MODELOS MISTOS NORMAIS ASSIMÉTRICOS EM DADOS DE MICROARRAYS ORIGINADOS DE PEDIGREES COMPLEXOS

DANIELA CARINE RAMIRES DE OLIVEIRA

2009

DANIELA CARINE RAMIRES DE OLIVEIRA

MODELOS MISTOS NORMAIS ASSIMÉTRICOS EM DADOS DE MICROARRAYS ORIGINADOS DE PEDIGREES COMPLEXOS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para obtenção do título de "Doutor".

Orientador

Prof. Júlio Sílvio de Sousa Bueno Filho

LAVRAS

MINAS GERAIS-BRASIL

2009

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

Oliveira, Daniela Carine Ramires de.
Modelos mistos normais assimétricos em dados de *microarrays* originados de pedigrees complexos / Daniela Carine Ramires de Oliveira. – Lavras : UFLA, 2009.
106 p. : il.

Tese (Doutorado) – Universidade Federal de Lavras, 2009.
Orientador: Júlio Sívio de Sousa Bueno Filho.
Bibliografia.

1. Simulação Monte Carlo via cadeias de Markov. 2. Modelo aditivo-dominante. 3. Distribuição normal assimétrica multivariada. 4. Inferência bayesiana. I. Universidade Federal de Lavras. II. Título.

CDD – 519.282

DANIELA CARINE RAMIRES DE OLIVEIRA

MODELOS MISTOS NORMAIS ASSIMÉTRICOS EM DADOS DE MICROARRAYS ORIGINADOS DE PEDIGREES COMPLEXOS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para obtenção do título de "Doutor".

APROVADA em 17 de agosto de 2009.

Profa. Júlia Maria Pavan Soler	USP
Prof. Daniel Furtado Ferreira	UFLA
Prof. Lucas Monteiro Chaves	UFLA
Profa. Thelma Sáfadi	UFLA

Prof. Júlio Sílvio de Sousa Bueno Filho
UFLA
(Orientador)

LAVRAS
MINAS GERAIS - BRASIL

A Deus,

que me protegeu e deu forças

em todos os momentos,

DEDICO

Agradecimentos

⋆ Ao meu marido Marcos, que teve uma participação muito especial neste trabalho, dando-me força nos momentos mais difíceis, contribuindo com seu conhecimento para a melhora da escrita e da parte computacional, e também em minha vida, transmitindo seu amor, carinho e paciência e cuidando do Lucas.

⋆ Ao meu filho Lucas (um pedacinho de mim), por ter sido paciente em vários momentos que não pude estar com ele.

⋆ Aos meus pais, Odair e Maria Luiza e às minhas irmãs Alessandra, Fabiana e Luciana, que mesmo distantes de mim, me incentivaram, me deram força e torceram por mim.

⋆ Ao professor Júlio Sílvio de Sousa Bueno Filho, pela orientação competente, pela amizade, pela confiança depositada na realização desse trabalho, pelos ensinamentos teóricos, pela descontração, pela ajuda e compreensão que muito contribuíram para a elaboração desse trabalho.

⋆ Ao amigo Fábio Mathias Corrêa, pelo grande apoio computacional.

⋆ Aos professores do IME/USP, pelos valiosos ensinamentos recebidos ao longo das disciplinas de doutorado, em especial, à Profa. Júlia Maria Pavan Soler por me introduzir na linha de pesquisa Genética Quantitativa e aos Professores Carlos Alberto de Bragança Pereira e Márcia D´Elia Branco por me introduzirem nas linhas de pesquisa Bayesiana e Normal Assimétrica, respectivamente.

⋆ Aos professores do DEX/UFLA, em especial, ao Prof. Daniel Furtado Ferreira, à Profa. Thelma Sáfadi e ao Prof. Lucas Monteiro Chaves pelas

valiosas contribuições para minha tese de doutorado e ao Prof. Augusto Ramalho de Morais por transmitir seus valiosos conhecimentos em planejamento de experimentos.

⋆ À Universidade Federal de São João del Rei, ao Departamento de Matemática, Estatística e Ciência da Computação pelo apoio, em especial, aos amigos de departamento Luciane, Andréa, Rejane, Fábio, Waliston, Flaviano, Flávia, Romélia, Toledo, Jorge, Carlos, Marcos e Guilherme, que assistiram meus seminários, e com suas discussões muito contribuíram para a realização dos mesmos e para o aprimoramento deste trabalho.

⋆ Novamente às amigas de departamento Luciane e Andréa, por cederem seus computadores para fazer parte da aplicação computacional do meu trabalho.

⋆ Aos meus amigos de São João del Rei, pelos momentos de descontração, diversão, companhia e força.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

OLIVEIRA, Daniela Carine Ramires de. **Modelos mistos normais assimétricos em dados de microarrays originados de pedigrees complexos**. 2009. 106 p. Tese (Doutorado em Estatística e Experimentação Agropecuária) − Universidade Federal de Lavras, Lavras, MG.*

Estimativas de herdabilidade para a expressão gênica são escassas e, em geral, provenientes de estruturas de famílias, em que para a variabilidade das respostas entre e dentro das famílias assume-se covariância uniforme para os indivíduos relacionados, ignorando o parentesco conhecido entre todos os indivíduos da genealogia. Para tais estimativas usa-se modelos lineares (mistos) Gauss-Markov normais, mas em estudos com microarrays é comum encontrar assimetria de resíduos ao analisar o ajuste de dados previamente normalizados. Isto por si só justificaria o uso de modelos assimétricos. Neste estudo, avaliou-se um delineamento proveniente de uma genealogia com famílias e indivíduos identificados, para os quais se mediu a expressão gênica. Através da genealogia é possível estimar componentes da variância aditivos e dominantes e é razoável assumir que cada um dos efeitos correspondentes nos indivíduos (efeitos aditivos e dominantes) possa apresentar distribuição normal assimétrica. Neste sentido, este trabalho trata do desenvolvimento e implementação computacional do modelo aditivo-dominante normal assimétrico para a análise de microarrays, permitindo assimetria nas distribuições de todos os efeitos aleatórios. Através do método de Monte Carlo via cadeias de Markov, geram-se amostras das condicionais completas a posteriori de todos os parâmetros em modelos com ou sem parâmetros de assimetria, para cada efeito aleatório. Para as inferências, foram calculados os fatores de Bayes, para a seleção dos melhores modelos e intervalos de credibilidade de máxima densidade a posteriori, para a estimação dos parâmetros. Foram apresentados os resultados dos ajustes dos modelos para duas das sondas estudadas. Para estas sondas, houve maior evidência em favor de modelos que consideraram a distribuição normal assimétrica para os efeitos aleatórios. Os modelos aditivos-dominantes normais assimétricos considerados neste trabalho tenderam a confundir as estimativas de componentes da variância e de parâmetros de assimetria, possivelmente devido à estrutura familiar considerada. No entanto, esses modelos são os mais prováveis e têm melhores distribuições para os resíduos do que os modelos simétricos correspondentes.

*Orientador: Júlio Sílvio de Sousa Bueno Filho − UFLA.

ABSTRACT

OLIVEIRA, Daniela Carine Ramires de. **Skew normal mixed models in microarray data generated from complex pedigrees**. 2009. 106 p. Thesis (Doctor in Statistics and Agricultural Experimentation) − Federal University of Lavras, Lavras, MG. *

Estimates of heritability for gene expression are scarce and commonly originated from family structures, in which the variability of responses among and within families are provided under a uniform covariance structure for related individuals, ignoring the known relationship among all individuals in the pedigree. Gauss-Markov normal mixed models are the usual choice for such estimates, but in microarrays studies it is common to find asymmetry in residuals of the adjustment of data previously normalized. This, by itself, justifies the use of skew models. In this study it was analyzed a family based pedigree with gene expression measured by microarrays for all individuals. From this pedigree it is possible to estimate additive and dominance variance components and it is reasonable to assume that each of the corresponding individual effects (additive and dominance effects) may have a skew normal distribution. Thus, this work deals with the development and computational implementation of skew normal additive-dominance model for the analysis of microarrays, that allows skewness in all distributions of random effects. Through the MCMC method, it was generated samples from conditional posteriori distributions for all parameters in models with or without skewness parameters for each random effect. It was calculated the Bayes factors for the selection of the best models and HPD intervals for marginal estimates. Results are shown for two of the analyzed probes. For these probes, there was more evidence in favor of models that considered the skew normal distribution for the random effects. The skew normal additive-dominance models considered in this work tended to confound variance components and skewness parameters estimates, possibly due to pedigree limitations. However, these models are the most probable ones and have better residual behavior than their symmetric counterparts.

*Adviser: Júlio Sílvio de Sousa Bueno Filho − UFLA.

1 INTRODUÇÃO

A concentração relativa de RNA mensageiro de um determinado gene em células de um tecido é, em geral, um indicativo do quanto esse gene está sendo expresso, isto é, do quanto a célula está investindo do seu maquinário bioquímico para produzir a proteína codificada pelo gene. Com isso, pesquisadores de diversas áreas voltaram suas atenções ao desenvolvimento de tecnologias, visando medir tal concentração relativa em diversos tecidos. Uma das principais ferramentas para este tipo de estudo são os microarrays (Saraiva et al., 2007; Speed, 2003).

A tecnologia de microarrays possibilita a avaliação simultânea da expressão de milhares de genes, em diferentes tecidos de um determinado organismo e em diferentes estágios de desenvolvimento ou condições ambientais. Esta tecnologia tem sido largamente utilizada em experimentos de genômica funcional em diversas espécies animais e vegetais. No entanto, os experimentos com microarrays ainda são consideravelmente caros e trabalhosos e, como consequência, são geralmente conduzidos com tamanhos amostrais relativamente pequenos. Tais experimentos envolvem uma série de procedimentos laboratoriais, os quais introduzem diferentes fontes de variação aos dados. Desta maneira, a condução de ensaios com microarrays requer cuidados no delineamento experimental e na análise dos dados (Rosa et al., 2007; Kerr & Churchill, 2001).

Em estudos com microarrays é comum encontrar assimetria e alta variabilidade nos resíduos, ao analisar o ajuste de dados previamente normalizados (Durbin et al., 2002; Ritz & Edén, 2008). Isto por si só justificaria o uso de outros tipos de modelos nos resíduos em tais dados, para capturar e ajustar de maneira mais robusta essas características (assimetria e superdispersão). Além disso, são poucos os delineamentos para microarrays que envolvem famílias e indivíduos e, em geral,

nestes delineamentos prevalece a manifestação do caráter em estudo, o que justificaria o uso de um modelo em que tanto os erros quanto os efeitos genéticos aleatórios tenham uma distribuição mais robusta que a normal.

Na maioria das vezes em que se aplica os modelos lineares mistos para dados de microarrays, trabalha-se com a suposição de que tanto os erros como os efeitos aleatórios do modelo possuem distribuição normal (simétrica). Apesar deste modelo oferecer uma grande flexibilidade para modelar estes efeitos, este sofre da mesma falta de robustez para fugas de normalidade como outros modelos estatísticos baseados na distribuição normal e pode ser demasiadamente restritivo ao fornecer pouca flexibilidade para a representação adequada da estrutura que está presente nos dados. Por exemplo, apesar deste trabalho não ter explorado este tema, Arellano-Valle et al. (2007) apresentam que uma densidade assimétrica estimada para estes tipos de efeitos pode sugerir a exclusão de importantes covariáveis do modelo.

Do ponto de vista prático, o método mais adotado para alcançar a normalidade é a transformação de variáveis que funciona bem em muitos casos. Embora tal método possa dar resultados empíricos razoáveis, deve ser evitado se um modelo mais robusto puder ser encontrado. Azzalini & Capitanio (1999) apresentam algumas razões para evitar esse procedimento:

1. A transformação não fornece informação útil para entender o mecanismo de geração dos dados.

2. A transformação de variáveis dificulta a interpretação, especialmente quando cada variável é transformada usando diferentes funções.

3. A transformação para um conjunto de dados pode frequentemente não ser aplicável para outros conjunto de dados.

4. Quando a suposição de homocedasticidade é necessária, algumas vezes a transformação requerida difere da transformação para alcançar a normalidade.

Assim, considerável esforço tem sido direcionado para relaxar a suposição de normalidade e, conjuntamente, estimar a densidade dos efeitos aleatórios e parâmetros do modelo.

Muitos autores, como Azzalini & Capitanio (1999), Sahu et al. (2003), Cancho et al. (2008), entre outros, estudaram modelos de regressão com distribuições assimétricas. Genton (2004) apresenta a teoria de modelos mistos considerando a distribuição dos efeitos aleatórios pertencente à classe das distribuições elípticas assimétricas. Jara et al. (2008) também apresenta modelos mistos com distribuições elípticas assimétricas, considerando a análise dos parâmetros do modelo no enfoque bayesiano. Arellano-Valle et al. (2007) apresentam uma versão da distribuição normal assimétrica multivariada para ser utilizada na distribuição dos efeitos aleatórios em modelos lineares mistos. Esta distribuição tem como caso particular a distribuição normal multivariada, quando o parâmetro de assimetria for uma matriz composta de zeros. Os autores utilizam a abordagem bayesiana na estimação dos parâmetros do modelo, pois oferece a vantagem de fornecer estimadores e algoritmos mais eficientes computacionalmente para a realização das inferências nos parâmetros do modelo comparado com o uso da abordagem frequentista. Von Rohr & Hoeschele (2002) propuseram o uso de distribuições assimétricas somente para os resíduos em modelos utilizados no contexto de melhoramento animal. Varona et al. (2008) apresentam o uso da distribuição normal assimétrica proposta por Sahu et al. (2003), somente para os resíduos em modelos mistos com efeitos aleatórios aditivos. Leiva et al. (2009) apresentam a distribuição glog-normal, suas propriedades e o seu ajuste em dados de microarrays.

Este trabalho apresenta um modelo clássico da genética quantitativa, conhecido como modelo aditivo-dominante, para ajustar dados de microarrays, oriundo da plataforma Affymetrix, com a seguinte modificação: suposição de normalidade assimétrica para os efeitos aleatórios. Basicamente, esse modelo aditivo-dominante normal assimétrico é uma adaptação do modelo misto normal assimétrico proposto por Arellano-Valle et al. (2007). Os dados de microarrays utilizados nesse trabalho foram previamente analisados por Morley et al. (2004), com o ajuste de um modelo misto Gauss-Markov com efeito aleatório de família e estrutura de covariância uniforme entre os indivíduos relacionados e posteriormente fornecido no Genetic Analysis Workshoop 15 (GAW 15), em 2006. Diferentes formas de análise desses dados foram apresentadas no GAW 15, em que se destaca o uso de modelos bayesianos hierárquicos para as médias e covariâncias dos dados de expressão gênica dentro de famílias, o uso do modelo de mistura de normais para a nálise de todos os genes conjuntamente, dentre outros.

Nesse trabalho, foram utilizadas três estratégias de análise: (i) modelo misto com efeito aleatório de família, em que a variabilidade das respostas entre e dentro das famílias são comparadas sob uma estrutura de covariância uniforme para as respostas de indivíduos relacionados; (ii) modelo misto com efeito aleatório aditivo, em que a covariância entre indivíduos é dada em função do grau de parentesco que os relaciona; (iii) modelo misto com efeito aleatório aditivo e efeito aleatório dominante, também considerando a estrutura das famílias nas matrizes de covariâncias. Foram apresentadas todas as configurações possíveis de ajustes (assimetria apenas no efeito aleatório, assimetria apenas no resíduo e assimetria em ambos os efeitos) com estes três tipos de modelos, utilizando a distribuição normal assimétrica proposta por Arelano-Valle et al. (2007). As estimativas dos parâmetros dos modelos foram realizadas sob o enfoque bayesiano.

A inferência bayesiana sobre os modelos e seus parâmetros foi utilizada com os objetivos de:

1. Selecionar o melhor modelo;

2. Estudar os tipos de assimetrias presentes nos efeitos aleatórios e

3. Obter a densidade a posteriori das herdabilidades, que são medidas de extrema importância em genética, referentes à porção herdável da variação de um caráter e, também, são medidas muito escassas para estes tipos de dados, principalmente em casos em que há fuga de normalidade (devido à assimetria).

Para a implementação computacional, foi utlizado o programa R[1], por ser um *software* estatístico gratuito e por fornecer uma estrutura amigável, de modo que modelos complexos possam ser facilmente manipulados.

Este trabalho está organizado em 5 capítulos. O Capítulo 2 contém o referencial teórico, em que se apresentam uma introdução sobre dados de microarrays, a distribuição normal assimétrica multivariada com algumas propriedades e demonstrações, o modelo misto normal assimétrico proposto por Arellano-Valle et al. (2007), uma revisão dos principais conceitos sobre valores genéticos (fenotípicos e genotípicos), o modelo aditivo-dominante normal assimétrico e alguns conceitos de inferência bayesiana. O Capítulo 3 traz uma descrição do conjunto de dados reais, o ajuste do modelo aditivo-dominante normal assimétrico para essa aplicação, a modelagem bayesiana para todos os parâmetros do modelo e os detalhes da implementação computacional. O Capítulo 4 apresenta de forma detalhada os resultados do fator de Bayes, para a seleção do melhor modelo, a descrição

[1] $http://www.r-project.org/$

dos parâmetros de interesse do melhor modelo, através do fator de Bayes e uma discussão sobre um estudo de simulação realizado. No quinto e último capítulo finaliza-se esta tese com as conclusões sobre o estudo e algumas perspectivas para trabalhos futuros.

2 REFERENCIAL TEÓRICO

2.1 Dados de microarrays

Esta seção foi baseada em Rosa et al. (2007), que apresentam de forma detalhada em seu trabalho o delineamento, a análise e aplicações na pesquisa zootécnica com os dados de microarrays.

Todos os seres vivos guardam em uma molécula de DNA (ácido desoxirribonucléico) informações vitais para o funcionamento do organismo. O processo de expressão gênica é responsável por transcrever pequenas porções do DNA (genes) nas moléculas de RNA.

O funcionamento adequado de células e tecidos depende de os genes serem expressos de forma correta. Qualquer expressão alterada ou realizada em momento incorreto, pode ocasionar patologias, por isso, foi identificada a necessidade de analisar a expressão gênica em diversos tecidos biológicos a fim de entender e poder diagnosticar essas patologias. Neste trabalho, não será esse o foco, o objetivo principal aqui é verificar se a expressão gênica é uma característica herdável.

A análise de microarrays é uma tecnologia capaz de medir a expressão gênica de milhares de genes simultaneamente. Microarray é uma lâmina com uma matriz de pontos (mais conhecidos como *spots*) impressa sobre ela. Cada *spot* contém uma quantidade ínfima de DNA devidamente imobilizada, denominada sonda (ou *probe*). Cada uma destas sondas tende a se ligar apenas à sua sequência complementar de nucleotídeos, mediante processo chamado hibridização. Essa seqüência complementar, normalmente um DNA complementar (cDNA) produzido à partir de um RNA mensageiro (mRNA), representa apenas um único gene do genoma e é chamada "alvo" ou target. A hibridização de cada sonda de DNA com o seu correspondente alvo (cDNA) é um processo baseado na complementaridade das

cadeias de nucleotídeos, ou seja, na propriedade que duas cadeias homólogas têm de parear suas bases complementares (A com T e C com G), mediante a formação de pontes de hidrogênio.

A plataforma sólida mais utilizada na confecção dos microarrays é a lâmina de vidro, do tipo usado em microscopia, que depois de ter as sondas imobilizadas na sua superfície, é normalmente referida como slide ou simplesmente lâmina de microarray. Basicamente existem duas formas de distribuir as sondas em um microarray de DNA. A primeira, mais simples, é feita por robôs de alta precisão que utilizam agulhas especiais para depositar as amostras de DNA na superfície de uma lâmina de vidro, processo que também é conhecido como "impressão do slide". Normalmente essas amostras são constituídas de oligonucleotideos pré-sintetizados, cDNAs produzidos em projetos de sequenciamento, ou ainda produtos de amplificação por PCR (reação em cadeia de polimerase). A segunda, mais complexa, utiliza processos especiais (fotolitografia, por exemplo) para realizar a síntese química de oligonucleotídeos diretamente sobre a superfície da lâmina de vidro (Walsh & Henderson, 2004).

Desta maneira, a tecnologia de microarrays consiste na utilização de um slide (lâmina ou microarranjo) no qual as sondas (amostras de DNA) foram imobilizadas em quantidades e posições precisamente definidas (*spots*), para se fazer a hibridização com um pool de mRNAs extraídos de amostras biológicas (targets), que foram previamente marcados com fluoróforos (marcadores florescentes). Como as moléculas de mRNA são bastante instáveis quando manipuladas, a maioria dos protocolos laboratoriais utiliza o processo de transcrição reversa para convertê-las nos correspondentes cDNAs durante o processo de marcação.

Após o processo de hibridização, cada lâmina é lavada para remoção dos "alvos" excedentes (que não se ligaram às sondas) e, em seguida, exposta à ação de

raios laser que excitam os fluoróforos que foram incorporados aos “alvos”, fazendo com que estes emitam luz (fluorescência). Em princípio, quanto maior for a expressão de um determinado gene, maior será a quantidade de “alvos” marcados com o fluoróforo e, consequentemente, maior será a intensidade da fluorescência do complexo alvo-sonda após a hibridização. Assim, a tecnologia de microarrays fornece uma medida indireta do nível de expressão gênica, mediante quantificação da abundância dos RNAs transcritos.

Para delinear um estudo com microarrays, há várias combinações possíveis de lâminas que podem ser utilizadas. As duas lâminas mais utilizadas são a Affymetrix® e os microarrays de cDNA.

Do ponto de vista de delineamento e análise estatística de experimentos de microarray, a distinção mais importante entre as diferentes tecnologias existentes refere-se ao número de amostras hibridizadas em cada lâmina. Neste sentido, os tipos de tecnologias de microarrays podem ser divididos em dois grupos básicos: sistema de uma cor (single-color ou single-channel microarray), utilizada pela lâmina Affymetrix® e sistema de duas cores (two-color microarray), utilizada pela lâmina de microarrays de cDNA. Neste trabalho foi utilizado o sistema de uma cor. Na lâmina do tipo Affymetrix®, cada amostra de RNA é marcada e hibridizada individualmente numa lâmina. Uma vantagem deste tipo de tecnologia é que a condução dos experimentos é geralmente mais simples, mas por outro lado, variações naturais entre as lâminas ficam de certa maneira confundidas com as diferenças entre as amostras. Este problema, no entanto, é minimizado se as diferenças entre as lâminas não forem importantes.

A análise estatística de dados de microarray geralmente envolve três componentes: 1) obtenção dos dados e eliminação de valores espúrios; 2) pré-ajuste dos dados para efeitos sistemáticos (este ajuste é conhecido por normalização dos da-

dos); e 3) análise estatística propriamente dita.

A obtenção dos dados refere-se à análise de imagem de cada lâmina para a extração dos valores de intensidade fluorescente em cada *spot*, os quais são medidas indiretas da abundância de transcritos de RNA dos genes representados pelas sondas. Existe uma série de procedimentos e programas computacionais disponíveis para a leitura das imagens para diferentes tipos de lâminas. Para o sistema de duas cores (lâminas de cDNA), os valores de intensidade são geralmente obtidos a partir da classificação de cada *pixel* da imagem como pertencente a determinado *spot* (isto é, sonda) ou a espaços vazios entre *spots*, denominados *foreground* e *background*, respectivamente. Posteriormente, para cada lâmina, os *pixels* de *foreground* relativos a cada *spot* são combinados em alguma medida resumo, como média, mediana ou intensidade total, as quais muitas vezes são ajustadas para valores de *background*. Este mesmo procedimento é efetuado tanto para as intensidades relativas à uma cor (chamada de Cy3) quanto aquelas relativas à outra cor (denominada de Cy5), de maneira que para cada *spot* têm-se duas medidas de intensidade, das quais obtém-se uma medida da expressão relativa de cada gene nas duas amostras hibridizadas em cada lâmina.

Em lâminas do tipo Affymetrix® cada gene é representado por um grupo (geralmente de 11 a 20 pares) de sondas de cadeias curtas de oligonucleotídeos (de 25 bases). Cada par inclui uma sonda com sequência nucleotídea idêntica ao gene (chamada *perfect match*, PM), e outra sonda com uma mudança nucleotídea na 13^a base (chamada *mismatch*, MM). De maneira similar às lâminas de hibridização competitiva, os valores de intensidade observados para cada gene são geralmente combinados numa única medida resumo para expressar o nível de abundância de transcritos de RNA, como por exemplo, utilizando-se a média das diferenças entre PM e MM para cada gene, dada por:

$$AvDiff_g = \frac{1}{K}\sum_{i=1}^{K}(PM_{gi} - MM_{gi})$$

onde $AvDiff_g$ é a medida de expressão relativa ao gene g, PM_{gi} e MM_{gi} são as intensidades PM e MM relativas ao j-ésimo par de sondas ($j = 1, 2,...,K$) do gene g.

A medida resumo $AvDiff$ foi inicialmente proposta pela Affymetrix®, mas hoje já existem outras metodologias alternativas, supostamente melhores, para a sumarização das intensidades observadas para cada gene, como por exemplo o procedimento MAS5.0 da própria Affymetrix®, o MBEI (Multiplicative Model-Based Expression Index; proposto por Li & Wong, 2001) e o RMA (Robust Multi-array Average; proposto por Irizarry et al., 2003).

Após a obtenção dos dados e eliminação de valores espúrios em decorrência de possíveis problemas na fixação das sondas, marcação das amostras, hibridização etc., um ajuste geral dos dados é geralmente necessário antes de uma análise estatística mais formal dos mesmos. Este processo de correção dos dados é geralmente denominado normalização, e considera ajustes para diferenças entre lâminas (em termos de média ou mediana e variância), efeito de marcação, etc. Alguns procedimentos de normalização dos dados baseiam-se somente em alguns genes presentes nas lâminas (como genes controles ou genes com expressão supostamente constante nos diversos grupos experimentais), outros baseiam-se em todos os genes e utilizam procedimentos estatísticos robustos, com a suposição de que a maioria dos genes é não diferencialmente expresso entre os grupos experimentais.

Um procedimento comumente utilizado para a normalização de dados de microarray em hibridização competitiva utiliza uma metodologia de regressão não-paramétrica robusta (denominada *LOWESS*) para estabilizar a relação entre o logaritmo da razão de intensidades e a média do logaritmo das intensidades em cada

lâmina (Yang et al., 2002). Esta metodologia é também utilizada para dados de lâminas do tipo Affymetrix®, mas neste caso o procedimento *LOWESS* é aplicado sucessivamente para cada par de lâminas, e a normalização final para cada lâmina é dada pela média geral dos resultados de cada um de seus pareamentos. O procedimento é repetido até a convergência, isto é, até que pareamentos adicionais não alterem a normalização dos dados, e é denominado *LOWESS* cíclico. Outro procedimento bastante comum de normalização é denominado normalização quantílica (Bolstad et al., 2003). Este procedimento faz com que todas as lâminas apresentem mesma distribuição empírica dos valores de intensidade, de maneira que elas são coincidentes em termos de locação (incluindo medidas de centralidade e percentis) e escala ou variabilidade.

Após a análise das imagens e normalização dos dados é que se inicia este trabalho.

Duas características naturais desses dados após o processo de normalização são, em geral, alta variabilidade e assimetria. Por essa razão, houve motivação para utilizar um modelo, que será detalhado posteriormente, que é baseado na distribuição normal assimétrica, que engloba além dos parâmetros de posição e escala, o parâmetro de assimetria.

