Influência de Funções de Covariâncias Sobre o Modelo Fatorial Latente Esparso com Interações

Erick da Conceição Amorim

Departamento de Estatística - ICEx - UFMG

Fevereiro de 2016

Influência de Funções de Covariâncias Sobre o Modelo Fatorial Latente Esparso com Interações

Erick da Conceição Amorim

Orientador: Vinícius Diniz Mayrink

Dissertação submetida ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como parte dos requisitos necessários à obtenção do grau de Mestre em Estatística.

Departamento de Estatística
Instituto de Ciências Exatas
Universidade Federal de Minas Gerais

Belo Horizonte, MG - Brasil
Fevereiro de 2016

Aos meus pais.

Agradecimentos

Agradeço aos meus queridos e amados pais, Marly Rose, Silvio César e a toda minha família: irmã, tia Marcia, tia Wera, vó Creuza e vó Fátima, por sempre estarem ao meu lado.

À todos os professores da UFPA principalmente ao prof. Héliton Tavares pela paciência, incentivo e apoio na graduação e durante minha vinda para UFMG.

Ao meu orientador Vinícius, pelo incentivo, paciência e por todo o conhecimento passado.

Aos colegas que fiz em especial: Guilherme Lopes, Jéssica Assunção, Rafael Procópio, Danielle Resende, Douglas Mateus, Ana Cláudia (Baldini) e Ramona Paula. Pelos momentos alegres e divertidos nos bares.

E por fim, agradeço a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio durante o mestrado através de uma bolsa de estudos.

Resumo

A análise fatorial é uma ferramenta estatística bastante utilizada para identificar um número reduzido de fatores que explicam o relacionamento entre diversas variáveis em um conjunto de dados. Neste trabalho, exploramos essa técnica com uma abordagem Bayesiana no contexto de análise de expressão de genes. Inicialmente, estudamos o modelo fatorial latente simples e verificamos seu ajuste a um conjunto de dados simulados. Em seguida, analisamos o modelo fatorial latente com interações juntamente com distribuições *a priori* esparsas para testar se os fatores, definidos para regiões com alteração do número de cópias, estariam afetando genes localizados em outras regiões do genoma. A interação não linear foi introduzida no modelo por meio de um Processo Gaussiano que apresenta em sua estrutura uma função de covariâncias que será o foco principal neste trabalho. O comportamento e desempenho do modelo fatorial latente esparso com interações será avaliado a partir de simulações utilizando diferentes funções de covariâncias: exponencial quadrática, como abordado em Mayrink e Lucas (2013), exponencial potência e funções da classe Matérn que se distinguem em termos da escolha dos parâmetros suavizadores. Uma análise de sensibilidade é realizada considerando estas configurações, os resultados indicam que algumas especificações fornecem melhores ajustes que outras. Para finalizar, uma aplicação a dados reais é mostrada considerando a configuração de covariâncias com indicação de melhores resultados no caso simulado.

Palavras-chave: Inferência Bayesiana, *Priori* Esparsa, Processo Gaussiano, Genoma.

Abstract

The factor analysis is an statistical tool widely used to identify a reduced number of factors supposed to explain the relationship between many variables in a data set. In this work, we explore this technique using the Bayesian approach in context of the analysis of gene expression. Initially, we study the simple latent factor model and verify its performance to fit simulated data. Next, we evaluate the latent factor model with interactions assuming sparse prior distributions to test whether the factors, defined for regions with copy number alterations, would affect genes located in other regions of the genome. The interaction was introduced in the model through a Gaussian process having in its structure a covariance function which is a key element in our study. The behavior and performance of the sparse latent factor model with interactions was evaluated through simulations using different covariances functions: quadratic exponential, as discussed in Mayrink and Lucas (2013), power exponential and some functions options in the Matern class that differ in terms of the choice of the smoothing parameters. A sensitivity analysis is made considering these settings and the results indicate that some specifications provide a better model fit than others. Finally, an application involving real data is presented considering the covariance function exhibiting the best results.

Keywords: Bayesian Inference, Sparsity prior, Gaussian Process, Genome.

