ANÁLISE BAYESIANA DE MODELOS MISTOS NORMAIS ASSIMÉTRICOS EM DADOS DE EXPRESSÃO GÊNICA ORIGINADOS DE PEDIGREES COMPLEXOS

Daniela Carine Ramires de OLIVEIRA[1]
Júlio Sílvio de Sousa BUENO FILHO[2]

▪ RESUMO: Estimativas de herdabilidade para a expressão gênica são escassas e, em geral, provenientes de estruturas de famílias, em que se assume covariância uniforme para os indivíduos relacionados. Para tais estimativas usam-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 as intensidades das expressões gênicas de 3554 sondas. Neste sentido, este trabalho trata do desenvolvimento e implementação computacional do modelo aditivo-dominante normal assimétrico para a análise dessas expressões gênicas, originadas de um pedigree complexo, permitindo assimetria nas distribuições de todos os efeitos aleatórios. 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 do modelo misto normal assimétrico.

▪ PALAVRAS-CHAVE: Simulação Monte Carlo via cadeias de Markov; modelos mistos; distribuição normal assimétrica multivariada; inferência bayesiana.

[1]Universidade Federal de São João del-Rei - UFSJ, Departamento de Matemática, Estatística e Ciências da Computação, CEP: 36307-352, São João del-Rei, MG, Brasil. E-mail: *ramires.daniela@gmail.com*

[2]Universidade Federal de Lavras - UFLA, Departamento de Ciências Exatas, Caixa Postal 37, CEP: 37200-000, Lavras, MG, Brasil. E-mail: *juliobuenof@gmail.com*

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 e 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 e 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.

Assim, considerável esforço tem sido direcionado para relaxar a suposição de normalidade e, conjuntamente, estimar a densidade dos efeitos aleatórios e parâmetros do modelo. 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. Rohr e 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 análise 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 Arellano-Valle et al. (2007). As estimativas dos parâmetros dos modelos foram realizadas sob o enfoque bayesiano.

O objetivo deste trabalho foi selecionar o melhor modelo; estudar os tipos de assimetrias presentes nos efeitos aleatórios e 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), utilizando a inferência bayesiana.

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 artigo está organizado em 3 seções. A seção 2 apresenta a distribuição normal assimétrica multivariada, uma descrição sobre os dados reais, os modelos mistos normais assimétricos, a modelagem bayesiana e a implementação computacional dos modelos. A seção 3 apresenta os resultados e as discussões dos mesmos e, por fim, as conclusões.

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

2 Material e métodos

2.1 Distribuição normal assimétrica multivariada

A primeira versão da distribuição normal assimétrica multivariada foi apresentada em Azzalini e 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 e 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 Oliveira (2009).

Para apresentar a densidade da normal assimétrica multivariada, proposta por Arelano-Valle et al. (2007) 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, $N_n(\mu,\Sigma)$, avaliada 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 $n \times n$.

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

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

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 (1) se reduz à usual distribuição normal multivariada, $N_n(\mu,\Sigma)$.

A seguir, serão apresentados um lema, duas proposições e um corolário. 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 $\Lambda = (\Omega^{-1}+A^\top\Sigma^{-1}A)^{-1}$.

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.*

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.

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

2.2 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).

O banco de dados contém as variáveis que permitem o cálculo do parentesco, isto é, o número do indivíduo (1, 2, $\cdots$, 194), seu pai e sua mãe, caso sejam conhecidos. Além disso, também contém o sexo dos indivíduos (1: masculino e 2: feminino) e 3554 colunas correspondentes a respostas das intensidades das expressões gênicas 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 e Spielman, 2007).

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, sendo essas 3554 sondas disponibilizadas no GAW 15 para serem analisadas. Diversos trabalhos foram realizados com essas 3554 sondas, tais como o estudo do efeito da normalização na análise de ligação, ajuste de modelos multivariados para a análise conjunta das sondas, entre outros.

2.3 Modelos mistos normais assimétricos

O seguinte modelo, também encontrado em Sorensen e Gianola (2002), é ajustado para cada uma das 3554 sondas do banco de dados

$$Y = X\beta + Za + Wd + \varepsilon, \tag{2}$$

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.