Antes de apresentar o modelo a ser utilizado, apresenta-se mais detalhadamente na Seção a seguir, a distribuição normal assimétrica multivariada, com algumas propriedades que serão utilizadas no decorrer do trabalho.

2.2 Distribuição normal assimétrica multivariada

Antes de se introduzir a distribuição normal assimétrica multivariada, será feito um breve relato histórico, baseado em Azzalini[2].

O primeiro pesquisador a publicar sobre a distribuição normal assimétrica uni-

[2] $http://azzalini.stat.unipd.it/SN/azzalini-sis2006.ps$

variada foi Fernando de Helguero, em Roma, em abril de 1908, no estudo da distribuição dos salários de trabalhadores. Infelizmente, em dezembro desse mesmo ano, Fernando de Helguero morreu no terremoto em Messina, na Itália.

Roberts (1966), do Instituto de Nutrição, na Guatemala, através de um estudo com gêmeos, aplica a distribuição normal assimétrica univariada ao estudo da distribuição das estatísticas de ordem (mínimo e máximo) de duas variáveis aleatórias com distribuição normal bivariada.

O'Hagan & Leonard (1976) propõem o uso de prioris assimétricas, incluindo dentre elas a distribuição normal assimétrica.

Até então, a distribuição apresentrada nesses trabalhos ainda não era chamada de normal assimétrica. Azzalini (1985) nomeou essa distribuição e apresentou diversas de suas propriedades no caso univariado.

O caso multivariado foi apresentado em Azzalini & Dalla-Valle (1996). Atualmente, existem várias versões da distribuição normal assimétrica multivariada. Considera-se neste trabalho um caso especial da distribuição normal assimétrica proposta por Arellano-Valle & Genton (2005) e que foi apresentada por Arellano-Valle et al. (2007). Esta versão generaliza a apresentada por Sahu et al. (2003), por causa da matriz de variâncias e covariâncias, que neste caso é assumida ser uma matriz positiva definida e em Sahu et al. (2003), uma matriz diagonal. A seguir, apresenta-se a definição desta versão e algumas propriedades, com suas respectivas demonstrações, que também podem ser encontradas em Arellano-Valle et al. (2007).

Para apresentar a densidade da normal assimétrica multivariada e algumas propriedades é necessário introduzir a notação que se segue.

Seja $\phi_n(y|\mu, \Sigma)$ a função densidade de probabilidade (fdp) e $\Phi_n(y|\mu, \Sigma)$ a função de distribuição acumulada (fda) da normal multivariada, $N_n(\mu, \Sigma)$, avali-

ada em y. Considere também as seguintes notações: $diag(c_1, \ldots, c_n)$, para representar uma matriz diagonal com elementos $c_1, \ldots, c_n$ na sua diagonal e I_n, para representar uma matriz identidade de dimensão $n \times n$.

Definição 1: Um vetor aleatório n-dimensional Y segue uma distribuição normal assimétrica multivariada (SN_n) com vetor de locação $\mu \in \Re^n$, matriz de dispersão Σ (uma matriz de dimensão $n \times n$ positiva definida) e matriz de assimetria $\Delta = diag(\delta_1, \ldots, \delta_n)$, com $\delta_k \in \Re, k = 1, \ldots, n$, se sua fdp é dada por

$$\begin{aligned} f(y|\mu, \Sigma, \Delta) &= 2^n \phi_n(y|\mu, \Sigma + \Delta^2) \times \\ &\quad \Phi_n(\Delta(\Sigma + \Delta^2)^{-1}(y - \mu)|0, (I_n + \Delta\Sigma^{-1}\Delta)^{-1}). \quad (2.1) \end{aligned}$$

Será utilizada a notação $Y \sim SN_n(\mu, \Sigma, \Delta)$. Note que quando Δ é uma matriz de zeros de dimensão $n \times n$, a equação (2.1) se reduz à usual distribuição normal multivariada, $N_n(\mu, \Sigma)$.

Para fins ilustrativos, serão apresentadas a seguir as Figuras 1 a 9 com a densidade normal assimétrica bivariada para diferentes tipos de assimetria. Essas figuras podem ser geradas facilmente acessando a página de Azzalini[3]. A versão da normal multivariada utilizada nesta página é apresentada em Azzalini & Dalla Valle (1996) e a parametrização utilizada no programa em R é apresentada em Azzalini & Capitanio (1999).

Será usada a notação $SN_2((\mu_1, \mu_2)^\top, diag(\sigma_1^2,\sigma_2^2), diag(\delta_1,\delta_2))$, mais especificamente,

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

[3] $http://azzalini.stat.unipd.it/SN/plot-SN2.html$

para diferentes variações de δ_1 e δ_2.

A Figura 1 apresenta como é a distribuição, quando as duas variáveis não são assimétricas, isto é, o caso da distribuição normal padrão bivariada. As Figuras 2, 3, 4 e 5 mostram como ficam as distribuições quando apenas um δ_i, $i = 1, 2$ é diferente de zero. Pode-se observar que os gráficos das densidades de (X, Y) apresentam mais massa para os valores do sinal de δ da variável que possui assimetria e as curvas de nível ficam voltadas para o lado do sinal de δ da variável assimétrica. As Figuras 6, 7, 8 e 9 mostram como ficam as distribuições quando as duas variáveis são assimétricas. As curvas de nível ficam voltadas para o quadrante correspondente aos sinais de δ_1 e δ_2, isto é, na Figura 6, as curvas de nível estão voltadas para os valores negativos de X e para os valores negativos de Y e, assim por diante para as demais figuras. O mesmo ocorre para o comportamento dos gráficos das funções densidades de (X, Y), com maior massa nos quadrantes correspondentes aos sinais de δ_1 e δ_2.

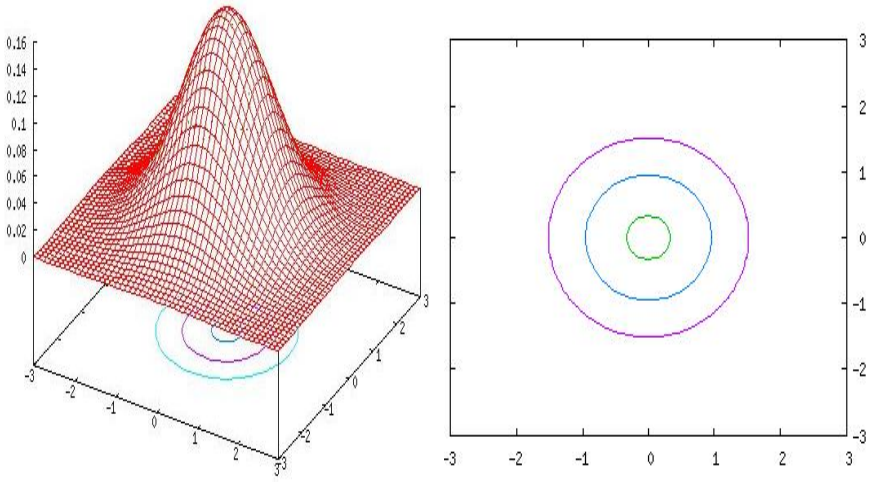

FIGURA 1 Distribuição $SN_2((0,0)^\top, diag(1,1), diag(0,0))$ com seu gráfico de contorno.

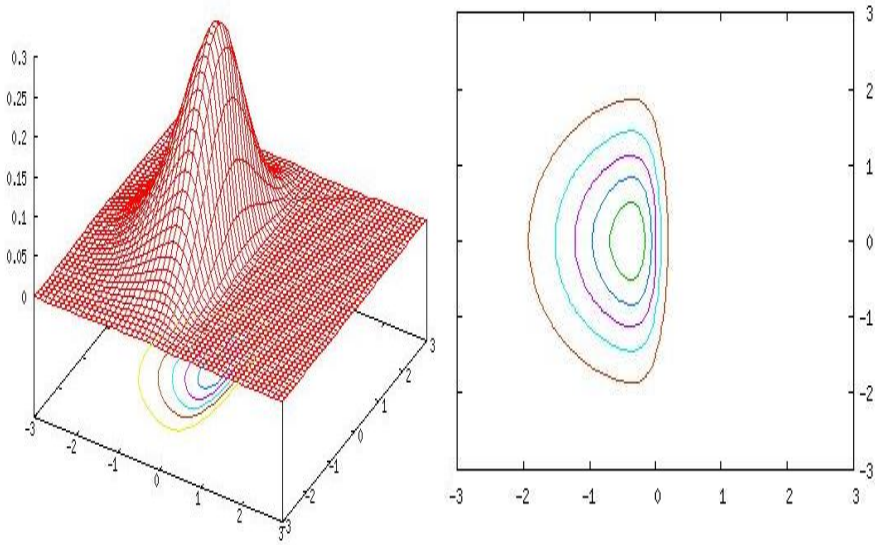

FIGURA 2 Distribuição $SN_2((0,0)^\top, diag(1,1), diag(-5,0))$ com seu gráfico de contorno.

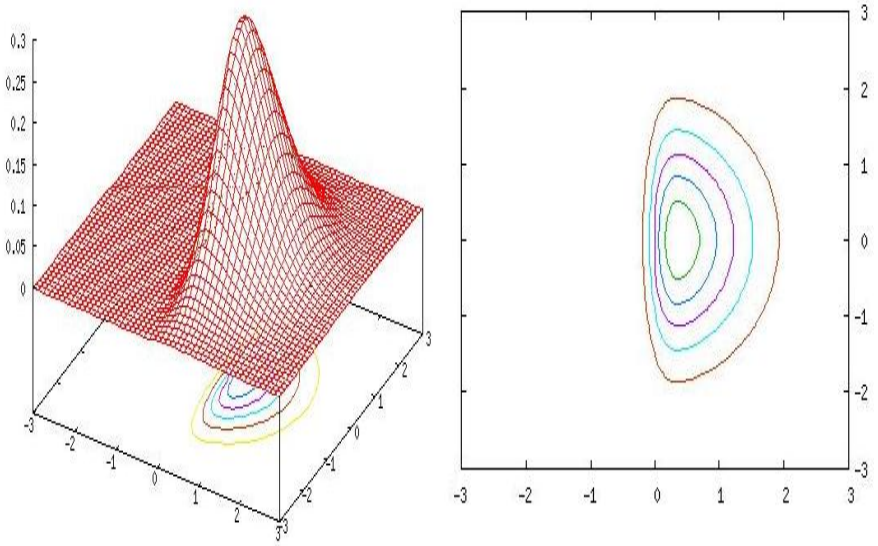

FIGURA 3 Distribuição $SN_2((0,0)^\top, diag(1,1), diag(5,0))$ com seu gráfico de contorno.

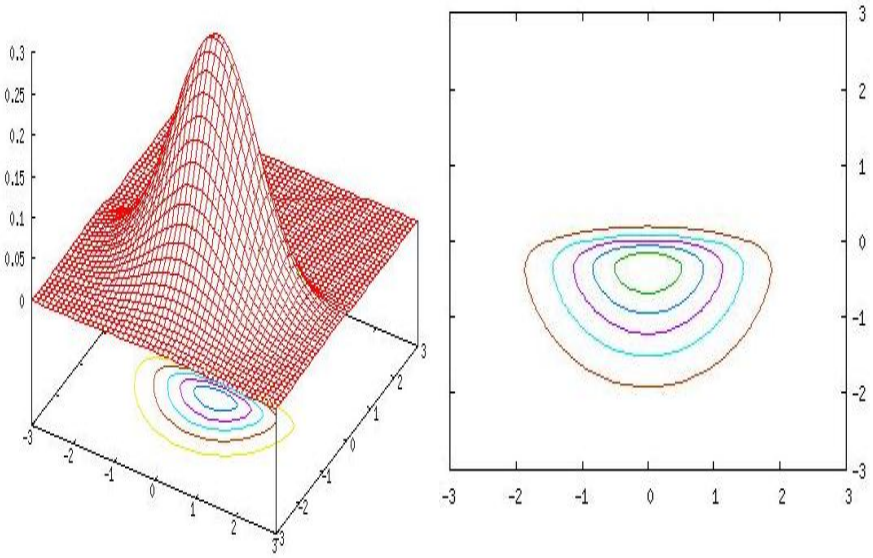

FIGURA 4 Distribuição $SN_2((0,0)^\top, diag(1,1), diag(0,-5))$ com seu gráfico de contorno.

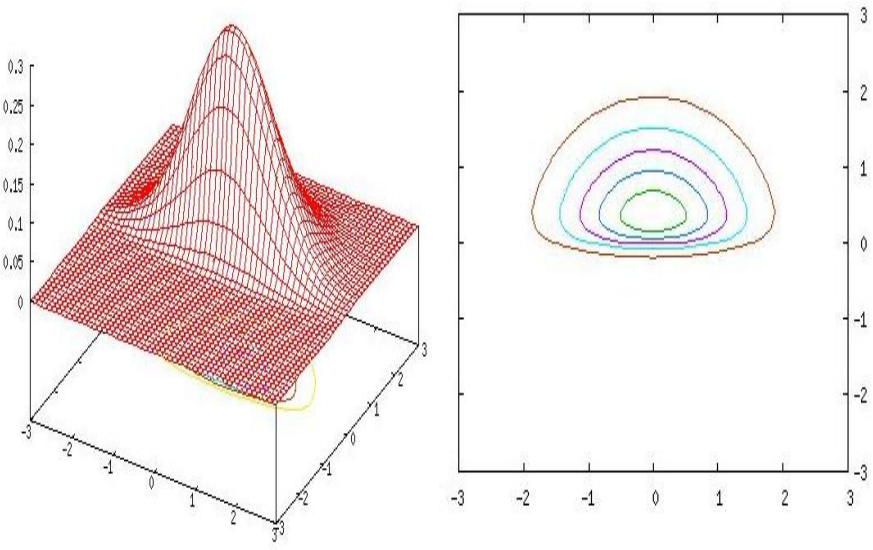

FIGURA 5 Distribuição $SN_2((0,0)^\top, diag(1,1), diag(0,5))$ com seu gráfico de contorno.

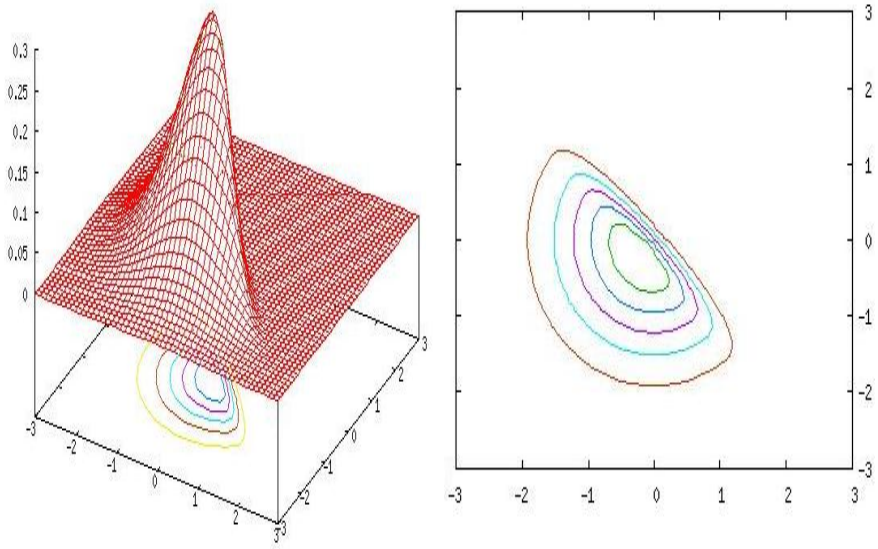

FIGURA 6 Distribuição $SN_2((0,0)^\top, diag(1,1), diag(-5,-5))$ com seu gráfico de contorno.

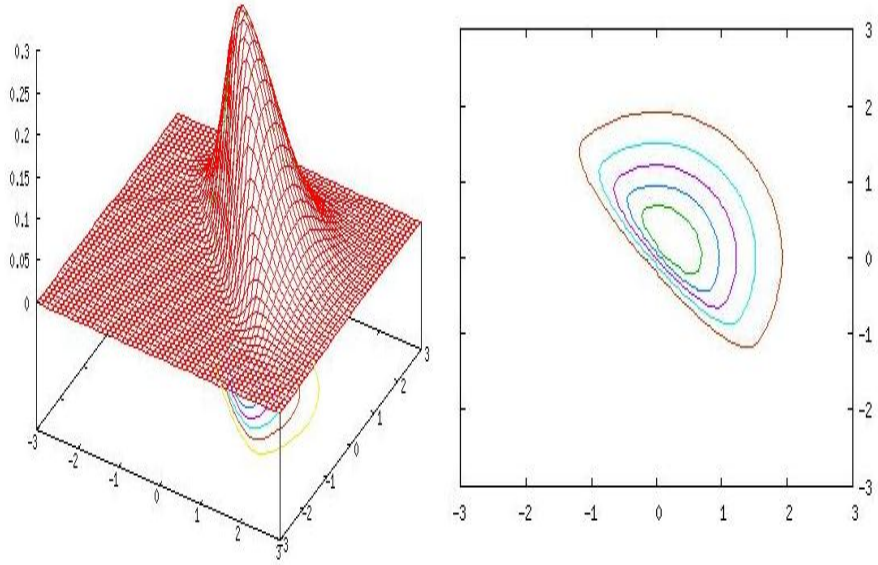

FIGURA 7 Distribuição $SN_2((0,0)^\top, diag(1,1), diag(5,5))$ com seu gráfico de contorno.

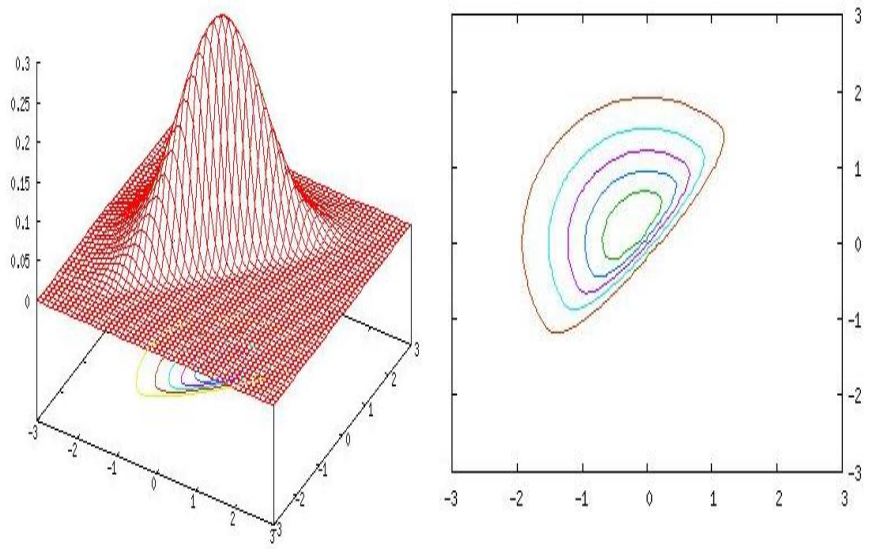

FIGURA 8 Distribuição $SN_2((0,0)^\top, diag(1,1), diag(-5,5))$ com seu gráfico de contorno.

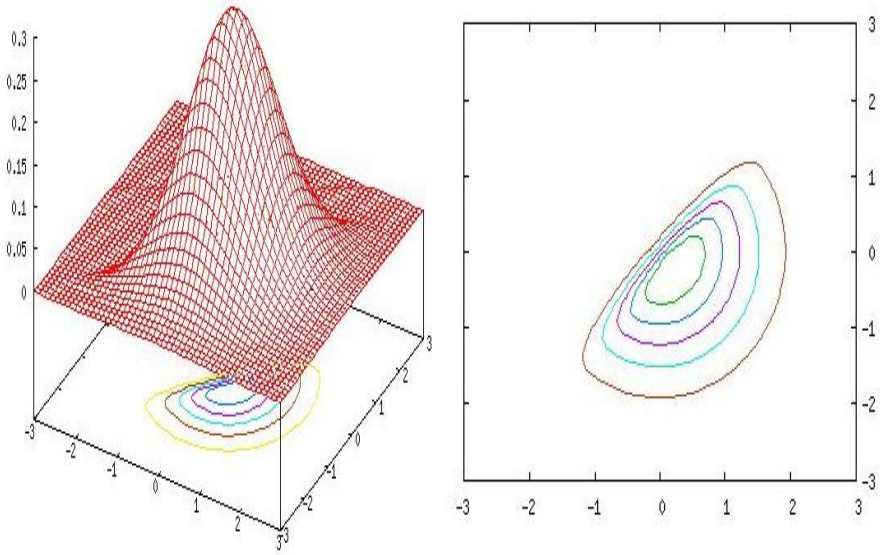

FIGURA 9 Distribuição $SN_2((0,0)^\top, diag(1,1), diag(5,-5))$ com seu gráfico de contorno.

A seguir, serão apresentados um lema, duas proposições e um corolário, com as respectivas demonstrações. Estes facilitaram o trabalho de inferência com a distribuição normal assimétrica multivariada.

Lema 1: Seja $Y|X = x \sim N_p(\mu + Ax, \Sigma)$ e $X \sim N_q(\eta, \Omega)$. Então,

$$\begin{aligned}\phi_p(y|\mu + Ax,\Sigma)\phi_q(x|\eta, \Omega) &= \phi_p(y|\mu + A\eta,\Sigma + A\Omega A^\top) \times \\ &\quad \phi_q(x|\eta + \Lambda A^\top \Sigma^{-1}(y - \mu - A\eta), \Lambda),\end{aligned}$$

em que $\Lambda = (\Omega^{-1} + A^\top \Sigma^{-1} A)^{-1}$.

Demonstração: Reescrevendo o lado esquerdo da igualdade tem-se

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

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

$$\frac{(2\pi)^{-p/2}}{\sqrt{|\Sigma + A\Omega A^\top|}} \exp\Big[-\frac{1}{2}(y-\mu-A\eta)^\top(\Sigma + A\Omega A^\top)^{-1}(y-\mu-A\eta)\Big] \times$$
$$\frac{(2\pi)^{-q/2}}{\sqrt{|\Lambda|}} \exp\Big\{ -\frac{1}{2}\big[x-\eta-\Lambda A^\top\Sigma^{-1}(y-\mu-A\eta)\big]^\top\Lambda^{-1} \times$$
$$\big[x-\eta-\Lambda A^\top\Sigma^{-1}(y-\mu-A\eta)\big]\Big\} =$$
$$\frac{(2\pi)^{-p/2}(2\pi)^{-q/2}}{\sqrt{|\Sigma + A\Omega A^\top||\Lambda|}} \exp\Big\{ -\frac{1}{2}\Big\{(y-\mu-A\eta)^\top(\Sigma + A\Omega A^\top)^{-1}(y-\mu-A\eta) +$$
$$\big[x-\eta-\Lambda A^\top\Sigma^{-1}(y-\mu-A\eta)\big]^\top\Lambda^{-1}\big[x-\eta-\Lambda A^\top\Sigma^{-1}(y-\mu-A\eta)\big]\Big\}\Big\},$$

com $\Lambda = (\Omega^{-1} + A^\top\Sigma^{-1}A)^{-1}$.

Para facilitar as manipulações algébricas, z representa a expresssão $y-\mu-A\eta$ e w a expressão $x-\eta$. Assim, o lado esquerdo da igualdade passa a ficar da seguinte forma

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

e o lado direito da igualdade fica

$$\frac{(2\pi)^{-p/2}(2\pi)^{-q/2}}{\sqrt{|\Sigma + A\Omega A^\top||\Lambda|}} \exp\Big\{ -\frac{1}{2}\Big\{z^\top(\Sigma + A\Omega A^\top)^{-1}z +$$
$$\big[w-\Lambda A^\top\Sigma^{-1}z\big]^\top\Lambda^{-1}\big[w-\Lambda A^\top\Sigma^{-1}z\big]\Big\}\Big\}.$$

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

são iguais, isto é,

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

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

Neste desenvolvimento serão utilizadas as seguintes identidades matriciais (Searle et al., 2006):

1. Transposta de matrizes:

$$(AB)^\top = B^\top A^\top;$$

2. Complemento de Schur:

$$(D - CA^{-1}B)^{-1} = D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1};$$

3. Determinante do produto de matrizes:

$$|AB| = |A|\,|B|.$$

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

$$z^\top(\Sigma + A\Omega A^\top)^{-1}z + (w - \Lambda A^\top\Sigma^{-1}z)^\top\Lambda^{-1}(w - \Lambda A^\top\Sigma^{-1}z) =$$
$$z^\top\Big[\Sigma^{-1} - \Sigma^{-1}A(\Omega^{-1} + A^\top\Sigma^{-1}A)^{-1}A^\top\Sigma^{-1}\Big]z + w^\top\Lambda^{-1}w$$
$$-w^\top\underbrace{\Lambda^{-1}\Lambda}_{I}A^\top\Sigma^{-1}z - (\Lambda A^\top\Sigma^{-1}z)^\top\Lambda^{-1}w + (\Lambda A^\top\Sigma^{-1}z)^\top\underbrace{\Lambda^{-1}\Lambda}_{I}A^\top\Sigma^{-1}z =$$
$$z^\top\Sigma^{-1}z - z^\top\Sigma^{-1}A\underbrace{(\Omega^{-1} + A^\top\Sigma^{-1}A)^{-1}}_{\Lambda}A^\top\Sigma^{-1}z + w^\top\Lambda^{-1}w$$
$$-(Aw)^\top\Sigma^{-1}z - z^\top\Sigma^{-1}A\underbrace{\Lambda\Lambda^{-1}}_{I}w + z^\top\Sigma^{-1}A\Lambda A^\top\Sigma^{-1}z =$$
$$z^\top\Sigma^{-1}z + w^\top\Lambda^{-1}w - (Aw)^\top\Sigma^{-1}z - z^\top\Sigma^{-1}Aw =$$
$$z^\top\Sigma^{-1}z + w^\top(\Omega^{-1} + A^\top\Sigma^{-1}A)w - (Aw)^\top\Sigma^{-1}z - z^\top\Sigma^{-1}Aw =$$
$$z^\top\Sigma^{-1}z + w^\top\Omega^{-1}w + (Aw)^\top\Sigma^{-1}Aw - (Aw)^\top\Sigma^{-1}z - z^\top\Sigma^{-1}Aw =$$
$$z^\top\Sigma^{-1}z - z^\top\Sigma^{-1}Aw - (Aw)^\top\Sigma^{-1}z + (Aw)^\top\Sigma^{-1}Aw + w^\top\Omega^{-1}w =$$
$$(z - Aw)^\top\Sigma^{-1}(z - Aw) + w^\top\Omega^{-1}w. \qquad \blacksquare$$

Parte II: Inicia-se de $\bullet$ para chegar em $\bullet\bullet$:

$$|\Sigma + A\Omega A^\top||\Lambda| = |(\Sigma + A\Omega A^\top)\Lambda| = |(\Sigma + A\Omega A^\top)(\Omega^{-1} + A^\top\Sigma^{-1}A)^{-1}| =$$
$$|(\Sigma + A\Omega A^\top)\big[\Omega - \Omega A^\top(\Sigma + A\Omega A^\top)^{-1}A\Omega\big]| =$$
$$|\Sigma\Omega - \Sigma\underbrace{\Omega A^\top(\Sigma + A\Omega A^\top)^{-1}A\Omega} + A\Omega A^\top\Omega - A\Omega A^\top\underbrace{\Omega A^\top(\Sigma + A\Omega A^\top)^{-1}A\Omega}| =$$
$$|\Sigma\Omega - (\Sigma + A\Omega A^\top)\Omega A^\top(\Sigma + A\Omega A^\top)^{-1}A\Omega + A\Omega A^\top\Omega| =$$
$$|\Sigma\Omega - A\Omega\underbrace{(\Sigma + A\Omega A^\top)^{-1}(\Sigma + A\Omega A^\top)}_{I}A^\top\Omega + A\Omega A^\top\Omega| =$$
$$|\Sigma\Omega - A\Omega A^\top\Omega + A\Omega A^\top\Omega| = |\Sigma\Omega| = |\Sigma|\,|\Omega|. \qquad \blacksquare$$

Proposição 1: Seja $Y \sim SN_n(\mu, \Sigma, \Delta)$. Então

$$Y \stackrel{d}{=} \Delta|X_0| + X_1,$$

em que $X_0 \sim N_n(0, I_n)$ e $X_1 \sim N_n(\mu, \Sigma)$. O vetor aleatório Y tem distribuição igual à de $\Delta|X_0| + X_1$, desde que os vetores aleatórios X_0 e X_1 sejam independentes.