Sumário

Capítulo 1

Introdução

Analisar dados com alta dimensão requer técnicas especiais tais como seleção de variáveis ou redução da dimensão. O modelo fatorial é uma ferramenta flexível e poderosa para a análise de dependência multivariada e para a verificação de padrões e associações nos dados. A principal função da análise fatorial é reduzir ou resumir a informação em uma grande quantidade de variáveis a um número pequeno de fatores, que são usados para identificar características subjacentes principais e associações entre as variáveis.

Com os avanços computacionais, muitos estudos tem usado o modelo fatorial com a abordagem Bayesiana, em particular, modelos Bayesianos têm levado a resultados interessantes para análise de expressão de genes. West (2003) introduziu o modelo fatorial latente esparso como uma extensão de um modelo de regressão esparso, e mostrou a capacidade do modelo fatorial em identificar e estimar padrões e grupos de genes relacionados a fenômenos biológicos. Lucas et al. (2006) também utiliza um modelo fatorial latente hierárquico com uma distribuição *a priori* esparsa para as cargas e obtém grandes melhorias na identificação de padrões complexos em termos de covariações entre genes. A complexa rede de dependência entre genes motivou Mayrink e Lucas (2013) a construir um modelo fatorial com interações, pois a interação é introduzida para explicar uma parte dessa dependência. Interações são bastante comuns no contexto de regressão, elas podem ser introduzidas através de um o produto de covariáveis. Neste caso, o modelo descreve uma relação multiplicativa entre duas covariáveis influenciando a variável resposta;

lembramos que outras formas de interações podem ser consideradas em regressão.

Modelos com interações não lineares também tem sido estudados por diversos pesquisadores e em muitos casos a não linearidade é introduzida no modelo a partir de uma especificação *a priori* de um Processo Gaussiano. Henao e Winther (2011) consideram variáveis latentes esparsas e modelos Bayesianos lineares para uma análise parcimoniosa de dados multivariados. O uso dessas ferramentas consiste em uma hierarquia completa em modelos esparsos usando uma distribuição *a priori* do tipo mistura com uma componente com ponto de massa em zero, fatores latentes não-Gaussianos e uma busca estocástica sobre a ordenação das variáveis. Os autores argumentam que o modelo é flexível de modo que ele pode ser estendido ao trocar a distribuição *a priori* estabelecida para o conjunto de variáveis latentes por um Processo Gaussiano, que permite a não linearidade entre as variáveis observadas.

Lawrence (2004) e Lawrence (2005) exploram a interação não linear entre fatores latentes propondo um modelo de regressão com Processo Gaussiano. Além disso, eles introduzem uma interpretação probabilística para a análise de componentes principais chamada de Análise de Componente Principal Probabilística Dual (DPPCA). A vantagem desse modelo estaria no fato de que o mapeamento linear do espaço latente para o espaço de dados pode ser normalizado pelo Processo Gaussiano que é assumido para as variáveis latentes. O DPPCA com um Processo Gaussiano introduz não linearidade no modelo a partir de uma função de covariâncias que define a associação entre as variáveis, com isso o modelo passa a ser chamado de Modelo de Variáveis Latentes com Processo Gaussiano (MVL-GP).

O uso do Processo Gaussiano é um tópico abordado por Mayrink e Lucas (2013) que estudaram o efeito de interação não linear no modelo fatorial latente esparso. Eles utilizaram a função de covariâncias exponencial quadrática e verificaram a forma do efeito de interação entre fatores por meio de um parâmetro de comprimento-escala. Este parâmetro tem um papel fundamental na estimação das interações, pois se aumentarmos o seu valor estaremos suavizando a superfície estimada que representaria a interação entre fatores.