Assume-se que os efeitos aleatórios do modelo apresentam as seguintes distribuições

$$a|\sigma_a^2, \delta_a \quad \sim \quad SN_{194}(0, \sigma_a^2 A, \delta_a I_{194}), \tag{3}$$

$$d|\sigma_d^2, \delta_d \quad \sim \quad SN_{194}(0, \sigma_d^2 D, \delta_d I_{194}) \tag{4}$$

$$\varepsilon|\sigma_\varepsilon^2, \delta_\varepsilon \quad \sim \quad SN_{194}(0, \sigma_\varepsilon^2 I_{194}, \delta_\varepsilon I_{194}). \tag{5}$$

Note que em (3) e (4), as matrizes de dispersão de a e d são $\sigma_a^2 A$ e $\sigma_d^2 D$, 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*. Lynch e Walsh (1998) apresentam os valores para o coeficiente de parentesco (ϕ), conforme o grau de parentesco entre os indivíduos, isto é, quando $i = j$, $\phi_{ij} = 1/2$, quando $i \neq j$ e os indivíduos i e j são irmãos ou pai e filho, então $\phi_{ij} = 1/4$, se os indivíduos i e j são avô e neto, então $\phi_{ij} = 1/8$, para os indivíduos i e j que não possuem relação familiar, então $\phi_{ij} = 0$.

A matriz D, em particular, de dimensão 194×194 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 e Walsh, 1998). Quando $i = j$, $D_{ij} = 1$. Se $i \neq j$ e os indivíduos i e j são irmãos, então $D_{ij} = 1/8$, para os outros relacionamentos entre os indivíduos i e j, $D_{ij} = 0$.

Através das suposições apresentadas em (3), (4) e (5) e utilizando a Proposição 2, o modelo (2) pode ser escrito da seguinte forma hierárquica:

$$\begin{array}{rcll} Y|\beta, a, d, \sigma_\varepsilon^2, \delta_\varepsilon & \sim & SN_{194}(X\beta + Za + Wd, \sigma_\varepsilon^2 I_{194}, \delta_\varepsilon I_{194}), & \\ a|\sigma_a^2, \delta_a & \sim & SN_{194}(0, \sigma_a^2 A, \delta_a I_{194}) \quad e & \\ d|\sigma_d^2, \delta_d & \sim & SN_{194}(0, \sigma_d^2 D, \delta_d I_{194}). & \end{array} \tag{6}$$

Note que a densidade condicional do vetor aleatório Y nos efeitos aleatórios (verossimilhança) é dada por (ver expressão (1))

$$\begin{aligned} f(Y|\beta, a, d, \sigma_\varepsilon^2, \delta_\varepsilon) \quad = \quad & 2^{194} \phi_{194}(Y|X\beta + Za + Wd, (\sigma_\varepsilon^2 + \delta_\varepsilon^2) I_{194}) \times \\ & \Phi_{194}\left(\frac{\delta_\varepsilon}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}(Y - X\beta - Za - Wd) \middle| 0, \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \delta_\varepsilon^2} I_{194} \right). \end{aligned} \tag{7}$$

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

2.4 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 vetor de parâmetros β de dimensão $p \times 1$. Para os parâmetros de escala, σ^2, considera-se a distribuição gama inversa, $GI(\frac{\tau}{2}, \frac{T}{2})$ e para os parâmetros de assimetria δ assume-se a distribuição normal univariada, isto é, $N(0, \gamma^2)$.

Considerando a distribuição condicional de Y apresentada de forma explícita em (7), as distribuições dos efeitos aleatórios apresentadas em (3) e (4) e as prioris especificadas, tem-se que a distribuição a posteriori conjunta de todas as quantidades envolvidas é dada por