Demonstração: Sejam $U = \Delta|X_0| + X_1$ e $T = |X_0| \sim N_n(0, I_n)\mathtt{I}_{t\geq 0}$, em que $\mathtt{I}$ é a função indicadora do domínio de T. A distribuição de T é conhecida como distribuição normal padrão truncada positiva (*Half Normal*).

Quando se condiciona $U|T$, tem-se uma relação direta entre a variável U e X_1, o que faz com que $U|T$ tenha uma distribuição normal multivariada. Mais especificamente, $U|T = t \sim N_n(\Delta t + \mu, \Sigma)$, pois

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

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

$$\begin{aligned} f_U(u) &= \int_{\Re_+^n} f(u,t)dt = \int_{\Re_+^n} f(u|t)f(t)dt \\ &= \int_{\Re_+^n} \phi_n(u|\mu + \Delta t, \Sigma)2^n\phi_n(t)dt = 2^n \int_{\Re_+^n} \phi_n(u|\mu + \Delta t, \Sigma)\phi_n(t)dt. \end{aligned}$$

Agora, utilizando o Lema 1, tem-se que

$$\begin{aligned} f_U(u) &= 2^n \int_{\Re_+^n} \phi_n(u|\mu, \Sigma + \Delta^2) \times \\ &\quad \phi_n(t|(I_n + \Delta\Sigma^{-1}\Delta)^{-1}\Delta\Sigma^{-1}(u-\mu), (I_n + \Delta\Sigma^{-1}\Delta)^{-1})dt. \end{aligned}$$

Arellano-Valle et al. (2007) apresenta que $(I_n + \Delta\Sigma^{-1}\Delta)^{-1}\Delta\Sigma^{-1} = \Delta(\Sigma + \Delta^2)^{-1}$. Assim, segue que

$$\begin{aligned} f_U(u) &= 2^n \phi_n(u|\mu, \Sigma + \Delta^2) \times \\ &\quad \int_{\Re_+^n} \phi_n(t|\Delta(\Sigma + \Delta^2)^{-1}(u-\mu), (I_n + \Delta\Sigma^{-1}\Delta)^{-1})dt = \\ &\quad 2^n \phi_n(u|\mu, \Sigma + \Delta^2) \times \\ &\quad \Phi_n(\Delta(\Sigma + \Delta^2)^{-1}(u-\mu)|0, (I_n + \Delta\Sigma^{-1}\Delta)^{-1}), \end{aligned}$$

ou seja, $U \stackrel{d}{=} Y \sim SN_n(\mu, \Sigma, \Delta)$. ■

Uma consequência direta da Proposição 1, relacionada com os momentos do vetor aleatório normal assimétrico é dada pelo seguinte Corolário.

Corolário 1: Seja $Y \sim SN_n(\mu, \Sigma, \Delta)$. Então

$$E[Y] = \mu + \sqrt{\frac{2}{\pi}}\,\delta \qquad e \qquad Var[Y] = \Sigma + \Big(1 - \frac{2}{\pi}\Big)\Delta^2,$$

em que $\delta = (\delta_1, \cdots, \delta_n)^\top$ é a diagonal da matriz Δ.

Demonstração: Sejam $T = |X_0|$ com distribuição normal truncada positiva e

$1_{n\times 1}$, um vetor de uns de dimensão $n \times 1$, então

$$E[T] = 1_{n\times 1}\sqrt{\frac{2}{\pi}} \qquad e \qquad Var[T] = I_n\Big(1 - \frac{2}{\pi}\Big).$$

Logo a média e a variância de Y são dadas por

$$\begin{aligned} E[Y] &= E(\delta|X_0| + X_1) = \delta E(T) + E(X_1) = \delta\sqrt{\frac{2}{\pi}} + \mu = \mu + \sqrt{\frac{2}{\pi}} \quad \delta \\ Var[Y] &= Var(\Delta|X_0| + X_1) = \\ & \Delta^2 Var(T) + Var(X_1) = \\ & \Delta^2 I_n\Big(1 - \frac{2}{\pi}\Big) + \Sigma = \\ & \Sigma + \Big(1 - \frac{2}{\pi}\Big)\Delta^2. \qquad \blacksquare \end{aligned}$$

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

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

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

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

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

$$\Phi_n(\Delta(\Sigma+\Delta^2)^{-1}(y-\mu)|0,(I_n+\Delta\Sigma^{-1}\Delta)^{-1}). \qquad \blacksquare$$

2.3 Modelo misto normal assimétrico

Segundo Searle et al. (2006) o modelo misto é expresso da seguinte forma

$$Y = X\beta + Zu + \varepsilon, \tag{2.2}$$

em que Y de dimensão $n \times 1$ é um vetor de respostas, X de dimensão $n \times p$ é uma matriz de delineamento dos efeitos fixos, β de dimensão $p \times 1$ é o vetor dos efeitos fixos, Z de dimensão $n \times q$ é uma matriz de delineamento dos efeitos aleatórios, u de dimensão $q \times 1$ é o vetor dos efeitos aleatórios e ε é o vetor de erros aleatórios ou resíduos de dimensão $n \times 1$. Tipicamente, assume-se que os efeitos aleatórios u e os erros aleatórios ε são independentes com

$$u \sim N_q(0, S) \qquad e \qquad \varepsilon \sim N_n(0, \Psi), \tag{2.3}$$

em que $S = S(\sigma_u^2)$ e $\Psi = \Psi(\sigma_\varepsilon^2)$ são matrizes de covariâncias entre os efeitos aleatórios $u_1, \ldots, u_q$ e os resíduos $\varepsilon_1, \ldots, \varepsilon_n$, respectivamente. Em geral, é considerado $\Psi = \sigma_\varepsilon^2 I_n$, em que I_n é uma matriz identidade $n \times n$.

Note que sob as suposições consideradas em (2.3) o modelo apresentado em (2.2) pode ser representado hierarquicamente como

$$Y|\beta, u, \sigma_\varepsilon^2 \sim N_n(X\beta + Zu, \sigma_\varepsilon^2 I_n) \quad e \quad u|\sigma_u^2 \sim N_q(0, S(\sigma_u^2)). \tag{2.4}$$

O principal interesse é fazer inferência sobre o vetor de parâmetros $\theta = (\beta^\top, u^\top, \sigma_u^2, \sigma_\varepsilon^2)$. Os métodos de estimação mais utilizados para o vetor de parâmetros θ são os de máxima verossimilhança (EMV) e os de máxima verossimilhança restrita (EMVR, Searle et al., 2006).

A metodologia de modelos mistos permite estimar β pelo procedimento de quadrados mínimos generalizados e predizer u pelo procedimento BLUP (Henderson, 1984). Para obtenção destas soluções, basta resolver o seguinte sistema de equações lineares

$$\begin{aligned}\begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} &= \begin{bmatrix} X^\top X & X^\top Z \\ Z^\top X & Z^\top Z + \frac{\sigma^2_\varepsilon}{\sigma^2_u} B^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X^\top Y \\ Z^\top Y \end{bmatrix} \\ &= \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} X^\top Y \\ Z^\top Y \end{bmatrix}, \end{aligned} \tag{2.5}$$

em que B representa uma matriz de correlação.

Atribuindo valores iniciais para os componentes de variâncias (σ^2_u e σ^2_ε) no sistema de equações lineares em (2.5), obtém-se a predição de u. Ao calcular as variâncias deste efeito predito ($\hat{u}$) e do resíduo (fazendo $\hat{\varepsilon} = Y - X\hat{\beta} - Z\hat{u}$), obtem-se as estimativas das variâncias $\hat{\sigma}^2_u$ e $\hat{\sigma}^2_\varepsilon$, as quais, provavelmente, serão diferentes dos valores iniciais utilizados no sistema de equações lineares, o que significa que os valores iniciais não foram verossímeis. Desta forma, deve-se resolver novamente o sistema em (2.5), usando estes componentes de variância calculados. Procedendo-se sucessivamente desta maneira, atinge-se a convergência para os componentes de variância, ou seja, tem-se que os valores utilizados no sistema em (2.5) equivalem às próprias variâncias dos efeitos aleatórios, o que significa que os valores utilizados no sistema em (2.5) passaram a ser verossímeis com o conjunto de dados, sendo estas estimativas as de máxima verossimilhança. (Farias Neto & Resende, 2001; Searle et al., 2006)

Segundo Resende (2002, p.391 e 392), com $\hat{\varepsilon} = Y - X\hat{\beta} - Z\hat{u}$, em que $\hat{\beta}$ e $\hat{u}$ são obtidos via o sistema em (2.5) apresentado anteriormente, os EMVR dos componentes de variância dos efeitos aleatórios, empregando-se o algoritmo EM

são dados por

$$\hat{\sigma}_{\varepsilon}^2 = \frac{\hat{\varepsilon}^{\top}\hat{\varepsilon}}{(n - r(X))} \qquad e \qquad \hat{\sigma}_{u}^2 = \frac{\hat{u}^{\top}B^{-1}\hat{u} + \hat{\sigma}_{\varepsilon}^2 tr(C_{22})}{q}, \tag{2.6}$$

em que $r(X)$ é o posto da matriz X e $tr(C_{22})$ é o traço da matriz C_{22} apresentada em (2.5).

Apesar desse modelo misto ser muito flexível e muito utilizado, a suposição de normalidade simétrica, que possui apenas dois parâmetros (posição e escala), em muitos casos, pode não ser realista, obscurecendo características importantes no modelo. A distribuição dos resíduos, quando se ajusta um modelo de análise de variância clássico aos dados de microarrays, em geral, apresenta assimetria e alta variabilidade, podendo implicar como resultado dos testes de normalidade, não possuir distribuição normal. Nessas situações, comumente os pesquisadores recorrem ao uso de transformações nos dados para ser possível ajustar este modelo misto gaussiano. Embora tal método (transformação dos dados) possa dar resultados empíricos razoáveis, deve ser evitado caso um modelo mais robusto possa ser encontrado (Azzalini & Capitanio, 1999).

Um enfoque alternativo à transformação de variáveis, para modelagem de dados de um ponto de vista paramétrico, apropriado para o tratamento de observações contínuas multivariadas, consiste em construir classes paramétricas flexíveis de distribuições multivariadas que exibam assimetria e curtose diferentes da distribuição normal. A classe das distribuições elípticas é provavelmente o exemplo mais conhecido desse enfoque. Embora modelos elípticos forneçam alternativas para o modelo normal, estes somente podem ser aplicados em situações práticas onde simetria parece razoável. Portanto, é de interesse prático estudar modelos estatísticos que sejam menos sensíveis do que a distribuição normal a certos desvios das suposições consideradas, construindo famílias paramétricas de distribuições

assimétricas que sejam analiticamente tratáveis, que possam acomodar valores de assimetria e curtose, e que incluam estritamente a distribuição normal, evitando assim, a necessidade de transformações. Nesse trabalho, será utilizada a distribuição normal assimétrica já discutida na Seção anterior.

Esta Seção está centrada na descrição do modelo misto com o uso da distribuição normal assimétrica multivariada apresentada em (2.1). Os autores utilizaram esta distribuição para representar a distribuição dos efeitos aleatórios, isto é,

$$u \sim SN_q(0, S(\sigma_u^2), \Delta_u) \qquad e \qquad \varepsilon \sim SN_n(0, \sigma_\varepsilon^2 I_n, \Delta_\varepsilon), \tag{2.7}$$

em que $\Delta_u \in \Re^{q\times q}$ e $\Delta_\varepsilon \in \Re^{n\times n}$ são matrizes diagonais com elementos $\delta_{u_1}, \ldots,$ δ_{u_q} e $\delta_{\varepsilon_1}, \ldots, \delta_{\varepsilon_n}$, repectivamente, correspondentes aos parâmetros de assimetria. Foi analisado o caso particular em que $\Delta_u = \delta_u I_q$ e $\Delta_\varepsilon = \delta_\varepsilon I_n$, com $\delta_u \in \Re$, $\delta_\varepsilon \in \Re$, I_q e I_n, matrizes identidades de dimensões $q \times q$ e $n \times n$, respectivamente.

Arellano-Valle et al. (2007) trataram a inferência dos parâmetros no enfoque bayesiano e denominaram este modelo como modelo misto normal assimétrico. Além disso, mencionaram que a inferência dos parâmetros no enfoque bayesiano oferece a vantagem de fornecer estimadores e algoritmos mais eficientes computacionalmente que no enfoque frequentista.

Conforme em (2.4) e utilizando a Proposição 2 pode-se escrever de forma hierárquica o modelo (2.2) da seguinte maneira

$$\begin{aligned} Y|\beta, u, \sigma_\varepsilon^2, \delta_\varepsilon &\sim SN_n(X\beta + Zu, \sigma_\varepsilon^2 I_n, \delta_\varepsilon I_n) \qquad e \\ u|\sigma_u^2, \delta_u &\sim SN_q(0, S(\sigma_u^2), \delta_u I_q). \end{aligned} \tag{2.8}$$

Esse modelo misto normal assimétrico proposto por Arellano-Valle et al. (2007) será adapatado para o contexto de microarrays e será denominado por modelo

aditivo-dominante normal assimétrico.

Antes de apresentar o modelo aditivo-dominante normal assimétrico, serão apresentados na Seção a seguir alguns conceitos que estarão incorporados no modelo aditivo-dominante normal assimétrico.

2.4 Valores genéticos

Os valores genéticos observados podem ser obtidos dos valores fenotípicos (Y), que se dividem em partes atribuídas a diferentes causas. São elas os valores genotípicos, efeitos que os genótipos produzem na expressão dessa característica e resíduos, todas as circunstâncias não genéticas que influenciam o valor fenotípico (Falconer & Mackay, 1996). A equação a seguir expressa este relacionamento

$$Y = \mu + g + \varepsilon, \tag{2.9}$$

em que Y representa os valores fenotípicos, μ é a média populacional da resposta, g é o valor ou efeito genotípico e ε é o resíduo. Assume-se μ como uma constante e, nesse trabalho, os valores g e ε como efeitos aleatórios.

Os valores Y, g e ε podem ser expressos em quaisquer unidades que representem uma propriedade biológica que possa ser medida de maneira contínua, tal como intensidade da expressão gênica, peso, altura, etc.

No modelo aditivo-dominante que será apresentado na Seção a seguir, supõe-se que infinitos locos (denominado poligenes) com alelos de efeito aditivo estejam segregando. Assim, os efeitos aditivos[4] são obtidos pela soma dos efeitos do par de alelos de todos os locos, que constitui o genótipo da sonda em questão, isto é, $a = \sum_{i=1}^{\infty} \alpha_i$. Além disso, supõe-se que infinitos locos com interação de tipo dominante estejam também segregando. Assim os efeitos dominantes são a soma

[4]também conhecido como valor genético aditivo, em inglês *breeding value*

dos desvios de dominância de cada loco i dados por $\delta_i = \mu_{A_iA_i} - (\mu_{A_iA_i} + \mu_{a_ia_i})/2$, ou seja, $d = \sum_{i=1}^{\infty} \delta_i$.

A Figura 10 apresenta uma idéia geométrica da composição de cada α_i e δ_i.

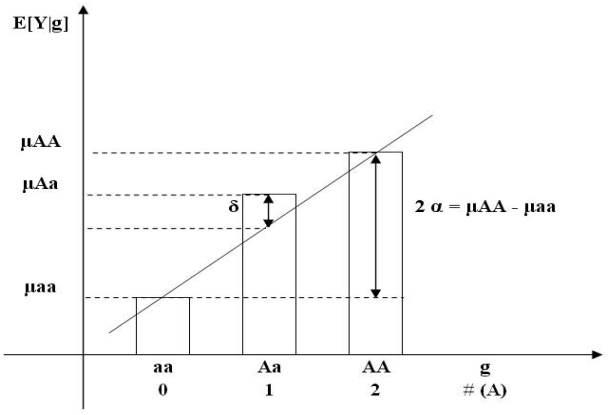

FIGURA 10 Representação geométrica dos valores genotípicos.

Supondo-se que a característica fenotípica medida (intensidade da expressão gênica) depende da composição genética de cada indivíduo (a soma dos efeitos a e d), como tem-se n indivíduos, tem-se a e d com dimensão $n \times 1$.

Usualmente, os efeitos aleatórios a e d possuem distribuição normal. Aqui, assume-se que esses efeitos possuem distribuição normal assimétrica, com vetor de médias iguais a zero ($0_{n\times 1}$); variâncias iguais a $\sigma_a^2 A$ e $\sigma_d^2 D$ e assimetrias iguais a $\Delta_a = \delta_a I_n$ e $\Delta_d = \delta_d I_n$, respectivamente.

A matriz A reflete as identidades alélicas entre indivíduos de mesma ascendência. A construção dessa matriz envolve o coeficiente de parentesco, uma matriz ϕ, também chamada de IBD (*Identity by Descent*), com elementos ϕ_{ij}, multiplicado por 2, isto é, $A = 2\phi$. Em particular, para os dados reais analisados neste trabalho, essa matriz tem dimensão 194×194 e é obtida computacionalmente através do pacote *kinship* do R. O coeficiente de parentesco (ϕ) é dado conforme o grau de

parentesco entre os indivíduos. A Tabela 1 apresenta o coeficiente de parentesco para alguns graus de parentesco, isto é, contém todos os elementos ϕ_{ij}, somente para $i \neq j$. Quando $i = j$, $\phi_{ij} = 1/2$.

TABELA 1 Coeficiente ϕ_{ij} para alguns graus de parentesco.

Relacionamento	ϕ_{ij}
Sem relação familiar	0
Indivíduo com ele mesmo	$\frac{1}{2}$
Irmãos	$\frac{1}{4}$
Pai e Filho	$\frac{1}{4}$
Avô e Neto	$\frac{1}{8}$

Fonte: Lynch & Walsh (1998, p. 145).

A matriz D, em particular, de dimensão 194×194 (devido aos dados reais considerados nesse trabalho) é uma matriz que contém a probabilidade esperada do par de indivíduos i e j compartilharem exatamente dois alelos IBD para um dado loco (Lynch & Walsh, 1998), conforme a Tabela 2, que contém todos os elementos D_{ij}, somente para $i \neq j$. Quando $i = j$, $D_{ij} = 1$.

A obtenção da matriz correta de parentescos entre os indivíduos dois a dois, de forma a obter uma relação adequada das covariâncias entre todos os indivíduos é um passo muito importante para a especificação do modelo aditivo-dominante.

TABELA 2 Construção da matriz D para alguns graus de parentesco.

Relacionamento	D_{ij}
Indivíduo com ele mesmo	1
Irmãos	$\frac{1}{8}$
Pai e Filho	0
Avô e Neto	0
Sem relação familiar	0

Fonte: Lynch & Walsh (1998, p. 145).

Outra suposição adotada é que a variação genética dos indivíduos de uma população será o resultado da variação genotípica devido aos efeitos de aditividade e dominância. Em termos de variância, tem-se

$$Var(Y) = Var(a) + Var(d) + Var(\varepsilon). \tag{2.10}$$

Para os programas de melhoramento genético, normalmente a variância genética aditiva é a mais importante, pois é aquela que traduz os efeitos de seleção herdáveis, como média dos pais. Esta variância é um reflexo da semelhança entre parentes. Desta forma, a covariância entre o efeito aditivo com os valores fenotípicos dividido pela variância fenotípica é a herdabilidade no sentido restrito (Kempthorne, 1973, p. 507; Falconer & Mackay, 1996, p. 123), representada por

$$h^2_{restrito} = \frac{Cov(a, Y)}{Var(Y)}, \tag{2.11}$$

em que $Cov(a, Y)$ é a covariância entre o efeito aditivo e os valores fenotípicos e $Var(Y)$ é como em (2.10). Esta medida também é conhecida como coeficiente de correlação intraclasse.

Além da herdabilidade no sentido restrito, também existe a herdabilidade no sentido amplo, representada por h^2_{amplo} (Falconer & Mackay, 1996, p. 123).

A herdabilidade no sentido amplo (h^2_{amplo}) expressa a proporção da variância total ($Var(Y)$) que é atribuível aos efeitos genotípicos ($a + d$). Em termos métricos, tem-se que

$$h^2_{amplo} = \frac{Cov(g, Y)}{Var(Y)}, \tag{2.12}$$

em que $Cov(g, Y)$ é a covariância entre os efeitos genotípicos e os valores fenotípicos e $Var(Y)$ é como em (2.10).

Segundo White & Hodge (1992), uma aproximação inicial do problema de predição é considerar a herdabilidade como coeficiente de predição. Para essa aplicação, o cálculo das herdabilidades nos sentidos restrito e amplo auxiliará a identificar quais sondas sinalizam presença de componente herdável. Cabe destacar que a herdabilidade é válida apenas para a população na qual ela foi medida, pois o seu cálculo depende da arquitetura genética e das frequências gênicas da população.

Após a introdução dos conceitos sobre valores genéticos, será apresentado na próxima Seção o modelo aditivo-dominante, que utiliza-se destes conceitos e é uma adaptação do modelo misto normal assimétrico apresentado por Arellano-Valle et al. (2007).

2.5 Modelo aditivo-dominante normal assimétrico

O modelo aditivo-dominante encontrado em Sorensen & Gianola (2002) é expresso da seguinte forma

$$Y = X\beta + Za + Wd + \varepsilon, \tag{2.13}$$

em que Y de dimensão $n \times 1$ é um vetor de respostas, X de dimensão $n \times p$ é uma matriz de incidência dos efeitos fixos, β de dimensão $p \times 1$ é o vetor dos efeitos fixos, Z de dimensão $n \times q_a$ é uma matriz de incidência dos efeitos aditivos, a de dimensão $q_a \times 1$ é o vetor dos efeitos aditivos, W de dimensão $n \times q_d$ é uma matriz de incidência dos efeitos dominantes, d de dimensão $q_d \times 1$ é o vetor dos efeitos dominantes e ε é o vetor de erros aleatórios ou resíduos de dimensão $n \times 1$. Tipicamente, assume-se que os efeitos aleatórios a, d e os erros aleatórios ε são independentes com

$$a \sim N_{q_a}(0, S_a), \qquad d \sim N_{q_d}(0, S_d) \qquad e \qquad \varepsilon \sim N_n(0, \Psi), \tag{2.14}$$

em que $S_a = \sigma_a^2 A$, $S_d = \sigma_d^2 D$ e $\Psi = \sigma_\varepsilon^2 I_n$ são matrizes de covariâncias entre os efeitos aleatórios $a_1, \ldots, a_{q_a}$, $d_1, \ldots, d_{q_d}$ e os resíduos $\varepsilon_1, \ldots, \varepsilon_n$, respectivamente, com A e D calculadas conforme as Tabelas 1 e 2, apresentadas na Seção anterior.

Note que sob as suposições consideradas em (2.14) o modelo apresentado em (2.13) pode ser representado hierarquicamente como

$$\begin{aligned} Y|\beta, a, d, \sigma_\varepsilon^2 &\sim N_n(X\beta + Za + Wd, \sigma_\varepsilon^2 I_n); \\ a|\sigma_a^2 &\sim N_{q_a}(0, S_a) \quad e \\ d|\sigma_d^2 &\sim N_{q_d}(0, S_d). \end{aligned} \tag{2.15}$$

A proposta é utilizar esse modelo hierárquico, considerando que os efeitos genéticos a e d e os resíduos tenham distribuição normal assimétrica multivariada (2.1) no contexto de microarrays. Basicamente, o modelo aditivo-dominante com o uso desta versão da distribuição normal assimétrica é um modelo adaptado do modelo apresentado por Arellano-Valle et al. (2007) e será denominado por modelo aditivo-dominante normal assimétrico (MADSN).

Como no modelo misto normal assimétrico, tem-se as seguintes suposições para o MADSN

$$a \sim SN_{q_a}(0, S_a, \Delta_a),\ d \sim SN_{q_d}(0, S_d, \Delta_d) \text{ e } \varepsilon \sim SN_n(0, \sigma_\varepsilon^2 I_n, \Delta_\varepsilon), \quad (2.16)$$

em que $\Delta_a \in \Re^{q_a \times q_a}$, $\Delta_d \in \Re^{q_d \times q_d}$ e $\Delta_\varepsilon \in \Re^{n \times n}$ são matrizes diagonais com elementos $\delta_{a_1}, \ldots, \delta_{a_{q_a}}$, $\delta_{d_1}, \ldots, \delta_{d_{q_d}}$ e $\delta_{\varepsilon_1}, \ldots, \delta_{\varepsilon_n}$, repectivamente, correspondentes aos parâmetros de assimetria. Assume-se que $\Delta_a = \delta_a I_{q_a}$, $\Delta_d = \delta_d I_{q_d}$ e $\Delta_\varepsilon = \delta_\varepsilon I_n$, com $\delta_a \in \Re$, $\delta_d \in \Re$ e $\delta_\varepsilon \in \Re$.

Conforme em (2.15) e utilizando a Proposição 2 pode-se escrever de forma hierárquica o modelo (2.13) da seguinte maneira

$$\begin{aligned} Y|\beta, a, d, \sigma_\varepsilon^2, \delta_\varepsilon &\sim SN_n(X\beta + Za + Wd, \sigma_\varepsilon^2 I_n, \delta_\varepsilon I_n), \\ a|\sigma_a^2, \delta_a &\sim SN_{q_a}(0, S_a, \delta_a I_{q_a}) \qquad e \\ d|\sigma_d^2, \delta_d &\sim SN_{q_d}(0, S_d, \delta_d I_{q_d}). \end{aligned} \quad (2.17)$$

Logo, a densidade condicional do vetor aleatório Y nos efeitos aleatórios (verossimilhança) é dada analiticamente por

$$\begin{aligned} f(y|\beta, a, d, \sigma_\varepsilon^2, \delta_\varepsilon) &= 2^n \phi_n(y|X\beta + Za + Wd, (\sigma_\varepsilon^2 + \delta_\varepsilon^2) I_n) \times \\ &\quad \Phi_n\left(\frac{\delta_\varepsilon}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}(y - X\beta - Za - Wd)|0, \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \delta_\varepsilon^2} I_n\right). \end{aligned} \quad (2.18)$$

O principal interesse é fazer inferências sobre o vetor de parâmetros $\theta = (\beta^\top, a^\top, d^\top, \sigma^2_\varepsilon, \sigma^2_a, \sigma^2_d, \delta_\varepsilon, \delta_a, \delta_d)$.

No Capítulo 3 será apresentado o modelo aditivo-dominante normal assimétrico adaptado ao conjunto de dados reais, a modelagem bayesiana para inferência nos parâmetros e a sua implementação computacional. Como a inferência sobre o vetor de parâmetros do modelo será feita numa perspectiva bayesiana, a Seção a seguir apresenta uma breve revisão dos principais conceitos teóricos que serão utilizados nesse trabalho sobre inferência bayesiana.

