando núcleo estimador (Kernel). Através da visualização deste gráfico, é possível se ter uma idéia da possível ou não convergência.

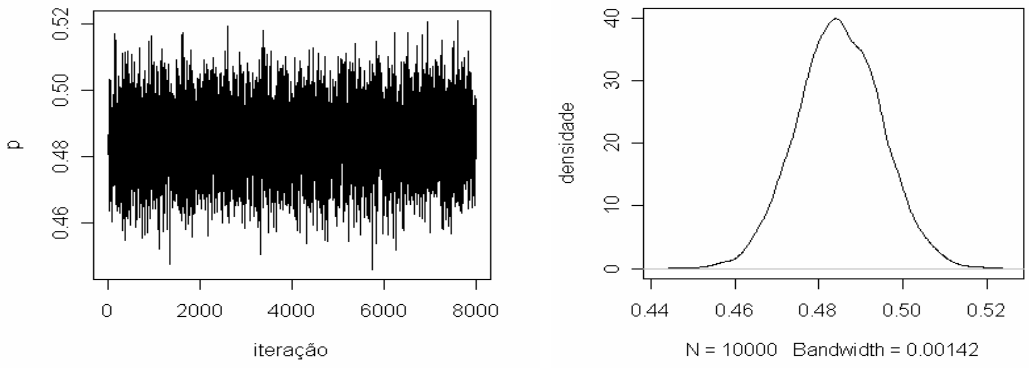

Figura 3.1: Visualização gráfica da convergência

4. GENÉTICA DE POPULAÇÕES

A Genética é uma ciência que estuda a hereditariedade e os mecanismos da evolução nos organismos. O termo Genética vem do grego "gerar". Esta ciência estuda como e porque acontecem as transferências de informações dos progenitores para sua prole ao longo de gerações.

A Genética de Populações é um ramo da genética de capital importância porque fornece subsídios para o melhoramento de plantas e animais e ainda, as bases necessárias à compreensão de como se processa a evolução (RAMALHO et al., 2004).

4.1. Conceitos de Genética

O cromossomo é uma longa seqüência de "fita" dupla de DNA. O material genético de um cromossomo consiste de um filamento ininterrupto de DNA que contém vários genes. O gene é a unidade funcional básica da hereditariedade. As formas alternativas de apresentar um gene são conhecidas como alelos, os quais estão num loco, que é a posição de um gene ao longo de um cromossomo.

As propriedades genéticas das populações são determinadas a partir do conhecimento de suas freqüências alélicas e genotípicas. As freqüências alélicas correspondem às proporções dos diferentes alelos de um determinado gene na população. As freqüências genotípicas, por sua vez, são as proporções dos diferentes genótipos para o gene considerado.

O Genótipo é a constituição genética de um organismo. No caso de organismos diplóides, cada organismo possui dois conjuntos gênicos, um herdado do pai e outro da mãe. Quando o genótipo tiver dois alelos em comum no mesmo loco, será chamado homozigoto (por exemplo, A1A1), e quando tiver dois alelos diferentes, será chamado de heterozigoto (por exemplo, A1A2).

4.2. Métodos Clássicos na Estimação dos Parâmetros

Endogamia, segundo a definição de FALCONER (1964) é o acasalamento entre indivíduos que são relacionados por ascendência, tendo como primeiro efeito uma mudança nas frequências genotípicas de Hardy-Weinberg devida a um aumento na freqüência de genótipos homozigóticos em detrimento das freqüências de genótipos heterozigóticos. Para o autor, o coeficiente de endogamia F, mede a probabilidade de dois genes, em qualquer loco de um indivíduo serem originados da cópia de apenas um gene numa geração anterior.

De acordo com COCKERHAM (1969), endogamia, variâncias de frequências alélicas, tamanho efetivo de população são termos comuns no estudo de genética de populações, sendo os conceitos e a maior parte da teoria foi introduzida pelos trabalhos clássicos de WRIGHT (1921) e FISHER (1949).

HARTL e CLARK (1989) afirmam que, em termos biológicos, o coeficiente de endogamia F mede a redução fracionária na heterozigosidade, em relação a uma população de acasalamento aleatório com a mesma freqüência alélica. No caso de um loco com dois alelos, as frequências genotípicas de AA, Aa e aa podem ser expressas em relação ao coeficiente de endogamia pela função aabixo:

$$(1\text{-}F)\ (p^2, 2pq, q^2) + F(p, 0, q).$$

Para os autores, esta expressão facilita o entendimento e a comparação das frequências genotípicas no princípio de Hardy-Weinberg. Quando F = 0, tem-se acasalamento aleatório, ou seja, não existe endogamia. Neste caso, as frequências genotípicas estarão de acordo com o equilíbrio de Hardy-Weinberg. Quando F = 1, tem-se endogamia total e a população só apresentará genótipos homozigóticos AA e aa, respectivamente nas frequências p e q.

WRIGHT (1965) discute o relacionamento entre gametas, mostrando o significado do coeficiente de endogamia e de parâmetros relacionados a pares de gametas em geral, realçando o fato de que eles podem ser interpretados tanto como coeficientes de correlação, quanto como probabilidade de identidade de genes de mesma origem. No caso de populações naturais hierarquicamente subdivididas, três parâmetros de descrição ou índices de fixação foram propostos em termos da população total (T), de sub-populações (S) e de indivíduos (I). Os parâmetros são os seguintes:

- F_{IT} - expressa a correlação entre os gametas que se unem para produzir os indivíduos em relação aos gametas da população total.

- F_{IS} - expressa a média das correlações, sendo cada uma delas proveniente dos gametas que se unem em cada sub-população em relação aos gametas desta sub-população.

- F_{ST} - expressa a correlação entre os gametas ao acaso dentro da sub-população em relação aos gametas da população total.

Segundo o autor, o parâmetro F_{IT} reúne informações dos outros dois, existindo a seguinte relação:

$$1 - F_{IT} = (1 - F_{IS})(1 - F_{ST}).$$

COCKERHAM e WEIR (1983) apontam os coeficientes de endogamia e de coancestria, bem como outras medidas de identidade por descendência dos genes, como parâmetros importantes em Genética Quantitativa e de Populações. Estes parâmetros são úteis para informar sobre homozigosidade, deriva, endogamia e variação quantitativa.

WEIR e COCKERHAM (1984) consideram, para o caso de um dos alelos de um loco, as seguintes definições e notações: F, a correlação entre os genes dentro de indivíduos ou endogamia; θ, a correlação entre os genes de diferentes indivíduos da mesma população ou coancestria e f, a correlação entre genes dentro de indivíduos dentro de populações. Os três parâmetros definidos pelos autores correspondem às estatísticas F de Wright da seguinte forma:

$$F = F_{IT}, \theta = F_{ST} \quad \text{e} \quad f = F_{IS}$$

estando relacionados por:

$$\mathbf{f = \frac{F - \theta}{1 - \theta}} \ .$$

Vários trabalhos existentes discutem a estimação de parâmetros associados à estrutura genética de populações utilizando inferência estatística clássica.

REYNOLDS, WEIR e COCKERHAM (1983), estudaram a estimação do coeficiente de coancestria θ, no caso de uma população com u alelos, propondo dois métodos. O primeiro considera a média das u estimativas obtidas nos diversos alelos e o segundo utiliza uma estimativa baseada na análise de variância conjunta envolvendo todos os alelos. WEIR e COCKERHAM (1984) compararam os dois métodos e concluíram que o segundo é melhor por apresentar resultados mais consistentes.

MUNIZ, BARBIN e VENCOVSKY (1996) estudaram as propriedades dos estimadores do coeficiente de endogamia e da taxa de fecundação cruzada obtidos pela análise de variância com dados de frequências alélicas em populações diplóides.

MUNIZ, VENCOVSKY e BARBIN (1997) compararam fórmulas para estimação da variância do estimador do coeficiente de endogamia obtido na análise de variância das frequências alélicas em uma população diplóide. Resultados de simulação mostraram que as três fórmulas propostas apresentam valores semelhantes e satisfatórios quando a freqüência alélica da população estiver entre 0,3 e 0,7 o coeficiente de endogamia da população for inferior a 0,5 e se trabalhar com pelo menos 30 indivíduos.

MUNIZ et al. (1999) estudando a estimação do coeficiente de endogamia em uma população diplóide, avaliaram a distribuição do quociente dos quadrados médios entre indivíduos e entre gene dentro de indivíduos, verificando que o teste F da análise de variância pode ser utilizado para testar a nulidade do coeficiente de endogamia quando a freqüência alélica estiver entre 0,3 e 0,7 trabalhando-se com 30 indivíduos, entre 0,25 e 0,75 com 50 indivíduos e entre 0,20 e 0,80 com 100 indivíduos. Estudo de simulação validou os resultados teóricos.

RAUFASTE e BONHOMME (2000) avaliaram as propriedades de estimadores do coeficiente de coancestria, através de estudo teórico e de simulação para o caso de alelos múltiplos. Expressões do viés e da variância dos estimadores foram desenvolvidas utilizando-se o método delta.

MUNIZ et al. (2001a) estudaram a distribuição do quociente entre quadrados médios na análise de variância de frequências alélicas de indivíduos em populações haplóides, avaliando o teste F para testar a nulidade do coeficiente de coancestria. Os autores observaram que o teste F pode ser usado quando se trabalha com pelo menos cinco populações com freqüência alélica média entre 0,3 e 0,7 utilizando-se no mínimo 50 indivíduos.

MUNIZ et al. (2001b) avaliaram as propriedades do estimador do coeficiente de coancestria em populações haplóides com dois alelos. Os autores desenvolveram expressões por série de Taylor e avaliaram as propriedades assintóticas. Um estudo de simulação confirmou os resultados teóricos.

WEIR (1996) apresenta uma discussão geral sobre os métodos de estimação de parâmetros genéticos com base em dados de freqüências alélicas. Entre os diversos métodos o autor destaca o método dos momentos, o método da máxima verossimilhança e

a análise de variâncias das freqüências alélicas. No caso da técnica da análise de variância, o autor aborda o caso de organismos haplóides bem como populações diplóides com modelos hierárquicos de até quatro níveis. O autor aborda ainda a possibilidade de usar técnicas Bayesianas na estimação dos parâmetros genéticos, uma vez que estas incorporam informações prévias ao procedimento de estimação, sendo úteis na descrição da estrutura genética de populações principalmente nas situações em que a estimação envolve a utilização de frequências alélicas.

Fecundação cruzada é a união entre gametas de indivíduos diferentes, mas da mesma espécie. Populações de indivíduos que apresentam fecundação cruzada têm maiores possibilidades de aumentar a variabilidade genética sem adição de genes novos (por mutação, por exemplo) do que populações de indivíduos com autofecundação.

A variabilidade genética é importante para a sobrevivência da espécie. Nesse sentido, até bissexuados desenvolveram, ao longo de sua evolução, vários mecanismos que dificultam a autofecundação e favorecem a fecundação cruzada , possibilitando desse modo , aumento na variabilidade maior . Através da recombinação genética, uma população pode aumentar sua variabilidade genética sem adição de genes novos, produzindo por mutação ou por imigração de indivíduos de outras populações.

Conforme já comentado, no que diz respeito a caracterizacão genética de populacões de espécies vegetais, a determinacão do sistema reprodutivo predominante na espécie fornece informações de grande importância. O sistema reprodutivo predominante de uma determinada espécie estabelece, em grande parte, todo um conjunto de propriedades genéticas de suas populações. Estas propriedades, em última instância, são as que sugerem as estratégias mais adequadas de conservação e melhoramento genético que devem ser utilizadas em cada caso.

Atualmente, a abordagem mais comumente utilizada na obtenção de estimativas de taxas de fecundação cruzada é a metodologia inicialmente proposta em RITLAND & JAIN (1981). Exemplos de aplicação desta metodologia podem ser encontrados em MILLAR et al. (2000) e BESSEGA el al. (2000).

A metodologia proposta por RITLAND & JAIN (1981) utiliza a informacão de múltiplos locos para se estimar a taxa de fecundação cruzada e possibilita ainda a obtenção de uma série de outros parâmetros indicadores do sistema reprodutivo.

Conforme sugerido inicialmente por FYFE & BAILEY (1951), dada a relação entre os valores do coeficiente de endogamia e os da taxa de fecundação cruzada, obtém-se o seguinte estimador para a taxa de fecundação cruzada:

$$t = \frac{1 - f}{1 + f}$$

Conforme apresentado em WEIR (1996), a progênie de indivíduos homozigóticos pode ser utilizada para a estimação da taxa de fecundação cruzada. O método se baseia no princípio de que a probabilidade de que um descendente de um indivíduo homozigótico seja heterozigótico é dada pelo produto das probabilidades de ocorrência de dois enventos independentes: o primeiro é a ocorrência de cruzamento, e o segundo é a ocorrência de um alelo diferente daquele presente no indivíduo homozigótico no gameta masculino que dá origem ao descendente.

O método proposto por RITLAND & JAIN (1981) constrói a função de verossimilhança para os parâmetros de interesse com base nas freqüências genotípicas observadas em um conjunto de progênies e obtém estimativas para estes parâmetros pelo método de máxima verossimilhança.

O grande desenvolvimento dos recursos computacionais ultimamente proporcionou aplicações da metodologia Bayesiana, que atualmente é utilizada em diversas áreas de pesquisa, como Genética, Melhoramento Animal e Vegetal, Física, Ecologia entre outras. A estimação do coeficiente de endogamia e da taxa de fecundação cruzada será feita considerando toda a abordagem envolvida nas técnicas Bayesianas.

4.3. Métodos Bayesianos na Estimação de Parâmetros

Na área de Genética de Populações trabalhos importantes têm sido desenvolvidos considerando técnicas Bayesianas. BALDING e NICHOLS (1997) utilizaram metodologia Bayesiana na estimação do coeficiente de coancestria de uma população. AYRES e BALDING (1998) aplicaram a metodologia Bayesiana no estudo do coeficiente de endogamia populacional. WILSON e BALDING (1998) aplicaram técnicas Bayesianas MCMC em estudo de amostragem de árvores para locus microsatélites. SORIA et al. (1998) desenvolveram técnicas Bayesianas para fazer inferência sobre parâmetros genéticos na cultura de Eucalyptus. Os autores não encontraram diferenças nos resultados

obtidos para as estimativas dos valores genéticos pelo método BLUP e pela técnica Bayesiana.

KARHU (2001) aborda a utilização de marcador de microsatélites para estimar o coeficiente de endogamia em populações de Pinus, por meio de técnicas Bayesianas usando MCMC (Cadeia de Markov e Simulação de Monte Carlo). O amostrador de Gibbs foi utilizado na obtenção da distribuição *a posteriori* dos parâmetros. MCGUIRE, DENHAM e BALDING (2001) implementaram o MCMC em amostragem de árvores filogenéticas usando modelos de substituição de nucleotídeos. AYRES e BALDING (2001) estimaram coeficientes do desequilíbrio de gametas por técnicas Bayesianas

LEUTENNEGER et al. (2003) avaliaram a estimação do coeficiente de endogamia pelo método da máxima verossimilhança levando-se em consideração a dependência do marcador molecular, que envolveu desenvolvimento de cadeias de Markov. O método foi avaliado com dados de manifestação autossômica recessiva obtidos de estudo envolvendo a doença dos dentes Charcot-Marie.

WILSON e RANNALA (2003) apresentaram uma abordagem Bayesiana para estimar a taxa de migração dentro de populações no caso genótipos multilocus. O método exige pressuposições de que os estimadores do fluxo gênico de longo prazo podem ser aplicados em populações não estacionárias que não estejam em equilíbrio genético Os parâmetros foram estimados usando o MCMC. Distribuições *a posteriori* foram usadas na inferência dos parâmetros.

HOLSINGER e WALLACE (2004) utilizaram uma extensão da análise Bayesiana para estudar a estrutura genética de populações. A distribuição beta foi utilizada como aproximação da distribuição *a posteriori* do coeficiente de endogamia. Para ilustrar o método, foram utilizados dados obtidos de amostragem de DNA de orquídeas.

ARMBORST (2005) relata um caso de estimação multi-paramétrico que pode ser tratado com o uso de técnicas Bayesianas e o método de MCMC, que é a estimação das proporções alélicas e da medida de endocruzamento f, já que em grande parte das situações há vários alelos num mesmo loco na população, e a estimação via Máxima Verossimilhança é complexa.

Os dados $y_1,...,y_n$, indicados pela freqüência do alelo j dentro do indivíduo i, representam uma amostra aleatória de uma população com densidade f, portanto são utilizados na análise Bayesiana através da função de verossimilhança $L(y_1,...,y_n / \theta)$, que é a densidade conjunta destes dados. De acordo com HOLSINGER (2005), em se tratando

de dados referentes a freqüências alélicas, a função de verossimilhança é representada por uma distribuição binomial, a qual é dada por:

$$L(\boldsymbol{\theta}|y)=\frac{n!}{y!(n-y)!}p^{y}(1-p)^{n-y}$$

Quando em determinado estudo o pesquisador tem pouca ou nenhuma informação para se incorporar a distribuição *a priori* considera-se *a priori* não-informativa, a qual é representada principalmente pela priori de Jeffreys (JEFFREYS, 1961). Em relação as distribuições *a priori* dos parâmetros de modelos utilizados em Genética de Populações, HOLSINGER (2005) relata as seguintes considerações sobre o espaço paramétrico: [0,1] pode-se utilizar as distribuições Uniforme e Beta; $[0,\infty)$ distribuições Gama, Gama Inversa e Lognormal; $(-\infty,\infty)$ Distribuições Normal e t.

5. SOFTWARE LIVRE R

R é uma linguagem e ambiente para computação estatística e gráfica. É um projeto GNU, que é um projeto iniciado por Richard Stallman em 1984, com o objetivo de criar um sistema operacional totalmente livre, que qualquer pessoa teria direito de usar e distribuir sem ter que pagar licenças de uso e este é similar à linguagem e ambiente S que foi desenvolvida no Bell Laboratories. R pode ser considerada como uma implementação diferente da S. Há algumas diferenças importantes, mas muito código para S funciona inalterado em R.

R fornece uma ampla variedade de técnicas estatísticas (modelagem linear e não linear, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento,...) e gráficos, e é altamente extensível.

A linguagem S é muitas vezes o veículo de escolha para pesquisa em metodologia estatística, e R fornece uma rota *Open Source* para participação naquela atividade.

Um dos pontos fortes de R é a facilidade com que gráficos bem-desenhados com qualidade para publicação podem ser produzidos, incluindo símbolos matemáticos e fórmulas quando necessário. Muitos cuidados têm sido feitos sobre as definições padrão para as menores escolhas em desenho, entretanto o usuário retém controle total. Abaixo, apresenta-se a tela inicial do software livre R.

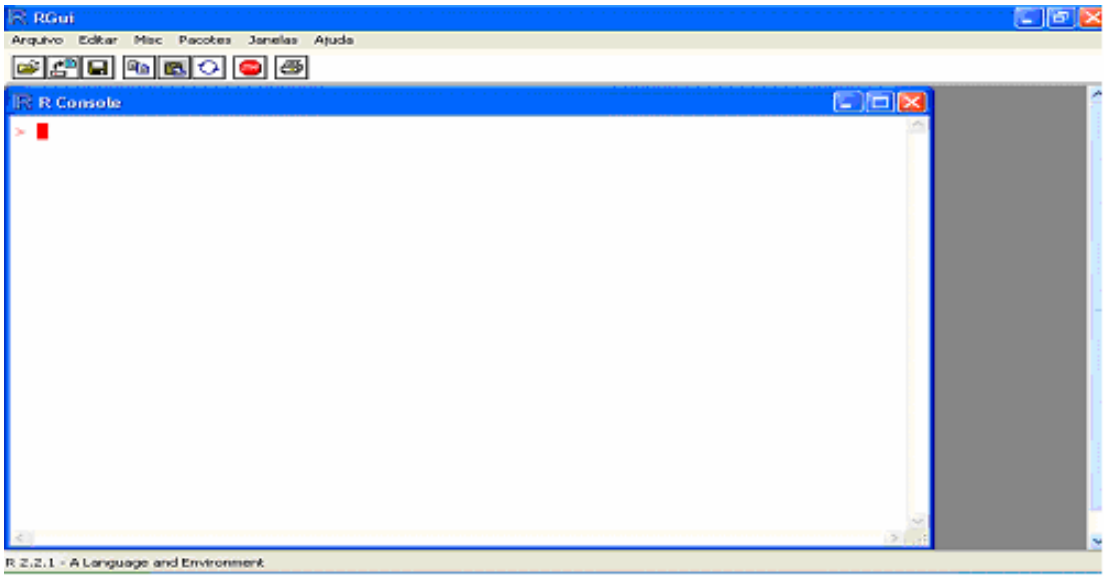

Figura 5.1: Tela principal do Software Livre R

R é disponível como Software Livre sob os termos da Licença Pública Geral GNU da Free Software Foundation na forma de código fonte. Ela compila e funciona em uma grande variedade de plataformas UNIX e sistemas similares (incluindo FreeBSD e Linux). Ele compila e funciona em Windows 9x/NT/2000 e MacOS.

5.1. O ambiente R

R é um conjunto integrado de facilidades de software para manipulação de dados, cálculo e visualização gráfica. Ele inclui uma facilidade efetiva para manipulação e armazenagem de dados, um conjunto de operadores para cálculos sobre quadros de dados, em particular as matrizes, uma grande e coerente coleção integrada de ferramentas intermediárias para análise de dados, facilidades gráficas para análise de dados e visualização na tela ou impressa, uma linguagem de programação bem desenvolvida, simples e efetiva que inclui condicionais, alças, funções recursivas definidas pelo usuário, e facilidades para entrada e saída.

O termo "ambiente" pretende caracterizar R como um sistema totalmente planejado e coerente, em vez de uma aglomeração de ferramentas muito específicas e inflexíveis, como é o caso com outros softwares de análise de dados.

R fornece um interface de entrada por linha de comando. Todos os comandos são digitados e o *mouse* é pouco usado. Pode parecer "antigo" ou até "pobre em recursos visuais", mas aí há o melhor recurso do R, a sua flexibilidade. Para usuários, a linguagem da R se torna clara e simples e a flexibilidade da interface de entrada por linha de comando permite que uns poucos comandos simples sejam juntados para criar funções poderosas. Além disso, a transparência das funções e a entrada de dados é altamente didática. O usuário é sempre consciente do que foi pedido através da interface. Isso contraste com outros pacotes em que uma interface bonita e sofisticada pode esconder a dinâmica dos cálculos, e potencialmente pode esconder erros.