$$
\begin{aligned}
&\pi(\beta, a, d, \sigma^2_\varepsilon, \sigma^2_a, \sigma^2_d, \delta_\varepsilon, \delta_a, \delta_d | y) \propto \phi_{194}(y | X\beta + Za + Wd, (\sigma^2_\varepsilon + \delta^2_\varepsilon) I_{194}) \\
&\times \Phi_{194}\left(\frac{\delta_\varepsilon}{\sigma^2_\varepsilon + \delta^2_\varepsilon}(y - X\beta - Za - Wd) \middle| 0, \frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon + \delta^2_\varepsilon} I_{194} \right) \\
&\times \phi_{194}(a | 0, \sigma^2_a A + \delta^2_a I_{194}) \\
&\times \Phi_{194}\left(\delta_a (\sigma^2_a A + \delta^2_a I_{194})^{-1} a \middle| 0, \left(I_{194} + \frac{\delta^2_a}{\sigma^2_a} A^{-1} \right)^{-1} \right) \\
&\times \phi_{194}(d | 0, \sigma^2_d D + \delta^2_d I_{194}) \\
&\times \Phi_{194}\left(\delta_d (\sigma^2_d D + \delta^2_d I_{194})^{-1} d \middle| 0, \left(I_{194} + \frac{\delta^2_d}{\sigma^2_d} D^{-1} \right)^{-1} \right) \\
&\times \exp\Big[-\frac{1}{2}(\beta - \beta_0)^\top S^{-1}_\beta (\beta - \beta_0) \Big] \\
&\times \left(\frac{1}{\sigma^2_\varepsilon} \right)^{\frac{\tau_\varepsilon}{2}+1} \exp\left[-\frac{T_\varepsilon}{2\sigma^2_\varepsilon} \right] \left(\frac{1}{\sigma^2_a} \right)^{\frac{\tau_a}{2}+1} \exp\left[-\frac{T_a}{2\sigma^2_a} \right] \left(\frac{1}{\sigma^2_d} \right)^{\frac{\tau_d}{2}+1} \exp\left[-\frac{T_d}{2\sigma^2_d} \right] \\
&\times \exp\Big[-\frac{1}{2}\left(\frac{\delta_\varepsilon}{\gamma_\varepsilon} \right)^2 \Big] \exp\Big[-\frac{1}{2}\left(\frac{\delta_a}{\gamma_a} \right)^2 \Big] \exp\Big[-\frac{1}{2}\left(\frac{\delta_d}{\gamma_d} \right)^2 \Big]. \qquad (8)
\end{aligned}
$$

Dada a forma algébrica da posteriori conjunta dada em (8), 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 (6) 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

$$\begin{aligned} Y|\beta,a,d,\sigma^2_\varepsilon,\delta_\varepsilon,w_\varepsilon &\sim N_n(X\beta+Za+Wd+\delta_\varepsilon w_\varepsilon,\sigma^2_\varepsilon I_n),\\ w_\varepsilon &\sim N_n(0,I_n)\mathtt{I}_{w_\varepsilon>0}, \end{aligned} \tag{9}$$

$$\begin{aligned} a|\sigma^2_a,\delta_a,w_a &\sim N_{q_a}(\delta_a w_a,\sigma^2_a A),\\ w_a &\sim N_{q_a}(0,I_{q_a})\mathtt{I}_{w_a>0}, \end{aligned} \tag{10}$$

$$\begin{aligned} d|\sigma^2_d,\delta_d,w_d &\sim N_{q_d}(\delta_d w_d,\sigma^2_d D),\\ w_d &\sim N_{q_d}(0,I_{q_d})\mathtt{I}_{w_d>0}, \end{aligned} \tag{11}$$

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.

Por meio do modelo completo especificado em (9)-(11) e as prioris apresentadas anteriormente, 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 em Oliveira (2009). A seguir são apresentados os resultados das condicionais completas.

$$\beta|a,d,\sigma^2_\varepsilon,\delta_\varepsilon,w_\varepsilon,Y \sim N_p(M_\beta^{-1}m_\beta,M_\beta^{-1}), \tag{12}$$

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

$$a|\beta,d,\sigma^2_\varepsilon,\sigma^2_a,\delta_\varepsilon,\delta_a,w_\varepsilon,w_a,Y \sim N_{q_a}(M_a^{-1}m_a,M_a^{-1}), \tag{13}$$

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

$$d|\beta,a,\sigma^2_\varepsilon,\sigma^2_d,\delta_\varepsilon,\delta_d,w_\varepsilon,w_d,Y \sim N_{q_d}(M_d^{-1}m_d,M_d^{-1}), \tag{14}$$