2.6 Inferência bayesiana

Segundo Pena (2006) "o filósofo Richard Price (1723-1791) apresentou à Royal Statistical Society em 1763 um artigo encontrado entre os papéis de seu amigo, o reverendo Thomas Bayes (1701-1761). Tal artigo foi nomeado "*An essay towards solving a problem in the doctrine of chances*" e continha a demonstração do teorema de Bayes. O matemático francês Pierre-Simon de Laplace (1749-1827) difundiu o artigo mundialmente. Harold Jeffreys, em meados do século XX, conseguiu dar ao bayesianismo um status lógico e solucionar problemas que não haviam sido resolvidos até aquele tempo com a utilização das noções bayesianas. A partir daí, cada vez mais surgiram estudiosos interessados na inferência bayesiana".

A inferência bayesiana leva em conta o conceito de probabilidade subjetiva, isto é, ela mede o grau de incerteza que se tem sobre a ocorrência de um determinado evento do espaço amostral, utilizando distribuições de probabilidades, denominadas distribuições a priori.

O conhecimento prévio que o pesquisador tem sobre o parâmetro θ é incorporado à análise por meio da densidade a priori $\pi(\theta)$, a qual deve representar a distribuição de probabilidades desse parâmetro antes da observação dos dados

(Box & Tiao, 1992). O parâmetro pode ser um escalar ou um vetor de parâmetros ($\theta = (\theta_1, \theta_2, \ldots, \theta_p)^\top$).

Os parâmetros indexadores das distribuições a priori são chamados de hiperparâmetros, distinguindo-os assim dos parâmetros de interesse. Inicialmente, os hiperparâmetros são considerados conhecidos e traduzem a informação que se tem sobre o parâmetro, antes da realização da amostra.

As distribuições a priori podem ser informativas ou não informativas. Quando o pesquisador tem alguma informação prévia sobre o parâmetro em questão, ele pode usar uma priori informativa, descrevendo uma densidade $\pi(\theta)$ que represente essa informação. Quando se tem pouca ou nenhuma informação sobre o parâmetro, pode-se usar uma priori não informativa. A idéia é pensar em todos os valores como igualmente prováveis, ou seja, com uma distribuição a priori uniforme.

Uma importante classe de prioris informativas são as prioris conjugadas, as quais conduzem a uma distribuição a posteriori que pertence à mesma classe de distribuições a priori, envolvendo para isso apenas uma atualização nos parâmetros. A tratabilidade analítica das expressões é uma grande vantagem decorrente do uso de prioris conjugadas.

A informação sobre θ, resumida pela distribuição a priori, pode ser aumentada observando-se uma quantidade aleatória y relacionada com θ, sendo o teorema de Bayes a regra de atualização dessa informação. Na análise bayesiana toda a inferência é feita a partir da distribuição a posteriori $\pi(\theta|y)$, obtida pelo teorema de Bayes conforme a expressão a seguir.

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

em que $L(\theta|y) = \prod_i f(y_i|\theta)$ é a função de verossimilhança.

Note que essa expressão do teorema de Bayes possui y, que é o vetor de observações, $L(\theta|y)$, que é a função de verossimilhança e $\pi(\theta)$, que é a densidade a priori de θ. O denominador da expressão funciona como uma constante normalizadora, já que não depende de θ. Assim, desconsiderando-se a constante, a distribuição a posteriori é proporcional ao produto entre priori e verossimilhança, isto é,

$$\pi(\theta|y) \propto L(\theta|y)\pi(\theta).$$

No caso de θ ser um vetor de parâmetros, $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^\top$, as distribuições marginais dos componentes θ_i, $(i = 1, 2, \ldots, p)$, a partir das quais as inferências para cada parâmetro serão feitas, podem ser obtidas integrando a densidade conjunta a posteriori $\pi(\theta|y)$ em relação aos demais parâmetros do vetor, ou seja,

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

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

A resolução analítica desta integral, em geral, não é trivial. Uma alternativa, nesse caso, é utilizar o método de Monte Carlo via Cadeias de Markov (MCMC), que é baseado em simulação estocástica e fornece aproximações que serão tanto melhores quanto maior for o número de valores gerados. A Seção a seguir fornece mais informações sobre o MCMC.

2.6.1 Método de Monte Carlo via cadeia de Markov

Esta Seção aborda o método de Monte Carlo baseado na simulação de Cadeias de Markov, cuja distribuição estacionária é a distribuição a posteriori de interesse. Os pioneiros no assunto foram Gelfand & Smith (1990). Para implementar este método, há necessidade de saber construir cadeias de Markov com distribuições de equilíbrio específicas. Metropolis et al. (1953) desenvolveram um algoritmo

para esse efeito, o qual foi mais tarde generalizado por Hastings (1970), sendo atualmente conhecido na literatura por "algoritmo de Metropolis-Hastings". A associação deste algoritmo para a simulação de distribuições com o método de Monte Carlo para a aproximação de integrais conduz ao chamado "Método de Monte Carlo via Cadeias de Markov" (MCMC). Um caso particular deste método é o método de amostragem Gibbs (ou amostrador de Gibbs), o qual foi introduzido por Geman & Geman (1984).

Uma cadeia de Markov é um processo estocástico, no qual o próximo estado da cadeia depende somente do estado atual e dos dados e não da história passada da cadeia. Além disso, é suposto que as primeiras cadeias são influenciadas pelo estado inicial e, por isso, são descartadas. Este período é conhecido como aquecimento da cadeia ou *burn-in*. Também é considerado existir uma dependência entre as observações subsequentes da cadeia e, para se obter uma amostra independente, as observações finais devem ser obtidas a cada J iterações, sendo este valor conhecido como salto, *thin*, *lag* ou intervalo de amostragem (Gamerman, 1997).

Basicamente, o método MCMC diz que a distribuição condicional completa de cada parâmetro, denotada por $\pi(\theta_i|\theta_{-i},y)$ é obtida considerando que, na densidade conjunta, os demais parâmetros θ_{-i} são constantes e, assim, podem ser desconsiderados, levando a uma expressão menos complexa.

Em alguns casos, a expressão da condicional completa tem a forma de uma densidade conhecida e é fácil de ser amostrada diretamente. Nestes casos, o método de simulação a ser utilizado é o amostrador de Gibbs. No entanto, existem casos em que sua expressão não tem forma de uma densidade conhecida ou fácil de ser amostrada. Em tais situações, destaca-se o algoritmo de Metropolis-Hastings.

Para esse trabalho as condicionais completas a posteriori possuem distribuições conhecidas, logo, será utilizado apenas o amostrador de Gibbs, que está descrito a

seguir.

Para mais detalhes sobre a teoria dos métodos MCMC, ver Gamerman (1997) e suas aplicações em genética quantitativa, ver Sorensen & Gianola (2002) e Sorensen (2009).

2.6.2 Amostrador de Gibbs

Segundo Paulino et al. (2003), Geman & Geman (1984) introduziram o amostrador de Gibbs como um algoritmo de simulação de distribuições multivariadas complexas e de dimensão elevada, que aparecem em problemas de reconstrução de imagens. Gelfand & Smith (1990) mostraram como esse algoritmo pode ser usado para simular distribuições a posteriori e, consequentemente, a sua importância na resolução de problemas de inferência bayesiana.

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

Cada um dos componentes θ_i pode ser um escalar ou um vetor. A distribuição não precisa ser uma distribuição a posteriori e o método pode ser aplicado em qualquer contexto de integração numérica para a obtenção da distribuição conjunta a partir de distribuições condicionais. No caso da inferência bayesiana, a distribuição $\pi(\theta|y)$ corresponde à distribuição a posteriori conjunta. O interesse aqui é gerar amostras da densidade conjunta, mas, sendo esta geração extremamente complicada, o algoritmo de Gibbs fornece uma forma alternativa de gerações por meio das distribuições condicionais completas.

Considerando-se, ainda, que as densidades condicionais completas a posteriori $\pi(\theta_i|\theta_{-i}, y)$, i = 1, ..., p estejam disponíveis, o algoritmo é descrito de acordo com os seguintes ítens:

1 - iniciar o contador de iterações da cadeia t = 1 e estabelecer valores iniciais $\theta^0 = (\theta_1^0, \theta_2^0, \ldots, \theta_p^0)$;

2 - obter um novo valor $\theta^t = (\theta_1^t, \theta_2^t, \ldots, \theta_p^t)^\top$ a partir de θ^{t-1} por meio de sucessivas gerações de valores:

$$\begin{aligned}
\theta_1^{(t)} &\sim \pi(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)} \ldots, \theta_p^{(t-1)}) \\
\theta_2^{(t)} &\sim \pi(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_p^{(t-1)}) \\
&\vdots \\
\theta_p^{(t)} &\sim \pi(\theta_p|\theta_1^{(t)}, \theta_2^{(t)} \ldots, \theta_{p-1}^{(t)})
\end{aligned}$$

3 - mudar o contador t para $t + 1$ e voltar ao passo 2 até que a convergência seja alcançada.

Após obter uma amostra de θ, há necessidade de averiguar a partir de que iteração se pode considerar que se atingiu convergência para a distribuição de equilíbrio. A Seção a seguir descreve apenas algumas das estratégias mais utilizadas na avaliação da covergência.

2.6.3 Diagnóstico de convergência

Um dos objetivos da amostragem via métodos MCMC é a convergência para a distribuição marginal dos parâmetros. Assim, torna-se muito importante o monitoramento dessa convergência, evitando um número excessivo ou insuficiente de iterações no processo de amostragem. Vários trabalhos são dedicados exclusivamente aos critérios de avaliação dessa convergência (Gelman & Rubin, 1992 a; Gelman & Rubin, 1992 b; Geweke, 1992; Raftery & Lewis, 1992; citados e avaliados por Nogueira et al., 2004).

Para avaliar a convergência, existem alguns testes implementados no pacote *coda*[5] (*"Output Analysis and Diagnostics for MCMC"*) que podem ser utiliza-

[5] http://cran.r-project.org/web/packages/coda/coda.pdf

dos. Este pacote pode ser instalado no software livre R e contém um conjunto de funções para a análise de convergência das cadeias geradas. Adota-se aqui o critério de Raftery & Lewis (1992), que se baseia na qualidade da estimativa do quantil. Assim, determina-se o número de iterações necessárias para se estimar o quantil q dentro de uma acurácia de $\pm r$, com probabilidade p e uma tolerância ζ. Portanto, além do número de iterações, este teste também apresenta como resultado o número de iterações iniciais que devem ser descartadas (*burn-in*) e o salto de uma iteração a outra (*thin*) para se obter uma amostra independente. Outra saída importante é o fator de dependência, que é responsável pelo acréscimo multiplicativo ao número de iterações necessárias para se alcançar a convergência, sendo essa atingida quando esse valor é menor que 5. (Nogueira et al., 2004)

Após obter-se amostras das condicionais completas, o próximo passo é fazer inferências sobre os parâmetros. Para este fim, serão apresentados o intervalo HPD e o fator de Bayes nas próximas Seções.

2.6.4 Intervalo HPD

Um resumo da distribuição a posteriori mais informativo que qualquer estimativa pontual é obtido de uma região do espaço paramétrico Θ que contenha uma parte substancial desta distribuição (Paulino et al., 2003). Assim, um intervalo (a, b) é chamado de intervalo de credibilidade, com um nível de credibilidade de $100(1-\alpha)\%$ para θ, se

$$\int_a^b \pi(\theta|y)d\theta = 1-\alpha,$$

com $0 \leq \alpha \leq 1$.

Dada a infinidade de intervalos de credibilidade com o mesmo nível de credibilidade $100(1-\alpha)\%$, interessa selecionar aquele com a menor amplitude possível.

Os intervalos de amplitude mínima são obtidos tomando-se os valores de θ com maior densidade a posteriori. Essa região é chamada de HPD (*Highest Probability Density Interval*) ou intervalo de máxima densidade a posteriori. Assim, quanto menor for o tamanho do intervalo, mais concentrada é a distribuição do parâmetro, pois a amplitude do intervalo também informa a dispersão de θ.

2.6.5 Fator de Bayes

Os intervalos de credibilidade são muitas vezes usados como meios de construir testes bayesianos de significância, pois após obter-se uma estimativa intervalar de determinado parâmetro, pode-se testar se realmente o modelo necessita do mesmo.

Em aplicações práticas, em geral, existem incertezas com relação ao modelo que melhor se ajusta ao conjunto de dados, sendo necessário algum método que auxilie a tomada de decisão em favor do modelo mais plausível. Um critério comumente utilizado para esse fim é o fator de Bayes (Kass & Raftery, 1995).

Na prática, podem existir K modelos de interesse. O fator de Bayes auxilia a selecionar entre dois modelos, qual seria o mais plausível para representar os dados. Para tanto, é necessário testar dois modelos dentre os K possíveis, depois outros 2 e assim por diante. Os modelos serão representados por H_0 e H_1. Em seguida, definir as prioris e posterioris para os parâmetros dos dois modelos H_0 e H_1, que serão denotadas por $\pi(H_0)$ e $\pi(H_1)$ e $\pi(H_0|y)$ e $\pi(H_1|y)$, respectivamente.

Assim, o fator de Bayes é definido por

$$FB_{H_0,H_1} = \frac{f(y|H_0)}{f(y|H_1)}, \tag{2.19}$$

em que $f(y|H_0)$ e $f(y|H_1)$ são as verossimilhanças marginais de cada modelo (ou

distribuições preditivas), obtidas da seguinte forma

$$f(y|H_k) = \int L_{H_k}(\theta|y)\pi(\theta|H_k)d\theta,\ k = 0, 1, \tag{2.20}$$

em que $L_{H_k}(\theta|y) = \prod_{i=1}^{n} f_{H_k}(y_i|\theta)$ é a função de verossimilhança para o modelo H_k e $\pi(\theta|H_k)$ é a função de densidade a priori para θ sob o modelo H_k, $k = 0, 1$.

Portanto, pode-se interpretar o fator de Bayes como sendo uma medida da evidência a favor de um dos modelos considerados. Kass & Raftery (1995), sugerem interpretar o fator de Bayes através de uma calibragem, dividindo os possíveis valores do fator de Bayes em quatro intervalos e considerando 2 vezes o logaritmo natural do fator de Bayes, conforme a Tabela 3.

TABELA 3 Interpretação do fator de Bayes.

Valor do FB_{H_0,H_1}	$2ln(FB_{H_0,H_1})$	Evidências a favor do modelo H_0
1 a 3	0 a 2	Fraca
3 a 20	2 a 6	Moderada
20 a 150	6 a 10	Forte
> 150	> 10	Muito Forte

Fonte: Kass & Raftery (1995).

Na Tabela 3, no primeiro intervalo (1 a 3), a evidência a favor do modelo H_0 é fraca. No segundo intervalo (3 a 20), a evidência a favor do modelo H_0 aumenta, favorecendo sua escolha. No terceiro intervalo (20 a 150), a escolha do modelo H_0 pode ser feita com mais confiança, pois há uma forte evidência a seu favor. E no quarto intervalo (> 150), a escolha do modelo H_0 deve ser feita.

Em algumas situações, o cálculo da integral em (2.20) não é trivial. Kass & Raftery (1995) descrevem vários métodos para aproximar a integral. Aqui utiliza-se a aproximação via método de Monte Carlo para aproximação de integrais.

Para simplificar, seja $I_k = f(y|H_k)$ e segundo Kass & Raftery (1995), I_k pode ser aproximada através do método de Monte Carlo fazendo

$$\hat{I}_k = \frac{1}{m}\sum_{i=1}^{m} L_{H_k}(\theta^{(i)}|y), \tag{2.21}$$

em que os $\theta^{(i)}$ são parâmetros gerados da distribuição a priori $\pi(\theta|H_k)$, m é a quantidade de $\theta^{(i)}$ gerados, $L_{H_k}(\theta^{(i)}|y)$ é o valor da função de verossimilhança, do modelo H_k, dado o valor gerado a priori $\theta^{(i)}$, para $k = 0, 1$ e $i = 1, 2, \ldots, m$.

Para Geweke (1989), a precisão do método de Monte Carlo pode ser melhorada pelo método de *Importance Sampling* que consiste em gerar $\theta^{(i)}$ de uma densidade a posteriori $\pi(\theta^{(i)}|y, H_k)$, para $i = 1, 2, \ldots, m$ e obter a verossimilhança $(L_{H_k}(\theta^{(i)}|y))$ ponderada através de pesos w_i. Dessa forma, I_k pode ser aproximada por

$$\hat{I}_k = \frac{\sum_{i=1}^{m} w_i L_{H_k}(\theta^{(i)}|y)}{\sum_{i=1}^{m} w_i}, \tag{2.22}$$

em que $w_i = \frac{\pi(\theta^{(i)}|H_k)}{\pi(\theta^{(i)}|y,H_k)}$ e

$$\pi(\theta^{(i)}|y, H_k) = \frac{L_{H_k}(\theta^{(i)}|y)\pi(\theta^{(i)}|H_k)}{I_k}. \tag{2.23}$$

Substituindo (2.23) em (2.22), tem-se que

$$
\begin{aligned}
\hat{I}_k &= \frac{\sum_{i=1}^{m} \frac{\pi(\theta^{(i)}|H_k)}{\pi(\theta^{(i)}|y,H_k)} L_{H_k}(\theta^{(i)}|y)}{\sum_{i=1}^{m} \frac{\pi(\theta^{(i)}|H_k)}{\pi(\theta^{(i)}|y,H_k)}} = \frac{\sum_{i=1}^{m} \frac{\pi(\theta^{(i)}|H_k) I_k}{L_{H_k}(\theta^{(i)}|y)\pi(\theta^{(i)}|H_k)} L_{H_k}(\theta^{(i)}|y)}{\sum_{i=1}^{m} \frac{\pi(\theta^{(i)}|H_k) I_k}{L_{H_k}(\theta^{(i)}|y)\pi(\theta^{(i)}|H_k)}} \\
&= \frac{\sum_{i=1}^{m} I_k}{\sum_{i=1}^{m} \frac{I_k}{L_{H_k}(\theta^{(i)}|y)}} = \frac{m I_k}{I_k \sum_{i=1}^{m} \frac{1}{L_{H_k}(\theta^{(i)}|y)}} = \frac{m}{\sum_{i=1}^{m} \left(L_{H_k}(\theta^{(i)}|y)\right)^{-1}} \\
&= \frac{1}{\frac{1}{m}\sum_{i=1}^{m} \left(L_{H_k}(\theta^{(i)}|y)\right)^{-1}} = \left[\frac{1}{m}\sum_{i=1}^{m} \left(L_{H_k}(\theta^{(i)}|y)\right)^{-1}\right]^{-1}, \qquad (2.24)
\end{aligned}
$$

em que os $\theta^{(i)}$ são parâmetros gerados da distribuição a posteriori $\pi(\theta^{(i)}|y,\, H_k)$, m é a quantidade de $\theta^{(i)}$ gerados, $L_{H_k}(\theta^{(i)}|y)$ é o valor da função de verossimilhança do modelo H_k, dado o valor gerado a posteriori $\theta^{(i)}$, para $k = 0, 1$ e $i = 1, 2, \ldots, m$.

Segundo Sorensen & Gianola (2002, p. 426 e 427), uma vantagem de utilizar esse estimador $\hat{I}_k$, também conhecido como estimador média harmônica das verossimilhanças, é que não precisa obter as distribuições marginais dos parâmetros a posteriori, basta utilizar as cadeias de Markov via métodos de Monte Carlo para fornecer amostras das condicionais completas a posteriori dos parâmetros de interesse. A desvantagem, porém, é a sua instabilidade numérica. A forma de (2.24) revela que os valores de $\theta^{(i)}$ com probabilidade muito pequena pode ter um forte impacto sobre o estimador. Kass & Raftery (1995) afirmam que, apesar da falta de estabilidade, o estimador é suficientemente preciso para a interpretação em uma escala logarítmica, sendo assim, uma possível estratégia poderia ser a seguinte. Seja

$$
\begin{aligned}
v_k &= \frac{1}{m}\sum_{i=1}^{m}\Big(L_{H_k}(\theta^{(i)}|y)\Big)^{-1} \\
&= \frac{1}{m}\sum_{i=1}^{m} S_k^{(i)}, \qquad (2.25)
\end{aligned}
$$

em que $S_k^{(i)} = \big(L_{H_k}(\theta^{(i)}|y)\big)^{-1}$. Obtenha o logaritmo de $S_k^{(i)}$ ($lnS_k^{(i)}$) e armazene esses valores. Então, uma vez que

$$
S_k^{(i)} = exp(lnS_k^{(i)}) = exp(lnS_k^{(i)} - c_k + c_k) = exp(lnS_k^{(i)} - c_k)\ exp(c_k),
$$

v_k pode ser escrito na forma

$$
v_k = \frac{1}{m}\sum_{i=1}^{m} exp\left(lnS_k^{(i)} - c_k\right) exp(c_k),
$$

em que c_k é o máximo de $(lnS_k^{(1)}, \ldots, lnS_k^{(m)})$.

Tomando o logaritmo de v_k tem-se que

$$
lnv_k = ln\left[\frac{1}{m}\sum_{i=1}^{m} exp\left(lnS_k^{(i)} - c_k\right)\right] + c_k. \qquad (2.26)
$$

Assim,

$$
ln[\hat{I}_k] = -lnv_k \Rightarrow \hat{I}_k = exp(-lnv_k). \qquad (2.27)
$$

Resumindo, o fator de Bayes seguirá os seguintes passos:

1. Gerar uma amostra de tamanho m de θ ($\theta^1, \ldots, \theta^m$) das condicionais completas a posteriori assumidas com o modelo H_0.

2. Substituir os parâmetros $\theta^{(i)}$ obtidos no Passo 1 na expressão $ln(S_0^{(i)}) = ln\left[\left(L_{H_0}(\theta^{(i)}|y)\right)^{-1}\right]$, para $i = 1, 2, \ldots, m$. Calcular c_0, que é o máximo de $\left(ln(S_0^{(1)}), \ldots, ln(S_0^{(m)})\right)$ e obter (2.26), para $k = 0$.

3. Gerar uma amostra de tamanho m de θ ($\theta^1, \ldots, \theta^m$) das condicionais completas a posteriori assumidas com o modelo H_1.

4. Substituir os parâmetros $\theta^{(i)}$ obtidos no Passo 3 na expressão $ln(S_1^{(i)}) = ln\left[\left(L_{H_1}(\theta^{(i)}|y)\right)^{-1}\right]$, para $i = 1, 2, \ldots, m$. Calcular c_1, que é o máximo de $\left(ln(S_1^{(1)}), \ldots, ln(S_1^{(m)})\right)$ e obter (2.26), para $k = 1$.

5. Calcular $\hat{I}_0$ e $\hat{I}_1$ através de (2.27) e obter o valor aproximado do fator de Bayes, dado por
$$\hat{FB}_{H_0,H_1} = \frac{\hat{I}_0}{\hat{I}_1}.$$

6. Para decidir qual o melhor modelo, olhar para o resultado de $\hat{FB}_{H_0,H_1}$, se for maior que 1, significa que o modelo H_0 é melhor que o modelo H_1.

No Capítulo 4 será apresentada a aplicação do fator de Bayes na escolha do melhor modelo para representar os dados reais.

3 MATERIAL E MÉTODOS

3.1 Descrição do conjunto de dados reais

O conjunto de dados reais a ser analisado foi fornecido no *Genetic Analysis Workshop* 15 (*GAW* 15), em 2006. Esses dados foram coletados em 2004 no estado de Utah nos Estados Unidos pelo Centro de Estudos de Polimorfismos Humano (CEPH - Centre d'Etude du Polymorphisme Humain). Este Centro de Estudos foi criado pelo Professor Jean Dausset em 1984 e a partir de 1993 tornou-se a Fundação Jean Dausset - CEPH, um instituto de pesquisa sem fins lucrativos, financiada em parte pela República da França.

O banco de dados contém dois arquivos, um denominado linkagePed e o outro linkagePhn. O primeiro é composto por 194 observações e 5 variáveis. As variáveis são: família (1333, 1340, 1341, 1345, 1346, 1347, 1362, 1408, 1416, 1418, 1421, 1423, 1424 e 1454), indivíduo (1, 2, $\cdots$, 194), pai, mãe e sexo (1: masculino e 2: feminino). Todas as famílias eram compostas de 14 indivíduos, exceto as famílias 1340 e 1345, com 13 indivíduos. A partir desse arquivo, pode-se ter uma idéia da genealogia (em inglês *pedigree*) dessas famílias, compostas por três gerações. Na Figura 11 ilustra-se a planilha de dados e a árvore genealógica da Família 1333, que contém 14 pessoas. Na representação, o círculo representa o sexo feminino e o quadrado, o sexo masculino.

O segundo arquivo contém 194 observações e 3556 variáveis. As variáveis são: família, indivíduo e 3554 sondas de células linfoblastóides do tipo B em lâminas do tipo *Affymetrix*®, isto é, são intensidades de expressões normalizadas utilizando o procedimento MAS da Affymetrix® (Cheung & Spielman, 2007).

FAMID	ID	FA	MO	SEX
1333	1	11	12	1
1333	2	13	14	2
1333	3	1	2	1
1333	4	1	2	1
1333	5	1	2	1
1333	6	1	2	2
1333	7	1	2	1
1333	8	1	2	1
1333	9	1	2	1
1333	10	1	2	1
1333	11	0	0	1
1333	12	0	0	2
1333	13	0	0	1
1333	14	0	0	2
...	...	...	...	...
1454	194	0	0	2

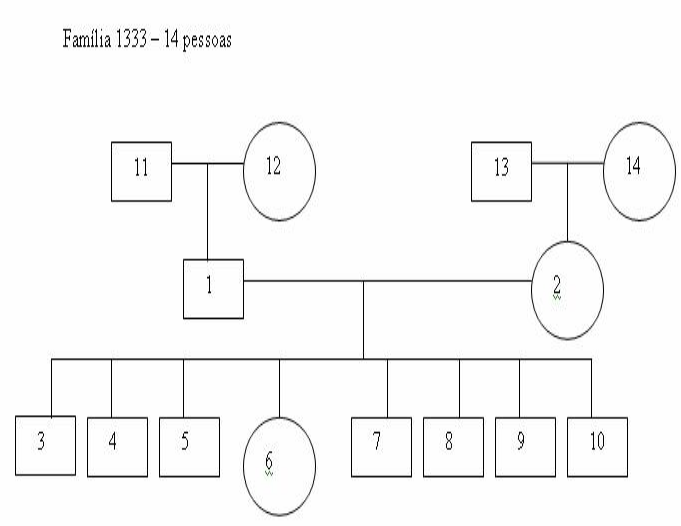

FIGURA 11 Arquivo de dados linkagePed do GAW 15 e ilustração da árvore genealógica da Família 1333.

Um dos problemas de pesquisa do GAW 15 foi investigar se a expressão gênica definida como fenótipo tem componente herdável. Por isso foram avaliados dados de famílias que não são comuns em experimentos com microarrays. Morley et al. (2004) usam um modelo de análise de variância clássico para comparar a variância dos níveis de expressão entre indivíduos não relacionados e entre réplicas do

mesmo indivíduo. Com base nesta metodologia simples estes autores conseguem reduzir de 8500 sondas para 3554 sondas informativas para a pesquisa de ligação dos níveis de expressão. Diversos trabalhos foram realizados com esse banco de dados, tais como o estudo do efeito da normalização na análise de ligação, ajuste de modelos multivariados para a análise conjunta das sondas, entre outros.