5.2. A Linguagem R

R, bem como S, é desenhada ao de uma verdadeira linguagem de computador, e permite aos usuários acrescentar funcionalidade adicional por definição de novas funções. Muito do sistema é escrita no dialeto R da S, que faz com que seja fácil para usuários seguir as escolhas algorítmicas feitas. Para tarefas computacionalmente-intensivas, C, C++

e código Fortran podem ser ligados e chamados na hora de calcular. Usuários avançados podem escrever código C para manipular objetos R diretamente.

Muitos usuários pensam em R como um sistema estatístico. Nós preferimos pensar nele como um ambiente dentro do qual técnicas estatísticas são implementadas. R pode ser estendido (facilmente) através de pacotes. Alguns pacotes já vêm fornecidos com a distribuição R e muitos outros são disponíveis através da família CRAN de sítios na Internet cobrindo uma ampla variedade de estatísticas modernas.

R tem seu próprio formato de documentação, parecido com LaTex, que é usado para fornecer documentação compreensiva, tanto on-line e em uma variedade de formatos, como impressa.

5.3. Software Livre

O R é um Software Livre (livre no sentido de liberdade) distribuído sob a Licença Pública Geral e pode ser livremente copiado e distribuído entre usuários, bem como pode ser instalado em diversos computadores livremente. Isso contraste com pacotes comerciais que têm licenças altamente restritivas e não permitem que sejam feitas cópias ou que seja instalado em mais de um computador sem a devida licença (e pagamento, claro!).

A grande maioria de Softwares Livres são grátis e R não é uma exceção, isso contrasta com os pacotes comerciais.

Sendo um Software Livre, os códigos fontes do R estão disponíveis e atualmente são gerenciados por um grupo chamado o *Core Development Team* (http://r-project.org/contributors). A vantagem de ter o código aberto é que falhas podem ser detectadas e corrigidas rapidamente e atualizações para Softwares Livres podem ser disponibilizadas em uma questão de dias. Esse sistema de revisão depende pesadamente da participação dos usuários. Em contraste, em muitos pacotes comerciais, as falhas não são corrigidas até o próximo lançamento que pode levar vários anos.

5.4. Aplicação na Inferência Bayesiana

Neste trabalho foram utilizados além dos pacotes que já vem com o R, outros pacotes necessários para sua implementação. Estes foram:

5.4.1. Pacote BOA

O BOA (Bayesian Analysis Program) é um pacote específico para métodos MCMC, que realiza as análises de convergência das simulações e também disponibiliza os gráficos destas análises. Várias funções já implementadas neste pacote foram utilizadas, visto que todos os métodos de convergência vistos anteriormente estão implementados neste pacote.

5.4.2. Pacote MCMCpack

O MCMCpack (Markov Chain Monte Carlo) é um pacote que contém funções para utilização na Inferência Bayesiana usando simulação *a posteriori* para um número de modelos estatísticos. Sua utilização neste trabalho foi através da utilização da distribuição gama inversa, que já está implementada neste pacote.

.

6. MATERIAL E MÉTODOS

Os passos para a execução deste trabalho estão localizados abaixo:

6.1. Material

O material usado neste trabalho foi um conjunto de dados simulados. Nesta simulação de dados para a aplicação do amostrador de Gibbs, foi considerado o número de alelos por indivíduo igual a 2. E assim foram simulados indivíduos de uma população cada um com seus alelos.

No intuito de testar a utilização da inferência bayesiana, foram simulados os dados, sob vários cenários que diferiram;

1. pelo tamanho da amostra de indivíduos na população, n =10, 50,100 e 200.
2. pela probabilidade p de um alelo A no indivíduo da população, p= 0,1;0,5;0,9.

Com as restrições acima, perfaz-se um total de 12 cenários amostrados.

Baseado nesses dados simulados, através do modelo de COCKERHAM, considerado sob vários cenários estimou-se o coeficiente de endogamia e a taxa de fecundação cruzada através da inferência bayesiana.

6.2. Métodos

A seqüência realizada para alcance dos resultados foi:

6.2.1. Modelo de COCKERHAM

No caso de uma amostra de indivíduos de uma população diplóide, a variável indicadora da presença de um determinado alelo é descrita pelo modelo

$$\mathbf{y_{ij} = p + a_i + g_{(j)i}}$$

em que:

$\mathbf{y_{ij}}$ a freqüência do alelo j dentro do indivíduo i, que assume o valor um na presença do alelo e zero caso contrário;

p a freqüência paramétrica do alelo A na população;

$\mathbf{a_i}$ o efeito do indivíduo i, com i = 1, 2, ..., n;

$\mathbf{g_{(j)i}}$ o efeito do alelo j dentro do indivíduo i, com j = 1, 2.

Admitindo modelo aleatório conforme COCKERHAM (1969) tem-se a seguinte análise de variância:

Causa de Variação	GL	QM	E(QM)
Indivíduos	n-1	QMI	p(1 - p)[(1 - F) +
Alelos de Indivíduos	n	QMA	p(1 - p)[(1 - F)]

De acordo com demonstrações realizadas por MUNIZ et al. (1996) o coeficiente de endogamia (F), segundo o modelo considerado, pode ser estimado por:

$$\hat{F} = \frac{QMI - QMG}{QMI + QMG}$$

No processo de amostragem de indivíduos de uma população endogâmica com dois alelos, de acordo com o modelo de COCKERHAM (1969), a taxa de fecundação cruzada, t, se relaciona com o coeficiente de endogamia pela seguinte expressão:

$$F = \frac{1 - t}{1 + t},$$

que pode ser resolvida para t, fornecendo o seguinte resultado:

$$t = \frac{1 - F}{1 + F}.$$

Foi utilizada esta metodologia para geração dos valores iniciais para os parâmetros da metodologia bayesiana.

Assim, todo o desenvolvimento da teoria Bayesiana obtido na estimação do coeficiente de endogamia será utilizado no estudo da estimação da taxa de fecundação cruzada.

Propõe-se aqui a estimação dos parâmetros do modelo seja feita utilizando técnicas da inferência bayesiana. A derivação das expressões encontra-se descrita nas seções seguintes.

De acordo com o amostrador de gibbs, foram calculadas as distribuições condicionais completas *a posteriori* que foram usados na implementação deste algoritmo.

A seqüência para se chegar a estas distribuições foram:

6.2.2. Distribuição *a priori* para os parâmetros

Neste trabalho consideram-se as seguintes distribuições *a priori* para os parâmetros:

1) Distribuição *a priori* para p:

Devido à característica do parâmetro p, a distribuição *a priori* escolhida foi a distribuição beta, com valores entre 0 e 1 para a freqüência alélica.

$$p \sim Beta(\alpha, \beta)$$

2) Distribuição *a priori* para a_i:

Esse parâmetro assume uma distribuição normal.

$$a_i \sim N(0, \sigma_a^2)$$

3) Distribuição *a priori* para σ_a^2:

O parâmetro assume uma distribuição gama inversa, de acordo com as observações do modelo.

$$\sigma_a^2 \sim IG(a_1, b_1)$$

4) Distribuição *a priori* para σ_g^2:

O parâmetro também assume uma distribuição gama inversa, de acordo com as observações do modelo.

$$\sigma_g^2 \sim IG(a_2, b_2)$$

6.2.3. Função de verossimilhança dos dados

Devido á característica dos dados, que assume valores de 1 quando o alelo A está presente e 0 quando a alelo a está presente, assume-se uma distribuição de Bernoulli.

$$y \sim p^T (1-p)^{2n-T}$$

Na expressão n representa número de indivíduos amostrados e T corresponde ao:

$$\sum_{i=1}^{n} \sum_{j=1}^{2} y_{ij}$$

6.2.4. Distribuição conjunta *a posteriori*

A função de verossimilhança junto com as distribuições a *priori* dos parâmetros formam a distribuição a *posteriori* conjunta dos dados observados e dos parâmetros:

$$p(p, \sigma_a^2, \sigma_g^2, a_i \mid y) \propto L(p, \sigma_a^2, \sigma_g^2, a_i \mid y) p(p) p(\sigma_a^2) p(\sigma_g^2) p(a_i)$$

De forma equivalente tem-se:

$$p(p, \sigma_a^2, \sigma_g^2, a_i \mid y) \propto p^T (1-p)^{2n-T} Beta(\alpha, \beta) IG(a_1, b_1) IG(a_2, b_2) N(0, \sigma_a^2)$$

6.2.5. Distribuições condicionais completas a posteriori

Como a distribuição a *posteriori* conjunta não é tratável algebricamente, a inferência é baseada em amostras obtidas através das distribuições condicionais completas a *posteriori*, usando algoritmos MCMC (amostrador de Gibbs para distribuiçõoes condicionais completas a *posteriori* com forma fechada e Metropolis-Hastings para distribuições condicionais completas a *posteriori* desconhecidas).

1) Distribuição condicional completa a *posteriori* para p:

A distribuição condicional completa a *posteriori* para p é uma distribuição beta reparametrizada por:

$$p(p \mid \sigma_a^2, \sigma_g^2, a_i, y) \propto Beta(\alpha, 2(n-T)+\beta)$$

2) Distribuição condicional completa a *posteriori* para a_i:

$$p(a_i \mid p, \sigma_a^2, \sigma_g^2, y) \propto N(0, \sigma_a^2)$$

3) Distribuição condicional completa a *posteriori* para σ_a^2:

A distribuição condicional completa a *posteriori* para σ_a^2 é uma distribuição gama inversa reparametrizada por:

$$p(\sigma_a^2 \mid p, a_i, \sigma_g^2, y) \propto IG(n/2+a_1, \sum_{i=1}^{n} a_i^2/2+b_1)$$

4) Distribuição condicional completa a *posteriori* para σ_g^2:

A distribuição condicional completa a *posteriori* para σ_g^2 é uma distribuição gama inversa reparametrizada por:

$$p(\sigma_g^2 \mid p, a_i, \sigma_a^2, y) \propto IG(-T + a_2, b_2)$$

As distribuições condicionais completas a *posteriori* possuem forma fechada para os parâmetros $p, a_i, \sigma_a^2, \sigma_g^2$, ou seja, são distribuições conhecidas (normal, gama inversa e beta), portanto, pode-se utilizar o amostrador de Gibbs para amostrar dessas distribuições.

Os valores para os hiperparâmetros foram testados sendo escolhidos aqueles valores que possuíssem uma distribuição que abrangesse maior número de dados.

6.2.6. Simulação do conjunto de dados

Para a simulação dos dados são considerados dois argumentos: o número de indivíduos amostrados e a freqüência do alelo A. Como o número de alelos por indivíduo é fixo e igual a 2, os passos abaixo definem a simulação. Abaixo é apresentada uma seqüência dos passos executados na implementação da função para simulação dos dados.

Função: Simulação dos dados

Entrada: *Define-se o número de indivíduos amostrados n e a probabilidade p do alelo A nos indivíduos amostrados;*
Saída: *Uma matriz de n linhas e 2 colunas contendo os pares de alelos por indivíduo e a proporção do alelo A;*
Início
Número de alelos nAlelo fixo e igual a 2;
Para *i=1,...,n* ***faça***
Para *j=1,...,nAlelo* ***faça***
Gerar uma distribuição de Bernoulli com parâmetro p
Cálculo da proporção do Alelo A
Fim

6.2.5. Implementação do Amostrador de Gibbs

Foram consideradas nesta simulação 10.000 iterações sendo a 2000 primeiras descartadas para o aquecimento da cadeia.

As funções implementadas para obtenção de uma amostra da distribuição conjunta a *posteriori* , e, conseqüentemente, para as distribuições marginais de $p, a_i, \sigma_a^2, \sigma_g^2$ foram implementadas no programa estatístico R e consta basicamente dos seguintes passos:

Função: Amostrador de Gibbs

Entrada: *Definem-se os hiperparâmetros das distribuições condicionais completas, os valores iniciais dos parâmetros analisados, a matriz de dados, número de iterações(n) e número de iterações descartadas;*

Saída: *Uma matriz de n (número de iterações) e 1 colunas para cada um dos parâmetros analisados;*

Início

Cálculos para reparametrização das distribuições analisadas;

Valores iniciais dos parâmetros

Para *i=2,...,n* ***faça***

Gerar uma distribuição Beta para parâmetro p *;*

Gerar uma distribuição Normal para parâmetro a_i *;*

Gerar uma distribuição Gama Inversa para parâmetro σ_a^2 *;*

Gerar uma distribuição Gama Inversa para parâmetro σ_g^2 *;*

Gerar uma ditribuição para coeficiente de endogamia(f)

Gerar uma distribuição para taxa de fecundação cruzada(t);

Fim

7. RESULTADOS E DISCUSSÃO

O conjunto de dados simulado foi analisado pelo amostrador de Gibbs, sendo todos os parâmetros analisados por este algoritmo, não havendo a necessidade de utilização do algoritmo Metropolis-Hastings.

Efetuo-se um processo com 10.000 iterações, sendo descartadas as 2000 iniciais, para o período de aquecimento da cadeia ("burn-in").

7.1. Resultados Alcançados e Análises

As estimativas bayesianas apresentam o resumo *a posteriori* para cada parâmetro, representadas pela média da distribuição marginal *a posteriori*, o desvio-padrão e seu intervalo de credibilidade.

Foram analisados 12 cenários, considerando sempre o número de indivíduos amostrados e a freqüência do alelo A. Os cenários foram:

a) 10 indivíduos amostrados com freqüência do alelo A baixa;
b) 10 indivíduos amostrados com freqüência do alelo A média;
c) 10 indivíduos amostrados com freqüência do alelo A alta;
d) 50 indivíduos amostrados com freqüência do alelo A baixa;
e) 50 indivíduos amostrados com freqüência do alelo A média;
f) 50 indivíduos amostrados com freqüência do alelo A alta;
g) 100 indivíduos amostrados com freqüência do alelo A baixa;
h) 100 indivíduos amostrados com freqüência do alelo A média;
i) 100 indivíduos amostrados com freqüência do alelo A alta;
j) 200 indivíduos amostrados com freqüência do alelo A baixa;
k) 200 indivíduos amostrados com freqüência do alelo A média;
l) 200 indivíduos amostrados com freqüência do alelo A alta;

Abaixo são apresentados os resultados obtidos da utilização da inferência bayesiana assim como os gráficos representativos do processo iterativo sob todos os cenários.

Tabela 7.1: Estimativas da média, desvio padrão e intervalo de credibilidade de 95% para os parâmetros considerando n=10 e freqüência alélica baixa, média e alta.

p	Parâmetros	Valores Iniciais	Média	d.p.	2,5%	97,5%
0,1	**p**	0,1500	0,1499	0,0099	0,1312	0,1704
	σ_a^2	0,1166	0,1171	0,0100	0,0985	0,1385
	σ_g^2	0,1500	0,1501	0,0101	0,1317	0,1708
	f	0,0000	0,0012	0,0041	0,0000	0,0118
	t	1,0000	0,99760	0,0080	0,9765	1,0000
0,5	**p**	0,4000	0,4000	0,0099	0,3805	0,4199
	σ_a^2	0,2000	0,2003	0,0099	0,1821	0,2211
	σ_g^2	0,3000	0,2998	0,0099	0,2809	0,3202
	f	0,0000	0,0009	0,0030	0,0000	0,0098
	t	1,0000	0,9980	0,0059	0,9804	1,0000
0,9	**p**	0,8500	0,8499	0,0101	0,8296	0,8690
	σ_a^2	0,2277	0,2282	0,0100	0,2092	0,2487
	σ_g^2	0,0500	0,0497	0,0099	0,0342	0,0723
	f	0,6400	0,6435	0,0588	0,5132	0,7414
	t	0,2195	0,2184	0,0448	0,1484	0,3216

Os resultados apresentados na tabela 7.1 mostram que os valores obtidos pelas médias da distribuição *a posteriori* estiveram bem próximos aos valores gerados inicialmente pela metodologia clássica.

Podemos verificar também que o desvio-padrão propiciou uma inferência muito segura, pois apresentou valores muito pequenos e também que a metodologia bayesiana mostrou-se eficiente para estimar os parâmetros, pois os valores paramétricos utilizados na simulação encontram-se dentro do intervalo de credibilidade de 95%.

Os gráficos representativos do processo iterativo envolvido no algoritmo Gibbs Sampler, bem como as distribuições marginais originadas, das quais se obtiveram as estimativas apresentadas na tabela, são mostrados nas figuras abaixo. São apresentados o histórico das iterações e o gráfico da densidade *a posteriori* dos parâmetros.

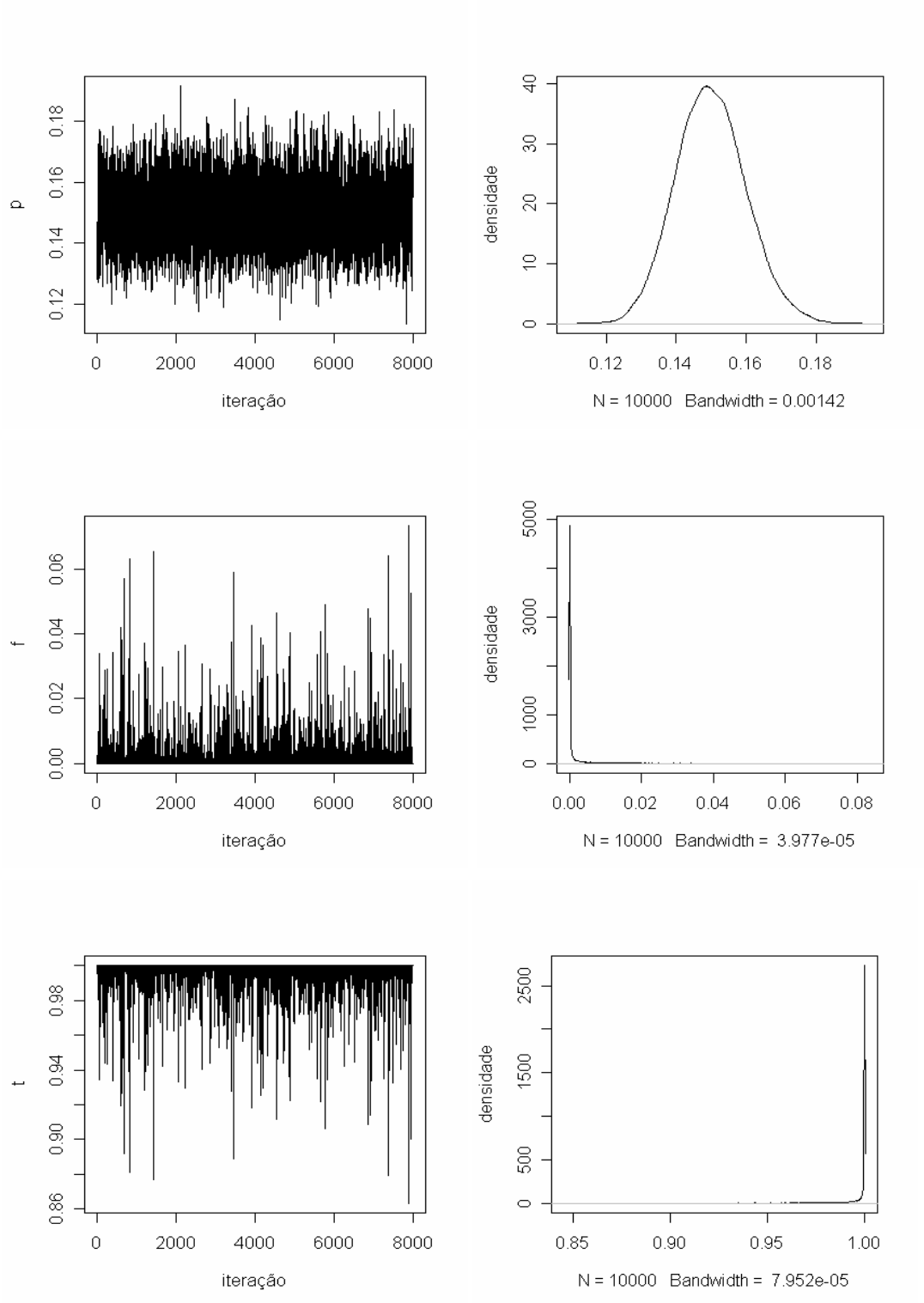

Figura 7.1: Representação gráfica do Gibbs Sampler e distribuição *a posteriori* para o parâmetros estimados para n=10 e p=0,1

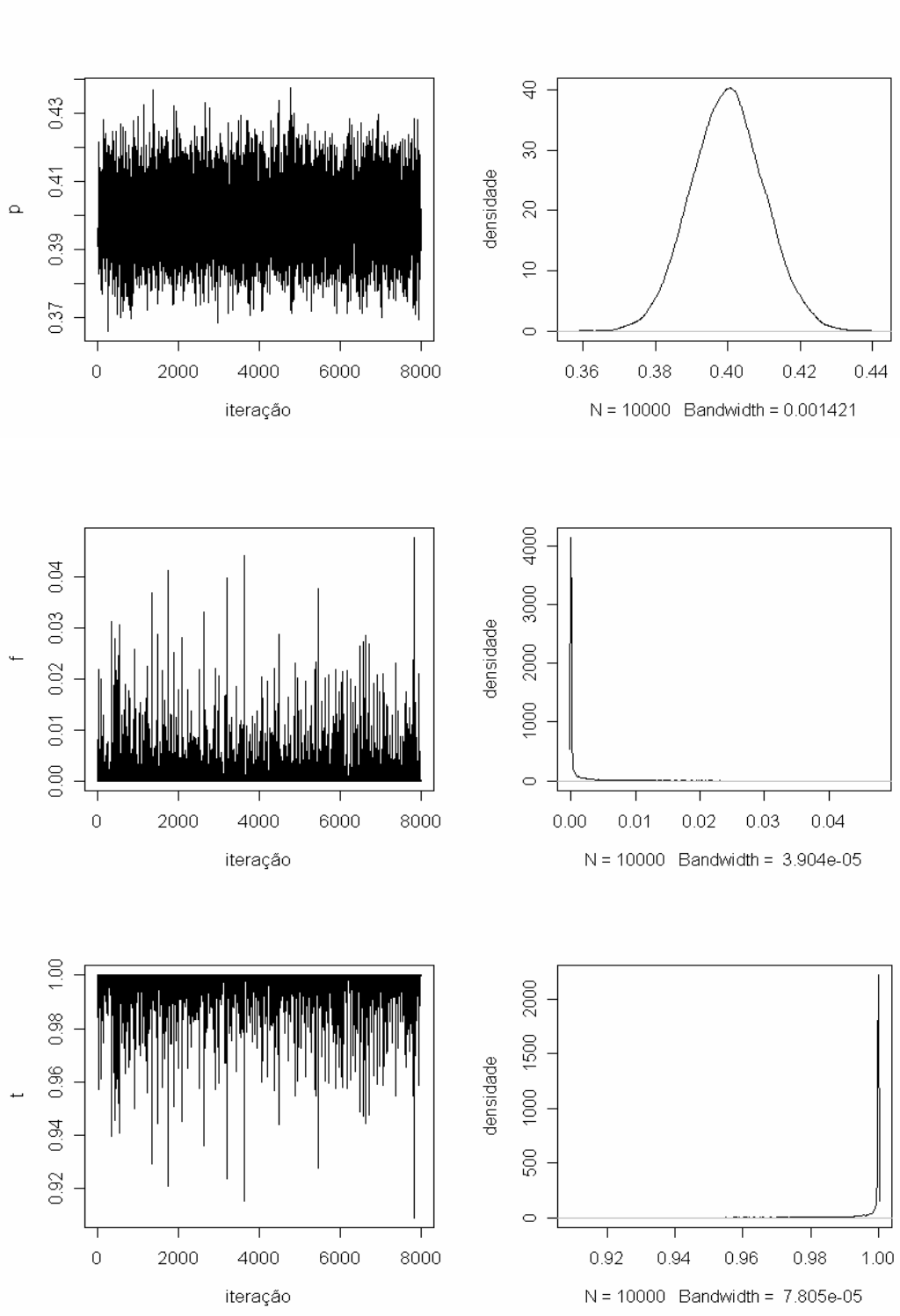

Figura 7.2: Representação gráfica do Gibbs Sampler e distribuição *a posteriori* para o parâmetros estimados para n=10 e p=0,5