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

$$w_\varepsilon|\beta,a,d,\sigma^2_\varepsilon,\delta_\varepsilon,Y \sim N_n(M_{w_\varepsilon}^{-1}m_{w_\varepsilon},M_{w_\varepsilon}^{-1})\mathtt{I}_{w_\varepsilon>0}, \tag{15}$$

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

$$w_a|a,\sigma^2_a,\delta_a \sim N_{q_a}(M_{w_a}^{-1}m_{w_a},M_{w_a}^{-1})\mathtt{I}_{w_a>0}, \tag{16}$$

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

$$w_d | d, \sigma_d^2, \delta_d \sim N_{q_d}(M_{w_d}^{-1} m_{w_d}, M_{w_d}^{-1}) \mathtt{I}_{w_d > 0}, \tag{17}$$

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

$$\sigma_\varepsilon^2 | \beta, a, d, \delta_\varepsilon, w_\varepsilon, Y \sim GI(\frac{n + \tau_\varepsilon}{2}, \frac{T_\varepsilon + \mu_{\sigma_\varepsilon}^\top \mu_{\sigma_\varepsilon}}{2}), \tag{18}$$

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

$$\sigma_a^2 | a, \delta_a, w_a \sim GI(\frac{n + \tau_a}{2}, \frac{T_a + \mu_{\sigma_a}^\top A^{-1} \mu_{\sigma_a}}{2}), \tag{19}$$

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

$$\sigma_d^2 | d, \delta_d, w_d \sim GI(\frac{n + \tau_d}{2}, \frac{T_d + \mu_{\sigma_d}^\top D^{-1} \mu_{\sigma_d}}{2}), \tag{20}$$

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

$$\delta_\varepsilon | \beta, a, d, \sigma_\varepsilon^2, w_\varepsilon, Y \sim N(M_{\delta_\varepsilon}^{-1} m_{\delta_\varepsilon}, M_{\delta_\varepsilon}^{-1}), \tag{21}$$

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

$$\delta_a | a, \sigma_a^2, w_a \sim N(M_{\delta_a}^{-1} m_{\delta_a}, M_{\delta_a}^{-1}), \tag{22}$$

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

$$\delta_d | d, \sigma_d^2, w_d \sim N(M_{\delta_d}^{-1} m_{\delta_d}, M_{\delta_d}^{-1}), \tag{23}$$

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

$$h^2_{amplo} = \frac{Cov(a + d, y)}{Var(y)} \quad \text{e} \quad h^2_{restrito} = \frac{Cov(a, y)}{Var(y)}. \tag{24}$$

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*. Os valores iniciais e os detalhes computacionais se encontram na seção a seguir.

2.5 Implementação computacional

Como foi apresentado anteriormente, 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 modelos.

Para selecionar essas sondas, primeiramente, foram ajustados modelos mistos normais usuais para as 3554 sondas, isto é, observe o modelo 2, foi considerado β representando o sexo, como efeito fixo de dimensão 2×1, a, as famílias, como efeitos aleatórios de dimensão 14×1, com $a \sim N(0, S_a)$, em que $S_a = I_{14\times 14}$ e o vetor d foi desconsiderado. Através dos ajustes destes modelos, foram obtidos os resíduos estimados (fazendo $\hat{\varepsilon} = y - X\hat{\beta} - Z\hat{a}$) 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 $\hat{h}_a^2 = \hat{\sigma}_a^2/\hat{\sigma}_y^2$, em que $\hat{\sigma}_a^2$ foi obtido utilizando o algoritmo EM e EMVR 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 foram consideradas como variáveis resposta as intensidades de expressão gênica já normalizadas e o efeito de sexo foi considerado 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