A função de covariâncias exponencial quadrática não é a única opção. Outras for-

mas de funções podem ser consideradas incluindo mais de um parâmetro que controla a suavização. Pretendemos explorar a função de covariâncias Exponencial Potência e funções da classe Matérn. Neste trabalho, usamos o modelo fatorial latente esparso com interações para identificar o efeito não linear de interação entre fatores sobre grupos de genes. Utilizamos inicialmente a função de covariâncias exponencial quadrática para confirmar os resultados obtidos por Mayrink e Lucas (2013), em seguida utilizamos as funções de covariâncias exponencial potência e algumas da classe Matérn para comparar e verificar através de dados simulados se há melhoras na estimação dos parâmetros do modelo. Uma aplicação envolvendo dados reais também foi desenvolvida no contexto do estudo de alteração do número de cópias em diferentes partes do genoma.

1.1 Organização da dissertação

Esta dissertação está organizada como segue. O Capítulo 2, apresentado a seguir, irá descrever alguns conceitos básicos de inferência Bayesiana, incluindo a análise conjugada e a descrição de dois dos principais métodos computacionais (algoritmo *Gibbs Sampling* e o Metropolis-Hastings) para amostragem indireta da distribuição *a posteriori*.

O Capítulo 3 descreve resumidamente a análise de expressão de genes e o pré-tratamento dos dados que serão utilizados e motivarão este trabalho. Além disso, fizemos um estudo simulado analisando o ajuste de um modelo fatorial latente simples.

O Capítulo 4 trata da aplicação de uma modelagem com interações entre fatores definidas para grupos de genes afetados por regiões do genoma apresentando uma alteração do número de cópias do DNA. Neste capítulo, simulamos um conjunto de dados contendo grupos de genes afetados por estas regiões e um grupo que não é afetado. Consideramos um modelo em que cada fator estará associado com cada grupo de gene afetado por estas regiões. Um grupo extra que não é afetado poderá ter relação com estes fatores e/ou com uma interação deles. Essa interação é introduzida por meio de um processo Gaussiano que apresenta a função de covariâncias exponencial quadrática. A existência do efeito de interação entre fatores é feita pelo uso de distribuições *a priori* esparsas.

No Capítulo 5 fizemos um estudo abordando as funções de covariâncias exponencial

potência e algumas da classe Matérn. Apresentamos alguns resultados de um estudo simulado onde comparamos os modelos com distintas funções de covariâncias. Nesta etapa, diferentes valores dos parâmetros de suavização são considerados.

No Capítulo 6 apresentamos uma aplicação do modelo com interações envolvendo um conjunto de dados reais representando expressões de genes para o câncer de mama. Utilizamos aqui uma função de covariâncias da classe Matérn que mostrou bons resultados no capítulo 5.

Finalmente, o Capítulo 7 fará um resumo de tudo aquilo que foi apresentado e discutido neste trabalho, além de apresentar algumas propostas de trabalhos futuros.

Capítulo 2

Conceitos Básicos de Inferência Bayesiana

Ao selecionar uma família $\Im$ de uma classe de distribuições de probabilidades que são indexadas por um parâmetro $\theta \in \Theta$ estamos atribuindo a um processo ou fenômeno físico, que gera observações, um modelo estatístico paramétrico. As observações $y = (y_1, y_2, ..., y_n)$ são uma amostra aleatória extraída da população que se distribui de acordo com a densidade $p(y \mid \theta) \in \Im$. O parâmetro θ representa a característica de interesse que se quer conhecer para poder ter uma descrição completa do modelo ou tomar decisões em algum procedimento. Por exemplo, considere um conjunto de medições y_i com $i = 1, 2, ..., n$, tal que $y_i \sim N(\theta, \sigma^2)$, com variabilidade σ^2 conhecida. Neste caso, $y_i = \theta + \epsilon_i$ é o modelo que será descrito pela seguinte função de verossimilhança assumindo independência condicional entre os y_i.

$$p(y \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \theta)^2\right\};$$

vale ressaltar que a quantidade θ é mais do que um parâmetro indexador, pois ela é a média das observações e estaríamos interessados na determinação do seu valor.