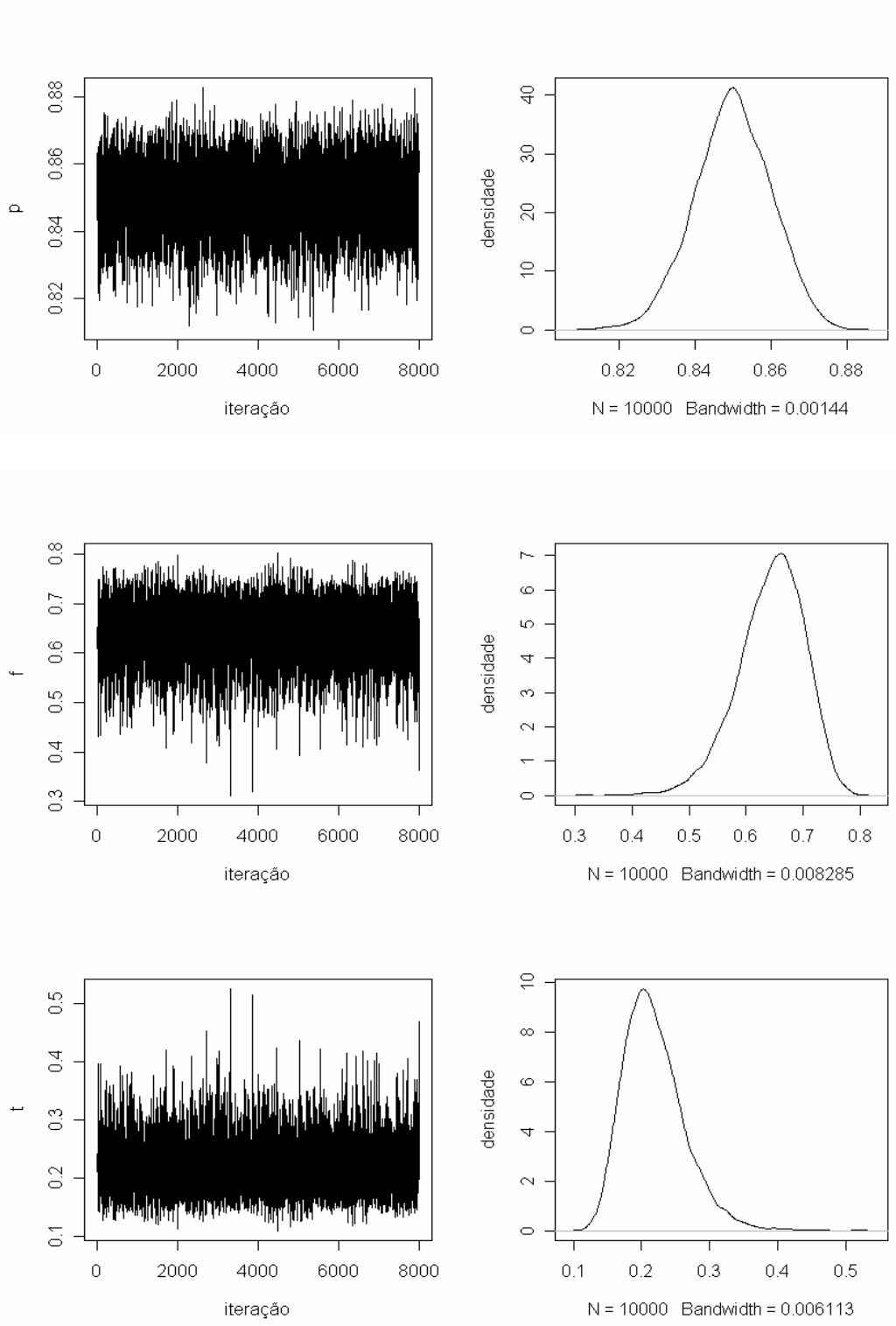

Figura 7.3: Representação gráfica do Gibbs Sampler e distribuição *a posteriori* para o parâmetros estimados para n=10 e p=0,9

Tabela 7.2: Média, desvio padrão e intervalo de credibilidade de 95% para os parâmetros considerando n=50 e freqüência alélica baixa, média e alta.

p	Parâmetros	Valores Iniciais	Média	d.p.	2,5%	97,5%
0,1	**p**	0,1000	0,0999	0,0099	0,0814	0,1203
	σ_a^2	0,0816	0,0844	0,0103	0,0665	0,1069
	σ_g^2	0,1000	0,1000	0,0100	0,0821	0,1214
	f	0,0000	0,0063	0,0185	0,0000	0,0688
	t	1,0000	0,9880	0,0339	0,8711	1,0000
0,5	**p**	0,5500	0,5498	0,0099	0,5302	0,5692
	σ_a^2	0,3112	0,3137	0,0099	0,2945	0,3341
	σ_g^2	0,1900	0,1900	0,0099	0,1719	0,2103
	f	0,2418	0,2456	0,0287	0,1884	0,3000
	t	0,6104	0,6064	0,0372	0,5384	0,6828
0,9	**p**	0,8600	0,8599	0,0101	0,8393	0,8791
	σ_a^2	0,1436	0,1462	0,0102	0,1274	0,1679
	σ_g^2	0,1000	0,1000	0,0098	0,0824	0,1210
	f	0,1792	0,1883	0,0576	0,0710	0,2970
	t	0,6960	0,6871	0,0826	0,5419	0,8674

Os resultados apresentados na tabela 7.2 mostram também que os valores obtidos pelas médias da distribuição *a posteriori* estiveram bem próximos aos valores gerados inicialmente pela metodologia clássica.

Podemos verificar também que o desvio-padrão propiciou uma inferência muito segura, pois apresentou valores muito pequenos, mas esse valor tendeu a aumentar com o número de indivíduos amostrados e também que a metodologia bayesiana mostrou-se eficiente para estimar os parâmetros, pois os valores paramétricos utilizados na simulação encontram-se dentro do intervalo de credibilidade de 95%.

Os gráficos representativos do processo iterativo envolvido no algoritmo Gibbs Sampler, bem como as distribuições marginais originadas, das quais se obtiveram as estimativas apresentadas na tabela, são mostrados nas figuras abaixo. São apresentados o histórico das iterações e o gráfico da densidade *a posteriori* dos parâmetros.

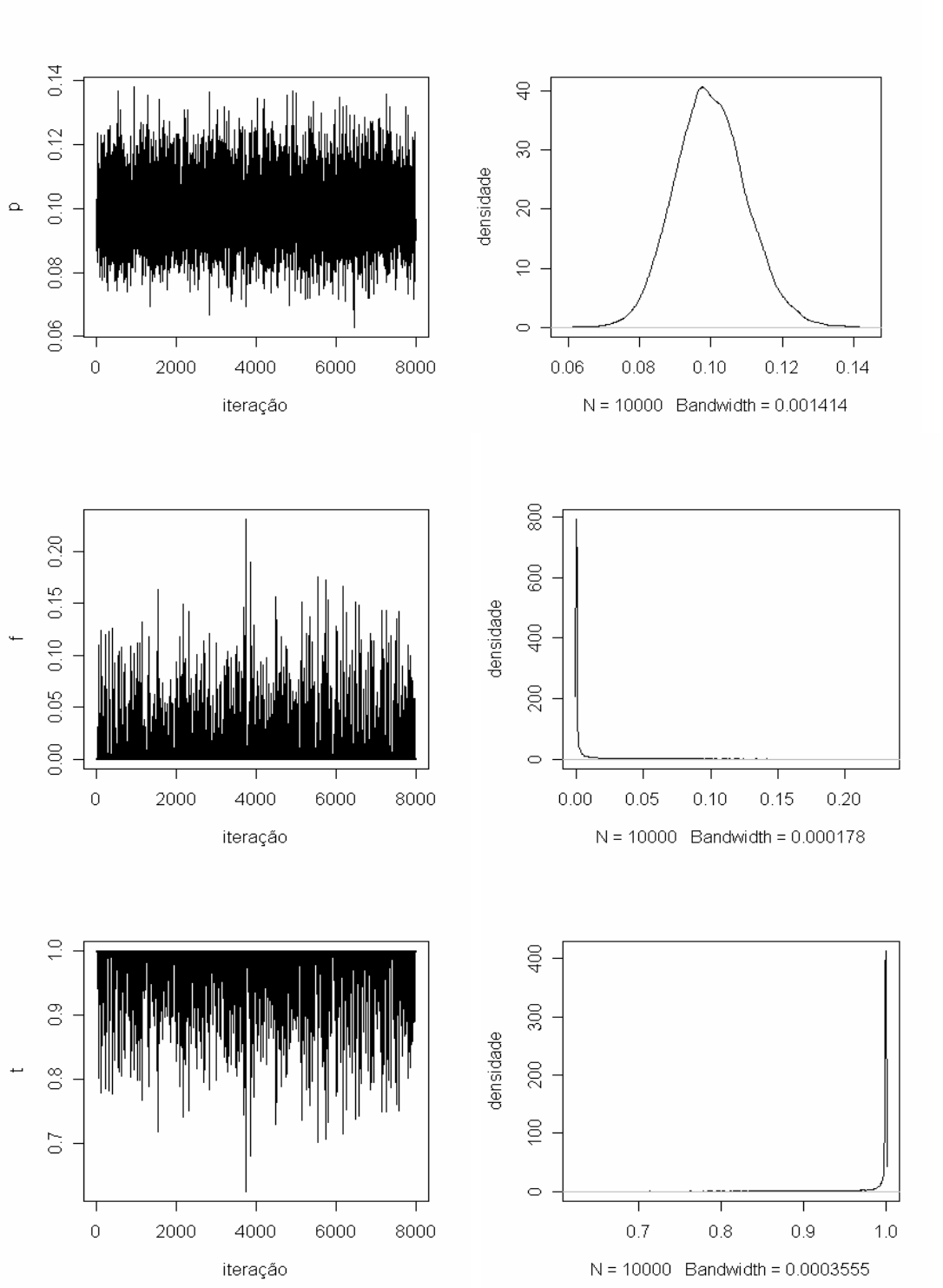

Figura 7.4: Representação gráfica do Gibbs Sampler e distribuição *a posteriori* para o parâmetros estimados para n=50 e p=0,1

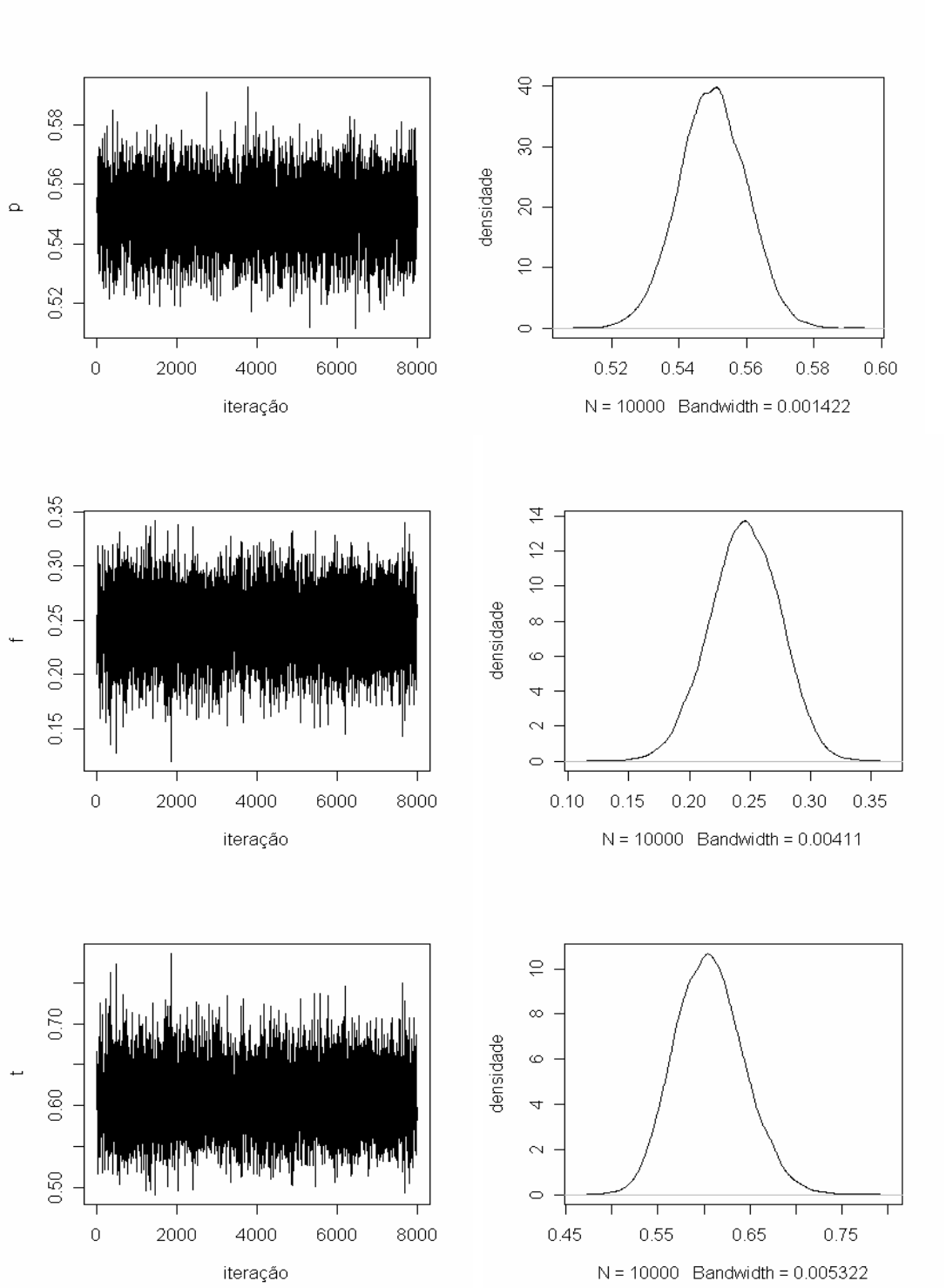

Figura 7.5: Representação gráfica do Gibbs Sampler e distribuição *a posteriori* para o parâmetros estimados para n=50 e p=0,5

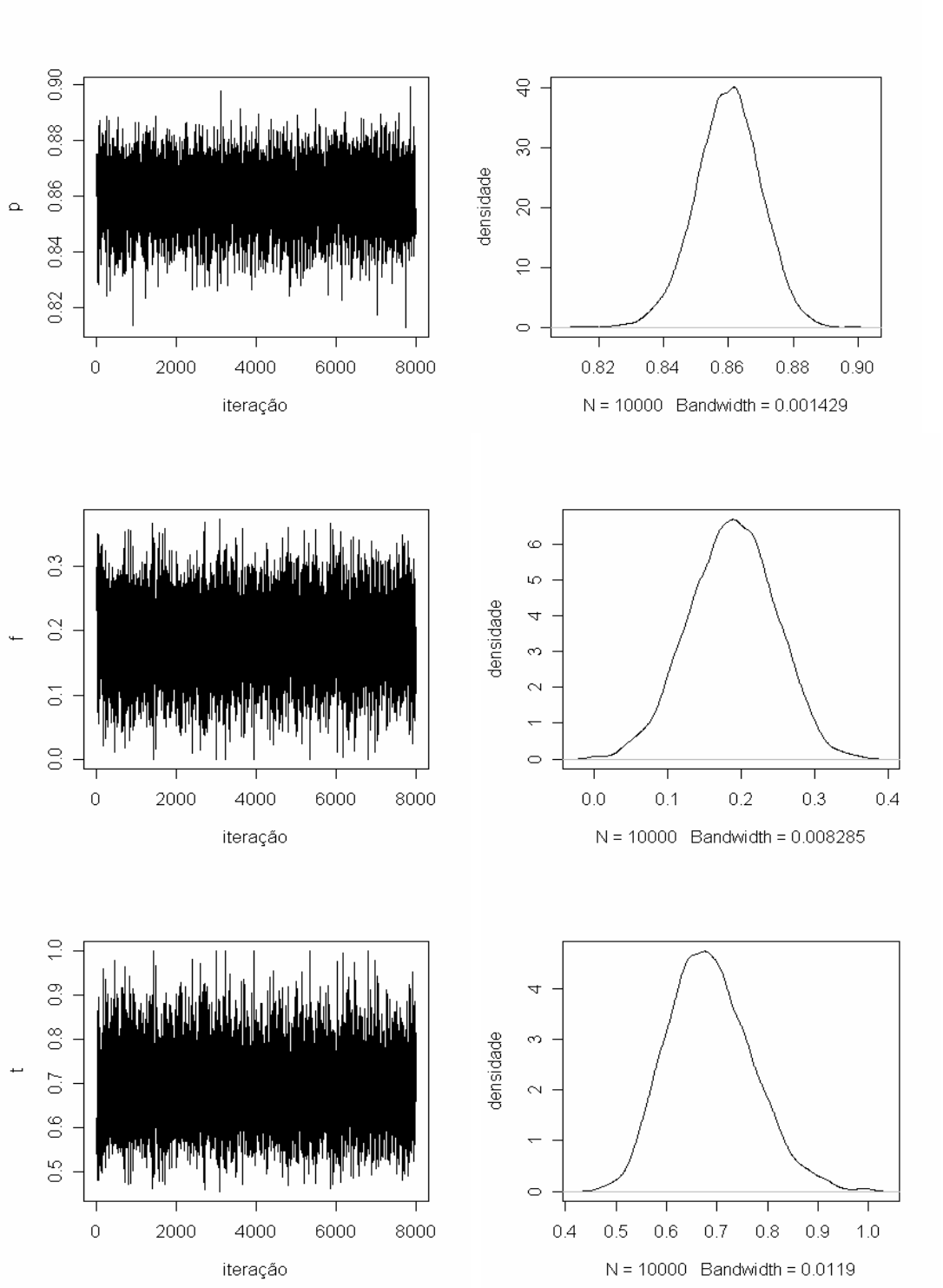

Figura 7.6: Representação gráfica do Gibbs Sampler e distribuição *a posteriori* para o parâmetros estimados para n=50 e p=0,9

Tabela 7.3: Média, desvio padrão e intervalo de credibilidade de 95% para os parâmetros considerando n=100 e freqüência alélica baixa, média e alta.

p	Parâmetros	Valores iniciais	Média	d.p.	2,5%	97,5%
0,1	**p**	0,1250	0,1249	0,0099	0,1062	0,1452
	σ_a^2	0,1047	0,1103	0,0104	0,0915	0,1327
	σ_g^2	0,1150	0,1151	0,0099	0,0967	0,1158
	f	0,0000	0,0166	0,0295	0,0000	0,1036
	t	1,0000	0,9688	0,0537	0,8121	1,0000
0,5	**p**	0,4650	0,4649	0,0099	0,4453	0,4846
	σ_a^2	0,2652	0,2704	0,0101	0,2512	0,2908
	σ_g^2	0,2350	0,2351	0,0100	0,2163	0,2561
	f	0,0603	0,0697	0,0280	0,0139	0,1249
	t	0,8861	0,8708	0,0492	0,7778	0,9724
0,9	**p**	0,9150	0,9152	0,0099	0,8950	0,9338
	σ_a^2	0,0813	0,0871	0,0107	0,0687	0,1108
	σ_g^2	0,0750	0,0747	0,0100	0,0574	0,0970
	f	0,0406	0,0867	0,0743	0,0000	0,2500
	t	0,9217	0,8487	0,1219	0,5999	1,0000

Os resultados apresentados na tabela 7.3 mostram também que os valores obtidos pelas médias da distribuição *a posteriori* estiveram bem próximos aos valores gerados inicialmente pela metodologia clássica.

Podemos verificar também que o desvio-padrão propiciou uma inferência muito segura, pois apresentou valores muito pequenos, mas esse valor tendeu a aumentar com o número de indivíduos amostrados e também que a metodologia bayesiana mostrou-se eficiente para estimar os parâmetros, pois os valores paramétricos utilizados na simulação encontram-se dentro do intervalo de credibilidade de 95%, mas esses intervalos tenderam a ficarem maiores com o aumento de indivíduos. Os gráficos representativos do processo iterativo envolvido no algoritmo Gibbs Sampler, bem como as distribuições marginais originadas, das quais se obtiveram as estimativas apresentadas na tabela, são mostrados nas figuras abaixo. São apresentados o histórico das iterações e o gráfico da densidade *a posteriori* dos parâmetros.