$$
\begin{aligned}
L_{MMFcafe}(\beta^\top, f^\top, \sigma_f^2, \sigma_\varepsilon^2, \delta_f, \delta_\varepsilon|y) &= 2^{194}\phi_{194}(y|X\beta + Zf, (\sigma_\varepsilon^2 + \delta_\varepsilon^2)I_{194}) \\
&\quad \times \Phi_{194}\left(\frac{\delta_\varepsilon}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}(y - X\beta - Zf)|0, \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}I_{194}\right); \\
L_{MMFcaf}(\beta^\top, f^\top, \sigma_f^2, \sigma_\varepsilon^2, \delta_f|y) &= \phi_{194}(Y|X\beta + Zf, \sigma_\varepsilon^2 I_{194}); \\
L_{MMFcae}(\beta^\top, f^\top, \sigma_f^2, \sigma_\varepsilon^2, \delta_\varepsilon|y) &= 2^{194}\phi_{194}(y|X\beta + Zf, (\sigma_\varepsilon^2 + \delta_\varepsilon^2)I_{194}) \times \\
&\quad \times \Phi_{194}\left(\frac{\delta_\varepsilon}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}(y - X\beta - Zf)|0, \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \delta_\varepsilon^2}I_{194}\right); \\
L_{MMFsa}(\beta^\top, f^\top, \sigma_f^2, \sigma_\varepsilon^2|y) &= \phi_{194}(Y|X\beta + Zf, \sigma_\varepsilon^2 I_{194}). \qquad (25)
\end{aligned}
$$

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 (25), 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) = \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) = \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) = \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); \\
&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}). \qquad (26)
\end{aligned}
$$

Note que para se fazer um estudo mais detalhado, foram formuladas para cada sonda 16 configurações de modelos. 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 e 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 Resultados e discussão

Foram selecionadas duas sondas dentre as 3554 para explorar com detalhe diversas configurações de ajustes de modelos mistos normais assimétricos, segundo o critério apresentado no início da seção 2.5.

Para esse conjunto de dados a sonda que mais se destacou pelos critérios mencionados anteriormente foi a 1950. Optou-se por selecionar também a sonda 2323, para a comparação dos resultados.

Na Tabela 1, apresenta-se uma análise descritiva dos valores das intensidades já normalizadas.

Tabela 1 - 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

3.1 Seleção de modelos

Como foi mencionado anteriormente, foi implementado 16 configurações de modelos mistos apresentadas na Tabela 2.

Tabela 2 - Configurações de modelos mistos para cada sonda

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

Para cada configuração considerada calculou-se o numerador do fator de Bayes (conforme Kass e Raftery, 1995), para as sondas 1950 e 2323, respectivamente (ver os resultados na Tabela 3), para finalmente, efetuar-se o cálculo do fator de Bayes.

Tabela 3 - Resultados do numerador do fator de Bayes (conforme Kass e Raftery, 1995), para as sondas 1950 e 2323, para o cálculo do fator de Bayes

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

Pode-se observar na Tabela 3 que para a sonda 1950 os maiores valores para o numerador 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 o numerador 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).

Foram apresentados para cada sonda os resultados nas Tabelas 4 e 5, 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 3. Foi aplicado essa transformação para os resultados serem interpretados conforme Kass e Raftery (1995). 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.

Todos os resultados apresentados na Tabela 4 a seguir indicam evidências muito fortes a favor dos modelos apresentados no numerador, pois os valores encontrados são maiores que 10, 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.

Novamente, todos os resultados apresentados na Tabela 5 a seguir indicam evidências muito fortes a favor dos modelos apresentados no numerador. Logo, para as 16 configurações consideradas, o melhor modelo através do FB para a sonda 2323 é o MMADcaa.

Tabela 4 - Logaritmo natural do fator de Bayes multiplicado por 2 para os modelos destacados para a sonda 1950

Modelos (1)	MMAsa	MMFcaf
MMADcad	82,2562	163,0770
MMAsa		80,8209

(1) 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 5 - Logaritmo natural do fator de Bayes multiplicado por 2 para os modelos destacados para a sonda 2323

Modelos (1)	MMAcaa	MMFcaf
MMADcaa	66,9367	297,6420
MMAcaa		230,7050

(1)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.

3.2 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 e Brant, 1988; Albert e Chib, 1995).

A Figura 1 a seguir 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 1-(c)) e MMADcaa (Figura 1-(f)) apresentam resíduos bem menores que os demais modelos.

A Figura 2 a seguir 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 2-(c)) e MMADcaa (Figura 2-(f)).