A proposta deste trabalho é utilizar três estratégias de análise: (i) o modelo misto com efeito aleatório de família, em que a variabilidade das respostas entre e dentro das famílias são comparadas sob uma estrutura de covariância uniforme para as respostas de indivíduos relacionados; (ii) o modelo misto com efeito aleatório aditivo, em que a covariância entre indivíduos é dada em função do grau de parentesco que os relaciona e (iii) o modelo misto com efeito aleatório aditivo e efeito aleatório dominante, também considerando a estrutura das famílias nas matrizes de covariâncias. Além disso, estudar todas as configurações possíveis de ajustes (assimetria apenas no efeito aleatório, assimetria apenas no resíduo e assimetria em ambos os efeitos) com estes três tipos de modelos, utilizando a distribuição normal assimétrica proposta por Arelano-Valle et al. (2007) numa perspectiva bayesiana para analisar esses dados de microarrays. As Seções a seguir apresentarão o ajuste desse modelo para os dados reais, a modelagem bayesiana para a realização das inferências nos parâmetros e, por fim, a implementação computacional.

3.2 Modelo aditivo-dominante normal assimétrico

Esta Seção apresenta o modelo aditivo-dominante apresentado no Capítulo 2, adaptado ao conjunto de dados reais. O seguinte modelo é ajustado para cada uma das 3554 sondas do banco de dados

$$Y = X\beta + Za + Wd + \varepsilon, \tag{3.1}$$

em que Y representa a intensidade da expressão gênica de dimensão 194×1, X é uma matriz que contém a incidência do efeito fixo (sexo) de dimensão 194×2, β é o vetor de efeitos fixos (sexo) de dimensão 2×1, Z é a matriz de incidência dos efeitos aleatórios (aditivos), sendo ela uma identidade de dimensão 194×194, a é o vetor de efeitos aleatórios aditivos de dimensão 194×1, W é a matriz de incidência dos efeitos aleatórios (dominantes), sendo ela uma identidade de dimensão 194×194, d é vetor de efeitos aleatórios dominantes, de dimensão 194×1 e ε, os resíduos de dimensão 194×1.

Conforme apresentado no Capítulo 2, assume-se que os efeitos aleatórios do modelo apresentam as seguintes distribuições

$$a|\sigma_a^2, \delta_a \quad \sim \quad SN_{194}(0, \sigma_a^2 A, \delta_a I_{194}), \tag{3.2}$$

$$d|\sigma_d^2, \delta_d \quad \sim \quad SN_{194}(0, \sigma_d^2 D, \delta_d I_{194}) \tag{3.3}$$

$$\varepsilon|\sigma_\varepsilon^2, \delta_\varepsilon \quad \sim \quad SN_{194}(0, \sigma_\varepsilon^2 I_{194}, \delta_\varepsilon I_{194}). \tag{3.4}$$

Apresenta-se a seguir como obter as matrizes A e D definidas no Capítulo 2.

Por exemplo, se houvesse duas famílias com a mesma estrutura familiar apre-

sentada na Figura 11, ou seja, 28 indivíduos, teríamos

$$A_{28\times 28} = \begin{bmatrix} A_{14\times 14} & 0_{14\times 14} \\ 0_{14\times 14} & A_{14\times 14} \end{bmatrix},$$

em que $0_{14\times 14}$ é uma matriz de zeros de dimensão 14×14 e a matriz $A_{14\times 14}$ de dimensão 14×14 é dada por

$$\begin{bmatrix}
1 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\
0 & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
\frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 1 & 0 & 0 & 0 \\
\frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 1 & 0 & 0 \\
0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 1 & 0 \\
0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 1
\end{bmatrix}.$$

Para exemplificar como obter a matriz D, considere duas famílias com a mesma estrutura familiar apresentada na Figura 11, ou seja, 28 indivíduos, assim tem-se que

$$D_{28\times 28} = \begin{bmatrix} D_{14\times 14} & 0_{14\times 14} \\ 0_{14\times 14} & D_{14\times 14} \end{bmatrix},$$

em que a matriz $D_{14\times 14}$ é dada por

$$\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{8} & 1 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{8} & \frac{1}{8} & 1 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 1 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 1 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 1 & \frac{1}{8} & \frac{1}{8} & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 1 & \frac{1}{8} & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}.$$

Note que o conjunto de dados reais é composto por 14 famílias, com um total de 194 indivíduos. Logo, essas matrizes serão de dimensão 194×194.

Através das suposições apresentadas em (3.2), (3.3) e (3.4) e utilizando a Proposição 2 apresentada no Capítulo 2, o modelo (3.1) pode ser escrito da seguinte

forma hierárquica:

$$\begin{aligned} Y|\beta, a, d, \sigma^2_\varepsilon, \delta_\varepsilon &\sim SN_{194}(X\beta + Za + Wd, \sigma^2_\varepsilon I_{194}, \delta_\varepsilon I_{194}), \\ a|\sigma^2_a, \delta_a &\sim SN_{194}(0, \sigma^2_a A, \delta_a I_{194}) \quad e \\ d|\sigma^2_d, \delta_d &\sim SN_{194}(0, \sigma^2_d D, \delta_d I_{194}). \end{aligned} \tag{3.5}$$

Note que a densidade condicional do vetor aleatório Y nos efeitos aleatórios (verossimilhança) é dada por (ver expressão (2.1))

$$\begin{aligned} f(Y|\beta, a, d, \sigma^2_\varepsilon, \delta_\varepsilon) = 2^{194}\phi_{194}(Y|X\beta + Za + Wd, (\sigma^2_\varepsilon + \delta^2_\varepsilon)I_{194}) \times \\ \Phi_{194}\left(\frac{\delta_\varepsilon}{\sigma^2_\varepsilon + \delta^2_\varepsilon}(Y - X\beta - Za - Wd)\middle| 0, \frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon + \delta^2_\varepsilon} I_{194}\right). \end{aligned} \tag{3.6}$$

O principal interesse é fazer inferências sobre o vetor de parâmetros $\theta = (\beta^\top, a, d, \sigma^2_\varepsilon, \sigma^2_a, \sigma^2_d, \delta_\varepsilon, \delta_a, \delta_d)^\top$, que será tratada na Seção a seguir, no enfoque bayesiano.

3.3 Modelagem bayesiana

Uma parte fundamental da análise bayesiana é especificar distribuições a priori para todos os parâmetros desconhecidos do modelo. Conforme Arellano-Valle et al. (2007), para garantir distribuições a posteriori próprias, adota-se o uso de prioris próprias para todas as quantidades desconhecidas do modelo. As prioris especificadas a seguir são análogas às especificadas por Arellano-Valle et al. (2007), exceto para os parâmetros de assimetria (δ), em que os mesmos sugeriram o uso da distribuição normal truncada positiva, mas na prática, dificilmente se saberá para qual lado se encontra a assimetria dos efeitos aleatórios e dos resíduos.

Assim, considera-se uma distribuição normal multivariada para a priori do ve-

tor de parâmetros β de dimensão $p \times 1$ com densidade

$$\pi(\beta|\beta_0, S_\beta) = \frac{1}{(2\pi)^{p/2}\sqrt{|S_\beta|}} \exp\Big[-\frac{1}{2}(\beta-\beta_0)^\top S_\beta^{-1}(\beta-\beta_0)\Big]. \quad (3.7)$$

Para os parâmetros de escala, σ^2, considera-se a distribuição gama inversa, $GI(\frac{\tau}{2}, \frac{T}{2})$, com densidade

$$\pi(\sigma^2|\tau, T) = \frac{\left(\frac{T}{2}\right)^{\frac{\tau}{2}}}{\Gamma\left(\frac{\tau}{2}\right)} \left(\frac{1}{\sigma^2}\right)^{\frac{\tau}{2}+1} \exp\left[-\frac{T}{2\sigma^2}\right]. \quad (3.8)$$

Para os parâmetros de assimetria δ assume-se a distribuição normal, com densidade

$$\pi(\delta|\mu, \gamma^2) = \frac{1}{\sqrt{2\pi}\gamma} \exp\Big[-\frac{1}{2\gamma^2}(\delta-\mu)^2\Big],$$

fazendo $\mu = 0$, que é uma escolha natural para o parâmetro de locação da priori do parâmetro de assimetria, tem-se

$$\pi(\delta|\gamma^2) = \frac{1}{\sqrt{2\pi}\gamma} \exp\Big[-\frac{1}{2}\left(\frac{\delta}{\gamma}\right)^2\Big]. \quad (3.9)$$

Considerando a distribuição condicional de Y apresentada de forma explícita em (3.6), as distribuições dos efeitos aleatórios apresentadas em (3.2) e (3.3) e as prioris especificadas em (3.7), (3.8) e (3.9), tem-se que a distribuição a posteriori conjunta de todas as quantidades envolvidas é dada por

$$
\begin{aligned}
&\pi(\beta, a, d, \sigma^2_\varepsilon, \sigma^2_a, \sigma^2_d, \delta_\varepsilon, \delta_a, \delta_d | y) \propto \phi_{194}(y | X\beta + Za + Wd, (\sigma^2_\varepsilon + \delta^2_\varepsilon) I_{194}) \\
&\times \Phi_{194}\left(\frac{\delta_\varepsilon}{\sigma^2_\varepsilon + \delta^2_\varepsilon}(y - X\beta - Za - Wd) \middle| 0, \frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon + \delta^2_\varepsilon} I_{194} \right) \\
&\times \phi_{194}(a | 0, \sigma^2_a A + \delta^2_a I_{194}) \\
&\times \Phi_{194}\left(\delta_a (\sigma^2_a A + \delta^2_a I_{194})^{-1} a \middle| 0, \left(I_{194} + \frac{\delta^2_a}{\sigma^2_a} A^{-1} \right)^{-1} \right) \\
&\times \phi_{194}(d | 0, \sigma^2_d D + \delta^2_d I_{194}) \\
&\times \Phi_{194}\left(\delta_d (\sigma^2_d D + \delta^2_d I_{194})^{-1} d \middle| 0, \left(I_{194} + \frac{\delta^2_d}{\sigma^2_d} D^{-1} \right)^{-1} \right) \\
&\times \exp\Big[-\frac{1}{2}(\beta - \beta_0)^\top S^{-1}_\beta (\beta - \beta_0) \Big] \\
&\times \left(\frac{1}{\sigma^2_\varepsilon} \right)^{\frac{\tau_\varepsilon}{2}+1} \exp\left[-\frac{T_\varepsilon}{2\sigma^2_\varepsilon} \right] \\
&\times \left(\frac{1}{\sigma^2_a} \right)^{\frac{\tau_a}{2}+1} \exp\left[-\frac{T_a}{2\sigma^2_a} \right] \\
&\times \left(\frac{1}{\sigma^2_d} \right)^{\frac{\tau_d}{2}+1} \exp\left[-\frac{T_d}{2\sigma^2_d} \right] \\
&\times \exp\Big[-\frac{1}{2} \left(\frac{\delta_\varepsilon}{\gamma_\varepsilon} \right)^2 \Big] \\
&\times \exp\Big[-\frac{1}{2} \left(\frac{\delta_a}{\gamma_a} \right)^2 \Big] \\
&\times \exp\Big[-\frac{1}{2} \left(\frac{\delta_d}{\gamma_d} \right)^2 \Big]. \qquad (3.10)
\end{aligned}
$$

Dada a forma algébrica da posteriori conjunta dada em (3.10), para ser mais fácil obter uma amostra desta ou de distribuições marginais de interesse, foi implementado o esquema MCMC. A seguir apresenta-se os passos necessários para implementar o amostrador de Gibbs, que é um caso especial do MCMC e necessita da especificação das condicionais completas a posteriori para cada parâmetro.

A fim de especificar o modelo (3.5) em uma estrutura conveniente para implementar o procedimento MCMC, usa-se a representação estocática apresentada na Proposição 1, tal que essas distribuições assimétricas possam ser representadas hierarquicamente como segue

$$\begin{aligned} Y|\beta, a, d, \sigma^2_\varepsilon, \delta_\varepsilon, w_\varepsilon &\sim N_n(X\beta + Za + Wd + \delta_\varepsilon w_\varepsilon, \sigma^2_\varepsilon I_n), \\ w_\varepsilon &\sim N_n(0, I_n)\mathtt{I}_{w_\varepsilon > 0}, \end{aligned} \tag{3.11}$$

$$\begin{aligned} a|\sigma^2_a, \delta_a, w_a &\sim N_{q_a}(\delta_a w_a, \sigma^2_a A), \\ w_a &\sim N_{q_a}(0, I_{q_a})\mathtt{I}_{w_a > 0}, \end{aligned} \tag{3.12}$$

$$\begin{aligned} d|\sigma^2_d, \delta_d, w_d &\sim N_{q_d}(\delta_d w_d, \sigma^2_d D), \\ w_d &\sim N_{q_d}(0, I_{q_d})\mathtt{I}_{w_d > 0}, \end{aligned} \tag{3.13}$$

em que $n = q_a = q_d = 194$. As variáveis w são as variáveis latentes com distribuição normal truncada positiva e $\mathtt{I}$ é uma função indicadora do domínio de variação de w.

Assim, a especificação do modelo completo é dada pela representação especificada de (3.11) a (3.13) e pelas prioris

$$\begin{aligned} &\beta \sim N_2(\beta_0, S_\beta), \\ &\sigma^2_\varepsilon \sim GI\Big(\frac{\tau_\varepsilon}{2}, \frac{T_\varepsilon}{2}\Big), \quad \sigma^2_a \sim GI\Big(\frac{\tau_a}{2}, \frac{T_a}{2}\Big) \quad \text{e} \quad \sigma^2_d \sim GI\Big(\frac{\tau_d}{2}, \frac{T_d}{2}\Big), \\ &\delta_\varepsilon \sim N(\mu_\varepsilon, \gamma^2_\varepsilon), \quad \delta_a \sim N(\mu_a, \gamma^2_a) \quad \text{e} \quad \delta_d \sim N(\mu_d, \gamma^2_d). \end{aligned} \tag{3.14}$$

Por meio do modelo completo especificado em (3.11)-(3.14) as condicionais

completas são facilmente obtidas, pois as mesmas são proporcionais ao produto da verossimilhança com a priori dos parâmetros envolvidos. As manipulações algébricas para a obtenção das condicionais completas dos parâmetros do modelo se encontram no Apêndice A. A seguir são apresentados os resultados das condicionais completas.

$$\beta|a, d, \sigma_\varepsilon^2, \delta_\varepsilon, w_\varepsilon, Y \sim N_p(M_\beta^{-1} m_\beta, M_\beta^{-1}), \tag{3.15}$$

em que $M_\beta = S_\beta^{-1} + X^\top X/\sigma_\varepsilon^2$ e $m_\beta = \beta_0 S_\beta^{-1} + X^\top(Y - Za - Wd - \delta_\varepsilon w_\varepsilon)/\sigma_\varepsilon^2$.

$$a|\beta, d, \sigma_\varepsilon^2, \sigma_a^2, \delta_\varepsilon, \delta_a, w_\varepsilon, w_a, Y \sim N_{q_a}(M_a^{-1} m_a, M_a^{-1}), \tag{3.16}$$

com $M_a = A^{-1}/\sigma_a^2 + Z^\top Z/\sigma_\varepsilon^2$ e $m_a = \delta_a A^{-1} w_a/\sigma_a^2 + Z^\top(Y - X\beta - Wd - \delta_\varepsilon w_\varepsilon)/\sigma_\varepsilon^2$.

$$d|\beta, a, \sigma_\varepsilon^2, \sigma_d^2, \delta_\varepsilon, \delta_d, w_\varepsilon, w_d, Y \sim N_{q_d}(M_d^{-1} m_d, M_d^{-1}), \tag{3.17}$$

com $M_d = D^{-1}/\sigma_d^2 + W^\top W/\sigma_\varepsilon^2$ e $m_d = \delta_d D^{-1} w_d/\sigma_d^2 + W^\top(Y - X\beta - Za - \delta_\varepsilon w_\varepsilon)/\sigma_\varepsilon^2$.

$$w_\varepsilon|\beta, a, d, \sigma_\varepsilon^2, \delta_\varepsilon, Y \sim N_n(M_{w_\varepsilon}^{-1} m_{w_\varepsilon}, M_{w_\varepsilon}^{-1})\mathtt{I}_{w_\varepsilon > 0}, \tag{3.18}$$

em que $M_{w_\varepsilon} = [\frac{\delta_\varepsilon^2}{\sigma_\varepsilon^2} + 1]I_n$ e $m_{w_\varepsilon} = \frac{\delta_\varepsilon}{\sigma_\varepsilon^2}(Y - X\beta - Za - Wd)$.

$$w_a|a, \sigma_a^2, \delta_a \sim N_{q_a}(M_{w_a}^{-1} m_{w_a}, M_{w_a}^{-1})\mathtt{I}_{w_a > 0}, \tag{3.19}$$

em que $M_{w_a} = \frac{\delta_a^2}{\sigma_a^2} A^{-1} + I_{q_a}$ e $m_{w_a} = \frac{\delta_a}{\sigma_a^2} A^{-1} a$.

$$w_d | d, \sigma_d^2, \delta_d \sim N_{q_d}(M_{w_d}^{-1} m_{w_d}, M_{w_d}^{-1}) \mathtt{I}_{w_d > 0}, \tag{3.20}$$

em que $M_{w_d} = \frac{\delta_d^2}{\sigma_d^2} D^{-1} + I_{q_d}$ e $m_{w_d} = \frac{\delta_d}{\sigma_d^2} D^{-1} d$.

$$\sigma_\varepsilon^2 | \beta, a, d, \delta_\varepsilon, w_\varepsilon, Y \sim GI(\frac{n + \tau_\varepsilon}{2}, \frac{T_\varepsilon + \mu_{\sigma_\varepsilon^\top} \mu_{\sigma_\varepsilon}}{2}), \tag{3.21}$$

com $\mu_{\sigma_\varepsilon} = Y - X\beta - Za - Wd - \delta_\varepsilon w_\varepsilon$.

$$\sigma_a^2 | a, \delta_a, w_a \sim GI(\frac{n + \tau_a}{2}, \frac{T_a + \mu_{\sigma_a^\top} A^{-1} \mu_{\sigma_a}}{2}), \tag{3.22}$$

em que $\mu_{\sigma_a} = a - \delta_a w_a$.

$$\sigma_d^2 | d, \delta_d, w_d \sim GI(\frac{n + \tau_d}{2}, \frac{T_d + \mu_{\sigma_d^\top} D^{-1} \mu_{\sigma_d}}{2}), \tag{3.23}$$

em que $\mu_{\sigma_d} = d - \delta_d w_d$.

$$\delta_\varepsilon | \beta, a, d, \sigma_\varepsilon^2, w_\varepsilon, Y \sim N(M_{\delta_\varepsilon}^{-1} m_{\delta_\varepsilon}, M_{\delta_\varepsilon}^{-1}), \tag{3.24}$$

em que $M_{\delta_\varepsilon} = \frac{1}{\gamma_\varepsilon^2} + \frac{w_\varepsilon^\top w_\varepsilon}{\sigma_\varepsilon^2}$ e $m_{\delta_\varepsilon} = \frac{w_\varepsilon^\top (Y - X\beta - Za - Wd)}{\sigma_\varepsilon^2}$.

$$\delta_a | a, \sigma_a^2, w_a \sim N(M_{\delta_a}^{-1} m_{\delta_a}, M_{\delta_a}^{-1}), \tag{3.25}$$

com $M_{\delta_a} = \frac{1}{\gamma_a^2} + \frac{w_a^\top A^{-1} w_a}{\sigma_a^2}$ e $m_{\delta_a} = \frac{w_a^\top A^{-1} a}{\sigma_a^2}$.

$$\delta_d | d, \sigma_d^2, w_d \sim N(M_{\delta_d}^{-1} m_{\delta_d}, M_{\delta_d}^{-1}), \tag{3.26}$$

com $M_{\delta_d} = \frac{1}{\gamma_d^2} + \frac{w_d^\top D^{-1} w_d}{\sigma_d^2}$ e $m_{\delta_d} = \frac{w_d^\top D^{-1} d}{\sigma_d^2}$.

Os cálculos das herdabilidades no sentido amplo e no sentido restrito, serão baseadas nas amostras das condicionais completas a posteriori de a e d e na variância do vetor observado y, isto é, a cada iteração as herdabilidades nos sentidos amplo e restrito serão calculadas, conforme as expressões

$$h^2_{amplo} = \frac{Cov(a+d,y)}{Var(y)} \quad \text{e} \quad h^2_{restrito} = \frac{Cov(a,y)}{Var(y)}. \tag{3.27}$$

Com esses valores, tem-se uma distribuição para representar as herdabilidades e estatísticas descritivas desta distribuição permitem inferir a respeito das herdabilidades das sondas consideradas.

Para implementar esta metodologia é necessário atribuir valores iniciais para todas as variáveis do modelo e as iterações geram amostras das distribuições condicionais apresentadas anteriormente até alcançar a convergência, que pode ser verificada e estudada através do pacote *coda* no software estatístico *R*, em que são usados os critérios apresentados no Capítulo 2. Os valores iniciais e os detalhes computacionais se encontram na Seção a seguir.

3.4 Implementação computacional

Para avaliar a metodologia proposta, utiliza-se o conjunto de dados reais disponibilizado no GAW 15 e foi feita a implementação computacional utilizando o software R.

Como foi apresentado na Seção 3.1, o conjunto de dados reais contém 3554 sondas para serem analisadas. Como o modelo aditivo-dominante normal assimétrico exige grande esforço computacional, optou-se por selecionar poucas sondas, para se explorar detalhadamente esse modelo e também compará-lo com outros

mais simples.

Para selecionar essas sondas, primeiramente, foram ajustados modelos mistos normais usuais para as 3554 sondas, como os apresentados no início da Seção 2.3 com β representando o sexo, como efeito fixo de dimensão 2×1 e u, as famílias, como efeitos aleatórios de dimensão 14×1, com $u \sim N(0, S_u)$, em que $S_u = I_{14\times14}$. Através dos ajustes destes modelos, foram obtidos os resíduos estimados (fazendo $y - X\hat{\beta} - Z\hat{u}$) para cada sonda e estabelecido o seguinte critério: selecionar as sondas que apresentarem valores altos para a assimetria dos resíduos, calculada da seguinte forma

$$assimetria = \frac{\sum_{i=1}^{194}(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^3}{n\ \hat{\sigma}_{\varepsilon}^3}$$

e, conforme Morley et al. (2004), valores altos para $\hat{h}_u^2 = \hat{\sigma}_u^2/\hat{\sigma}_y^2$, em que $\hat{\sigma}_u^2$ foi obtido utilizando o algoritmo EM e EMVR, conforme foi apresentado na Seção 2.3 e $\hat{\sigma}_y^2$ é a estimativa da variância amostral dos valores observados de y. Essa escolha foi feita com o objetivo de tentar captar sondas que apresentassem assimetria pelo menos no resíduo e também que fossem candidatas a apresentar componente herdável.

Com cada uma das sondas selecionadas foi feito o ajuste de três tipos de modelos:

1. Modelo misto com efeitos aleatórios de famílias (MMF),

2. Modelo Misto com efeitos aleatórios aditivos (MMA) e

3. Modelo Misto com efeitos aleatórios aditivos e dominantes (MMAD).

Para os três tipos de modelos foi considerado como variável resposta as intensidades de expressão gênica já normalizadas e o sexo, como efeito fixo.

O MMF, mais especificamente, $Y = X\beta + Zf + \varepsilon$, com β representando o sexo de dimensão 2×1, f representando a família de dimensão 14×1, ε, o vetor de resíduos de dimensão 194×1, X a matriz de incidência do sexo, de dimensão 194×2 e Z a matriz de incidência das famílias, de dimensão 194×14, foi ajustado para quatro tipos de configurações:

1. com distribuição normal assimétrica para f e para ε (MMFcafe);
2. com distribuição normal assimétrica apenas em f (MMFcaf);
3. com distribuição normal assimétrica apenas em ε (MMFcae);
4. com distribuição normal simétrica (sem assimetria) para f e para ε (MMFsa).

As verossimilhanças para os MMFcafe, MMFcaf, MMFcae e MMFsa são dadas por

$$
\begin{aligned}
&L_{MMFcafe}(\beta^\top, f^\top, \sigma_f^2, \sigma_\varepsilon^2, \delta_f, \delta_\varepsilon | y) = \\
&2^{194}\phi_{194}(y|X\beta + Zf, (\sigma_\varepsilon^2 + \delta_\varepsilon^2)I_{194}) \times \\
&\Phi_{194}\left(\frac{\delta_\varepsilon}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}(y - X\beta - Zf)|0, \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}I_{194}\right);
\end{aligned}
$$

$$
\begin{aligned}
&L_{MMFcaf}(\beta^\top, f^\top, \sigma_f^2, \sigma_\varepsilon^2, \delta_f | y) = \\
&\phi_{194}(Y|X\beta + Zf, \sigma_\varepsilon^2 I_{194});
\end{aligned}
$$

$$
\begin{aligned}
&L_{MMFcae}(\beta^\top, f^\top, \sigma_f^2, \sigma_\varepsilon^2, \delta_\varepsilon | y) = \\
&2^{194}\phi_{194}(y|X\beta + Zf, (\sigma_\varepsilon^2 + \delta_\varepsilon^2)I_{194}) \times \\
&\Phi_{194}\left(\frac{\delta_\varepsilon}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}(y - X\beta - Zf)|0, \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}I_{194}\right);
\end{aligned}
$$

$$L_{MMFsa}(\beta^\top, f^\top, \sigma_f^2, \sigma_\varepsilon^2|y) =$$
$$\phi_{194}(Y|X\beta + Zf, \sigma_\varepsilon^2 I_{194}). \tag{3.28}$$

O MMA, mais especificamente, $Y = X\beta + Za + \varepsilon$, com β representando o sexo de dimensão 2×1, a representando os efeitos aditivos de dimensão 194×1, ε, o vetor de resíduos de dimensão 194×1, X a matriz de incidência do sexo, de dimensão 194×2 e Z a matriz de incidência dos efeitos aditivos, de dimensão 194×194 também foi ajustado para as mesmas quatro configurações apresentadas para o MMF, a saber: MMAcaae, MMAcaa, MMAcae e MMAsa. Assim, as quatro verossimilhanças são iguais a (3.28), mas com Z de dimensão 194×194; no lugar de f, substituir por a de dimensão 194×1 e no lugar de δ_f substituir por δ_a.

O MMAD, mais especificamente, $Y = X\beta + Za + Wd + \varepsilon$ é como o MMA, com a adição dos elementos W, a matriz de incidência dos efeitos dominantes de dimensão 194×194 e d, representando os efeitos dominantes de dimensão 194×1. Para esse modelo, foram ajustadas 8 configurações:

1. com distribuição normal assimétrica em a, d e ε (MMADcaade);
2. com distribuição normal assimétrica em a e d (MMADcaad);
3. com distribuição normal assimétrica em a e ε (MMADcaae);
4. com distribuição normal assimétrica em d e ε (MMADcade);
5. com distribuição normal assimétrica em a (MMADcaa);
6. com distribuição normal assimétrica em d (MMADcad);

7. com distribuição normal assimétrica em ε (MMADcae);

8. com distribuição normal simétrica em a, d e ε (MMADsa).