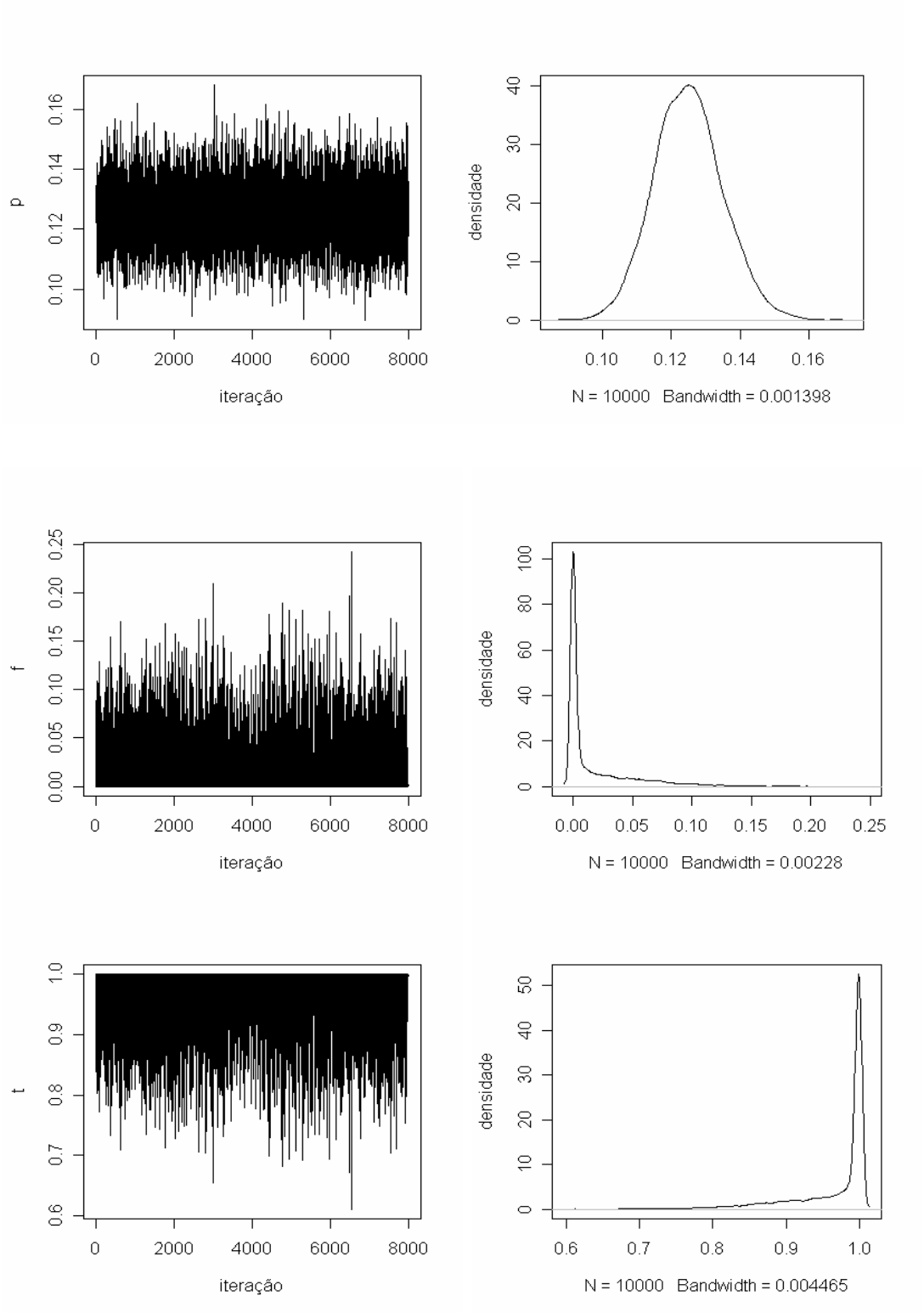

Figura 7.7: Representação gráfica do Gibbs Sampler e distribuição *a posteriori* para o parâmetros estimados para n=100 e p=0,1

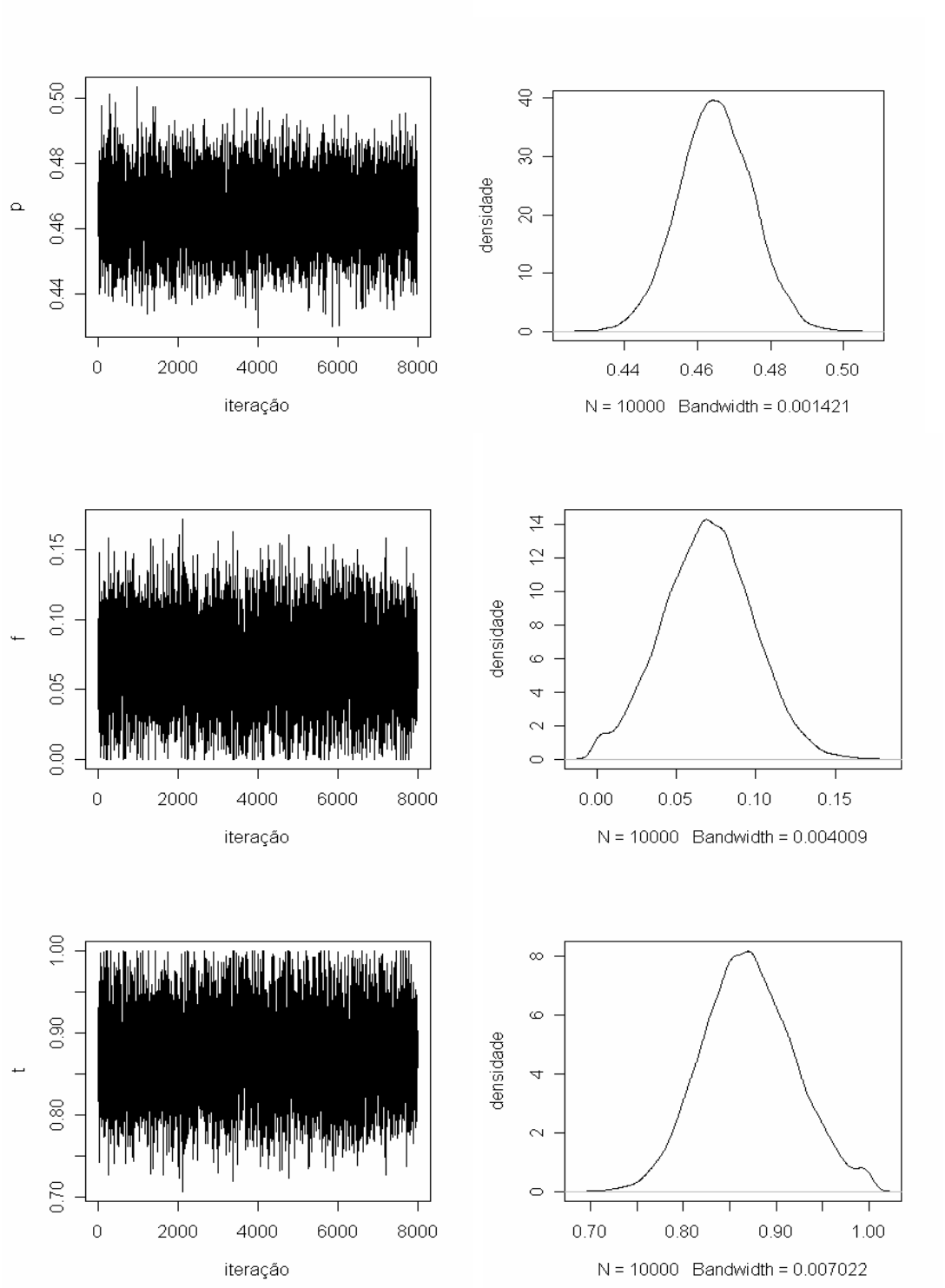

Figura 7.8: Representação gráfica do Gibbs Sampler e distribuição *a posteriori* para o parâmetros estimados para n=100 e p=0,5

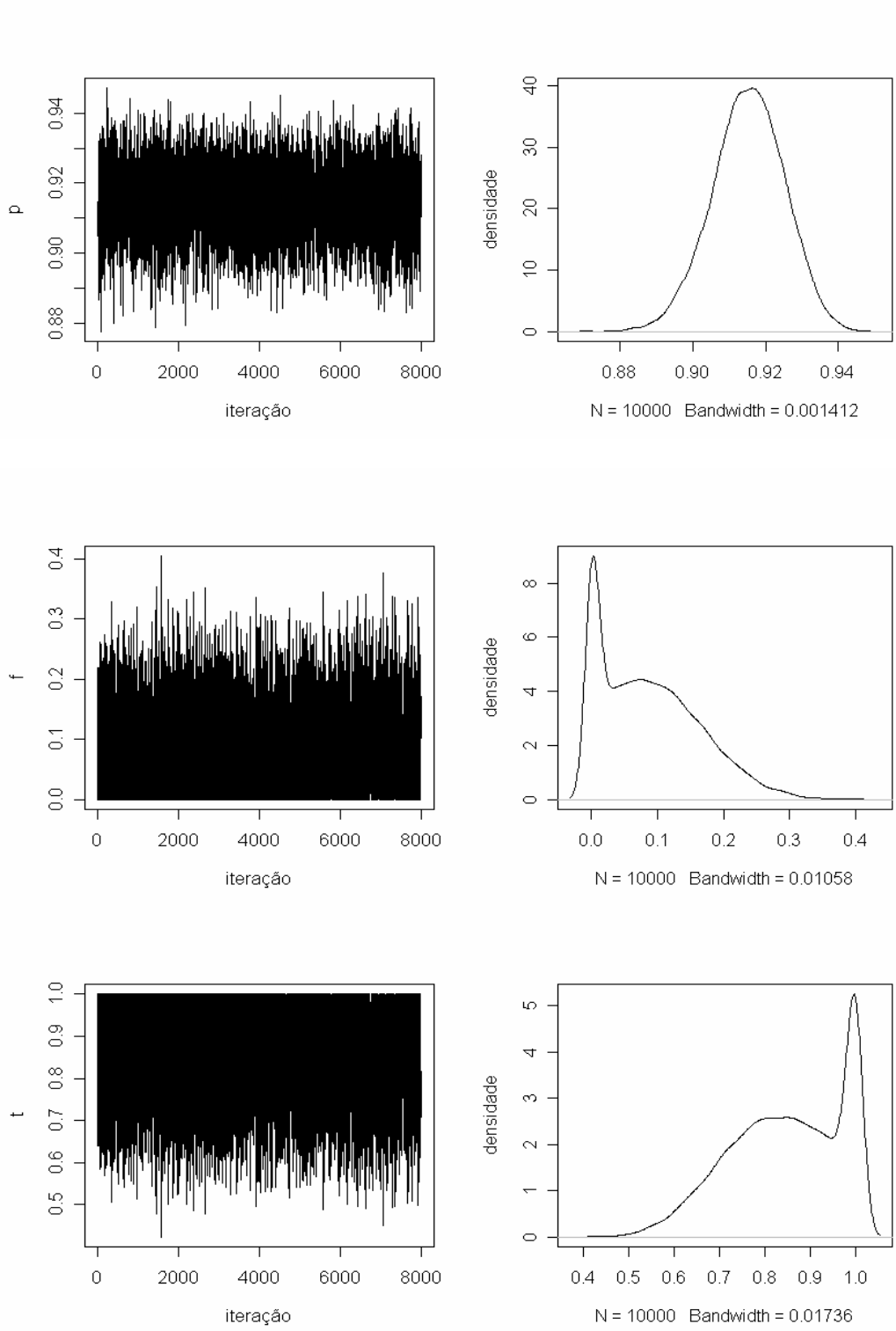

Figura 7.9: Representação gráfica do Gibbs Sampler e distribuição *a posteriori* para o parâmetros estimados para n=100 e p=0,9

Tabela 7.4: Média, desvio padrão e intervalo de credibilidade de 95% para os parâmetros considerando n=200 e freqüência alélica baixa, média e alta

p	Parâmetros	Valores iniciais	Média	d.p.	2,5%	97,5%
0,1	**p**	0,0850	0,0849	0,0100	0,0664	0,1053
	σ_a^2	0,0910	0,1039	0,0118	0,0833	0,1301
	σ_g^2	0,0650	0,0651	0,0100	0,0484	0,0874
	f	0,1666	0,2302	0,0889	0,0489	0,3966
	t	0,7142	0,6343	0,1207	0,4319	0,9065
0,5	**p**	0,4850	0,4848	0,0099	0,4653	0,5044
	σ_a^2	0,2508	0,2615	0,0107	0,2413	0,2834
	σ_g^2	0,2500	0,2500	0,0100	0,2314	0,2703
	f	0,0016	0,0262	0,0233	0,0000	0,0796
	t	0,9967	0,9498	0,0435	0,8524	1,0000
0,9	**p**	0,9050	0,9050	0,0099	0,8846	0,9237
	σ_a^2	0,0974	0,1103	0,0116	0,0895	0,1349
	σ_g^2	0,0750	0,0750	0,0101	0,0582	0,0975
	f	0,1301	0,1911	0,0804	0,0235	0,3435
	t	0,7697	0,6868	0,1156	0,4886	0,9540

Os resultados apresentados na tabela 7.3 mostram também que os valores obtidos pelas médias da distribuição *a posteriori* estiveram bem próximos aos valores gerados inicialmente pela metodologia clássica.

Podemos verificar também que o desvio-padrão propiciou uma inferência muito segura, pois apresentou valores muito pequenos, mas esse valor aumentou muito com o número de indivíduos amostrados e também que a metodologia bayesiana mostrou-se eficiente para estimar os parâmetros, pois os valores paramétricos utilizados na simulação encontram-se dentro do intervalo de credibilidade de 95%, mas esses intervalos tenderam a ficarem maiores com o aumento de indivíduos.

Os gráficos representativos do processo iterativo envolvido no algoritmo Gibbs Sampler, bem como as distribuições marginais originadas, das quais se obtiveram as estimativas apresentadas na tabela, são mostrados nas figuras abaixo. São apresentados o histórico das iterações e o gráfico da densidade *a posteriori* dos parâmetros.

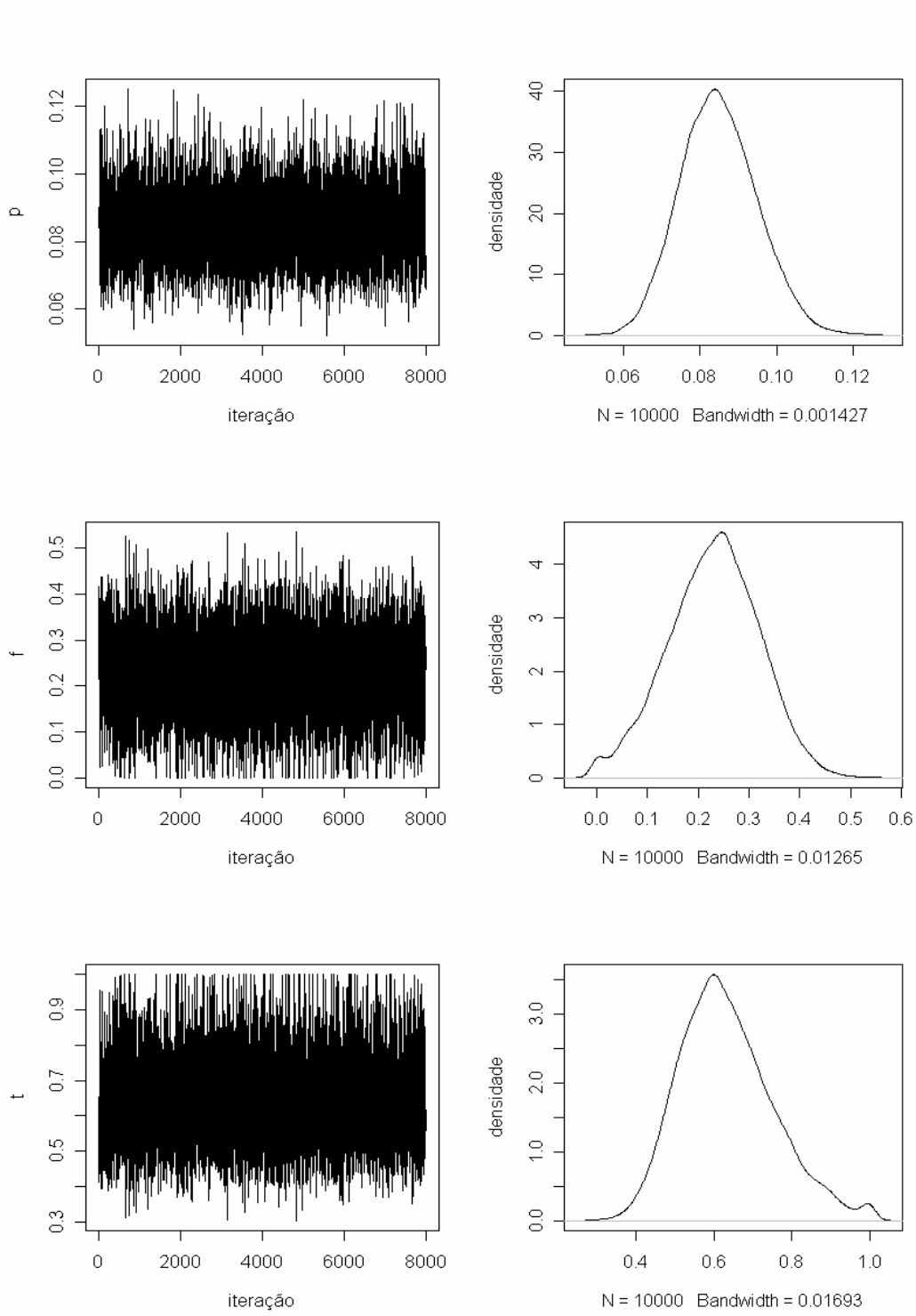

Figura 7.10: Representação gráfica do Gibbs Sampler e distribuição *a posteriori* para o parâmetros estimados para n=200 e p=0,1

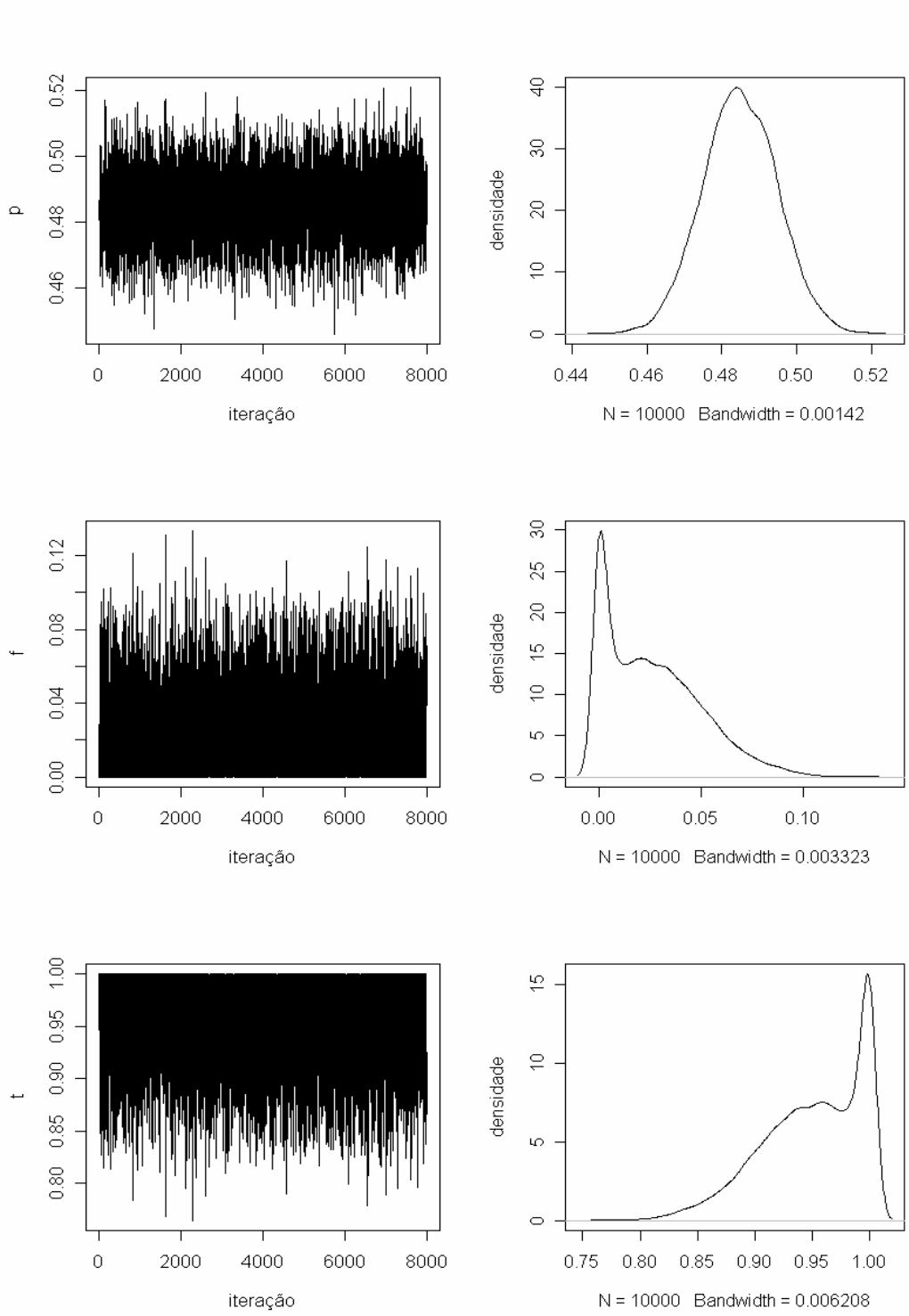

Figura 7.11: Representação gráfica do Gibbs Sampler e distribuição *a posteriori* para o parâmetros estimados para n=200 e p=0,5

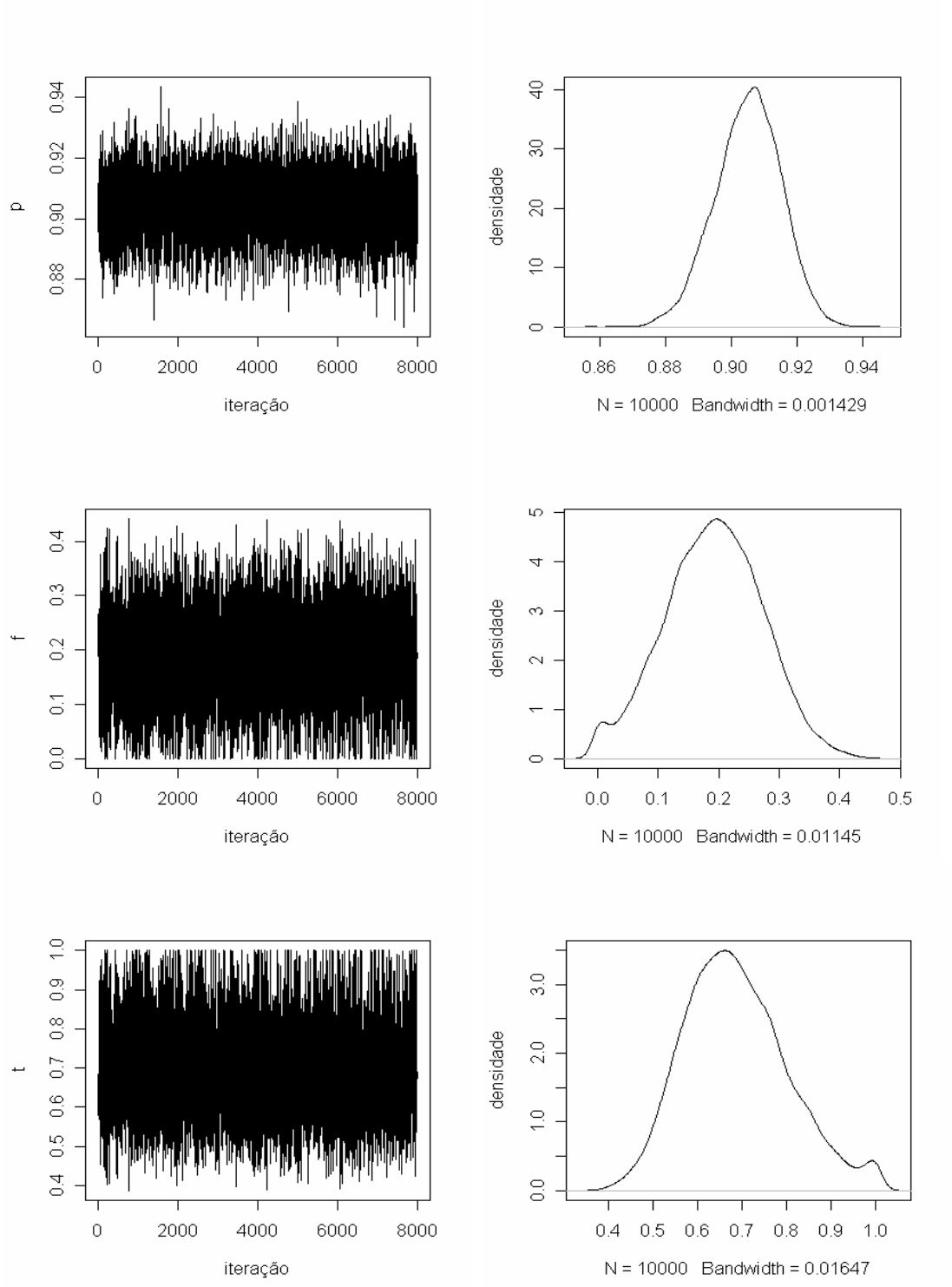

Figura 7.12: Representação gráfica do Gibbs Sampler e distribuição *a posteriori* para o parâmetros estimados para n=200 e p=0,9

Podemos dizer que à medida que o número de indivíduos amostrados aumenta, existe uma tendência a essas estimativas se distanciarem dos valores inicias, mas os valores ainda continuam muito próximos aos valores iniciais.