Foram apresentados na Figura 3 os histogramas das amostras a posteriori dos parâmetros β (sexo), σ^2 (variância) e δ (assimetria, caso seja suposta no

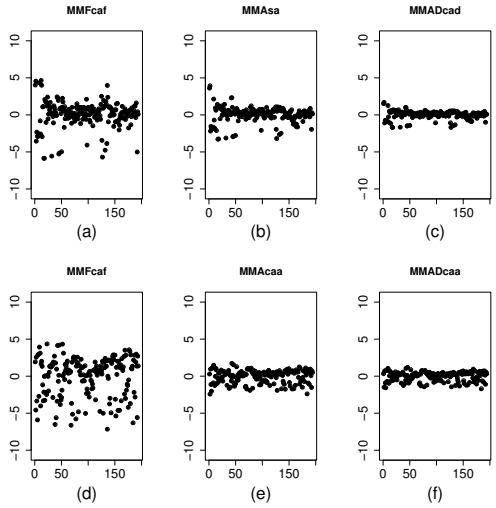

Figura 1 - Índice das observações no eixo das abscissas versus os resíduos preditos no eixo das ordenadas. Os gráficos (a), (b) e (c) apresentam os resultados para a sonda 1950 e os gráficos (d), (e) e (f), para a sonda 2323.

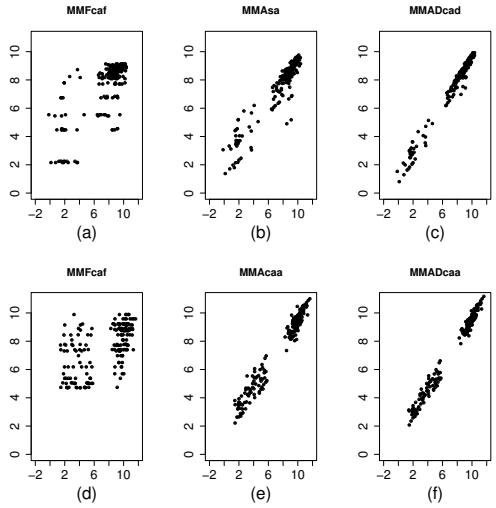

Figura 2 - Valores observados no eixo das abscissas versus valores preditos no eixo das ordenadas. Os gráficos (a), (b) e (c) apresentam os resultados para a sonda 1950 e os gráficos (d), (e) e (f), para a sonda 2323.

modelo) para os três melhores modelos através do FB, relativos às 16 configurações consideradas para a sonda 1950.