As verossimilhanças para todos os modelos considerados são dadas por

$$
\begin{aligned}
&L_{MMADcaade}(\beta^\top, a^\top, d^\top, \sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2, \delta_a, \delta_d, \delta_\varepsilon | y) = \\
&\quad 2^{194}\phi_{194}(y|X\beta + Za + Wd, (\sigma_\varepsilon^2 + \delta_\varepsilon^2)I_{194}) \times \\
&\quad \Phi_{194}\left(\frac{\delta_\varepsilon}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}(y - X\beta - Za - Wd)|0, \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}I_{194}\right); \\
&L_{MMADcaad}(\beta^\top, a^\top, d^\top, \sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2, \delta_a, \delta_d | y) = \\
&\quad \phi_{194}(y|X\beta + Za + Wd, \sigma_\varepsilon^2 I_{194}); \\
&L_{MMADcaae}(\beta^\top, a^\top, d^\top, \sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2, \delta_a, \delta_\varepsilon | y) = \\
&\quad 2^{194}\phi_{194}(y|X\beta + Za + Wd, (\sigma_\varepsilon^2 + \delta_\varepsilon^2)I_{194}) \times \\
&\quad \Phi_{194}\left(\frac{\delta_\varepsilon}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}(y - X\beta - Za - Wd)|0, \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}I_{194}\right); \\
&L_{MMADcade}(\beta^\top, a^\top, d^\top, \sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2, \delta_d, \delta_\varepsilon | y) = \\
&\quad 2^{194}\phi_{194}(y|X\beta + Za + Wd, (\sigma_\varepsilon^2 + \delta_\varepsilon^2)I_{194}) \times \\
&\quad \Phi_{194}\left(\frac{\delta_\varepsilon}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}(y - X\beta - Za - Wd)|0, \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}I_{194}\right); \\
&L_{MMADcaa}(\beta^\top, a^\top, d^\top, \sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2, \delta_a | y) = \\
&\quad \phi_{194}(y|X\beta + Za + Wd, \sigma_\varepsilon^2 I_{194}); \\
&L_{MMADcad}(\beta^\top, a^\top, d^\top, \sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2, \delta_d | y) = \\
&\quad \phi_{194}(y|X\beta + Za + Wd, \sigma_\varepsilon^2 I_{194}); \\
&L_{MMADcae}(\beta^\top, a^\top, d^\top, \sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2, \delta_\varepsilon | y) = \\
&\quad 2^{194}\phi_{194}(y|X\beta + Za + Wd, (\sigma_\varepsilon^2 + \delta_\varepsilon^2)I_{194}) \times \\
&\quad \Phi_{194}\left(\frac{\delta_\varepsilon}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}(y - X\beta - Za - Wd)|0, \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}I_{194}\right);
\end{aligned}
$$

$$L_{MMADsa}(\beta^\top, a^\top, d^\top, \sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2|y) =$$
$$\phi_{194}(y|X\beta + Za + Wd, \sigma_\varepsilon^2 I_{194}). \quad (3.29)$$

Note que para se fazer um estudo mais detalhado, foram formuladas para cada sonda 16 configurações de modelos: 1) MMFcafe, 2) MMFcaf, 3) MMFcae, 4) MMFsa, 5) MMAcaae, 6) MMAcaa, 7) MMAcae, 8) MMAsa, 9) MMADcaade, 10) MMADcaad, 11) MMADcaae, 12) MMADcade, 13) MMADcaa, 14) MMADcad, 15) MMADcae e 16) MMADsa.

Para cada uma dessas 16 configurações, o primeiro passo foi atribuir valores iniciais a todos os parâmetros. Para o efeito fixo β foi atribuído um vetor de dimensão 2×1 com a média dos valores de Y repetido 2 vezes; para os σ^2 gerou-se um valor da $N(0,1)$ e elevou-se ao quadrado; para os δ gerou-se de uma distribuição normal com média 0 e variância 10.000 e, finalmente, uma normal multivariada $N_q(0,I_q)$, para gerar os vetores u com $q = 14$ e para a e d, com $q = 194$. Para os hiperparâmetros do modelo tomou-se $\beta_0 = \beta$, S_β uma matriz 2×2 com a diagonal igual a variância de Y e o restante igual a zero, $\tau_\varepsilon = \tau_d = \tau_a = 5$, $T_\varepsilon = T_d = T_a = 10$ e $\gamma_\varepsilon = \gamma_d = \gamma_a = 1000$.

O amostrador de Gibbs foi implementado no software R. Foram usadas 50000 iterações e determinou-se através do pacote *coda* do R um *burn-in* igual a 1000 e um *jump* igual a 25, determinado pelo critério de Raftery & Lewis (1992). Em seguida, foram utilizadas 101000 iterações, com o *burn-in* e o *jump* mencionados, totalizando uma cadeia com 4000 iterações (pontos amostrais) para cada parâmetro do modelo. Foi observado que não houve problemas de convergência nas cadeias.

3.5 Um estudo simulado

Com o objetivo de investigar os resultados fornecidos pelo fator de Bayes, gerou-se a situação a seguir.

Foram consideradas as mesmas matrizes Z, W, A e D apresentadas em (3.1), (3.2) e (3.3). Para as variâncias, fixou-se $\sigma_a^2 = 4$, $\sigma_d^2 = 1$ e $\sigma_\varepsilon^2 = 1$. Para os parâmetros de assimetria, considerou-se $\delta_a = 30$, $\delta_d = 20$ e $\delta_\varepsilon = 50$.

Gerou-se a, utilizando a Proposição 1, isto é, gerou-se um vetor de dimensão 194×1 de $\delta_a|N_{194}(0, I_{194})|$, outro vetor também de dimensão 194×1 de $N_{194}(0, \sigma_a^2 A)$ e os dois foram somados, com isso, foi obtido um vetor a com distribuição $SN_{194}(0, \sigma_a^2 A, \delta_a I_{194})$. O mesmo foi feito para obter d com distribuição $SN_{194}(0, \sigma_d^2 D, \delta_d I_{194})$ e ε com distribuição $SN_{194}(0, \sigma_\varepsilon^2 I_{194}, \delta_\varepsilon I_{194})$. Assim, o vetor y foi obtido fazendo a soma de a, d e ε.

Finalmente, as 16 configurações foram ajustadas para esse vetor y gerado com os mesmos valores iniciais para todos os parâmetros dos modelos considerados, conforme foi descrito na Seção anterior.

No Capítulo 4 serão apresentados os resultados e as discussões.

4 RESULTADOS E DISCUSSÃO

Conforme foi apresentado no Capítulo 3, optou-se por selecionar algumas sondas dentre as 3554 para explorar com detalhe diversas configurações de ajustes de modelos mistos normais assimétricos.

A Tabela 4 resume as dimensões das variáveis e dos parâmetros utilizados nos ajustes dos modelos.

TABELA 4 Dimensões das variáveis e dos parâmetros considerados nos ajustes.

Variáveis/Parâmetros	Dimensão
Sonda	3554
Sexo	2
Família	14
Indivíduos	194
Efeito Aditivo	194
Efeito Dominante	194

Para a seleção das sondas, foi seguido o critério apresentado na Seção 3.4, isto é, primeiramente foram ajustados modelos mistos Gauss-Markov normais

$$Y = X\beta + Zu + \varepsilon,$$

em que $Y_{194\times1}$ representa a intensidade da expressão normalizada de cada sonda, $\beta_{2\times1}$ representa o sexo, $u_{14\times1} \sim N_{14}(0, \sigma_u^2 I_{14})$ representa o efeito de famílias e $\varepsilon \sim N_{194}(0, \sigma_\varepsilon^2 I_{194})$, os resíduos. Em seguida, conforme apresentados na Seção 3.4, foram calculados os coeficientes de assimetria dos resíduos estimados e $\hat{h}_u^2$ para cada sonda. Por fim, foram selecionadas as sondas que apresentassem maior valor absoluto para essas duas quantidades.

Para facilitar essa identificação construiu-se um gráfico de dispersão com os valores de $\hat{h}_u^2$ versus os coeficientes de assimetria dos resíduos estimados para as

3554 sondas. O resultado é apresentado na Figura 12.

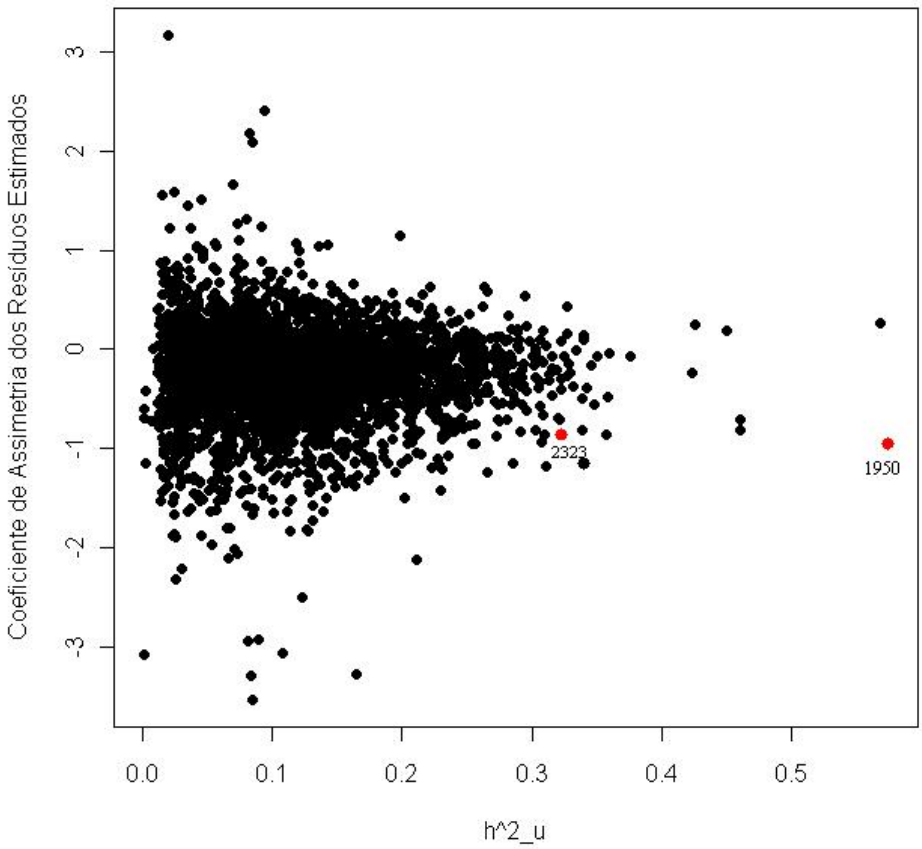

FIGURA 12 Diagrama de dispersão com os valores de $\hat{h}_u^2$ e os coeficientes de assimetria dos resíduos estimados para as 3554 sondas.

Pode-se observar na Figura 12 que a maior parte das sondas possui $\hat{h}_u^2$ inferiores a 0,4 e coeficientes de assimetria para os resíduos estimados entre -1 e 1.

Para esse conjunto de dados e com o ajuste desse modelo mais simples, a sonda que mais se destacou pelos critérios mencionados anteriormente foi a 1950.

Optou-se por selecionar também a sonda 2323, conforme destacado na Figura 12 para a comparação dos resultados.

A seguir, apresenta-se uma análise descritiva dos valores observados das intensidades das expressões, o desenho esquemático desses valores e o histograma dos resíduos estimados através dos modelos mistos com efeitos aleatórios de famílias e covariância uniforme entre indivíduos relacionados, para as duas sondas.

TABELA 5 Medidas descritivas das intensidades das expressões das sondas 1950 e 2323.

Medidas Descritivas	Sonda 1950	Sonda 2323
Mínimo	-0,152	1,433
Quartil 1	7,230	4,295
Mediana	8,525	9,343
Média	7,476	7,568
Quartil 3	9,138	10,110
Máximo	10,360	11,620
Variância	7,644	10,267

A Tabela 5 mostra que os valores observados das intensidades já normalizados para a sonda 1950 oscilam de -0,152 a 10,360 e para a sonda 2323 oscilam de 1,433 a 11,620. Nota-se também uma assimetria dos valores observados das duas sondas pela média ser diferente da mediana, como também pelo fato dos valores dos quartis 1 serem mais distantes dos valores da mediana do que os valores da mediana com os valores dos quartis 3, principalmente para a sonda 2323, em que o quartil 1 foi igual a 4,295, a mediana foi igual a 9,343 e o quartil 3 foi igual a 10,110. Essa assimetria também é constatada na Figura 13.

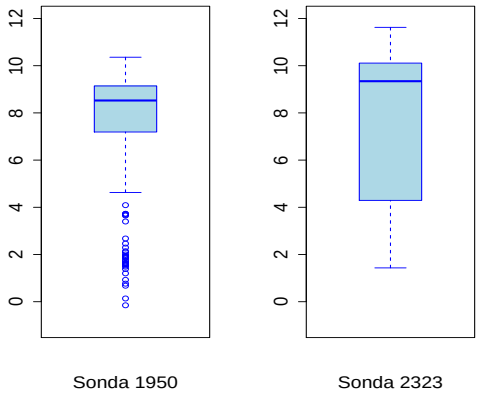

FIGURA 13 Desenhos esquemáticos dos valores observados das intensidades das expressões para as sondas 1950 e 2323, respectivamente.

A Figura 14 apresenta os histogramas dos resíduos estimados, por meio dos modelos mistos com efeitos aleatórios de famílias e covariância uniforme entre indivíduos relacionados, para as duas sondas. Por meio dos histogramas pode-se observar também o comportamento assimétrico nos resíduos.

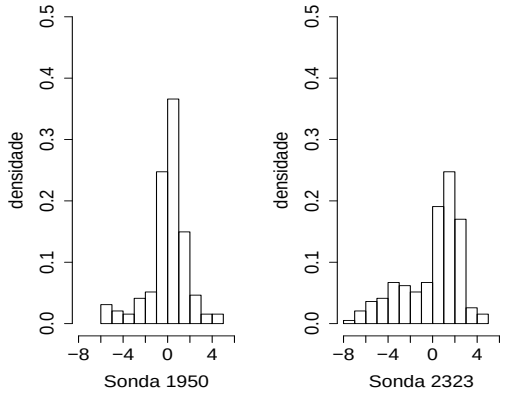

FIGURA 14 Histogramas para os resíduos do ajuste do modelo misto usual para as sondas 1950 e 2323, respectivamente.

4.1 Seleção de modelos

Como foi mencionado na Seção 3.4, foi implementado 16 configurações de modelos mistos apresentadas na Tabela 6.

TABELA 6 Configurações de modelos mistos para cada sonda.

Modelos	Parâmetros
1) MMFcafe	$\beta, f, \sigma_f^2, \sigma_\varepsilon^2, \delta_f, \delta_\varepsilon$
2) MMFcaf	$\beta, f, \sigma_f^2, \sigma_\varepsilon^2, \delta_f$
3) MMFcae	$\beta, f, \sigma_f^2, \sigma_\varepsilon^2, \delta_\varepsilon$
4) MMFsa	$\beta, f, \sigma_f^2, \sigma_\varepsilon^2$
5) MMAcaae	$\beta, a, \sigma_a^2, \sigma_\varepsilon^2, \delta_a, \delta_\varepsilon$
6) MMAcaa	$\beta, a, \sigma_a^2, \sigma_\varepsilon^2, \delta_a$
7) MMAcae	$\beta, a, \sigma_a^2, \sigma_\varepsilon^2, \delta_\varepsilon$
8) MMAsa	$\beta, a, \sigma_a^2, \sigma_\varepsilon^2$
9) MMADcaade	$\beta, a, d, \sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2, \delta_a, \delta_d, \delta_\varepsilon$
10) MMADcaad	$\beta, a, d, \sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2, \delta_a, \delta_d$
11) MMADcaae	$\beta, a, d, \sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2, \delta_a, \delta_\varepsilon$
12) MMADcade	$\beta, a, d, \sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2, \delta_d, \delta_\varepsilon$
13) MMADcaa	$\beta, a, d, \sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2, \delta_a$
14) MMADcad	$\beta, a, d, \sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2, \delta_d$
15) MMADcae	$\beta, a, d, \sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2, \delta_\varepsilon$
16) MMADsa	$\beta, a, d, \sigma_a^2, \sigma_d^2, \sigma_\varepsilon^2$

Para cada configuração considerada calculou-se o valor de $\hat{I}_k$ (calculado conforme a expressão (2.27)), para $k = 1, 2, 3, \ldots, 16$, para as sondas 1950 e 2323, respectivamente (ver os resultados na Tabela 7), para finalmente, efetuar-se o cálculo do fator de Bayes. Pode-se observar na Tabela 7 que para a sonda 1950 os maiores valores para $\hat{I}_k$ foram com os modelos MMFcaf (modelo misto com efeito aleatório de família e assimetria no efeito de família), MMAsa (modelo misto com efeito aleatório aditivo sem assimetria nos efeitos aleatórios) e MMADcad (modelo misto com efeitos aleatórios aditivo e dominante e com assimetria no efeito dominante). Já para a sonda 2323 os maiores valores para $\hat{I}_k$ foram com os modelos MMFcaf, MMAcaa (modelo misto com efeito aleatório aditivo com assimetria

no efeito aditivo) e MMADcaa (modelo misto com efeitos aleatórios aditivo e dominante e com assimetria no efeito aditivo).

TABELA 7 Resultados de $\hat{I}_k$ (calculado conforme a expressão (2.27)), para $k = 1, 2, 3, \ldots, 16$, para as sondas 1950 e 2323, para o cálculo do fator de Bayes.

k	Modelos	NFB1950	NFB2323
01	MMFcafe	7,9471e-221	7,8066e-284
02	MMFcaf	**3,0774e-220**	**2,8400e-281**
03	MMFcae	1,5173e-272	1,4302e-295
04	MMFsa	6,2271e-272	4,4766e-298
05	MMAcaae	8,1825e-208	8,4647e-235
06	MMAcaa	1,5132e-204	**3,5172e-231**
07	MMAcae	2,3003e-268	0,0000e+000
08	MMAsa	**1,1004e-202**	9,4682e-234
09	MMADcaade	0,0000e+000	0,0000e+000
10	MMADcaad	1,2620e-199	1,1923e-229
11	MMADcaae	0,0000e+000	0,0000e+000
12	MMADcade	0,0000e+000	0,0000e+000
13	MMADcaa	9,6390e-194	**1,2131e-216**
14	MMADcad	**7,9952e-185**	6,2196e-226
15	MMADcae	2,4495e-307	0,0000e+000
16	MMADsa	6,0012e-187	5,8030e-226

Foi apresentado para cada sonda os resultados nas Tabelas 8 e 9, repectivamente, do logaritmo natural do fator de Bayes (FB) multiplicado por 2 para selecionar o melhor modelo entre os três modelos considerados para cada uma delas, por meio dos resultados apresentados na Tabela 7. Foi aplicado essa transformação para os resultados serem interpretados conforme a Tabela 3. Os modelos apresentados na primeira coluna são os modelos considerados no numerador do FB, e os modelos que estão na primeira linha são os modelos considerados no denominador do FB.

TABELA 8 Logaritmo natural do fator de Bayes multiplicado por 2 para os modelos destacados para a sonda 1950. Os modelos apresentados na primeira coluna são os modelos considerados no numerador do FB, e os modelos que estão na primeira linha são os modelos considerados no denominador do FB.

Modelos	MMAsa	MMFcaf
MMADcad	82,2562	163,0770
MMAsa		80,8209

Todos os resultados apresentados na Tabela 8 indicam evidências muito fortes a favor dos modelos apresentados no numerador, pois os valores do logaritmo natural dos resultados do FB multiplicados por 2 são maiores que 10 (conforme a Tabela 3), isto é, MMAsa é melhor que o MMFcaf e MMADcad é melhor que o MMAsa e MMFcaf. Logo, para as 16 configurações consideradas, o melhor modelo através do FB para a sonda 1950 é o MMADcad.

TABELA 9 Logaritmo natural do fator de Bayes multiplicado por 2 para os modelos destacados para a sonda 2323. Os modelos apresentados na primeira coluna são os modelos considerados no numerador do FB, e os modelos que estão na primeira linha são os modelos considerados no denominador do FB.

Modelos	MMAcaa	MMFcaf
MMADcaa	66,9367	297,6420
MMAcaa		230,7050

Novamente, todos os resultados apresentados na Tabela 9 indicam evidências muito fortes a favor dos modelos apresentados no numerador, isto é, MMAcaa é melhor que o MMFcaf e MMADcaa é melhor que o MMAcaa e MMFcaf. Logo, para as 16 configurações consideradas, o melhor modelo através do FB para a sonda 2323 é o MMADcaa.

4.2 Descrição dos melhores modelos

Para fins de comparação, foi analisado a média dos resíduos a posteriori para os 3 melhores modelos para cada sonda (Chaloner & Brant, 1988; Albert & Chib, 1995).

A Figura 15 contém o índice das observações no eixo das abscissas e os resíduos preditos no eixo das ordenadas. Os gráficos (a), (b) e (c) apresentam os resíduos preditos para os modelos MMFcaf, MMAsa e MMADcad, respectivamente, para a sonda 1950 e os gráficos (d), (e) e (f) apresentam os resíduos para os modelos MMFcaf, MMAcaa e MMADcaa, respectivamente, para a sonda 2323. Fica muito claro que os modelos MMADcad (Figura 15-(c)) e MMADcaa (Figura 15-(f)) apresentam resíduos bem menores que os demais modelos.

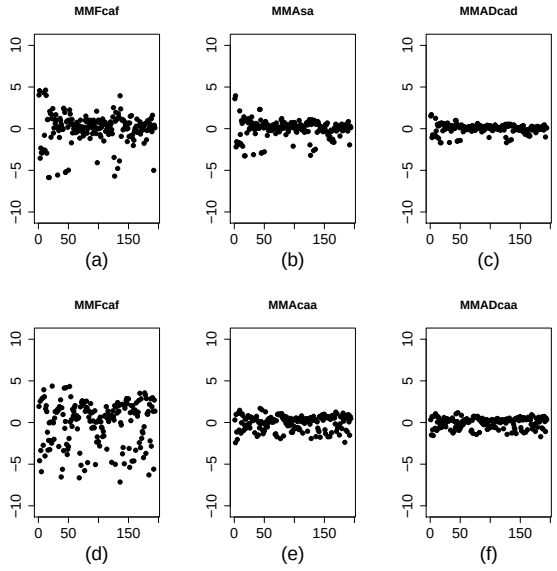

FIGURA 15 Índice das observações no eixo das abscissas versus os resíduos preditos no eixo das ordenadas. Os gráficos (a), (b) e (c) apresentam os resultados para a sonda 1950 e os gráficos (d), (e) e (f), para a sonda 2323.

A Figura 16 contém os valores observados das intensidades da expressão gênica

no eixo das abscissas e os valores preditos no eixo das ordenadas. Os gráficos (a), (b) e (c) foram construídos com base nos modelos MMFcaf, MMAsa e MMADcad, respectivamente, para a sonda 1950 e os gráficos (d), (e) e (f) foram construídos com base nos modelos MMFcaf, MMAcaa e MMADcaa, repectivamente, para a sonda 2323. Novamente, houve melhor ajuste com os modelos MMADcad (Figura 16-(c)) e MMADcaa (Figura 16-(f)).

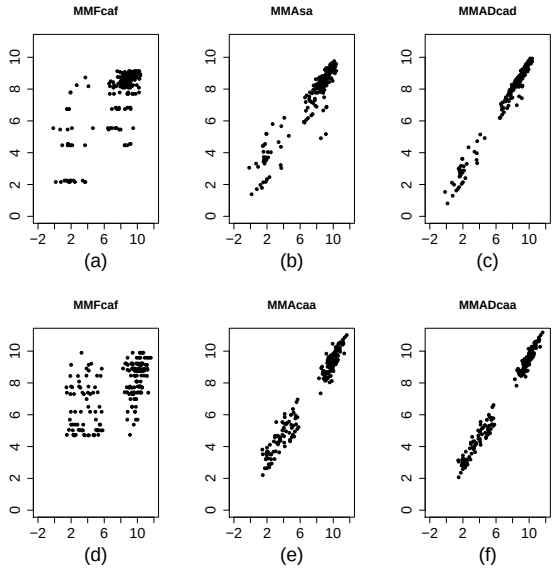

FIGURA 16 Valores observados no eixo das abscissas versus valores preditos no eixo das ordenadas. Os gráficos (a), (b) e (c) apresentam os resultados para a sonda 1950 e os gráficos (d), (e) e (f), para a sonda 2323.

Foram apresentados na Figura 17 os histogramas das amostras a posteriori dos parâmetros β (sexo), σ^2 (variância) e δ (assimetria, caso seja suposta no modelo) para os três melhores modelos através do FB, relativos às 16 configurações consideradas para a sonda 1950.

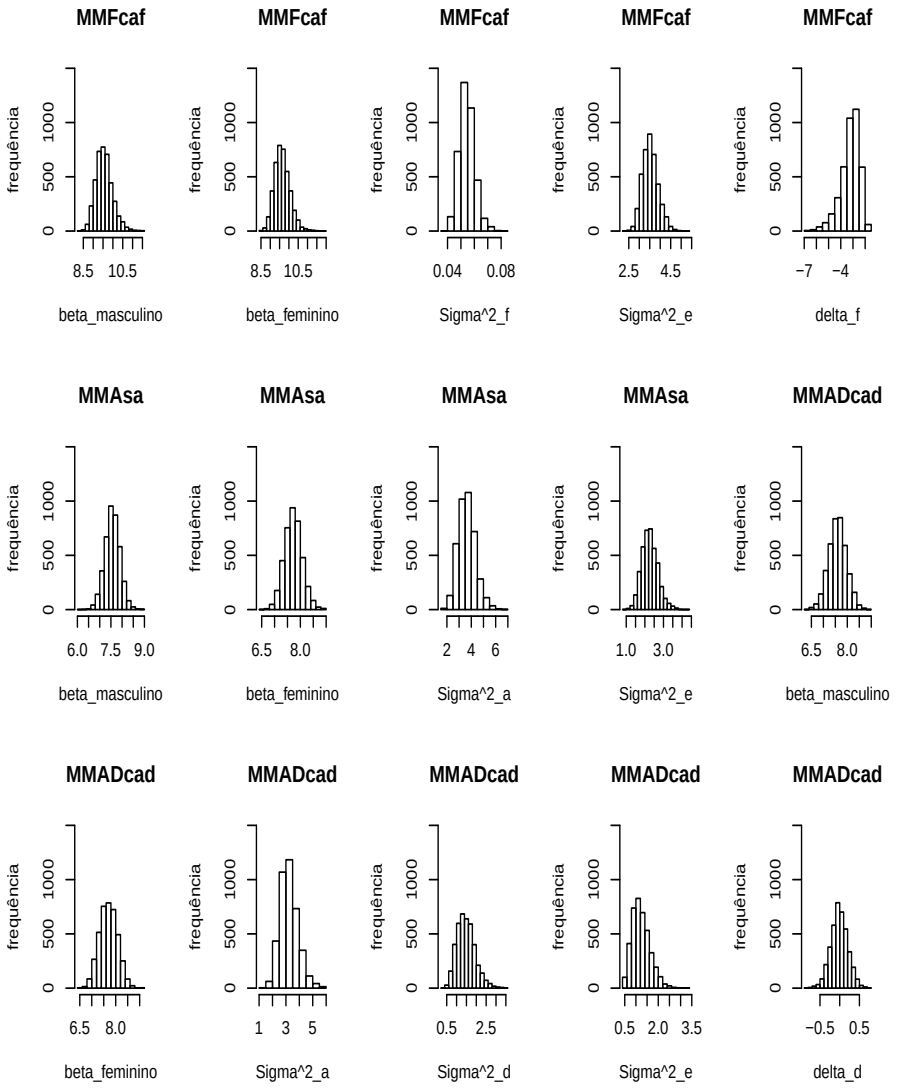

FIGURA 17 Histogramas das amostras a posteriori dos parâmetros β (sexo), σ^2 (variância) e δ (assimetria, caso seja suposta no modelo) para os três melhores modelos através do FB, relativos às 16 configurações consideradas para a sonda 1950.