7.2. Análise de Convergência dos Parâmetros

Depois de analisadas as estatísticas descritivas, fazem-se necessário todo uma análise de convergência dos parâmetros. Neste caso, foram analisados os principais parâmetros de interesse desse trabalho, sendo então avaliados a freqüência alélica (p), o coeficiente de endogamia (f) e a taxa de fecundação cruzada (t).

Para todos os cenários foram realizados os seguintes testes e resultados:

a) Diagnóstico de convergência de Heidelberger e Welch: Em todos os cenários, a cadeia passou pelo teste de estacionariedade, sendo em alguns casos descartadas algumas iterações, mas estas representaram muito pouco o total, e passaram também pelo teste de Halfwidth.

b) Diagnóstico de convergência de Raftery e Lewis: Em todos os cenários, a cadeia mostrou que 10000 iterações foi um número muito grande, sendo que para que ocorresse a convergência necessitaria em torno de umas 3500 iterações em média.

c) Diagnóstico de convergência de Gelman e Rubin: Com a divisão da cadeia, esse teste mostrou que o início e o fim da cadeia possuem valores próximos.

d) Função de autocorrelação: Como será mostrado pelas figuras abaixo, esta apresentou valores pequenos que representam a convergência da cadeia.

e) Diagnóstico de convergência de Geweke: Através do resultados apresentados abaixo através de figuras, pode observar que o teste aplicado se mostrou nos níveis de uma convergência.

As figuras abaixo representam os gráficos do diagnóstico de convergência de Geweke e da função de autocorrelação para todos os cenários.

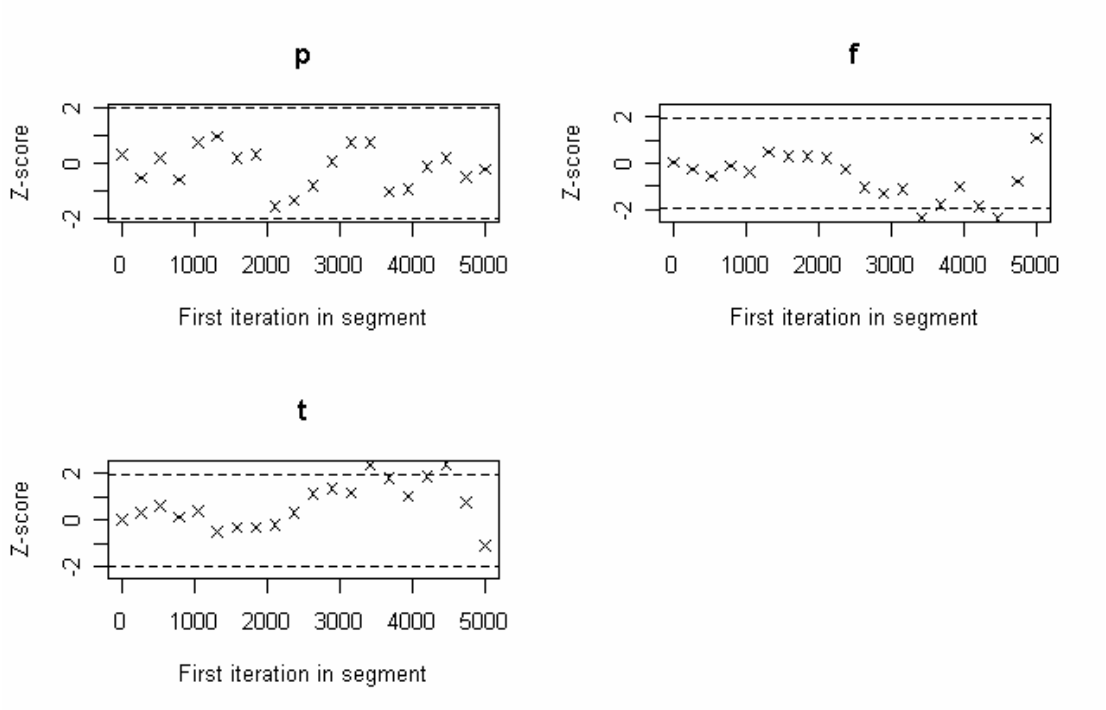

Figura 7.13: Representação gráfica do diagnóstico de convergência Geweke para os parâmetros estimados sendo n=10 e p=0,1

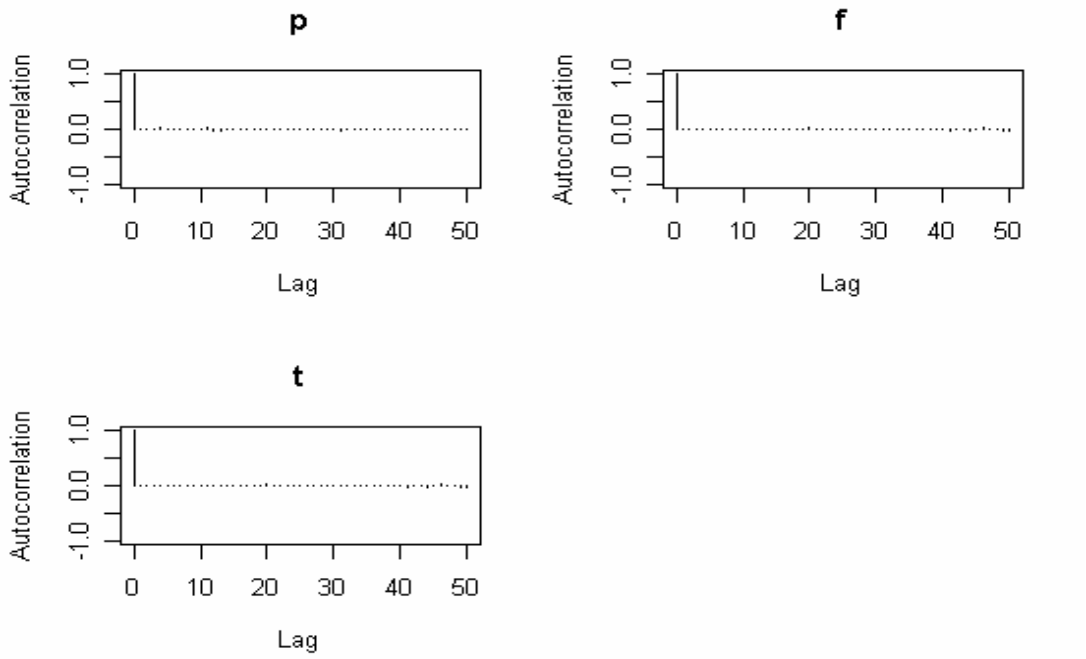

Figura 7.14: Representação gráfica da função de autocorrelação para os parâmetros estimados para n=10 e p=0,1

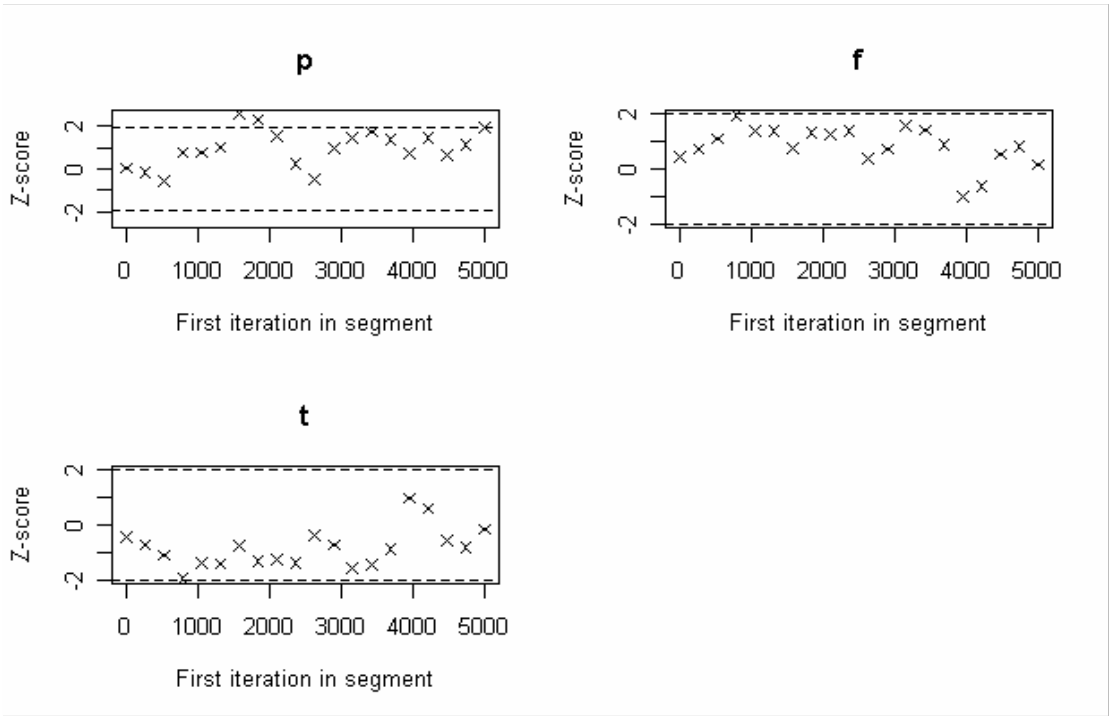

Figura 7.15: Representação gráfica do diagnóstico de convergência Geweke para os parâmetros estimados sendo n=10 e p=0,5

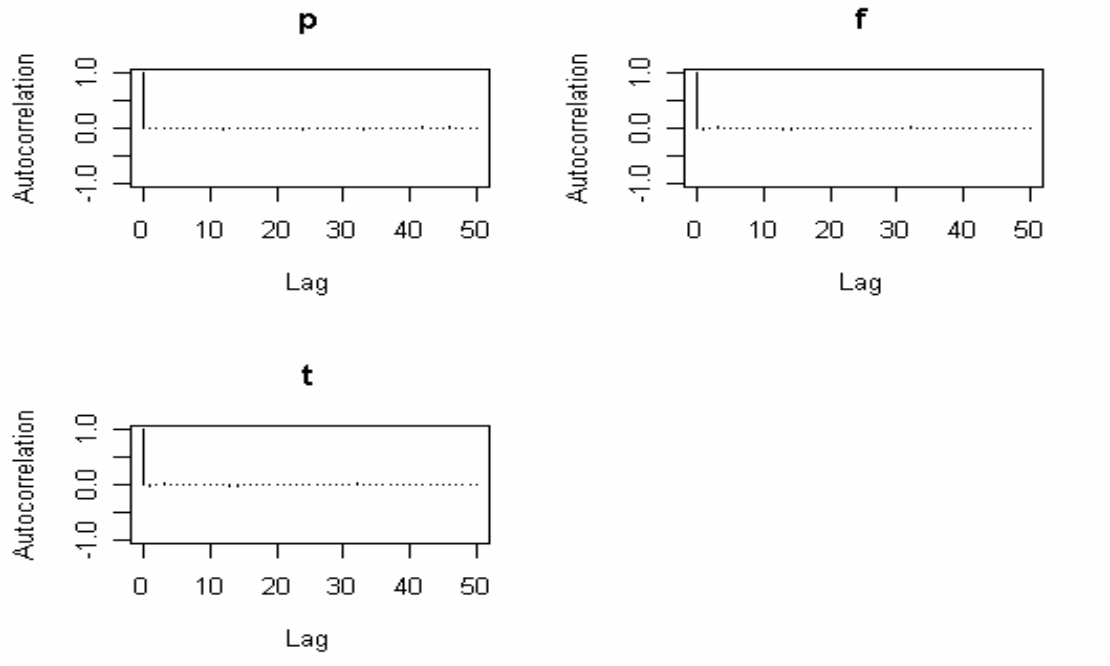

Figura 7.16: Representação gráfica da função de autocorrelação para os parâmetros estimados para n=10 e p=0,5

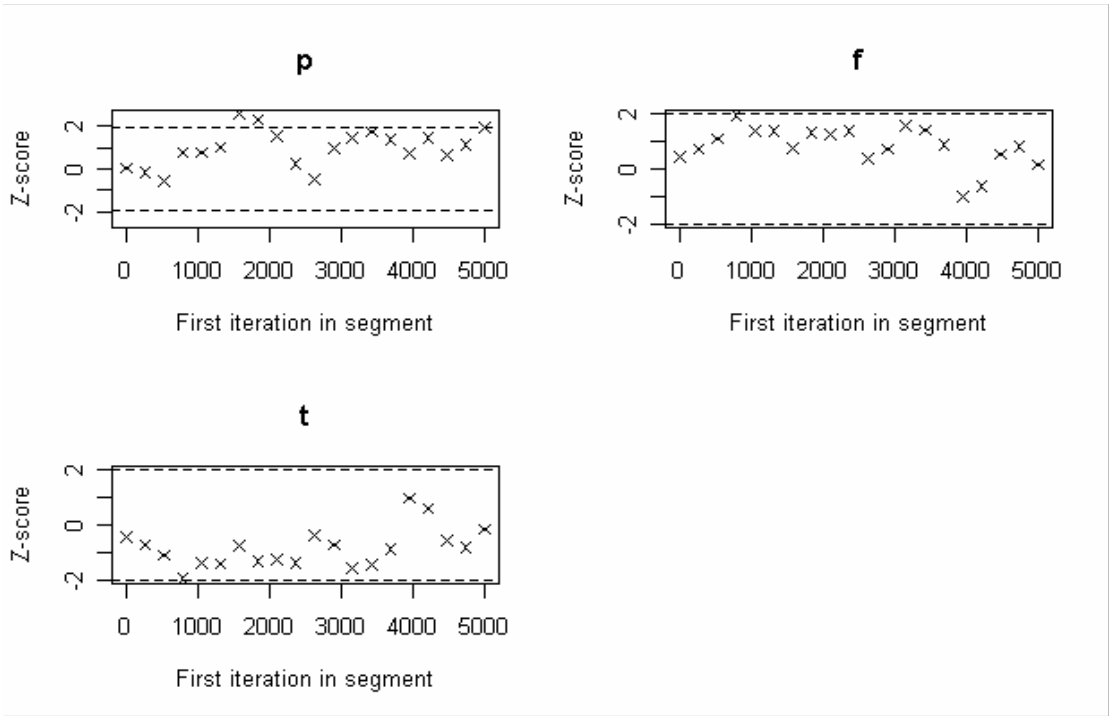

Figura 7.17: Representação gráfica do diagnóstico de convergência Geweke para os parâmetros estimados sendo n=10 e p=0,9

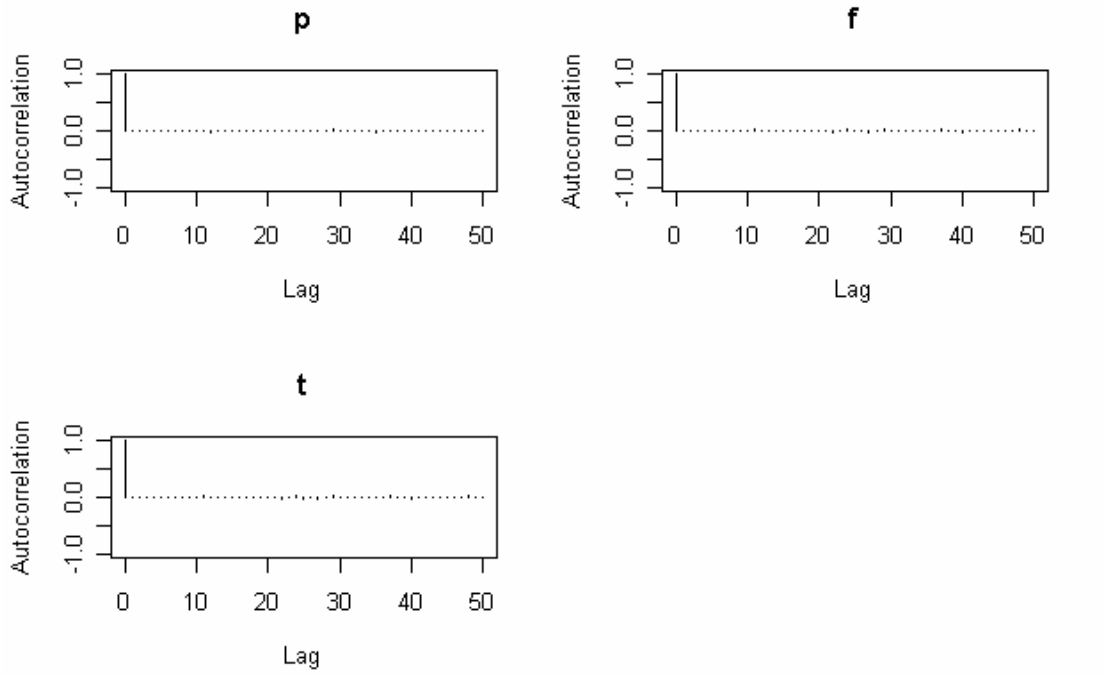

Figura 7.18: Representação gráfica da função de autocorrelação para os parâmetros estimados para n=10 e p=0,9

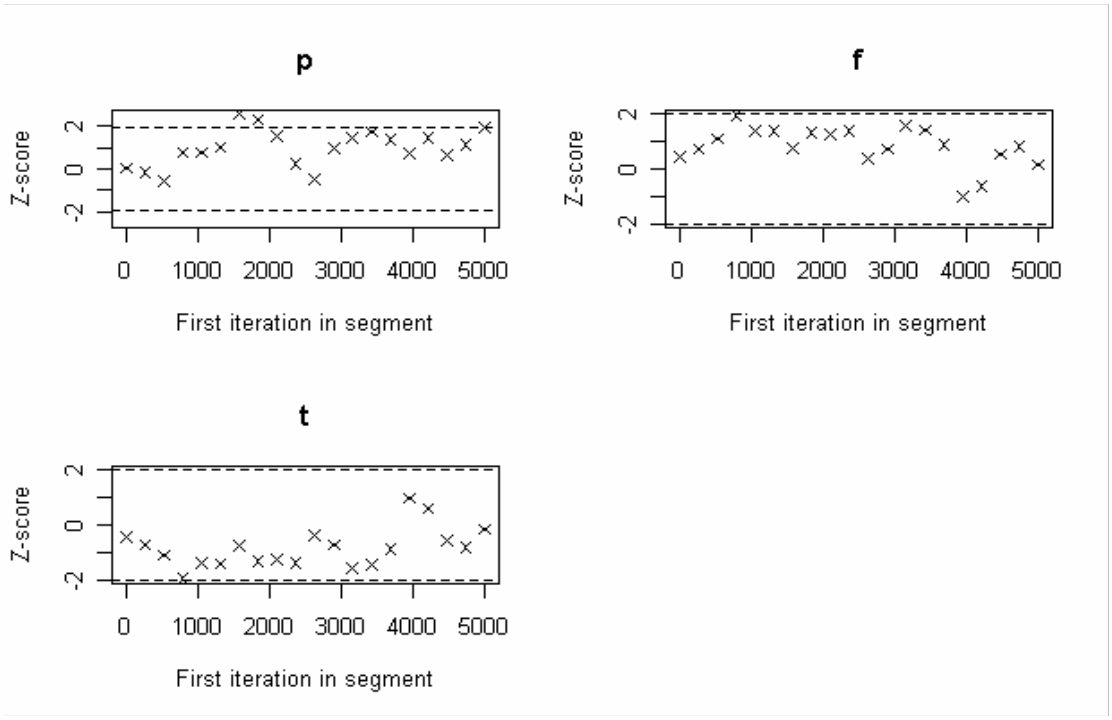

Figura 7.19: Representação gráfica do diagnóstico de convergência Geweke para os parâmetros estimados sendo n=50 e p=0,1