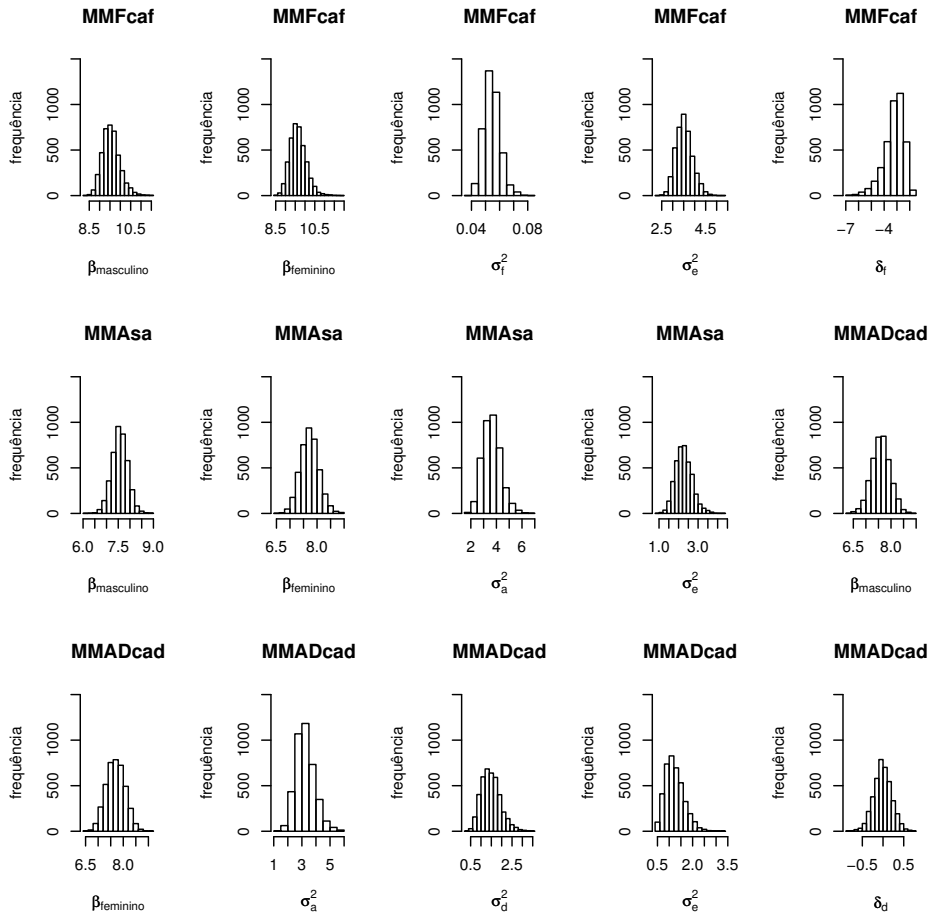

Figura 3 - Histogramas das amostras a posteriori dos parâmetros β (sexo), σ^2 (variância) e δ (assimetria, caso seja suposta no modelo) para os três melhores modelos através do FB, relativos às 16 configurações consideradas para a sonda 1950.

Foram apresentados na Tabela 6 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.

Por meio da Tabela 6 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

Tabela 6 - 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]

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 4 a seguir contém os histogramas das amostras a posteriori dos parâmetros β, σ^2 e δ para os três melhores modelos para essa sonda.

Foram apresentados na Tabela 7 a seguir 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.

Novamente, os resultados apresentados na Tabela 7 a seguir 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 6 e 7), 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.