Foram apresentados na Tabela 10 os resultados da média, do desvio padrão e do HPD com 95% de credibilidade dos parâmetros β, σ^2 e δ dos 3 melhores modelos para a sonda 1950.

TABELA 10 Média, desvio padrão e HPD com 95% de credibilidade dos parâmetros β, σ^2 e δ dos modelos MMFcaf, MMAsa e MMADcad, para a sonda 1950.

Modelos	Parâmetros	Média	DP	HPD
MMFcaf	$\beta_{masculino}$	9,5580	0,4238	[8,7909 ; 10,4378]
	$\beta_{feminino}$	9,6460	0,4189	[8,8707 ; 10,4461]
	σ^2_f	0,0545	0,0057	[0,0440 ; 0,0657]
	σ^2_ε	3,5251	0,3711	[2,8655 ; 4,2809]
	δ_f	-3,2286	0,7825	[-4,9170 ; -1,9868]
MMAsa	$\beta_{masculino}$	7,5640	0,3372	[6,9042 ; 8,2211]
	$\beta_{feminino}$	7,7250	0,3413	[7,0308 ; 8,3764]
	σ^2_a	3,6380	0,7009	[2,3566 ; 5,0672]
	σ^2_ε	2,2670	0,4310	[1,4986 ; 3,1589]
MMADcad	$\beta_{masculino}$	7,5940	0,3708	[6,8232 ; 8,3037]
	$\beta_{feminino}$	7,6970	0,3725	[6,9891 ; 8,4027]
	σ^2_a	3,2316	0,6738	[2,0069 ; 4,5857]
	σ^2_d	1,4794	0,4609	[0,6500 ; 2,3878]
	σ^2_ε	1,2323	0,4003	[0,5021 ; 1,9793]
	δ_d	-0,0103	0,2173	[-0,4236 ; 0,4220]

Por meio da Tabela 10 pode-se notar que não houve diferença entre os $\beta_{masculino}$ e o $\beta_{feminino}$. Para o modelo MMFcaf é obervada uma assimetria negativa para o efeito de família. Para o MMADcad, embora tenha a suposição de assimetria para o efeito dominante, o HPD com 95% de credibilidade contém o zero, ou seja, o parâmetro de assimetria do efeito dominante não é relevante.

O mesmo foi feito para a sonda 2323, isto é, a Figura 18 contém os histogramas das amostras a posteriori dos parâmetros β, σ^2 e δ para os três melhores modelos para ess sonda.

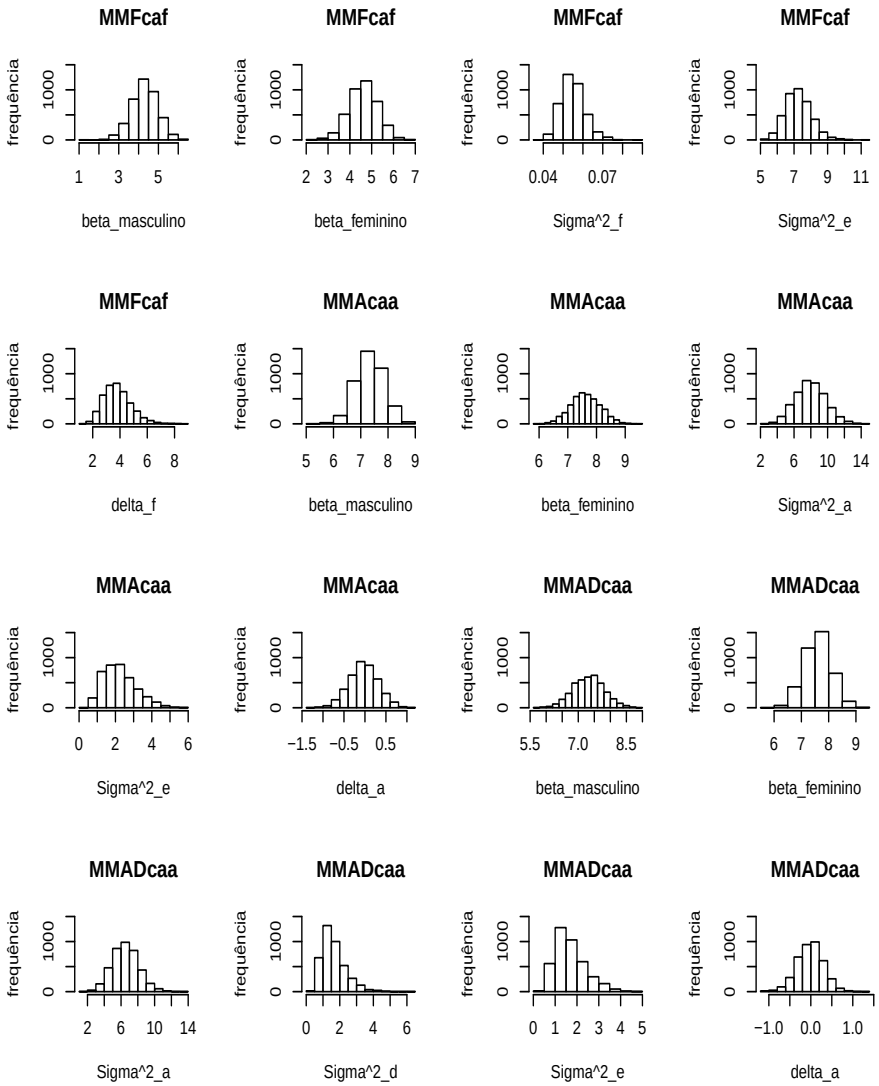

FIGURA 18 Histogramas das amostras a posteriori dos parâmetros β (sexo), σ^2 (variância) e δ (assimetria, caso seja suposta no modelo) para os três melhores modelos através do FB, relativos às 16 configurações consideradas para a sonda 2323.

Foram apresentados na Tabela 11 os resultados da média, do desvio padrão e do HPD com 95% de credibilidade dos parâmetros β, σ^2 e δ dos 3 melhores modelos para a sonda 2323.

TABELA 11 Média, desvio padrão e HPD com 95% de credibilidade dos parâmetros β, σ^2 e δ dos modelos MMFcaf, MMAcaa e MMADcaa, para a sonda 2323.

Modelos	Parâmetros	Média	DP	HPD
MMFcaf	$\beta_{masculino}$	4,3020	0,6525	[3,0616 ; 5,6399]
	$\beta_{feminino}$	4,6160	0,6581	[3,3788 ; 5,9350]
	σ^2_f	0,0549	0,0059	[0,0446 ; 0,0672]
	σ^2_ε	7,2567	0,7729	[5,8122 ; 8,7906]
	δ_f	3,8227	1,0206	[2,0189 ; 5,8264]
MMAcaa	$\beta_{masculino}$	7,3380	0,5066	[6,3451 ; 8,2825]
	$\beta_{feminino}$	7,6170	0,5131	[6,5866 ; 8,5658]
	σ^2_a	7,9841	1,7963	[4,3789 ; 11,3766]
	σ^2_ε	2,2375	0,8925	[0,7911 ; 4,0443]
	δ_a	-0,0354	0,3449	[-0,6945 ; 0,6373]
MMADcaa	$\beta_{masculino}$	7,3160	0,4866	[6,3424 ; 8,25139]
	$\beta_{feminino}$	7,5950	0,4922	[6,6270 ; 8,54210]
	σ^2_a	6,5226	1,5657	[3,5748 ; 9,71320]
	σ^2_d	1,6337	0,7087	[0,5649 ; 3,08911]
	σ^2_ε	1,6907	0,6926	[0,5683 ; 3,10289]
	δ_a	-0,0168	0,3121	[-0,6513 ; 0,58085]

Novamente, os resultados apresentados na Tabela 11 revelaram que não houve diferença entre os $\beta_{masculino}$ e o $\beta_{feminino}$. Para o modelo MMFcaf é obervado uma assimetria positiva para o efeito de família. Para os modelos MMAcaa e MMADcad, embora tenham a suposição de assimetria para os efeitos aditivos, o HPD com 95% de credibilidade para ambos os modelos contêm o zero, ou seja, também não são estatisticamente diferentes de zero.

Note que, embora os intervalos HPD para os parâmetros de assimetrias δ_d e δ_a para os modelos aditivos-dominantes normais assimétricos (MMADcad e MMADcaa) contenham o zero (ver Tabelas 10 e 11), esses modelos apresentaram ser melhores que todas as demais configurações consideradas nesse trabalho, tanto em

termos do fator de Bayes, quanto do comportamento dos resíduos do modelo.

Como esse resultado chamou a atenção, foi feita uma análise de resíduos e o gráfico dos valores observados versus os valores preditos com os modelos MMAD com assimetria (no efeito dominante para a sonda 1950 e no efeito aditivo para a sonda 2323) e os modelos MMAD sem assimetria para as duas sondas consideradas. Observou-se que os resíduos com os modelos MMAD com assimetria são muito parecidos com os resíduos com os modelos sem assimetria, sendo, no entanto, consistentemente menores para os modelos com assimetria. Já para as estimativas dos parâmetros, verificou-se que retirando a assimetria, as estimativas dos β e σ_a^2 são muito parecidas com os modelos que possuem assimetria, mas os componentes de variâncias dos efeitos dominantes e dos resíduos aumentaram com os modelos MMAD sem assimetria. Isto indica que os parâmetros de assimetria modificam as estimativas de componentes da variância e podem levar a conclusões diferentes sobre a herdabilidade das sondas, sendo necessários estudos mais detalhados (por exemplo, simulação extensiva) para verificar quais as relações entre componentes da variância e parâmetros de assimetria. No entanto, os valores preditos com os modelos com assimetria e com os modelos sem assimetria também foram semelhantes.

Para as estimativas das herdabilidades foram obtidas as densidades das expressões apresentadas em (3.27), através das médias das amostras a posteriori de f, com os modelos MMFcaf, de a, com os modelos MMAsa e MMAcaa e, de a e d, com os modelos MMADcad e MMADcaa, para as sondas 1950 e 2323, respectivamante. Os resultados da média, do desvio padrão e dos HPD de 95% de credibilidade das herdabilidades se encontram nas Tabelas 12 e 13.

Pode-se observar, com os resultados apresentados nas Tabelas 12 e 13, que os valores para as herdabilidades no sentido amplo com os modelos aditivos-

dominantes são ligeiramente maiores que os encontrados pelos demais modelos, indicando maior "acurácia" da predição dos valores genéticos (a e d), pois segundo White e Hodge (1992), a herdabilidade tem uma relação direta com a acurácia das estimativas. Além disso, a herdabilidade para a seleção de pais (sentido restrito) caiu do modelo MMA para o MMAD como era de se esperar, pois foi estimada uma variância de dominância não nula.

TABELA 12 Resultados da média, desvio padrão e HPD de 95% de credibilidade das herdabilidades com os modelos MMFcaf, MMAsa e MMADcad, para a sonda 1950.

Modelos	Parâmetros	Média	DP	HPD
MMFcaf	$Cov(\hat{f}, y_{observado})/Var(y_{observado})$	0,54	0,04	[0,46 ; 0,61]
MMAsa	$Cov(\hat{a}, y_{observado})/Var(y_{observado})$	0,70	0,06	[0,58 ; 0,81]
MMADcad	$Cov(\hat{a}, y_{observado})/Var(y_{observado})$	0,64	0,07	[0,50 ; 0,76]
	$Cov((\hat{a}+\hat{d}), y_{observado})/Var(y_{observado})$	0,84	0,05	[0,73 ; 0,94]

TABELA 13 Resultados da média, desvio padrão e HPD de 95% de credibilidade das herdabilidades com os modelos MMFcaf, MMAsa e MMADcad, para a sonda 2323.

Modelos	Parâmetros	Média	DP	HPD
MMFcaf	$Cov(\hat{f}, y_{observado})/Var(y_{observado})$	0,28	0,04	[0,20 ; 0,36]
MMAcaa	$Cov(\hat{a}, y_{observado})/Var(y_{observado})$	0,78	0,09	[0,59 ; 0,93]
MMADcaa	$Cov(\hat{a}, y_{observado})/Var(y_{observado})$	0,68	0,09	[0,49 ; 0,84]
	$Cov((\hat{a}+\hat{d}), y_{observado})/Var(y_{observado})$	0,83	0,07	[0,69 ; 0,96]

Portanto, para as duas sondas consideradas, concluiu-se que o modelo aditivo-dominante com assimetria apresentou melhores resultados para os resíduos, ajuste e acurácia das estimativas dos valores genéticos, tendo inclusive absorvido a assimetria do modelo misto com efeito aleatório de famílias e assimetria no efeito aleatório de família(MMFcaf).

4.3 Um estudo de simulação

Esta Seção apresenta os resultados do estudo de simulação que foi mencionado na Seção 3.5, em que se gerou a, utilizando a Proposição 1, isto é, a com distribuição $SN_{194}(0, \sigma_a^2 A, \delta_a I_{194})$, d com distribuição $SN_{194}(0, \sigma_d^2 D, \delta_d I_{194})$ e ε com distribuição $SN_{194}(0, \sigma_\varepsilon^2 I_{194}, \delta_\varepsilon I_{194})$. Foram consideradas as mesmas matrizes Z, W, A e D apresentadas em (3.1), (3.2) e (3.3). Para as variâncias, fixou-se $\sigma_a^2 = 4$, $\sigma_d^2 = 1$ e $\sigma_\varepsilon^2 = 1$. Para os parâmetros de assimetria, considerou-se $\delta_a = 30$, $\delta_d = 20$ e $\delta_\varepsilon = 50$. Assim, o vetor y foi obtido fazendo a soma de a, d e ε, ou seja, considerou-se $\beta = 0$. Além disso, foi feito o ajuste das 16 configurações para esse vetor y gerado com os mesmos valores iniciais para todos os parâmetros dos modelos considerados, conforme foi descrito na Seção 3.4.

Os melhores modelos relativos aos 16 considerados, utilizando o FB foram os modelos MMADcaade (modelo misto aditivo-dominante com assimetria nos efeitos aditivos, dominantes e resíduos) e MMADcade (modelo misto aditivo-dominante com assimetria nos efeitos dominantes e resíduos), confirmando que quando os dados apresentam assimetrias (como foi simulado) é necessário utilizar um modelo que leva em consideração parâmetros de assimetria.

TABELA 14 Comparação dos verdadeiros valores simulados com os estimados dos parâmetros do modelo misto aditivo-dominante com assimetria nos efeitos aditivos, dominantes e resíduos.

Parâmetros	Verdadeiro Valor	Média	DP	HPD
$\beta_{masculino}$	0	47,4463	19,8441	[25,2125 ; 94,7376]
$\beta_{feminino}$	0	42,8042	20,6411	[19,3863 ; 89,9581]
σ_a^2	4	7,8325	22,5343	[0,4582 ; 20,5101]
σ_d^2	1	164,2947	307,2831	[0,3643 ; 869,3678]
σ_ε^2	1	436,6074	596,1146	[0,3838 ; 1546,8330]
δ_a	30	1,3639	6,1544	[-6,8756 ; 21,0173]
δ_d	20	15,1585	25,1820	[-6,1022 ; 72,2764]
δ_ε	50	98,9573	504,7883	[-8,5786 ; 73,2705]

Pode-se observar que todos os intervalos HPD apresentados na Tabela 14 con-

têm os verdadeiros valores dos parâmetros, exceto os HPD dos β e do δ_a. Note que, mesmo o modelo correto (MMADcaade) sendo privilegiado pelo fator de Bayes, a simulação de valores altos para os parâmetros de assimetria ($\delta_a = 30$, $\delta_d = 20$ e $\delta_\varepsilon = 50$) implicaram em estimativas superestimadas para a média dos β, dos σ^2 e δ_ε e estimativas subestimadas para a média de δ_a e δ_d.

5 CONCLUSÕES

Por meio dos resultados apresentados no Capítulo 4, verificamos em um contexto geral, que para a sonda 1950, o modelo aditivo-dominante com assimetria apenas no efeito dominante mostrou ser o melhor modelo. No entanto, para a sonda 2323, o melhor modelo foi o modelo aditivo-dominante com assimetria apenas no efeito aditivo.

Com esse trabalho, notou-se que o modelo aditivo-dominante normal assimétrico mostrou ser uma alternativa eficiente para a análise de dados de microarrays, pois por meio deste pôde-se:

- ter o modelo aditivo-dominante usual como caso particular;
- incorporar informações de genealogia no cálculo das matrizes de identidade alélica (associada aos efeitos aditivos) e genotípica (associada aos efeitos de dominância);
- usar o modelo para qualquer fenótipo que apresente distribuição assimétrica e para qualquer estrutura de pedigree;
- obter o melhor modelo através do fator de Bayes com as 16 configurações possíveis de ajustes (assimetria apenas no efeito aleatório, assimetria apenas no resíduo e assimetria em ambos os efeitos);
- investigar os tipos de assimetrias nos efeitos aleatórios;
- notar que através do fator de Bayes e com os modelos analisados que a assimetria do efeito aleatório de família, ficou melhor descrita pelo modelo aditivo-dominante;

- obter as densidades a posteriori das herdabilidades no sentido restrito e amplo;

- fazer predições dos valores genéticos com maior acurácia.

O mesmo também tem suas limitações, isto é, mesmo sendo privilegiado pelo fator de Bayes, os modelos tendem a subestimar os efeitos de assimetria, apresentando evidências estatísticas (através dos intervalos HPD) que esses parâmetros não são significativos. Além disso, o ajuste desse modelo requer grande esforço computacional (utilizando as rotinas do R).

Perspectivas para Pesquisas Futuras

1. Evolução histórica da distribuição normal assimétrica no caso univariado e multivariado;

2. Implementação de um modelo misto normal assimétrico para várias sondas conjuntamente;

3. Otimização das rotinas computacionais no modelo misto família-aditivo-dominante normal assimétrico;

4. Diagnóstico no modelo misto normal assimétrico no enfoque bayesiano;

5. Modelo família-aditivo-dominante t assimétrico.

REFERÊNCIAS BIBLIOGRÁFICAS

ALBERT, J.; CHIB, S. Bayesian residual analysis for binary response regression models. **Biometrika**, London, v. 82, n. 4, p. 747-759, Dec. 1995.

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

ARELLANO-VALLE, R. B.; GENTON, M. G. On fundamental skew distributions. **Journal of Multivariate Analysis**, New York, v. 96, n. 1, p. 93-116, Sept. 2005.

AZZALINI, A. A class of distributions which includes the normal ones. **Scandinavian Journal of Statistics**: theory and applications, Stockholm, v. 12, n. 2, p. 171-178, Apr./June 1985.

AZZALINI, A.; CAPITANIO, A. Statistical applications of the multivariate skew normal distributions. **Journal of the Royal Statistical Society**, London, v. 61, n. 3, p. 579-602, 1999.

AZZALINI, A.; DALLA-VALLE, A. The multivariate skew-normal distribution. **Biometrika**, London, v. 83, n. 4, p. 715-726, Dec. 1996.

BOLSTAD, B. M.; IRIZARRY, R. A.; ASTRAND, M.; SPEED, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. **Bioinformatics**, Oxford, v. 19, n. 2, p. 185-193, 2003.

BOX, G. E.; TIAO, G. C. **Bayesian inference in statistical analysis.** New York: Wiley, 1992. 360 p.

CANCHO, V. G.; LACHOS, V. H.; ORTEGA, E. M. M. A nonlinear regression model with skew-normal errors. **Statistical Papers**, Berlin, May 2008. Disponível em: <http://www.springerlink.com/content/310883003h604586/fulltext.pdf>. Acesso em: 21 jul. 2008.

CHALONER, K.; BRANT, R. A. Bayesian approach to outlier detection and residual analysis. **Biometrika**, London, v. 75, n. 4, p. 651-659, Dec. 1988.

CHEUNG, V. G.; SPIELMAN, R. S. Data for genetic analysis workshop (GAW) 15: problem 1: genetics of gene expression variation in humans. **BMC Proceedings**, Flórida, Dec. 2007. Supllement. Disponível em: <http://www.biomedcentral.com/content/pdf/1753-6561-1-S1-S2.pdf>. Acesso em: 18 set. 2008.

DURBIN, B. P.; HARDIN, J. S.; HAWKINS, D. M.; ROCKE, D. M. A variance stabilizing transformation for gene-expression microarray data. **Briefings in Bioinformatics**, London, v. 18, p. 105-110, 2002. Supllement.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to quantitative genetics.** Longman: Scientific & Technical, 1996. 464 p.

FARIAS NETO, J. T.; RESENDE, M. D. V. Aplicação da metodologia de modelos mistos (REML/BLUP) na estimação de componenentes de variância e predição de valores genéticos em pupunheira (Bactris gasipaes). **Revista Brasileira de Fruticultura**, Cruz das Almas, v. 23, n. 2, p. 320-324, maio/ago. 2001.

GAMERMAN, D. **Markov chain Monte Carlo: stochastic simulation for bayesian inference.** Londres: Chapman-Hall, 1997. 245 p.

GELFAND, A. E.; SMITH A. F. M. Sampling-based approaches to calculating marginal densities. **Journal of the American Statistical Association**, New York, v. 85, n. 410, p. 398-409, June 1990.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, Hayward, v. 7, n. 4, p. 457-511, Nov. 1992a.

GELMAN, A.; RUBIN, D. B. A single series from the gibbs sampler provides a false sense of segurity. In: BERNARDO, J. M.; BERGER, J. O.; DAWID, A. P.; SMITH, A. F. M. (Ed.). **Bayesian Statistics**. 4. ed. Oxford: University, 1992b. p. 625-631.

GEMAN, S.; GEMAN, D. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. **IEEE Transactions on pattern analysis and machine intelligence**, New York, v. 6, n. 6, p. 721-741, June 1984.

GENETIC ANALYSIS WORKSHOOP. **Southwest foundation for biomedical research.** San Antonio: GAWs, 2006. Diponível em: <http://www.gaworkshop.org/>. Acesso em: 12 out. 2008.

GENTON, M. **Skew-elliptical distributions and their application: a journey beyond normality.** Boca Raton: CRC, 2004. 416 p.

GEWEKE, J. Bayesian inference in econometric models using Monte Carlo integration. **Econometrica**, Chicago, v. 57, n. 6, p. 1317-1339, Nov./Dec. 1989.

GEWEKE, J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: BERNARDO, J. M.; BERGER, J. O.; DAWID, A. P.; SMITH, A. F. M. (Ed.). **Bayesian Statistics**. 4. ed. Oxford: University, 1992. p. 169-193.

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

HENDERSON, C. R. **Applications of linear models in animal breeding.** Guelph: University of Guelph, 1984. 423 p.

IRIZARRY, R. A.; HOBBS, B.; COLLIN, F.; BEAZER-BARCLAY, Y. D.; ANTONELLIS, K. J.; SCHERF, U.; SPEED, T. P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. **Biostatistics**, Oxford, v.4, n. 2, p. 249-264, Apr. 2003.

JARA, A.; QUINTANA, F.; MARTÍN, E. S. Linear mixed models with skew elliptical distributions: a bayesian approach. **Computational Statistics and Data Analysis**, Amsterdam, v. 52, n. 10, 5033-5045, Oct. 2008.

KASS, R. E.; RAFTERY, A. E. Bayes factors. **Journal of the American Statistical Association**, New York, v. 90, n. 430, p. 773-795, Apr./June 1995.

KEMPTHORNE, O. **An introduction to genetics statistics.** Ames: Iowa State University, 1973. 545 p.

KERR, M. K.; CHURCHILL, G. A. Experimental design for gene expression microarrays. **Biostatistics**, Oxford, v. 2, n. 2, p. 183-201, June 2001.

LEIVA, V.; SANHUEZA, A.; KELMANSKY, D. M.; MARTÍNEZ, E. J. On the glog-normal distribution and its application to the gene expression problem. **Computational Statistics and Data Analysis**, Amsterdam, v. 53, n. 5, p. 1613-1621, Mar. 2009.

LI, C.; WONG, W. H. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. **Proceedings of the National Academy of Sciences**, Washington, v. 98, n. 1, p. 31-36, Jan. 2001.

LYNCH, M.; WALSH, B. **Genetics and analysis of quantitative traits.** Hardcover: Sinauer, 1998. 980 p.

METROPOLIS, N.; ROSEMBLUT, A. W.; ROSEMBLUT, M. N.; TELLER, A. H.; TELLER, E. Equations of state calculations by fast computing machines. **Journal of Chemical Physics**, New York, v. 21, n. 12, p. 1087-1092, Dec. 1953.

MORLEY, M.; MOLONY, C. M.; WEBER, T. M.; DEVLLN, J. L.; EWENS, K. G.; SPLELMAN, R. S.; CHEUNG, V. G. Genetic analysis of genome-wide variation in human gene expression. **Nature**, London, v. 430, n. 7001, p. 743-747, Aug. 2004.

NOGUEIRA, D. A.; SÁFADI, T.; FERREIRA, D. F. Avaliação de critérios de convergência para o método de Monte de Carlo via cadeias de Markov. **Revista Brasileira de Estatística**, Rio de Janeiro, v. 65, n. 224, p. 59-88, jan./mar. 2004.

O'HAGAN, A.; LEONARD, T. Bayes estimation subject to uncertainty about parameter constraints. **Biometrika**, London, v. 63, n. 2, p. 201-203, Dec. 1976.

PAULINO, D. C.; TURKAMAN, A. A.; MURTEIRA, B. **Estatística bayesiana.** Lisboa: Fundação Calouste Gulbenkian, 2003. 429 p.

PENA, S. D. Thomas Bayes: o "cara": raciocínio proposto por pároco do século 18 revolucionou vários campos do conhecimento. **Ciência Hoje**, São Paulo, v. 38, n. 228, jul. 2006.

R Development Core Team. **R:** A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2009. Disponível em: <http://www.r-project.org>. Acesso em: 15 apr. 2009.

RAFTERY, A. E.; LEWIS, S. How many iterations in the Gibbs sampler?. In: BERNARDO, J. M.; BERGER, J. O.; DAWID, A. P.; SMITH, A. F. M. (Ed.). **Bayesian Statistics**. 4. ed. Oxford: University, 1992. p. 763-773.

RESENDE, M. D. V. **Genética biométrica e estatística no melhoramento de plantas perenes.** Brasília: Embrapa Informação Tecnológica, 2002. 975 p.

RITZ, C.; EDÉN, P. Accounting for one-channel depletion improves missing value imputation in 2-dye microarray data. **BMC Genomics**, London, v. 9, n. 25, Jan. 2008. Disponível em: <http://www.biomedcentral.com/1471-2164/9/25>. Acesso em: 15 set. 2008.

ROBERTS, C. A correlation model useful in study of twins. **Journal of the American Statistical Association**, New York, v. 61, n. 316, p. 1184-1190, Oct./Dec. 1966.

ROHR, P. von; HOESCHELE, I. Bayesian QTL mapping using skewed Student t distributions. **Genetics Selection Evolution**, Paris, v. 34, n. 1, p. 1-21, 2002.

ROSA, G. J. M.; ROCHA, L. B.; FURLAN, L. R. Estudos de expressão gênica utilizando-se microarrays: delineamento, análise, e aplicações na pesquisa zootécnica. **Revista Brasileira de Zootecnia**, Viçosa, MG, v. 36, n. 4, p. 186-209, jul. 2007.