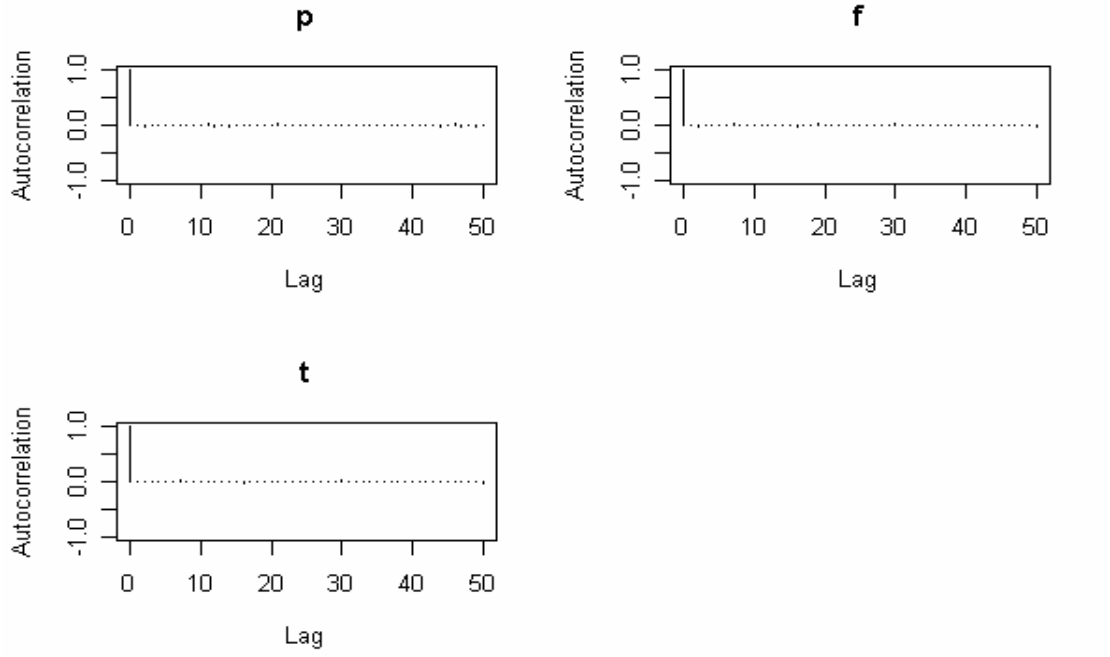

Figura 7.20: Representação gráfica da função de autocorrelação para os parâmetros estimados para n=50 e p=0,1

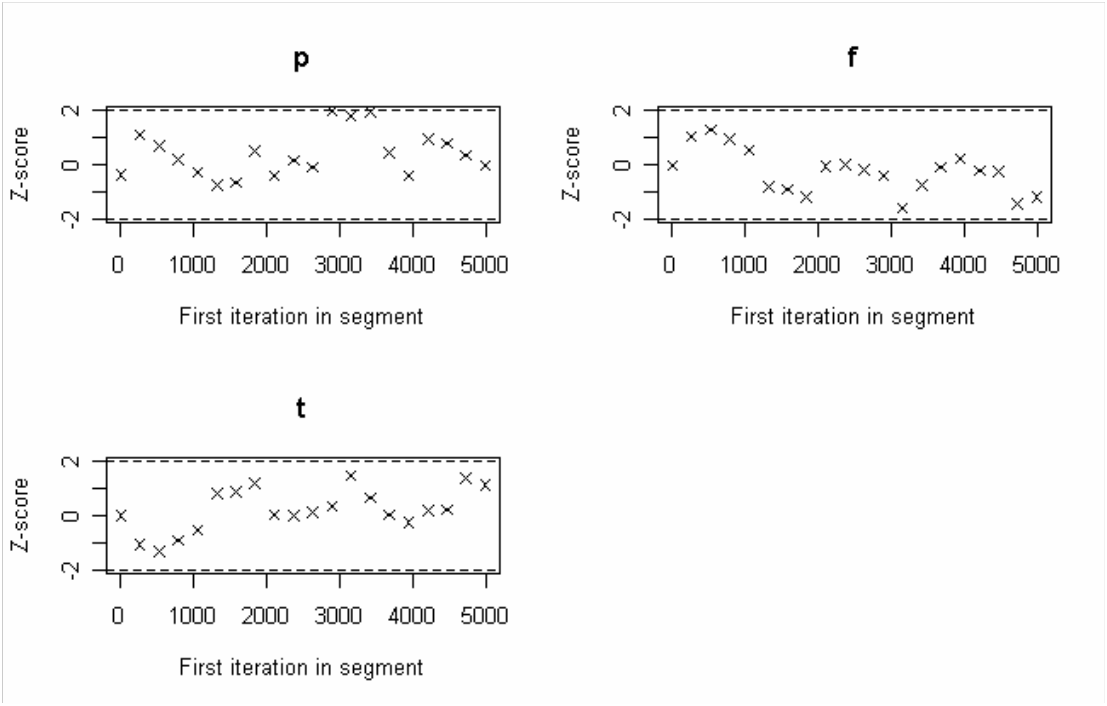

Figura 7.21: Representação gráfica do diagnóstico de convergência Geweke para os parâmetros estimados sendo n=50 e p=0,5

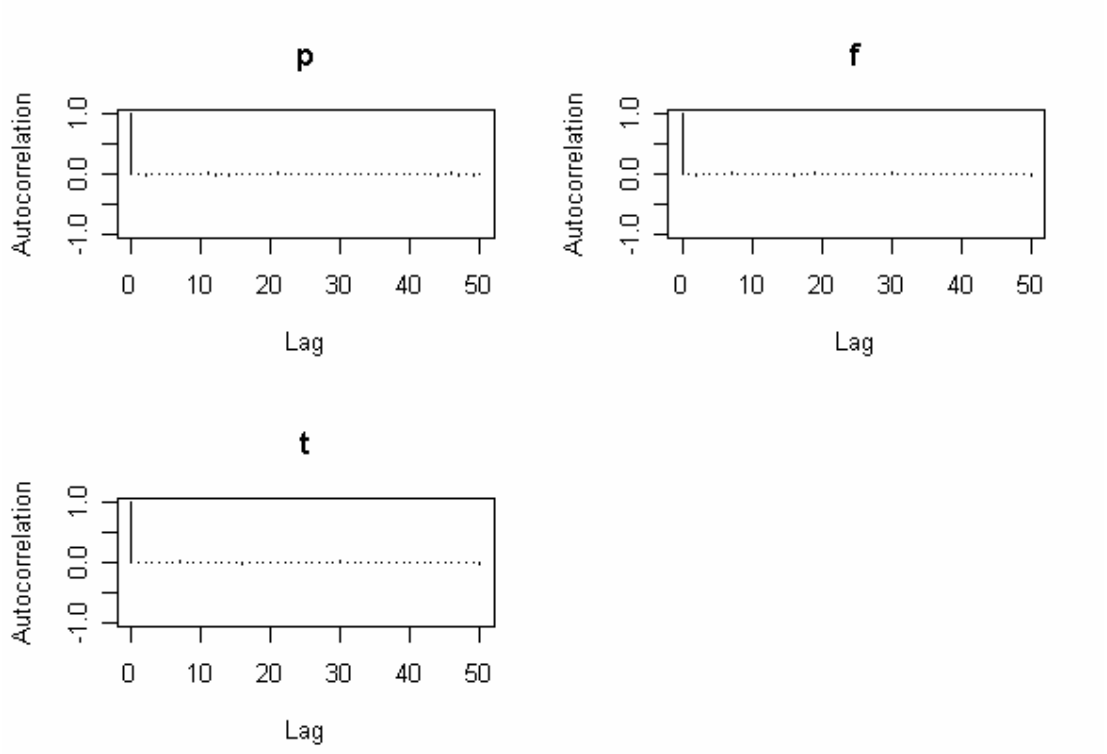

Figura 7.22: Representação gráfica da função de autocorrelação para os parâmetros estimados para n=50 e p=0,5

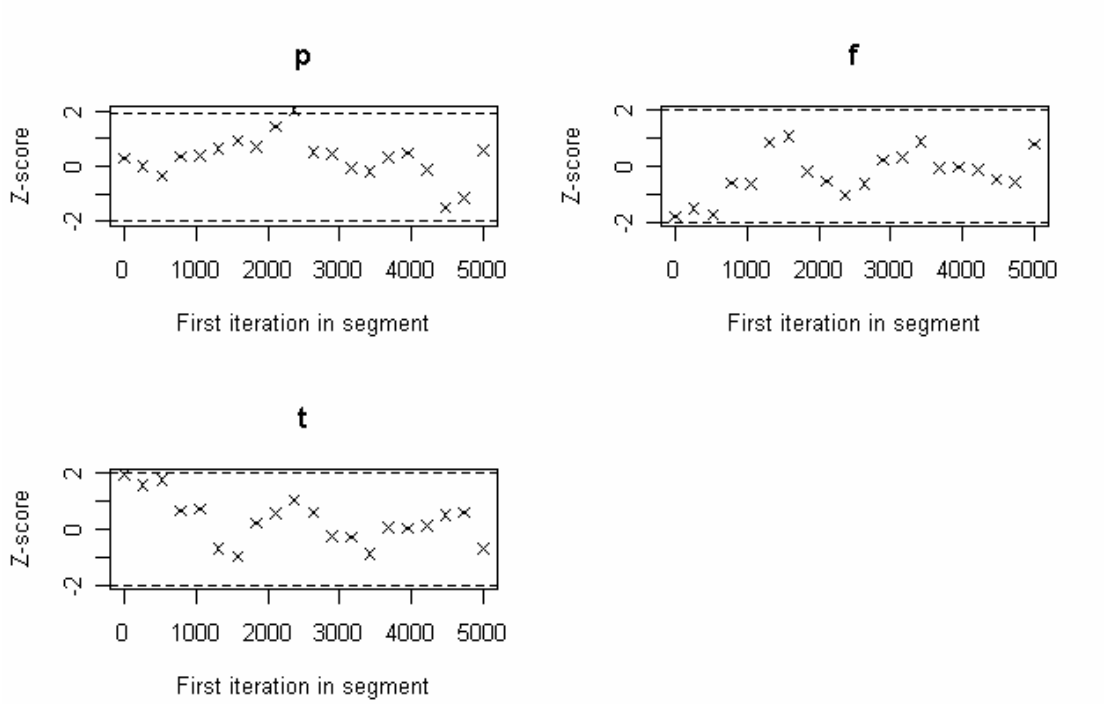

Figura 7.23: Representação gráfica do diagnóstico de convergência Geweke para os parâmetros estimados sendo n=50 e p=0,9

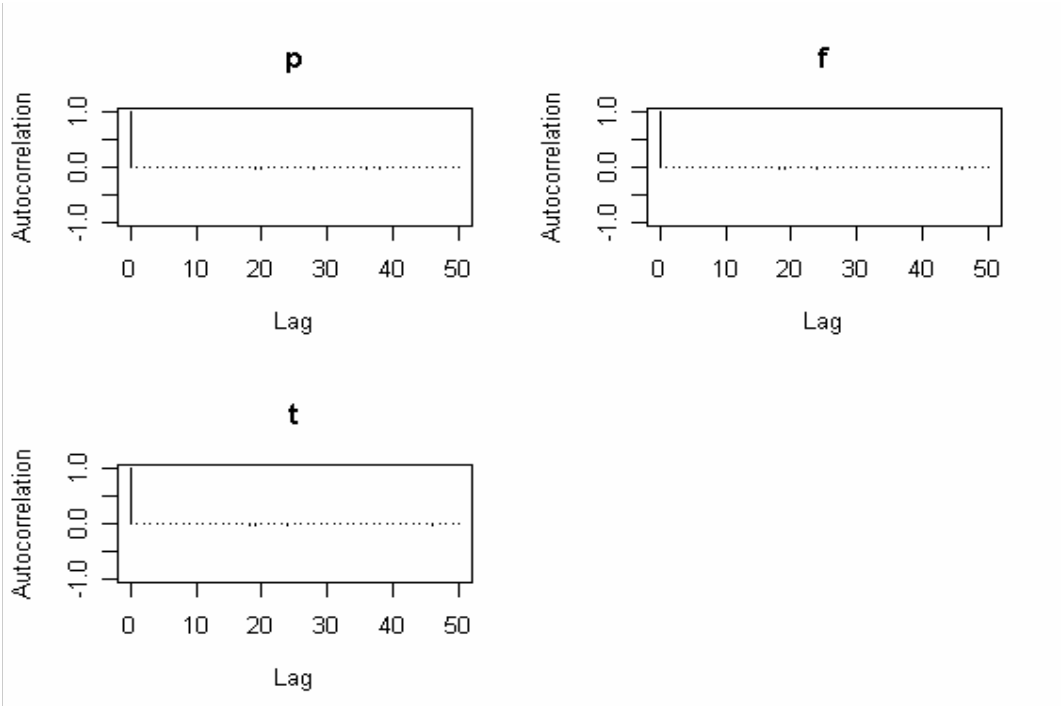

Figura 7.24: Representação gráfica da função de autocorrelação para os parâmetros estimados para n=50 e p=0,9

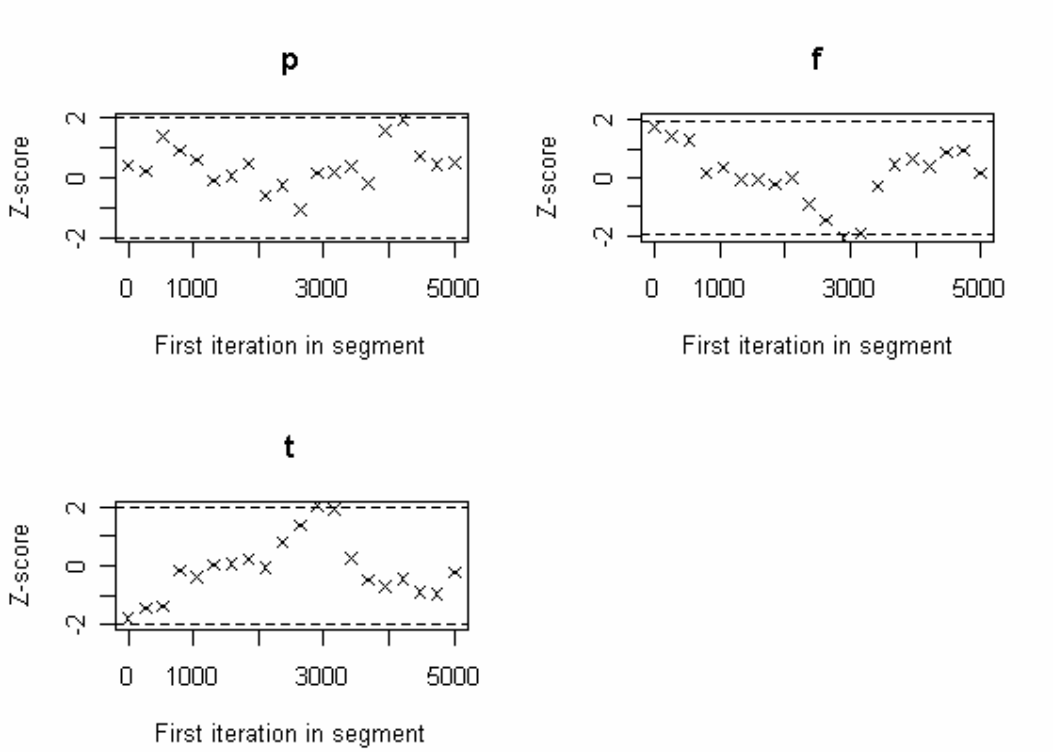

Figura 7.25: Representação gráfica do diagnóstico de convergência Geweke para os parâmetros estimados sendo n=100 e p=0,1

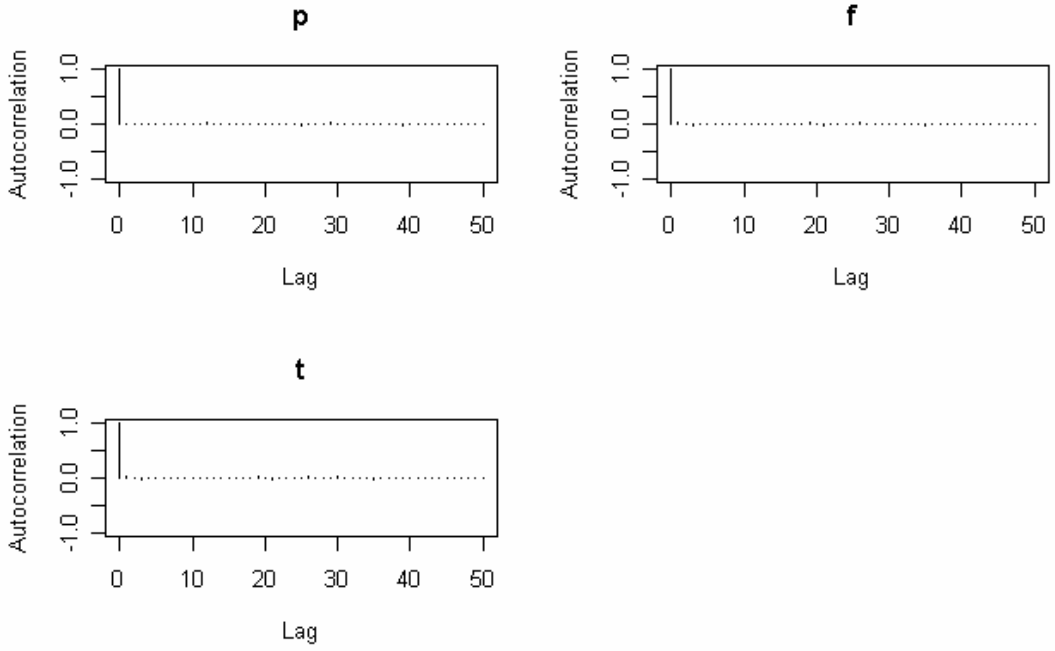

Figura 7.26: Representação gráfica da função de autocorrelação para os parâmetros estimados para n=100 e p=0,1

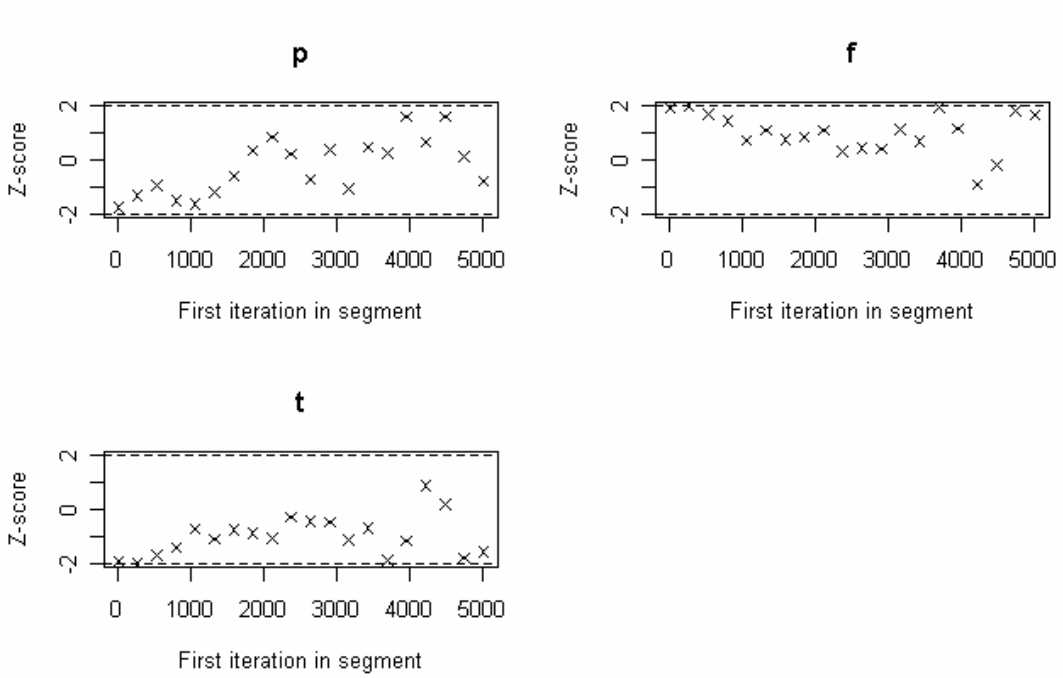

Figura 7.27: Representação gráfica do diagnóstico de convergência Geweke para os parâmetros estimados sendo n=100 e p=0,5

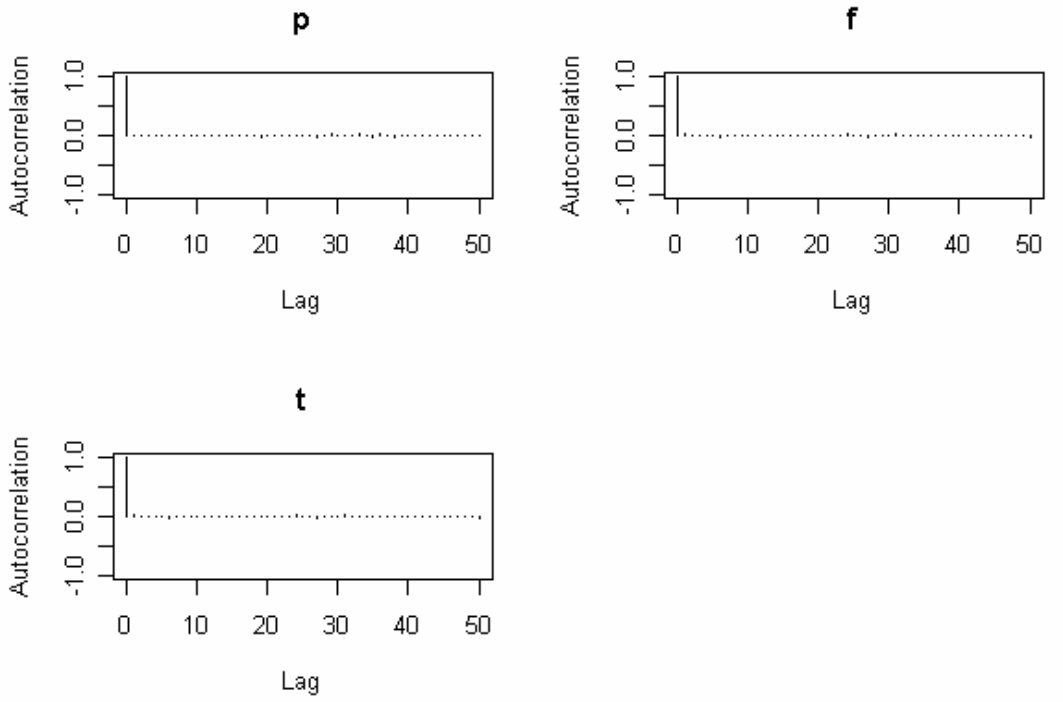

Figura 7.28: Representação gráfica da função de autocorrelação para os parâmetros estimados para n=100 e p=0,5