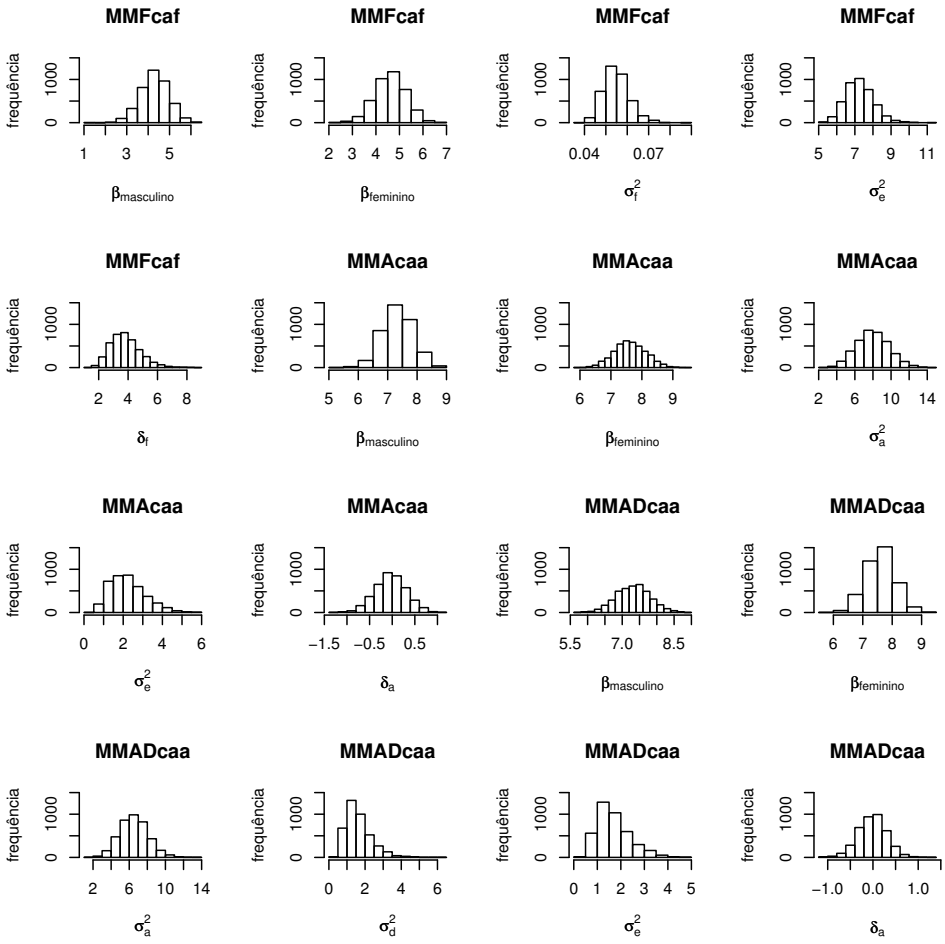

Figura 4 - Histogramas das amostras a posteriori dos parâmetros β (sexo), σ^2 (variância) e δ (assimetria, caso seja suposta no modelo) para os três melhores modelos através do FB, relativos às 16 configurações consideradas para a sonda 2323.

Tabela 7 - 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]

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^2_a 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 (24), 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 8 e 9.

Pode-se observar, com os resultados apresentados nas Tabelas 8 e 9, 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) (segundo White e Hodge, 1992).

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 8 - 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 9 - 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).

Conclusões

O modelo aditivo-dominante com assimetria mostrou-se mais provável para as duas sondas consideradas e através deste trabalho pode-se concluir que, apesar de relativamente trabalhoso do ponto de vista computacional, o modelo aditivo-dominante normal assimétrico é 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 e fazer predições dos valores genéticos com maior acurácia.

OLIVEIRA, D. C. R. de; BUENO FILHO, J. S. S. Bayesian analysis of skew normal mixed models in gene expression data from a complex pedigree. *Rev. Bras. Biom.,* São Paulo, v.28, n.2, p.137-160, 2010.

▪ *ABSTRACT: Estimates of heritability for gene expression are scarce and commonly originated from family structures, in which the variability of responses among and within families is provided under a uniform covariance structure for related individuals. 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. Thus, this work deals with the development and computational implementation of skew normal additive-dominance model for the analysis of microarrays by complex pedigrees, that allows skewness in all distributions of random effects. 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 skew normal additive-dominance model.*

▪ *KEYWORDS: MCMC; mixed models; multivariate skew normal distribution; Bayesian inference.*

Referências

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

ARELLANO-VALLE, R. B.; BOLFARINE, H.; LACHOS, V. H. Bayesian inference for skew-normal linear mixed models. *J. Appl. Stat.*, Abingdon, v.34, n.6, p.663-682, 2007.

ARELLANO-VALLE, R. B.; GENTON, M. G. On fundamental skew distributions. *J. Multiv. Anal.*, New York, v.96, n.1, p.93-116, 2005.

AZZALINI, A.; DALLA-VALLE, A. The multivariate skew-normal distribution. *Biometrika*, London, v.83, n.4, p.715-726, D1996.

CHALONER, K.; BRANT, R. A. Bayesian approach to outlier detection and residual analysis. *Biometrika*, London, v.75, n.4, p.651-659, 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, 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. *Brief. Bioinform.*, London, v.18, p.105-110, 2002. Supllement.

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

KASS, R. E.; RAFTERY, A. E. Bayes factors. *J. Am. Stat. Assoc.*, New York, v.90, n.430, p.773-795, 1995.

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

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

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

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, 2004.

OLIVEIRA, D. C. R. *Modelos mistos normais assimétricos em dados de microarrays originados de pedigrees complexos*, 2009. 106f. Tese (Doutorado em Estatística e Experimentação Agronômica) - Universidade Federal de Lavras, Lavras, 2009.

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.

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, 2008. Disponível em: $< http://www.biomedcentral.com/1471-2164/9/25 >$. Acesso em: 15 set. 2008.

ROHR, P. von; HOESCHELE, I. Bayesian QTL mapping using skewed Student t distributions. *Genet. Sel. Evol.*, 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. *Rev. Bras. Zootec.*, Viçosa, v.36, n.4, p.186-209, 2007.

SARAIVA, E. F.; MILAN, L. A.; DIAS, T. C. M. Métodos estatísticos aplicados à análise da expressão gênica. *Bol. ISBrA*, São Paulo, v.1, n.3, p.5-8, 2007.

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

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

SPEED, T. P. *Statistical analysis of gene expression microarray data.* Boca Raton: CRC, 2003. 12 p.

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

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

Recebido em 15.12.2009.

Aprovado após revisão em 13.04.2010.