SARAIVA, E. F.; MILAN, L. A.; DIAS, T. C. M. Métodos Estatísticos Aplicados à Análise da Expressão Gênica. **Boletim ISBrA**, São Paulo, v. 1, n. 3, p. 5-8, abr. 2007.

SAHU, S. K.; DEY, D. K.; BRANCO, M. D. A new class of multivariate distributions with applications to Bayesian regression models. **The Canadian Journal of Statistics**, Toronto, v. 31, n. 2, p. 129-150, June 2003.

SEARLE, S. R.; MCCULLOCH, C. E.; CASELLA, G. **Variance Components.** New York: J. Wiley, 2006. 536 p.

SORENSEN, D. Developments in Statistical analysis in quantitative genetics. **Genetica**, Dordrecht, v. 136, n. 2, p. 319-332, Jun. 2009.

SORENSEN, D.; GIANOLA, D. **Likelihood, Bayesian, and MCMC Methods in quantitative genetics.** New York: Springer Verlag, 2002. 740 p.

SPEED, T. P. **Statistical analysis of gene expression microarray data.** Boca Raton: CRC, 2003. 127 p.

VARONA, L.; IBAÑEZ-ESCRICHE, N.; QUINTANILLA, R.; NOGUERA, J. L.; CASELLAS, J. Bayesian analysis of quantitative traits using skewed distributions. **Genetics Research**, Cambridge, v. 90, p. 179-190, Apr. 2008.

YANG, Y. H.; DUDOIT, S.; LUU, P., LIN, D. M.; PENG, V.; NGAI, J.; SPEED, T. P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. **Nucleic Acids Research**, Oxford, v. 30, n. 4, p. 15, Jan. 2002.

WALSH, B.; HENDERSON, D. Microarrays and beyond: what potential do current and future genomics tools have for breeders?. **Journal of Animal Science**, Champaign, v. 82, n. 13, p. E292-E299, Jan. 2004. Supplement.

WHITE, T. L.; HODGE, G. R. **Predicting breeding values with applications in forest tree improvement.** 2. ed. Kluwer: Dema, 1992. 367 p.

ANEXOS

ANEXO A

Formas de Obtenção das Condicionais Completas

Este Anexo A contém as demonstrações das formas de obtenção das condicionais completas apresentadas no Capítulo 3.

A especificação do modelo completo é dada como segue

$$
\begin{aligned}
Y|\beta, a, d, \sigma^2_\varepsilon, \delta_\varepsilon, w_\varepsilon &\sim N_n(X\beta + Za + Wd + \delta_\varepsilon w_\varepsilon, \sigma^2_\varepsilon I_n), && (5.1)\\
w_\varepsilon &\sim N_n(0, I_n)\mathtt{I}_{w_\varepsilon > 0}, && (5.2)\\
a|\sigma^2_a, \delta_a, w_a &\sim N_{q_a}(\delta_a w_a, \sigma^2_a A), && (5.3)\\
w_a &\sim N_{q_a}(0, I_{q_a})\mathtt{I}_{w_a > 0}, && (5.4)\\
d|\sigma^2_d, \delta_d, w_d &\sim N_{q_d}(\delta_d w_d, \sigma^2_d D), && (5.5)\\
w_d &\sim N_{q_d}(0, I_{q_d})\mathtt{I}_{w_d > 0}, && (5.6)\\
\beta &\sim N_p(\beta_0, S_\beta), && (5.7)\\
\sigma^2_\varepsilon &\sim GI\Big(\frac{\tau_\varepsilon}{2}, \frac{T_\varepsilon}{2}\Big), && (5.8)\\
\sigma^2_a &\sim GI\Big(\frac{\tau_a}{2}, \frac{T_a}{2}\Big), && (5.9)\\
\sigma^2_d &\sim GI\Big(\frac{\tau_d}{2}, \frac{T_d}{2}\Big), && (5.10)\\
\delta_\varepsilon &\sim N(\mu_\varepsilon, \gamma^2_\varepsilon), && (5.11)\\
\delta_a &\sim N(\mu_a, \gamma^2_a), && (5.12)\\
\delta_d &\sim N(\mu_d, \gamma^2_d). && (5.13)
\end{aligned}
$$

onde GI representa uma distribuição gama inversa. As variáveis w são as variáveis latentes e $\mathtt{I}$ é uma função indicadora do domínio de variação de w.

Utilizando o Teorema de Bayes e por meio do modelo completo especificado de (5.1) a (5.13), todas as condicionais completas são especificadas e demonstradas a seguir.

A primeira a ser demonstrada é a condicional completa referente ao parâmetro β, dada por

$$\beta|a, d, \sigma_\varepsilon^2, \delta_\varepsilon, w_\varepsilon, Y \sim N_p(M_\beta^{-1} m_\beta, M_\beta^{-1}), \tag{5.14}$$

onde $M_\beta = S_\beta^{-1} + X^\top X/\sigma_\varepsilon^2$ e $m_\beta = \beta_0 S_\beta^{-1} + X^\top (Y - Za - Wd - \delta_\varepsilon w_\varepsilon)/\sigma_\varepsilon^2$.

Demonstração: A posteriori de β é obtida fazendo o produto de (5.1) com (5.7) e utilizando o Lema 1, chamando de $\mu = Za + Wd + \delta_\varepsilon w_\varepsilon$, $A = X$, $x = \beta$, $\Sigma = \sigma_\varepsilon^2 I_n$, $\eta = \beta_0$ e $\Omega = S_\beta$, isto é,

$$\begin{aligned}
& \pi(\beta|a, d, \sigma_\varepsilon^2, \delta_\varepsilon, w_\varepsilon, Y) \propto f(Y|\beta, a, d, \sigma_\varepsilon^2, \delta_\varepsilon, w_\varepsilon)\pi(\beta) \\
=^{Lema1} \quad & \phi_n(Y|X\beta + Za + Wd + \delta_\varepsilon w_\varepsilon, \sigma_\varepsilon^2 I_n)\phi_p(\beta|\beta_0, S_\beta) \\
= \quad & \phi_n(Y|Za + Wd + \delta_\varepsilon w_\varepsilon + X\beta_0, \sigma_\varepsilon^2 I_n + X S_\beta X^\top) \times \\
& \phi_p(\beta|\beta_0 + (S_\beta^{-1} + X^\top(\sigma_\varepsilon^2 I_n)^{-1} X)^{-1} X^\top (\sigma_\varepsilon^2 I_n)^{-1} \times \\
& (Y - Za - Wd - \delta_\varepsilon w_\varepsilon - X\beta_0), (S_\beta^{-1} + X^\top(\sigma_\varepsilon^2 I_n)^{-1} X)^{-1}) \\
\propto \quad & \phi_p\Bigg(\beta|\beta_0 + \left(S_\beta^{-1} + \frac{X^\top X}{\sigma_\varepsilon^2}\right)^{-1} X^\top \times \\
& (Y - Za - Wd - \delta_\varepsilon w_\varepsilon - X\beta_0)/\sigma_\varepsilon^2, \left(S_\beta^{-1} + \frac{X^\top X}{\sigma_\varepsilon^2}\right)^{-1}\Bigg).
\end{aligned}$$

Seja $M_\beta = S_\beta^{-1} + \frac{X^\top X}{\sigma_\varepsilon^2}$, então

$$\pi(\beta|a,d,\sigma^2_\varepsilon,\delta_\varepsilon,w_\varepsilon,Y) \propto \phi_p\left(\beta|\underbrace{\beta_0 + M_\beta^{-1}X^\top(Y - Za - Wd - \delta_\varepsilon w_\varepsilon - X\beta_0)/\sigma^2_\varepsilon}_{\dagger}, M_\beta^{-1}\right).$$

Simplificando a expressão †, tem-se

$$\begin{aligned}
& \beta_0 + M_\beta^{-1}X^\top(Y - Za - Wd - \delta_\varepsilon w_\varepsilon - X\beta_0)/\sigma^2_\varepsilon \\
= \;& M_\beta^{-1}X^\top(Y - Za - Wd - \delta_\varepsilon w_\varepsilon)/\sigma^2_\varepsilon + \beta_0 - M_\beta^{-1}X^\top X\beta_0/\sigma^2_\varepsilon \\
= \;& M_\beta^{-1}X^\top(Y - Za - Wd - \delta_\varepsilon w_\varepsilon)/\sigma^2_\varepsilon + \beta_0\Big[1 - M_\beta^{-1}\frac{X^\top X}{\sigma^2_\varepsilon}\Big] \\
= \;& M_\beta^{-1}X^\top(Y - Za - Wd - \delta_\varepsilon w_\varepsilon)/\sigma^2_\varepsilon + \\
& \beta_0\Big[1 - \Big(S_\beta^{-1} + \frac{X^\top X}{\sigma^2_\varepsilon}\Big)^{-1}\frac{X^\top X}{\sigma^2_\varepsilon}\Big] \\
= \;& M_\beta^{-1}X^\top(Y - Za - Wd - \delta_\varepsilon w_\varepsilon)/\sigma^2_\varepsilon + \\
& \beta_0\Big[1 - \Big(\frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon S_\beta^{-1} + X^\top X}\Big)\frac{X^\top X}{\sigma^2_\varepsilon}\Big] \\
= \;& M_\beta^{-1}X^\top(Y - Za - Wd - \delta_\varepsilon w_\varepsilon)/\sigma^2_\varepsilon + \beta_0\Big[1 - \Big(\frac{X^\top X}{\sigma^2_\varepsilon S_\beta^{-1} + X^\top X}\Big)\Big] \\
= \;& M_\beta^{-1}X^\top(Y - Za - Wd - \delta_\varepsilon w_\varepsilon)/\sigma^2_\varepsilon + \beta_0\Big[\frac{\sigma^2_\varepsilon S_\beta^{-1}}{\sigma^2_\varepsilon S_\beta^{-1} + X^\top X}\Big] \\
= \;& M_\beta^{-1}X^\top(Y - Za - Wd - \delta_\varepsilon w_\varepsilon)/\sigma^2_\varepsilon + \beta_0\Big[\frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon S_\beta^{-1} + X^\top X}S_\beta^{-1}\Big] \\
= \;& M_\beta^{-1}X^\top(Y - Za - Wd - \delta_\varepsilon w_\varepsilon)/\sigma^2_\varepsilon + \beta_0\Big[M_\beta^{-1}S_\beta^{-1}\Big] \\
= \;& M_\beta^{-1}\Big[X^\top(Y - Za - Wd - \delta_\varepsilon w_\varepsilon)/\sigma^2_\varepsilon + \beta_0 S_\beta^{-1}\Big].
\end{aligned}$$

Seja $m_\beta = X^\top(Y - Za - Wd - \delta_\varepsilon w_\varepsilon)/\sigma^2_\varepsilon + \beta_0 S_\beta^{-1}$, então

$$\pi(\beta|a,d,\sigma^2_\varepsilon,\delta_\varepsilon,w_\varepsilon,Y) \propto \phi_p\left(\beta|M_\beta^{-1}m_\beta, M_\beta^{-1}\right).$$

Portanto, a posteriori de β é como em (5.14). ■

As demonstrações das condicionais completas de a, d, δ_ε, δ_a e δ_d seguem de maneira análoga à precedente, onde são feitas as devidas multiplicações de densidades normais (para a: (5.1) com (5.3), para δ_ε: (5.1) com (5.11), para δ_a: (5.3) com (5.12)) e para δ_d: (5.5) com (5.13)), utiliza-se o Lema 1 e obtém-se seus respectivos M e m. O mesmo ocorre para w_ε, w_a e w_d, tendo cautela com as multiplicações de (5.1) com (5.2), (5.3) com (5.4) e (5.5) com (5.6), respectivamente, pois envolvem multiplicações de densidades normais com normais truncadas, que apresentam a função indicadora do domínio de variação dos w ($\mathtt{I}_{w>0}$), que afeta a posteriori de ambos, isto é, a posteriori continua sendo uma normal truncada positiva, com os parâmetros atualizados (dados em função de M e m).

Já as posterioris de σ^2_ε, σ^2_a e σ^2_d envolvem a multiplicação de (5.1) com (5.8), (5.3) com (5.9) e (5.5) com (5.10), respectivamente. Nestes casos, tem-se o produto da densidade normal com a densidade da gama inversa. Vê-se a seguir a demonstração de σ^2_ε, pois as de σ^2_a e σ^2_d decorrem de forma análoga. Note que a multiplicação de uma densidade normal com uma densidade gama inversa resulta numa densidade gama inversa com os parâmetros atualizados.

$$\sigma^2_\varepsilon|\beta,a,d,\delta_\varepsilon,w_\varepsilon,Y \sim GI\left(\frac{n+\tau_\varepsilon}{2}, \frac{T_\varepsilon + \mu_{\sigma_\varepsilon}{}^\top \mu_{\sigma_\varepsilon}}{2}\right), \tag{5.15}$$

com $\mu_{\sigma_\varepsilon} = Y - X\beta - Za - Wd - \delta_\varepsilon w_\varepsilon$.

Demonstração: Pelo Teorema de Bayes, tem-se que a posteriori de σ^2_ε é dada por

$$
\begin{aligned}
& \pi(\sigma^2_\varepsilon|\beta,a,d,\delta_\varepsilon,w_\varepsilon,Y) \propto f(Y|\beta,a,d,\sigma^2_\varepsilon,\delta_\varepsilon,w_\varepsilon)\pi(\sigma^2_\varepsilon) \\
= \quad & \left(\frac{1}{(2\pi)^{n/2}\sqrt{|\sigma^2_\varepsilon I_n|}}\right)^n exp\Big[-\frac{1}{2}(Y-X\beta-Za-Wd-\delta_\varepsilon w_\varepsilon)^\top \times \\
& \qquad (\sigma^2_\varepsilon I_n)^{-1}(Y-X\beta-Za-Wd-\delta_\varepsilon w_\varepsilon)\Big] \times \\
& \qquad \frac{\left(\frac{T_\varepsilon}{2}\right)^{\tau_\varepsilon/2}}{\Gamma\left(\frac{\tau_\varepsilon}{2}\right)}\left(\frac{1}{\sigma^2_\varepsilon}\right)^{(\frac{\tau_\varepsilon}{2})+1} exp\left(-\frac{1}{2}\,\frac{T_\varepsilon}{\sigma^2_\varepsilon}\right).
\end{aligned}
$$

Para simplificar a expressão, seja

$$\mu_{\sigma_\varepsilon} = Y - X\beta - Za - Wd - \delta_\varepsilon w_\varepsilon.$$

Assim,

$$
\begin{aligned}
\pi(\sigma^2_\varepsilon|\beta,a,d,\delta_\varepsilon,w_\varepsilon,Y) \quad \propto \quad & \left(\frac{1}{\sigma^2_\varepsilon}\right)^{\frac{n}{2}} exp\Big[-\frac{1}{2\sigma^2_\varepsilon}\mu^\top_{\sigma_\varepsilon}\mu_{\sigma_\varepsilon}\Big] \times \\
& \left(\frac{1}{\sigma^2_\varepsilon}\right)^{(\frac{\tau_\varepsilon}{2})+1} exp\left(-\frac{1}{2\sigma^2_\varepsilon}\,T_\varepsilon\right) \\
= \quad & \left(\frac{1}{\sigma^2_\varepsilon}\right)^{\left(\frac{n+\tau_\varepsilon}{2}\right)+1} exp\Big[-\Big(\frac{T_\varepsilon+\mu^\top_{\sigma_\varepsilon}\mu_{\sigma_\varepsilon}}{2}\Big)\frac{1}{\sigma^2_\varepsilon}\Big],
\end{aligned}
$$

que caracteriza uma gama inversa conforme a descrita em (5.15). ■

ANEXO B

Programas em R

Este Anexo B contém o programa feito em R utilizado no ajuste do modelo aditivo-dominante normal assimétrico e também um exemplo do cálculo do numerador do fator de Bayes para a sonda 1950 e para uma configuração.

```
# theta  <- c(t(beta), t(a), t(d), t(yhat), Va, Vd, Ve, deltaa,
# deltad, deltae)
# Dimensão de theta = 2 + 194 + 194 + 194 + 1 + 1 + 1 + 1 + 1 + 1 = 590%
rm(list=ls(all=T))
library(MASS)
library(MCMCpack)      # Para gerar da gama inversa
library(matrixcalc)    # Para realização dos cálculos matriciais
# Funções gerais para atualizar os parâmetros do modelo
# Variaveis latentes
atwe <- function(deltae,Ve,n,y,X,beta,Z,a,W,d) {
  Mwei <- ginv((1 + deltae^2 / Ve) * diag(n))
  mwe  <- deltae * (y - X %*% beta - Z %*% a - W %*% d) / Ve
  abs(mvrnorm(1,Mwei %*% mwe, Mwei))
}
atwa   <- function(deltaa,Va,qa,Ainv,a) {
  Mwai <- solve((deltaa^2 / Va) * Ainv + diag(qa))
  mwa  <- (deltaa / Va) * Ainv %*% a
  abs(mvrnorm(1,Mwai %*% mwa, Mwai))
}
atwd   <- function(deltad,Vd,qd,Dinv,d) {
  Mwdi <- solve((deltad^2 / Vd) * Dinv + diag(qd))
  mwd  <- (deltad / Vd) * Dinv %*% d
  abs(mvrnorm(1,Mwdi %*% mwd, Mwdi))
}
# beta - "efeitos fixos"
atbeta <- function(Sbetai,X,Ve,beta0,y,Z,a,W,d,deltae,we) {
  Mbetai <- ginv(Sbetai + (crossprod(X))/Ve)
  mbeta  <- Sbetai %*% beta0 + t(X) %*%
            (y - Z %*% a - W %*% d -deltae * we)/Ve
  mvrnorm(1,Mbetai %*% mbeta,Mbetai)
}
# a - "efeitos aleatórios" - efeito aditivo
ata <- function(Z,Ve,Va,Ainv,y,X,beta,W,d,deltae,we,deltaa,wa) {
  Mai <- solve(crossprod(Z)/Ve + (1 / Va) * Ainv)
  ma  <- t(Z) %*% (y - X %*% beta - W %*% d - deltae * we) / Ve
         +  (deltaa / Va) * Ainv %*% wa
  mvrnorm(1,Mai%*%ma,Mai)
}
# d - "efeitos aleatórios" - efeito dominante
atd <- function(W,Ve,Vd,Dinv,y,X,beta,Z,a,deltae,we,deltad,wd) {
  Mdi <- solve(crossprod(W)/Ve + (1 / Vd) * Dinv)
  md  <- t(W) %*% (y - X %*% beta - Z %*% a - deltae * we) / Ve
         +  (deltad / Vd) * Dinv %*% wd
  mvrnorm(1,Mdi%*%md,Mdi)
}
# Parâmetros de assimetria
```

```
atdeltae <- function(gamae,we,Ve,mue,y,X,beta,Z,a,W,d){
  vardeltae <- (as.real((1/gamae^2) + we%*%we/Ve))^(-1)
  mudeltae  <- vardeltae*((mue/gamae^2) + t(we) %*%
               (y - X%*%beta - Z%*%a - W %*% d)/Ve)
  rnorm(1,mudeltae,sqrt(vardeltae))
}
atdeltaa <- function(gamaa,Va,wa,Ainv,mua){
  Mdeltaai <- 1/as.real((1/gamaa^2) + (1 / Va) * t(wa) %*% Ainv %*% wa)
  mdeltaa  <- (mua / gamaa^2)  + (1 / Va) * t(wa) %*% Ainv %*% a
  rnorm(1,Mdeltaai %*% mdeltaa , sqrt(Mdeltaai))
}
atdeltad <- function(gamad,Vd,wd,Dinv,mud){
  Mdeltadi <- 1/as.real((1/gamad^2) + (1 / Vd) * t(wd) %*% Dinv %*% wd)
  mdeltad  <- (mud / gamad^2)  + (1 / Vd) * t(wd) %*% Dinv %*% d
  rnorm(1,Mdeltadi %*% mdeltad , sqrt(Mdeltadi))
}
# Leitura dos dados
gawphn<-read.table("linkagePhn.txt",sep=",",header=TRUE)
dim(gawphn)
n <- length(gawphn[,1])
y <- gawphn[,(2+1950)]
rm(gawphn)
# Tamanho da cadeia MCCM
B    <- 1000 # burn-in
J    <- 25   # jump
nef  <- 4000 # número de cadeias efetivas
nsMC <- B + nef*J
# Delineamento
pedig  <- read.table("genealogia.txt",header=TRUE)
attach(pedig)
# Matriz IBD
library(kinship)
A      <- 2*kinship(ID, FA, MO)
Ainv <- ginv(A)
# Matriz D
D      <- diag(n) for(i in 1:n){
  for(j in (i+1):n){
  if(FA[i]==FA[j] && MO[j]==MO[j] && FA[i]!=0){
    D[i,j] <- 1/8
    D[j,i] <- D[i,j]
    }
  }
}
Dinv   <- solve(D)
# Efeitos fixos
X <- model.matrix(lm(y~-1+factor(SEX)))
p <- ncol(X)
# Efeitos aditivos
Z  <- diag(n)
qa <- ncol(Z)
# Efeitos dominantes
W  <- diag(n)
qd <- ncol(W)
#####################################
# Análise
```

```
# Valores Iniciais dos parâmetros
beta   <- matrix(mean(y),p,1)
a      <- rnorm(qa)
d <- rnorm(qd)
Ve <- rnorm(1)^2
Va     <- rnorm(1)^2
Vd     <- rnorm(1)^2
deltae <- rnorm(1,0,10000)
deltaa <- rnorm(1,0,10000)
deltad <- rnorm(1,0,10000)
yhat   <- y
theta  <- c(t(beta), t(a), t(d), t(yhat), Va, Vd, Ve, deltaa,
deltad, deltae)
# Hiperparâmetros
beta0  <- beta
Sbeta  <- diag(rep(var(y),p),p)
Sbetai <- solve(Sbeta)
Taue   <- 5
Te     <- 10
Taua   <- Taue
Ta <- Te
Taud <- Taue
Td <- Te
mue <- 0
gamae  <- 1000
mua <- 0
gamaa <- 1000
mud    <- 0
gamad  <- 1000
cont <- 0
while (cont <= nsMC) {
  # Atualizando as variáveis latentes
  we <- atwe(deltae,Ve,n,y,X,beta,Z,a,W,d)
  wa <- atwa(deltaa,Va,qa,Ainv,a)
  wd <- atwd(deltad,Vd,qd,Dinv,d)
  # Atualizando beta - "efeitos fixos"
  beta <- atbeta(Sbetai,X,Ve,beta0,y,Z,a,W,d,deltae,we)
  # Atualizando a - "efeitos aleatórios" - efeito aditivo
  a <- ata(Z,Ve,Va,Ainv,y,X,beta,W,d,deltae,we,deltaa,wa)
  # Atualizando d - "efeitos aleatórios" - efeito dominante
  d <- atd(W,Ve,Vd,Dinv,y,X,beta,Z,a,deltae,we,deltad,wd)
  # Atualizando deltae
  deltae <- atdeltae(gamae,we,Ve,mue,y,X,beta,Z,a,W,d)
  # Atualizando deltaa
  deltaa <- atdeltaa(gamaa,Va,wa,Ainv,mua)
  # Atualizando deltad
  deltad <- atdeltad(gamad,Vd,wd,Dinv,mud)
  # Atualizando Ve
  yhat <- X%*%beta + Z%*%a + W %*% d + deltae*we
  ee   <- y - yhat
```

```
  Ve   <- rinvgamma(1,as.real((n+Taue)/2) , as.real((Te+t(ee)%*%ee)/2))
  # Atualizando Va
  Va <- rinvgamma(1,as.real((n+Taua)/2) ,
        as.real((Ta+t(a - deltaa*wa)%*%Ainv%*%(a - deltaa*wa))/2))
  # Atualizando Vd
  Vd <- rinvgamma(1,as.real((n+Taud)/2) ,
        as.real((Td+t(d - deltad*wd)%*%Dinv%*%(d - deltad*wd))/2))
  # Atualizar matriz de parâmetros
  if(cont%%J==0 && cont>B)
    {
    theta  <- c(t(beta),t(a),t(d),t(yhat), Va, Vd, Ve,
                deltaa, deltad, deltae)
    write(theta,"cadeiaMMADcaade1950.txt",length(theta),append=TRUE)
    }
  cont <- cont + 1
}
# Estimativa dos parâmetros de assimetria e da herdabilidade
dados <- read.table("cadeiaMMADcaade1950.txt")
dim(dados)
library(coda)
mcdelta <- mcmc(dados[,588:590])
Va      <- dados[,585]
Vd      <- dados[,586]
Ve      <- dados[,587]
h2rest  <- Va/(Va+Vd+Ve)
h2amp   <- (Va+Vd)/(Va+Vd+Ve)
mch2rest <- mcmc(h2rest)
mch2amp  <- mcmc(h2amp)
HPDinterval(mcdelta)
summary(mcdelta)
plot(mcdelta)
HPDinterval(mch2rest)
summary(mch2rest)
HPDinterval(mch2amp)
summary(mch2amp)
plot(mch2rest)
plot(mch2amp)

############################################################
# Cálculo do Numerador do Fator de Bayes
############################################################
rm(list=ls(all=T))
# Leitura de dados
gawphn <- read.table("linkagePhn.txt",sep=",",header=TRUE)
n      <- length(gawphn[,1])
y1     <- gawphn[,(2+1950)]   # Sonda 1950
y2 <- gawphn[,(2+2323)]       # Sonda 2323
pedig  <- read.table("genealogia.txt",header=TRUE)
attach(pedig)
names(pedig)
# Efeitos fixos
X   <- model.matrix(lm(y1~-1+factor(SEX)))
Z0  <- model.matrix(lm(y1~-1+factor(FAMID)))  # Efeitos de familia
```

```
Z1  <- diag(n)                                  # Efeitos aditivos
rm(gawphn)
rm(pedig)

############################################################
#                       Modelos MMAD                       #
############################################################
# Análise da Cadeia MMADcaade
dados  <- read.table("cadeiaMMADcaade1950.txt")
# theta  <- c(t(beta), t(a), t(d), t(yhat), Va, Vd, Ve, deltaa,
# deltad, deltae)
# Dimensão theta = 2 + 194 + 194 + 194 + 1 + 1 + 1 + 1 + 1 + 1 = 590
y      <- y1
Z <- Z1
W      <- Z1
beta   <- as.matrix(dados[,1:2])
a <- as.matrix(dados[,3:196])
d <- as.matrix(dados[,197:390])
Ve <- dados[,587]
deltae <- dados[,590]
# Variáveis auxiliares
densi  <- 0 * c(1:4000) # densidade
pcum   <- 0 * c(1:4000) # prob. acumulada
logS   <- 0 * c(1:4000) # log do inverso da verossimilhança
# Cálculo do fator de Bayes
mediaNM <- X%*%t(beta) + Z%*%t(a) + W%*%t(d)
desviNM <- sqrt(Ve^2 + deltae^2) # Não se altera
upper <- (deltae/(Ve^2 + deltae^2))*(y - X%*%t(beta) - Z%*%t(a)
         - W%*%t(d))
sigma <- sqrt(as.real(Ve^2/(Ve^2 + deltae^2))) # Não se altera
for (i in 1:4000){
  densi[i]  <- (sum(log(dnorm(y,mediaNM[,i] ,desviNM[i]))))
  pcum[i]   <- (sum(log(pnorm(upper[,i],0,sigma[i]))))
  logS[i]   <- -(log(2^194) + (densi[i]) + (pcum[i]))
}
c      <- max(logS)
lnvk   <- log(mean(exp(logS-c))) + c
NFB    <- exp(-lnvk)
```