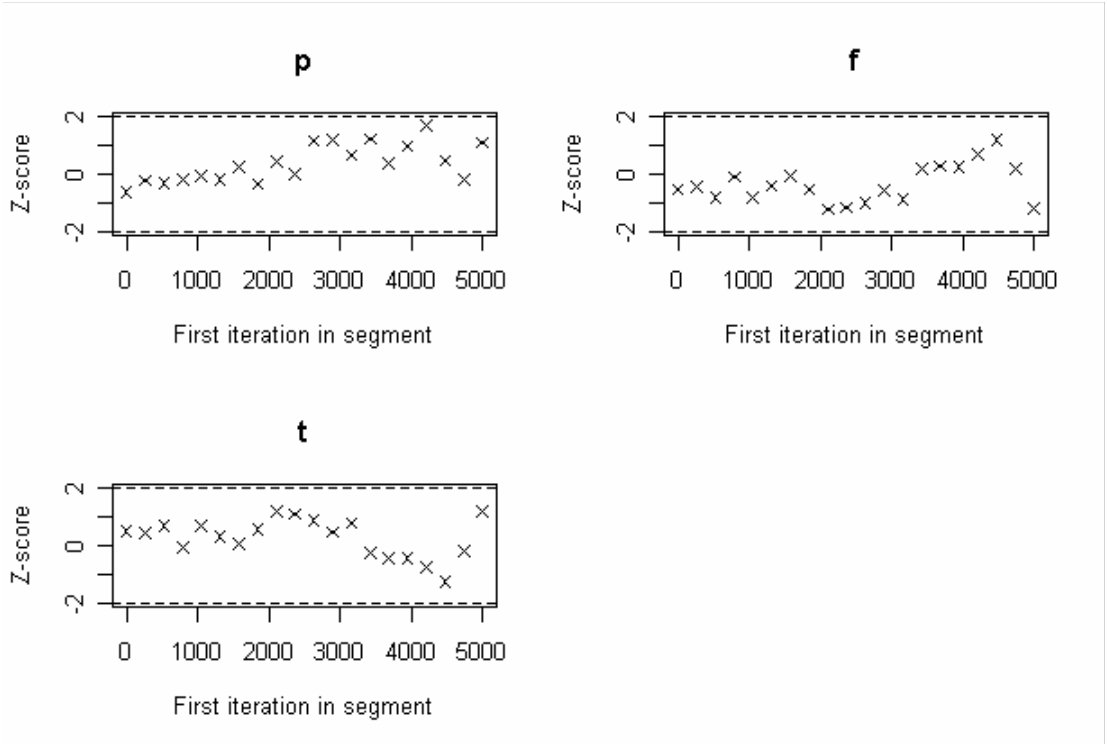

Figura 7.29: Representação gráfica do diagnóstico de convergência Geweke para os parâmetros estimados sendo n=100 e p=0,9

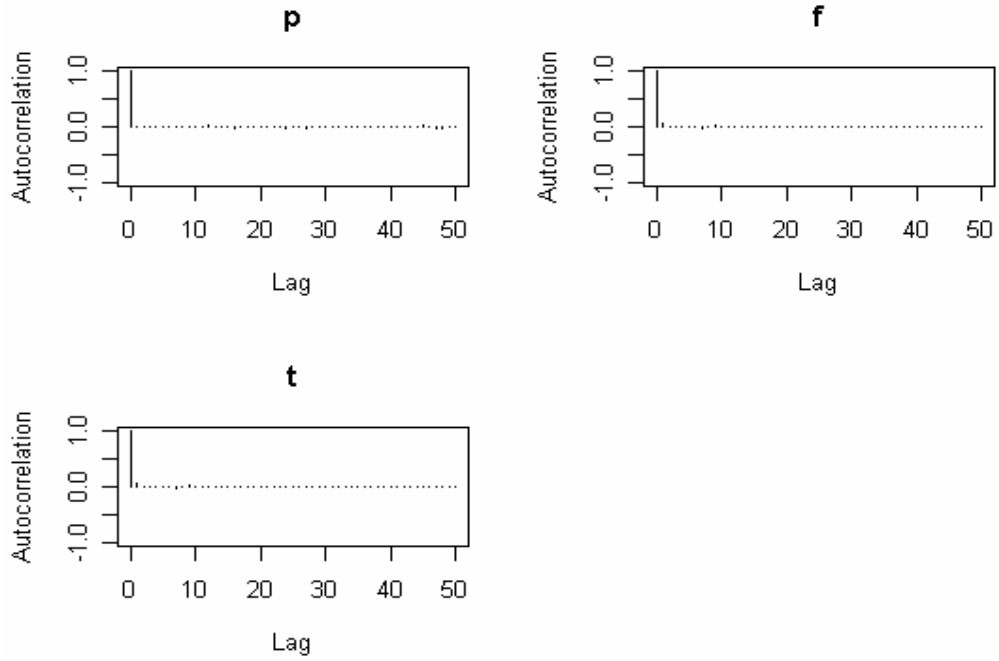

Figura 7.30: Representação gráfica da função de autocorrelação para os parâmetros estimados para n=100 e p=0,9

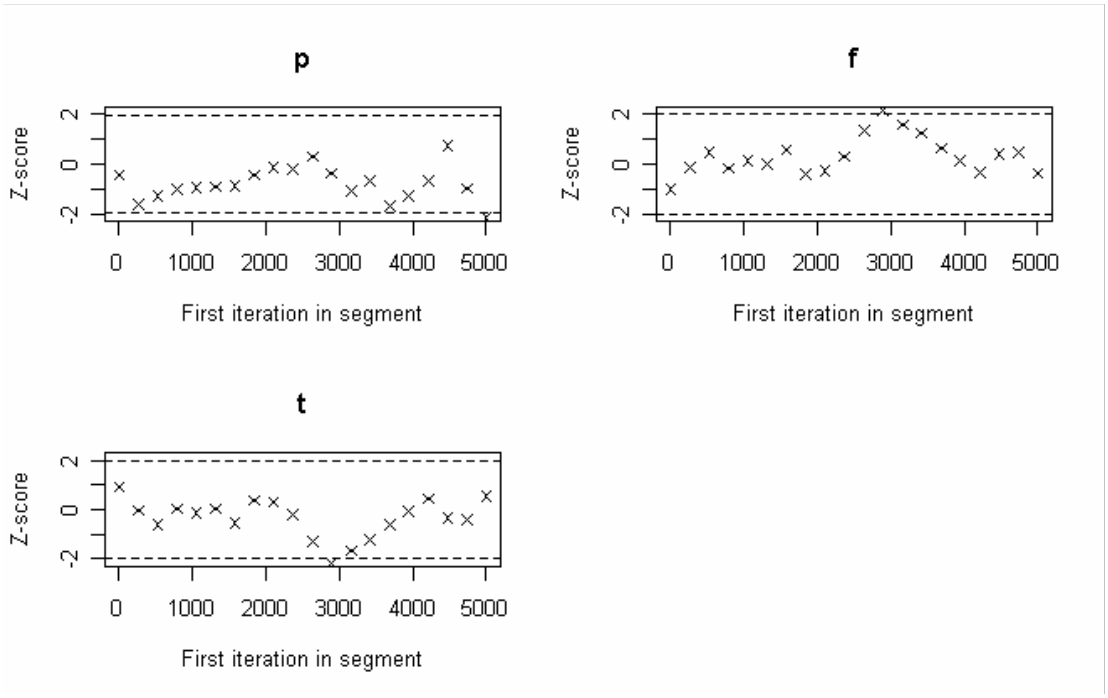

Figura 7.31: Representação gráfica do diagnóstico de convergência Geweke para os parâmetros estimados sendo n=200 e p=0,1

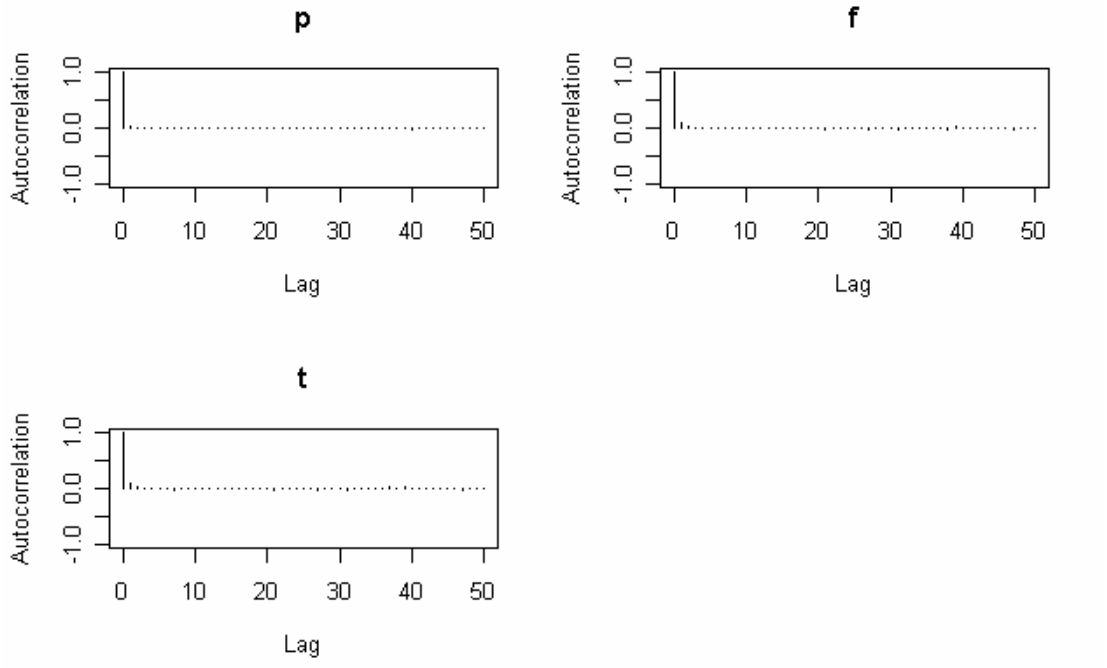

Figura 7.32 : Representação gráfica da função de autocorrelação para os parâmetros estimados para n=200 e p=0,1

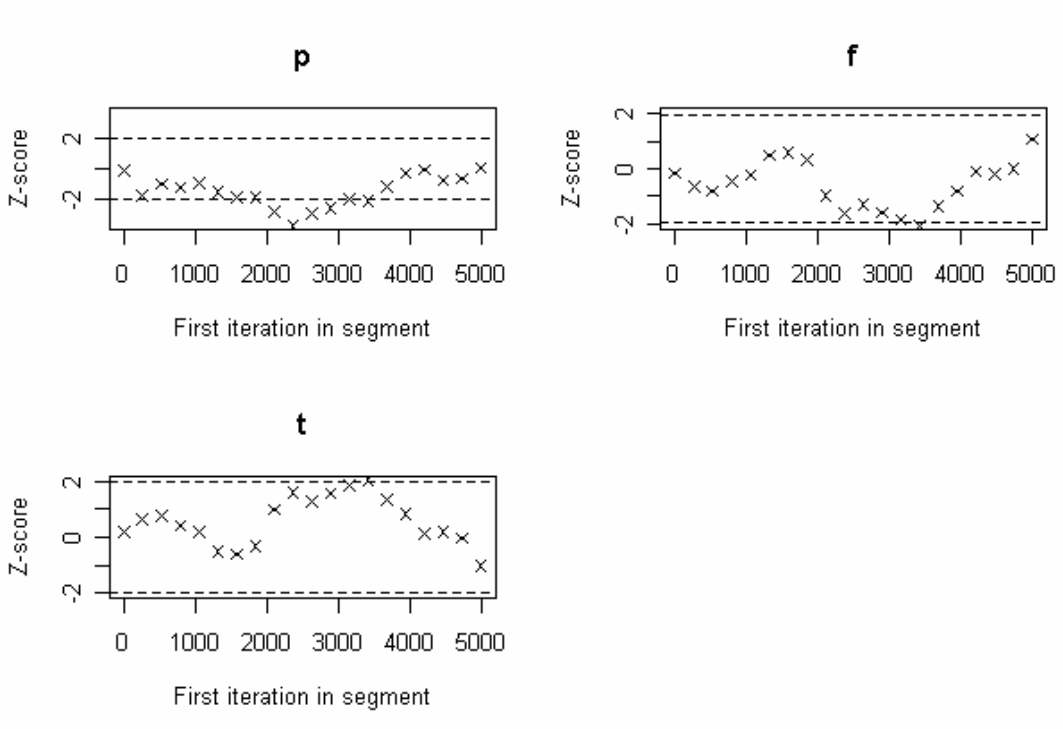

Figura 7.33: Representação gráfica do diagnóstico de convergência Geweke para os parâmetros estimados sendo n=200 e p=0,5

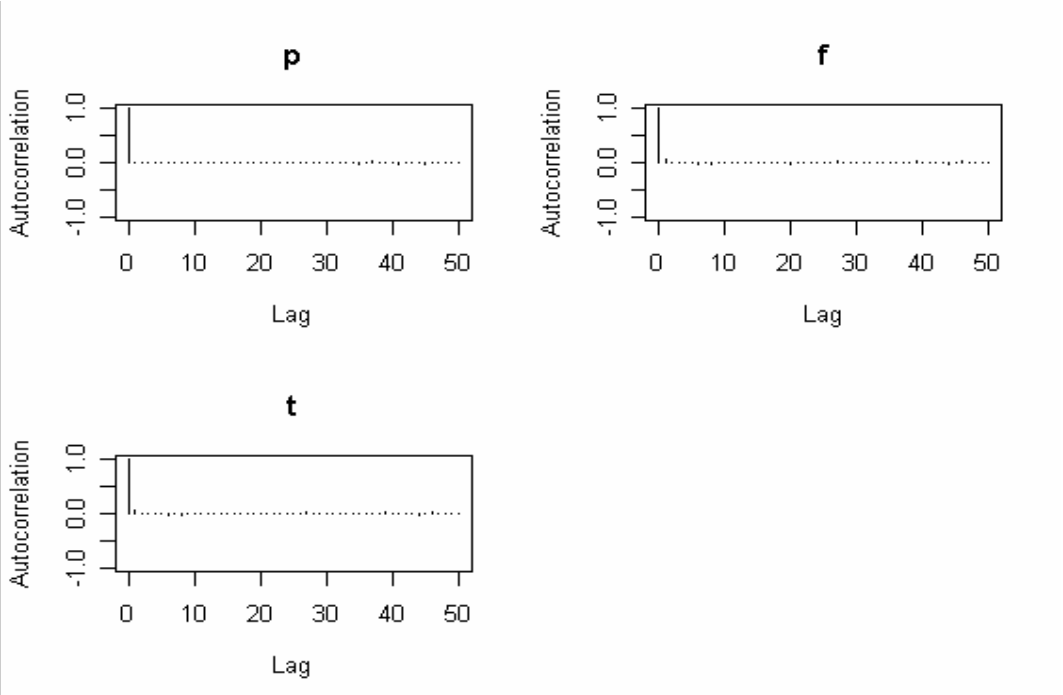

Figura 7.34: Representação gráfica da função de autocorrelação para os parâmetros estimados para n=200 e p=0,5

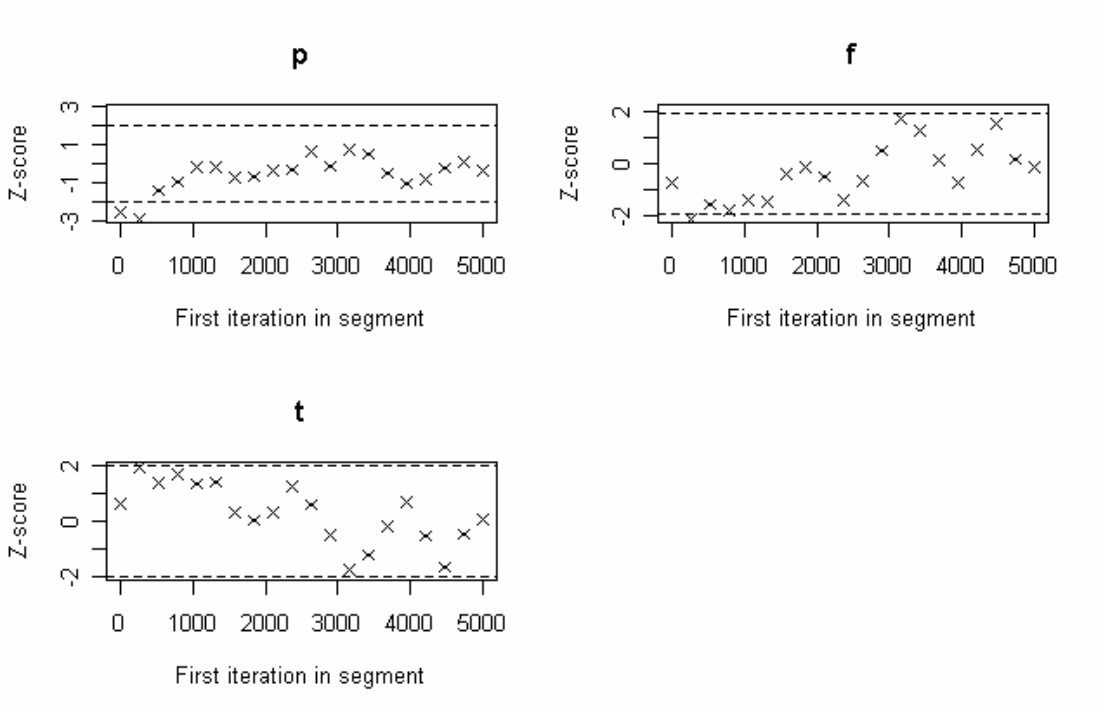

Figura 7.35: Representação gráfica do diagnóstico de convergência Geweke para os parâmetros estimados sendo n=200 e p=0,9

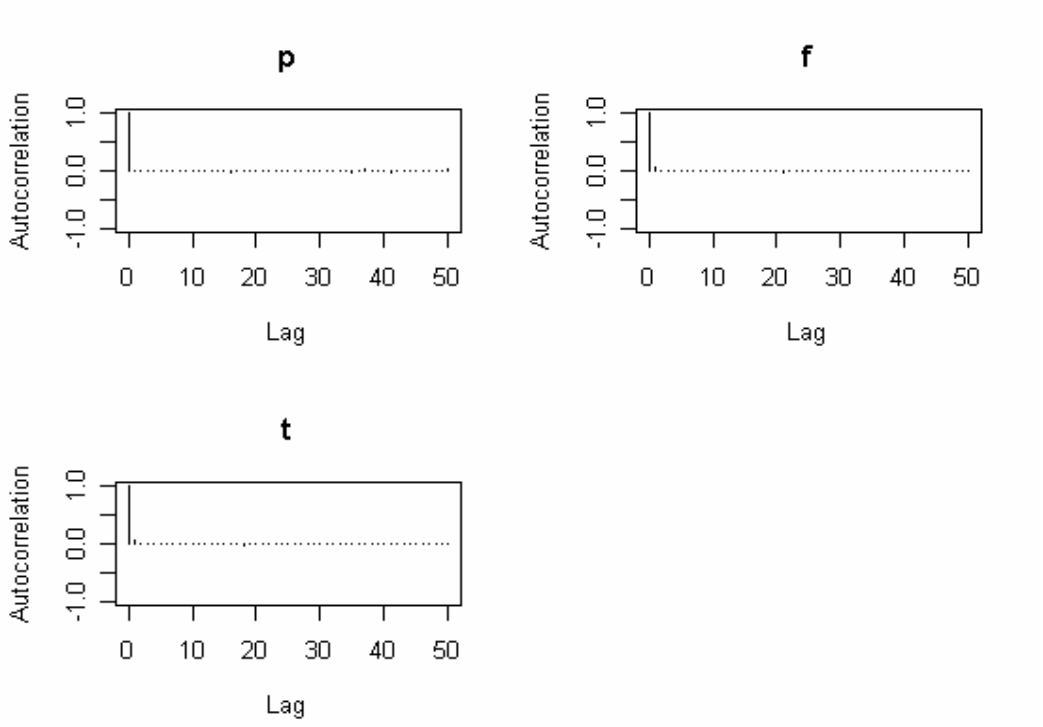

Figura 7.36: Representação gráfica da função de autocorrelação para os parâmetros estimados para n=200 e p=0,9

8. CONCLUSÕES

Nas condições do presente trabalho, e de acordo com os resultados obtidos, pode se concluir que os valores estimados para o coeficiente de endogamia e para a taxa de fecundação cruzada representaram bem uma população real através da utilização da simulação de dados.

A utilização de freqüências alélicas se tornou viável nesta estimação, visto que a informação extraída daí permite representar bem a estrutura genética de populações.

Podemos verificar também que todo o processo de inferência bayesiana condicionado através do modelo de COCKERHAM foi eficiente, pois propiciou resultados condizentes e que não houve muita diferença nas estimativas com o aumento do número de indivíduos.

A análise Bayesiana propicia resultados adicionais àqueles obtidos pela abordagem freqüentista, destacando-se os intervalos de confiança Bayesianos para as estimativas de parâmetros genéticos.

A facilidade de utilização do Software Livre R propiciou grande facilidade para a execução deste trabalho e mostrou que cada vez mais existe uma tendência para a substituição do software proprietário pelo software livre.

A inferência bayesiana é uma técnica que tem crescido muito nos últimos anos e que se tornará muito mais eficaz com sua crescente utilização na área da genética de populações pelos pesquisadores.

9. TRABALHOS FUTUROS

Com o advento deste trabalho, pretende-se ampliar esta análise com o aumento do número de alelos para análise, considerar coeficiente de endogamia negativo, o qual representa uma ausência de endogamia com um aumento da exogamia, aplicar redes bayesianas para estimação de parâmetros genéticos e em recuperação de informação e criar um pacote de funções e disponibilizá-las para utilização por pesquisadores interessados por essa área.

10. BIBLIOGRAFIA CITADA

ARMBORST, T. **Métodos para Medir o Desequilíbrio de Hardy-Weinberg através de Medidas de Endocruzamento**. Dissertação (Mestrado em Estatística) – Universidade Federal de Minas Gerais, Belo Horizonte, MG, 2005.

ARMBORST, T; SILVA, C.Q. **Métodos para Medir o Desequilíbrio de Hardy-Weinberg através de Medidas de Endocruzamento**. Anais :RBRAS-SEAGRO, Londrina-PR, 2005.

AYRES, K.L.; BALDING, D.J. **Measuring departures from Hardy-Weinberg: a Markov chain Monte Carlo method for estimating the inbreeding coefficient**. Heredity, v.80, p. 769-777, 1998.

AYRES, K.L.; BALDING, D.J. **Measuring gametic disequilibrium from multi-locus data. Genetics**, v.157, p.413-423, 2001.

BALDING, D.J.; NICHOLS, R.A. **Significant genetic correlations among Caucasian at forensic DNA loci**. Heredity, v.78, p. 583-589, 1997.