Deixando o exemplo da Normal acima e considerando o caso geral $p(y \mid \theta)$, o parâmetro θ em um modelo estatístico é uma quantidade desconhecida. O principal interesse em problemas de inferência é conhecê-lo ou caracterizá-lo. Por isso, é possível que a informação a respeito de θ seja incorporada a um modelo estatístico capaz de descrever

a incerteza inicial que se tem sobre θ como uma medida de probabilidade. Então, é necessário construir uma distribuição de probabilidade *a priori* para θ que irá resumir toda a informação disponível sobre θ previamente conhecida antes da realização do experimento, isto é, antes de observarmos a amostra. Após observarmos os dados o processo de inferência para θ será baseado na distribuição *a posteriori*, que é uma atualização da distribuição *a priori*. A inferência sobre θ é feita a partir da informação trazida pelos dados (amostra) juntamente com a informação que se tem de θ antes de observarmos os dados.

Em estatística Bayesiana a regra de Bayes é utilizada como mecanismo para a atualização da distribuição *a priori*. Ela é dada pela expressão:

$$p(\theta \mid y) = \frac{p(y, \theta)\ p(\theta)}{p(y)} = \frac{p(y \mid \theta)p(\theta)}{\int_\Theta p(y \mid \theta)p(\theta)d\theta}. \tag{2.1}$$

Os termos $p(\theta \mid y)$ e $p(\theta)$ em (2.1) medem, respectivamente, a incerteza sobre θ após observarmos os dados (distribuição *a posteriori*) e antes de observá-los (distribuição *a priori*). Vale ressaltar que o denominador em (2.1), no caso de distribuições discretas pode ser escrito como $p(y) = \sum_\theta\ p(y \mid \theta)\ p(\theta)$, além disso ele é visto como uma constante que não depende de θ o qual chamamos de distribuição preditiva. Essa distribuição tem grande importância no cálculo dos fatores de Bayes que são usados como um possível critério para comparar modelos. Com isso, a regra de Bayes pode ser escrita de forma mais resumida considerando apenas a parte que depende de θ, como segue:

$$p(\theta \mid y) \propto p(y \mid \theta)\ p(\theta).$$

No caso de $\theta = (\theta_1, \theta_2, ..., \theta_p)$ ser um vetor, as distribuições marginais e condicionais completas de cada θ_i podem ser obtidas a partir da distribuição *a posteriori* conjunta $p(\theta_1, \theta_2, ..., \theta_p \mid y)$. Então a distribuição marginal *a posteriori* de θ_i será:

$$p(\theta_i \mid y) = \int p(\theta_1, \theta_2, ..., \theta_p \mid y)d\theta_{-i};$$

sendo $\theta_{-i} = (\theta_1, \theta_2, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p)$ o vetor paramétrico sem a i-ésima componente. De maneira análoga, pode-se obter distribuições para subvetores de θ. Além do mais,

várias distribuições condicionais podem ser obtidas da seguinte forma:

$$p(\theta_i \mid \theta_{-i}, y) = \frac{p(\theta_1, \theta_2, ..., \theta_p \mid y)}{p(\theta_{-i} \mid y)}.$$

A distribuição de $(\theta_i \mid \theta_{-i}, y)$, é chamada de distribuição condicional completa *a posteriori* e elas são muito importantes em aplicações do algoritmo computacional *Gibbs Sampling*, um dos métodos Monte Carlo via Cadeias de Markov (MCMC), que será descrito adiante neste trabalho.

Este capítulo está organizado da seguinte maneira: A seção 2.1 irá descrever como propor uma distribuição *a priori* por meio de famílias de distribuições que facilitam a caracterização de modelos, as famílias conjugadas. Na seção 2.2 será apresentado uma breve descrição do algoritmo *Gibbs Sampling*, utilizado para gerar amostras da distribuição *a posteriori*. A seção 2.3 apresenta outro método de amostragem da distribuição *a posteriori* conhecido como Metropolis-Hastings, que também é um método MCMC. E a seção 2.4 finaliza este capítulo apresentando as principais conclusões.