BESSEGA, C.; FERREYRA, L.; JULIO, N.; MONTOYA, S.; SAIDMAN, B.; VILARDI, J.C. **Mating system parameters in species of genus Prosopis (Leguminosae)**. Hereditas, v.132, n.1, p.19-27, 2000.

BRIAN J. SMITH . **boa: Bayesian Output Analysis Program (BOA) for MCMC**. R package version 1.1.5-2, http://www.public-health.uiowa.edu/boa, 2005.

BOX, G. E. P.; TIAO, G.C. **Bayesian inference in statistical analysis**. New York: Jonh Wiley, 1973.

BROEMELING, L. D. **Bayesian analysis of linear models**. New York: M. Dekker, 1985. 454p.

COCKERHAM, C.C.; WEIR, B.S. **Variance of actual inbreeding.** Theoretical Population Biology, New York, v.23, p.85-109, 1983.

D. MARTIN, A.; M. QUINN, K. **MCMCpack: Markov chain Monte Carlo (MCMC) Package**. R package version 0.7-1. http://mcmcpack.wustl.edu, 2006.

COCKERHAM, C.C. **Variance of gene frequencies**. Evolution, Lancaster, v. 23, p. 72-84, 1969.

FALCONER, D.S**. Introduction to quantitative genetics**. New York, The Ronald Press Co, 365 p., 1964.

FISHER, R.A. **The theory of inbreeding**. Edinburg: Oliver and Boyd, 120p., 1949.

FYFE, J.L.; BAILEY, N.T.J. **Plant breeding studies in leguminous forage crops. I. Naturalcross-breeding in winter beans**. Journal of Agricultural Science, v.41, p.371-378, 1951.

GAMERMAN, D. **Simulação estocástica via cadeias de Markov**. XII SINAPE -ABE, Caxambu-MG, 196p, 1996.

GELFAND, A.E.; HILLS, S. E.; RACINE-POON, A.; SMITH, A.E.M. **Illustration of Bayesian inference in normal data models using Gibbs sampling**. Journal of the American Statistical Association, Alexandria, v.85, n.410, p.972-985, June 1990.

GELMAN, A.; RUBIN, D.B**. Inference from iterative simulation using multiple sequences**. Statistical Science, 7, 457 − 72, 1992.

GELMAN, A.; CARLIN, J.B.; STERN, H.S.; RUBIN, D.B. **Bayesian data analysis.** London: Chapman Hall, 526 p., 1997.

GEMAN, S.; GEMAN, D. **Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images**. IEEE Transactions on Pattern Analysis and Machine Intelligence, Washington, v.6, p. 721-741, 1984.

GEWEKE, J. **Evaluating the accuracy of sampling-based approaches to calcualting posterior moments**. Bayesian Statistics, 4, 1992.

HARTL, D.L.; CLARK, A. G. **Principles of population genetics**. Sunderland, Sinauer Associates Inc. Publishers, 468p., 1989.

HEIDELBERG, P.; WELCH, P. **Simulation run lenght control in the presence of an initial transient**. Operations Research, 31, 1109 – 44, 1983.

HOLSINGER, K.E.; WALLACE, L.E. **Bayesian approaches for the analysis of populatioa genetic structure: an example from Platanthera leucophaea (Orchidaceae)**. Molecular Ecology, v.13, n.4, p.887-896, 2004.

HOLSINGER, K.E. **Bayesian population genetic data analysis**. Department of Ecology & Evolutionary Biology, University of Connecticut, 2005.

JEFFREYS, H. **The times of P, S, and SKS, and the velocities Of P and S**; Monthly Notices R. A. S. Geophys. Suppl. p.498 – 533, 1939.

JEFFREYS, H. **Theory of Probability**. Oxford: Claredon Press, 1961. 325 p.

KARHU, A. **Estimation of inbreeding in radiata pine populations using microsatellites**. http://herkules.oulu.fi/isbn9514259246, University of Oulu, Finland, 2001.

LEUTENEGGER, A.L.; PRUM, B.; GENIN, E.; VERNY, C.; LEMAINQUE, A.; CLERGET-DARPOUX, F.; THOMPSON, E.A. **Estimation of inbreeding coefficient trough use of genomic data**. Am. J. Hum. Genet., v. 73, p.516-523, 2003.

McGUIRE, G.; DENHAM, M.C.; BALDING, D.J. **Models of evolution fro DNA sequences including gaps**. Molecular Biology and Evolution, v.18, p.481-490, 2001.

MILLAR, M.A.; BYRNE, M.; COATES, D.J.; STUKELY, M.J.C.; McCOMB, J.A. **Mating system studies in jarrah, Eucalyptus marginata (Myrtaceae).** Australian Journal of Botany, v.48, n.4, p.475-479, 2000.

MUNIZ, J.A.; VEIGA, R.D. **Avaliação das propriedades do estimador de freqüências alélicas em populações diplóides endogâmicas através de simulação**. Ciência e Agrotecnologia, Lavras, v. 20, n. 2, p. 143-150, 1996.

MUNIZ, J.A.; VENCOVSKY, R.; BARBIN, D. **A variância do estimador do coeficiente de endogamia obtido pelo método dos momentos em uma população diplóide**. Revista de Matemática e Estatística, São Paulo, v. 15, p.131-143, 1997.

MUNIZ, J.A.; VENCOVSKY, R.; BARBIN, D. **Estimação do coeficiente de endogamia através do método dos momentos em uma população diplóide com alelos múltiplos**. Ciência e Agrotecnologia, Lavras, v. 21, n. 2, p. 150-159, 1997.

MUNIZ, J.A.; BARBIN, D.; VENCOVSKY, R. **Properties of estimators of the inbreeding coefficient and the rate of cross fertilization obtained from gene frequency data in a diploid population**. Brazilian Journal of Genetics, Ribeirão Preto, v. 19, n. 3, p. 485-491, 1996.

MUNIZ, J.A.; BARBIN, D.; VENCOVSKY, R.; VEIGA, R.D. **Teste de hipótese sobre o coeficiente de endogamia de uma população diplóide**. Ciência e Agrotecnologia, Lavras, v.23, n.2, p. 410-420, abr./jun., 1999.

MUNIZ, J.A.; ITO, S.C.S.; VEIGA, R.D.; FERREIRA, D.F. (b) **Propriedades do estimador do coeficiente de coancestria com dados de frequências alélicas em populações haplóides**. Ciência e Agrotecnologia, Lavras, v.25, n.6, p.1396-1405, nov./dez.,2001.

MUNIZ, J.A.; ITO, S.C.S.; FERREIRA, D.F.; VEIGA, R.D. (a) **Teste de hipótese sobre o coeficiente de coancestria de populações haplóides**. Pesquisa Agropecuária Brasileira, Brasília, v.36, n.1, p.15-25, jan., 2001.

NEI, M.; CHESSER, R.K. **Estimation of fixation indices and gene diversities**. Ann. Hum. Genet., 47, 253 – 259, 1983.

PAULINO, C.D.; TURKMAN, M. A. A.; MURTEIRA,B. **Estatística Bayesiana**. Lisboa, 429p., 2003.

R DEVELOPMENT CORE TEAM . **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, 2005, URL http://www.R-project.org.

RAFTERY, A.L.; LEWIS, S. **How many iterations in the Gibbs sampler? Bayesian Statistics**, 4, 763 – 74, 1992B.

RAMALHO, M.A.P.; SANTOS, J.B. DOS; PINTO, C.A.B. **Genética na agropecuária**. 3ª edição revisada. UFLA, Lavras; 472p., 2004.

RITLAND, K.; JAIN, S.K. **A model for the estimation of outcrossing rate and gene frequencies using n independent loci**. Heredity, v.47, p.35-52, 1981.

RAUFASTE, N.; BONHOMME, F. **Properties of bias and variance of two multiallelic estimators of F_{ST}**. Theoritical Population Biology, New York, v.57, p.285-296, 2000.

REYNOLDS, J.; WEIR, B.S.; COCKERHAM, C.C. **Estimator of the coancestry coefficient: basis for short -term genetic distance**. Genetics, Baltimore, v. 105, p. 767-779, 1983.

ROBERTSON, A.; HILL, W.G. **Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients**. Genetics, 107, 703-718, 1984.

ROSA, G.J.M. **Análise Bayesiana de modelos lineares mistos robustos via Amostrador de Gibbs**. p.57. Tese (Doutorado em Estatística e Experimentação Agronômica) – Universidade de São Paulo, Piracicaba, SP, 1998.

SORENSEN, D. **Gibbs Sampling in quantitative genetic**. Copenhagen: Foulun, 186 p., 1996.

SORIA, F.; BASURCO, F.; TOVAL, G.; SILIÓ, L.; RODRIGUEZ, M.C.; TORO, M. **Na aplication of Bayesian techniques to the genetic evaluation of growth traits in Eucalyptus globulus.** Can. J. For. Res., v.28, n.9, p.1286-1294, 1998.

VENCOVSKY, R. **Análise de variância de freqüências alélicas**. Revista Brasileira de Genética. Suplemento 1, Ribeirão Preto, v. 15, n. 1, p. 53-60, 1992 (Suplemento).

WEIR, B.S. **Genetic Data Analysis II**. Methods for Discrete Population Genetic Data. Sinauer Associates, Inc. Sunderland, 445 p., 1996.

WEIR, B.S.; COCKERHAM, C.C. **Estimating F-statistics for the analysis if population structure**. Evolution, Lancaster, v. 38, n. 6, p. 1358-1370, 1984.

WILSON, G.A.; RANNALA, B. **Bayesian inference of recent migration rates usinga multilocus genotypes**. Genetics, v.163, p.1177-1191, 2003.

WILSON, I.J.; BALDING, D.J. **Genealogical inference from microsatellite data.** Genetics, v.150, p.499-510, 1998.

WRIGHT, S. **System of mating**. Genetics, Baltimore, v.6, p.111-178, 1921.

WRIGHT, S. **The interpretation of population structure by F-statistics with special regard to system of mating**. Evolution, Lancaster, v. 19, p. 395-420, 1965.

APÊNDICES

Abaixo está a relação de implementação das funções:

Apêndice A: Funções Implementadas

```
#-------------------------------------------------------------------------------------------------------
# Função: Geração das amostras de indivíduos de uma população
# Para executar essa função, precisa-se definir número de indivíduos(n) e
# um valor de probabilidade(p) na função logo abaixo. Exemplo:
# Ind <- IndAmostra(10,0.9)

IndAmostra = function (n,p) {

# Amostra de indivíduos de uma população
# n <- número de indivíduos amostrados
# nAlelo <- número de alelos amostrados
# p <- um valor de probabilidade
# y <- matriz de alelos

nAlelo <- 2

y <- matrix(0,n,2)

 for (i in 1:n) {

   for (j in 1:nAlelo) {

     y[i,j] <- rbinom(1,1,p)

   }
 }

# pA <- Proporção do alelo A
# Geração do valor inicial de p

pA <- sum(y)/length(y)

return(list(y=y,pA=pA))
}

# Execução da Função

Ind <- IndAmostra(10,1)


#-------------------------------------------------------------------------------------------------------
```

```
#-----------------------------------------------------------------------------------------------------------
# Função: Geração valores iniciais dos parâmetros

Inicial = function(y) {

# Tamanho da amostra
# Somatório de alelos A na amostra
# Auxiliar na soma de quadrados de alelos A
# Matriz média de tratamento de cada indivíduo

n <- length(y)/2
somA <- sum(y)
somquaA <- c(n,1)
mTrat <- matrix(0,1,n)

# Cálculo da soma de quadrados de indivíduos

 for (i in 1:n) {

   somquaA[i] <- sum(y[i,])^2

 }

# Fórmulas Auxiliares da Estatística Experimental
# Soma de quadrados de indivíduos
# Soma de quadrados de genes dentro de indivíduos

aux2SQ <- 0.5*sum(somquaA)
aux0SQ <- sum(y)
auxSQ <- (1/(2*n))*(aux0SQ^2)

# Cálculo das médias de tratamento dos indivíduos

 for (i in 1:n) {

   mTrat[1,i] <- (sum(y[i,])/2) - (aux0SQ/(n*2))

 }

# Soma dos quadrados dos indivíduos
# Soma dos quadrados dos alelos nos indivíduos
# Quadrado médio dos indivíduos
# Quadrado médio dos alelos indivíduos

sqInd <- aux2SQ - auxSQ
sqGen <- aux0SQ - aux2SQ
qmInd <- sqInd/(n-1)
qmGen <- sqGen/n
```

```
# Cálculo do Coeficiente de Endogamia (F)
# Cálculo da Taxa de Fecundação Cruzada (Tfc)

 if (qmInd+qmGen == 0) {

   F <- 1

 } else {

   F <-  (qmInd-qmGen)/(qmInd+qmGen)

 }

 if (qmInd<qmGen) {

   F <- 0

 }

Tfc <- (1-F)/(1+F)

return(list(n=n, qmInd=qmInd, qmGen=qmGen, F=F, Tfc=Tfc, mTrat=mTrat))

}

# Execução da função

Inicial <- Inicial(y)

#-----------------------------------------------------------------------------------------------------
# Função: Controle dos parâmetros da distribuição Beta e Gama inversa

Parametros = function(y,n,pA,qmInd,qmGen) {

# Distribuição Beta
# parâmetros Alfa e beta

Alfa <- -pA+10000*(pA^2)-10000*(pA^3)
Beta <- ((Alfa-pA*Alfa)/pA) -2*(n-sum(y))

# Controle para valores de pA iguais a 0 ou 1

 if (pA ==0) {

   Alfa <- 0.00001
   Beta <- 100

 }
```

```
 if (pA ==1) {

   Alfa <- 100
   Beta <- 0.00001+2*n

 }

# Distribuição Gama Inversa
# parâmetros variância a e variância g

# Parâmetros a1 e b1 da variância de a
# Algoritmo da solução equação 3º grau

A <- -10000*(qmInd^2)-4
B <- 20000*(qmInd^2)+3
C <- -10000*(qmInd^2)+2
p <- B-A*A/3
q <- C-A*B/3+2*A*A*A/27
D <- q*q/4+p*p*p/27

 if (D<0) {

   M <- sqrt(-D)
   r <- sqrt(q*q/4+M*M)
   t <- acos(-q/2.0/r)
   r1 <- 2*(r^(1/3))*cos(t/3)-A/3
   r2 <- 2*(r^(1/3))*cos((t+2*pi)/3)-A/3
   r3 <- 2*(r^(1/3))*cos((t+4*pi)/3)-A/3

 }

s1 <- qmInd*(r1-1)
s2 <- qmInd*(r2-1)
s3 <- qmInd*(r3-1)

# Definição dos valores de a1 e b1

 if ((r1>r2) && (r1>r3)) {

   a1 <- r1-(n/2)
   b1 <- s1

 }

 if ((r2>r1) && (r2>r3)) {

   a1 <- r2-(n/2)
   b1 <- s2

 }
```

```
  if ((r3>r2) && (r3>r1)) {

    a1 <- r3-(n/2)
    b1 <- s3

  }

# Parâmetros a2 e b2 da variância de g
# Algoritmo da solução equação 3º grau

A <- -10000*(qmGen^2)-4
B <- 20000*(qmGen^2)+3
C <- -10000*(qmGen^2)+2
p <- B-A*A/3
q <- C-A*B/3+2*A*A*A/27
D <- q*q/4+p*p*p/27

  if (D<0) {

    M <- sqrt(-D)
    r <- sqrt(q*q/4+M*M)
    t <- acos(-q/2.0/r)
    r1 <- 2*(r^(1/3))*cos(t/3)-A/3
    r2 <- 2*(r^(1/3))*cos((t+2*pi)/3)-A/3
    r3 <- 2*(r^(1/3))*cos((t+4*pi)/3)-A/3

  }

s1=qmGen*(r1-1)
s2=qmGen*(r2-1)
s3=qmGen*(r3-1)


# Definição dos valores de a2 e b2

  if ((r1>r2) && (r1>r3)) {

    a2 <- r1+sum(y)
    b2 <- s1

  }

  if ((r2>r1) && (r2>r3)) {

    a2 <- r2+sum(y)
    b2 <- s2

  }
```

```
  if ((r3>r2) && (r3>r1)) {

    a2 <- r3+sum(y)
    b2 <- s3

  }

# Controle para valores de qmInd e qmGen iguais a 0

  if (qmInd==0) {

    b1 <- 0.00001

  }

  if (qmGen==0) {

    b2 <- 0.00001
    a2 <- a1+2*n

  }

return(list(Alfa=Alfa, Beta=Beta, a1=a1, b1=b1, a2=a2, b2=b2))

}

# Execução da função

Par <- Parametros(y,n,pA,qmInd,qmGen)

#-----------------------------------------------------------------------------------------------------------
# Função: Algoritmo Gibbs Sampling

Gibbs = function(Alfa,Beta,a1,a2,b1,b2,qmInd,qmGen,mTrat,F,Tfc,y,niter,nburn=n/2) {

# Tamanho da amostra

N  <- length(y)/2

# Parâmetros a serem estimados

p <- matrix(0, nrow=niter)
f <-  matrix(0, nrow=niter)
t <- matrix(0, nrow=niter)
ai <- matrix(0, nrow=niter, ncol=N)
vara <- matrix(0, nrow=niter)
varg <- matrix(0, nrow=niter)
# Cálculos para reparametrização
```

```
somA <- sum(y)
Beta <- (2*(N-somA)) + Beta
a2 <- - somA + a2
a1 <- N/2 + a1

# Valores iniciais dos parametros
# Frequencia do alelo A
# Média de tratamento de indivíduo
# Variância entre indivíduos
# Variância dentro Indivíduos
# Coeficiente de endogamia
# Taxa de fecundação cruzada

p[1] <- pA
ai[1,] <- mTrat[1,]
vara[1] <- qmInd
varg[1] <- qmGen
f[1] <- F
t[1] <- Tfc

# Atualização dos parâmetros

 for (i in 2:niter) {

   p [i] <- rbeta(1, Alfa, Beta )
   sumAi <- 0

    for (l in 1:N) {

      ai [i,l] <- rnorm(1,0, vara[i-1])
      sumAi <- sumAi + (ai[i,l]^2)

    }

# reparametrização

b11 <- sumAi/2 + b1


vara[i] <-  rinvgamma(1, a1, b11)
varg [i] <-  rinvgamma(1, a2, b2)

# Controle para valores positivos de f

 if (vara[i]<varg [i]) {

  f[i] <- rbeta(1,0.1,100)

 } else {
```

```
   f[i] <- (vara[i]-varg[i])/(vara[i]+varg[i])

  }

  if (qmInd+qmGen == 0) {

   f[i] <- rbeta(1,100,0.1)

  }


t[i] <- (1-f[i])/(1+f[i])


  }

return(list(p=p,ai=ai, vara=vara, varg=varg,f=f,t=t))

}
#-------------------------------------------------------------------------------------------------------------
# Argumentos da função Gibbs

niter <- 10000
nburn <- 2000
par(mfrow = c(3,2))
r <- (nburn+1):niter
x <- 'iteração'

#-------------------------------------------------------------------------------------------------------------
# Execução da função

q <- Gibbs(Alfa,Beta,a1,a2,b1,b2,qmInd,qmGen,mTrat,F,Tfc,y,niter,nburn)

#-------------------------------------------------------------------------------------------------------------
```

Apêndice B: Gráficos e Análises Estatísticas

```
#-------------------------------------------------------------------------------------------------------
# Acessorios Graficos

plot(q$p[r], type='l', ylab=expression(p), xlab=x)
plot(density(q$p), main="", ylab="densidade")
plot(q$vara[r], type='l', ylab=expression(vara), xlab=x)
plot(density(q$vara), main="", ylab="densidade")
plot(q$varg[r], type='l', ylab=expression(varg), xlab=x)
plot(density(q$varg), main="", ylab="densidade")
plot(q$f[r], type='l', ylab=expression(f), xlab=x)
plot(density(q$f), main="", ylab="densidade")
plot(q$t[r], type='l', ylab=expression(t), xlab=x)
plot(density(q$t), main="", ylab="densidade")

# Gráficos das médias

means = cumsum(q$p[r])/(1:(niter-nburn))
plot(means,type='l',ylab=expression(bar(p)),xlab=x)

means = cumsum(q$vara[r])/(1:(niter-nburn))
plot(means,type='l',ylab=expression(bar(vara)),xlab=x)

means = cumsum(q$varg[r])/(1:(niter-nburn))
plot(means,type='l',ylab=expression(bar(varg)),xlab=x)

means = cumsum(q$f[r])/(1:(niter-nburn))
plot(means,type='l',ylab=expression(bar(f)),xlab=x)

means = cumsum(q$t[r])/(1:(niter-nburn))
plot(means,type='l',ylab=expression(bar(t)),xlab=x)

# Média e desvio-padrão

cat(mean(q$p[r]),sd(q$p[r]),"\n")

cat(mean(q$vara[r]),sd(q$vara[r]),"\n")

cat(mean(q$varg[r]),sd(q$varg[r]),"\n")

cat(mean(q$f[r]),sd(q$f[r]),"\n")

cat(mean(q$t[r]),sd(q$t[r]),"\n")

#-------------------------------------------------------------------------------------------------------
```

Apêndice C: Análise de Convergência

```
#-------------------------------------------------------------------------------------------------------
# Teste de Convergência e gráficos de convergência
# Intervalo de credibilidade

# Cria a matriz de análise

Conv <- matrix(0,niter,5,dimnames=list(c(1:niter),c("vara","varg","p","f","t")))
Conv1 <- matrix(0,niter,3,dimnames=list(c(1:niter),c("p","f","t")))

Conv[,1] <- q$vara[,1]
Conv[,2] <- q$varg[,1]
Conv[,3] <- q$p[,1]
Conv[,4] <- q$f[,1]
Conv[,5] <- q$t[,1]
Conv1[,1] <- q$p[,1]
Conv1[,2] <- q$f[,1]
Conv1[,3] <- q$t[,1]


# Função de autocorrelação

boa.acf(Conv,lags = c(0,1,5,10,50))

# Diagnóstico de convergência de Geweke

boa.geweke(Conv, 0.1, 0.5)

# Diagnóstico de convergência de Heidelberger e Welch

boa.handw(Conv, 0.1, 0.05)

# Diagnóstico de convergência de Raftery e Lewis

boa.randl(Conv, 0.025, 0.005, 0.95, 0.001)

Conv1 <- mcmc(Conv1)

# Gráficos do diagnóstico de convergência geweke

geweke.plot(Conv1,0.1, 0.5, 20, 0.05, auto.layout = TRUE, ask = TRUE)

# Gráficos da função de autocorrelação

autocorr.plot(Conv1, 50, auto.layout = TRUE,ask=TRUE)

#-------------------------------------------------------------------------------------------------------
```