2.1 Famílias Conjugadas

Uma maneira de escolher a distribuição *a priori* é por meio de conjugação. Ao construirmos famílias conjugadas devemos buscar uma família de uma classe de distribuições ζ versátil e ampla o bastante, para acomodar diferentes opiniões e informações *a priori*, permitindo uma boa interpretabilidade dos parâmetros e simplicidade no cálculo da distribuição *a posteriori*. Por definição, considere $\Im = \{p(y \mid \theta) : \theta \in \Theta\}$ uma família de distribuições relacionadas à amostra $y = (y_1, y_2, ..., y_n)$ e seja $\zeta = \{p(\theta) : \theta \in \Theta\}$ a classe de distribuições de probabilidade sobre θ. Dizemos que ζ e $\Im$ são famílias conjugadas se:

(i) A classe ζ é fechada por amostragem de $\Im$, isto é, o núcleo da distribuição $p(y \mid \theta)$ é proporcional a algum membro de ζ, e

(ii) A classe ζ é fechada com relação ao produto, ou seja, ao multiplicar $p(y \mid \theta)$ e $p(\theta)$ obtem-se uma distribuição *a posteriori* na mesma classe ζ da distribuição *a priori*.

Considere, por exemplo, $y = (y_1, y_2, ..., y_n)$ variáveis aleatórias que dado $\theta \in [0, 1]$ são independentes e identicamente distribuídas com distribuição Bernoulli. Então a verossimilhança é escrita da seguinte forma:

$$p(y \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta) = \theta^{\sum_{i=1}^{n} y_i} (1-\theta)^{n-\sum_{i=1}^{n} y_i}.$$

Uma candidata à família conjugada é dada por:

$$p(\theta) \propto \theta(1-\theta), \tag{2.2}$$

na qual as propriedades (i) e (ii) são satisfeitas, mas a distribuição em (2.2) não é flexível. Para colocar flexibilidade pode-se introduzir hiperparâmetros, assim a equação (2.2) pode ser escrita como:

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}; \tag{2.3}$$

então tem-se que (2.3) representa o núcleo de uma distribuição Beta com hiperparâmetros $a > 0$ e $b > 0$. E por construção, essa distribuição é fechada por amostragem, logo, basta verificar se a mesma é fechada por produto. Com isso tem-se:

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta)p(\theta) = \theta^{\sum_{i=1}^{n} y_i} (1-\theta)^{n-\sum_{i=1}^{n} y_i} \theta^{a-1}(1-\theta)^{b-1} \\ &\propto \theta^{a+\sum_{i=1}^{n} y_i - 1}(1-\theta)^{b+n-\sum_{i=1}^{n} y_i - 1}. \end{aligned} \tag{2.4}$$

Assim, a expressão em (2.4) apresenta o núcleo de uma Beta$(a + \sum_{i=1}^{n} y_i; b + n - \sum_{i=1}^{n} y_i)$, logo a classe de distribuições Beta é uma família conjugada para o modelo Bernoulli.

Outro exemplo importante que envolve conjugação é o caso Normal. Considere $y = (y_1, y_2, ..., y_n)$ uma amostra aleatória que dado θ tem distribuição $N(\mu, \theta)$, sendo μ conhecido e $\theta > 0$. Note que a função de verossimilhança será:

$$\begin{aligned} p(y \mid \theta) &= \prod_{i=1}^{n} p(y_i \mid \theta) \\ &= (2\pi)^{-\frac{n}{2}} \left(\frac{1}{\theta}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\theta}\sum_{i=1}^{n}(y_i - \mu)^2\right\}. \end{aligned} \tag{2.5}$$

Uma família de distribuições de probabilidade sobre θ que é conjugada em relação a expressão em (2.5) será:

$$p(\theta) = \frac{b^a}{\Gamma(a)} \left(\frac{1}{\theta}\right)^{a+1} \exp\left\{-\frac{1}{\theta}b\right\};\ a > 0 \text{ e } b > 0. \tag{2.6}$$

Em (2.6) temos a densidade da distribuição Gama-Invertida com parâmetros de escala $b > 0$ e de forma $a > 0$. A densidade *a posteriori* é obtida via regra de Bayes como segue:

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta)p(\theta) \\ &\propto \left(\frac{1}{\theta}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\theta}\sum_{i=1}^{n}(y_i - \mu)^2\right\} \left(\frac{1}{\theta}\right)^{a+1} \exp\left\{-\frac{1}{\theta}b\right\} \\ &\propto \left(\frac{1}{\theta}\right)^{\frac{n}{2}+a+1} \exp\left\{-\frac{1}{\theta}\left(b + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right)\right\}. \end{aligned}$$

Portanto, $(\theta \mid y)$ tem distribuição Gama-Inversa com parâmetros de forma $\frac{n}{2} + a$ e escala $b + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2$.

A conjugação é sempre válida para classes ζ muito abrangentes como, por exemplo, a classe de todas as distribuições contínuas, e quase nunca para uma classe contendo apenas uma única família de distribuição. Assim, o bom senso deve ser usado para utilizar classes de distribuições que não sejam tão amplas, mas para as quais a propriedade de conjugação é válida.

Uma vez obtida a distribuição *a posteriori*, se ela tiver forma fechada, podemos realizar de forma direta a descrição e inferência sobre o parâmetro de interesse. Entretanto, na maioria das vezes não é trivial a obtenção da distribuição *a posteriori*, isso ocorre devido a dificuldades imposta pela integral do denominador da regra de Bayes. Em muitos casos essa constante normalizadora não pode ser calculada e o que se tem é apenas o núcleo da distribuição alvo. Para contornar este problema e obter uma amostra da distribuição *a posteriori*, serão utilizados alguns métodos computacionais que são baseados em simulações estocásticas via cadeias de Markov. Nesse caso, pode-se gerar uma amostra da distribuição *a posteriori*, cuja forma não é conhecida, a qual será usada para construir, histogramas e estatísticas (estimativas) que irão resumir a informação sobre o parâmetro.

2.2 *Gibbs Sampling*

O *Gibbs Sampling* foi a primeira classe de esquemas largamente empregada para simulação estocástica via cadeia de Markov, ele foi proposto originalmente dentro do contexto de reconstrução de imagens por Geman e Geman (1984). Estes autores, propuseram um esquema de amostragem da distribuição de Gibbs explorando justamente suas condicionais completas, por meio de um algoritmo iterativo que define uma cadeia de Markov. Mais tarde, Gelfand e Smith (1990) compararam esse amostrador com outros esquemas de simulação estocástica e mostraram que o *Gibbs Sampling* poderia ser utilizado, em muitos casos, para amostrar de distribuições *a posteriori*.

Para implementar o *Gibbs Sampling* devemos ser capazes de gerar observações a partir de cada distribuição condicional completa $p(\theta_i \mid \theta_{-i}, y)$. As amostras podem ser geradas de maneira direta, se as distribuições condicionais completas tem forma conhecida; ou de maneira indireta, quando as condicionais completas não tem forma conhecida. Nesse último caso, alternativas para gerar da condicional completa *a posteriori* são, por exemplo, os algoritmos *Adaptive Rejection Sampling* (ARS) de Gilks e Wild (1992), o *Slice Sampling* de Neal (2003) ou o Metroplis-Hastings de Metropolis et al. (1953) e Hastings (1970), que será visto adiante, como um passo dentro do *Gibbs Sampling*.

O *Gibbs Sampling* é configurado da seguinte forma:

Passo 1 Escolha os valores iniciais $\theta^{(0)} = (\theta_1^{(0)}, \theta_1^{(0)}, ..., \theta_p^{(0)})$ e inicialize o contador de iterações $t = 1$.

Passo 2 Obtenha os novos valores $\theta^{(t)} = (\theta_1^{(t)}, \theta_1^{(t)}, ..., \theta_p^{(t)})$ a partir de sucessivas gerações:

gerar $\theta_1^{(t)}$ de $p(\theta_1 \mid \theta_2^{(t-1)}, \theta_3^{(t-1)}, ..., \theta_p^{(t-1)}, y)$;

gerar $\theta_2^{(t)}$ de $p(\theta_2 \mid \theta_1^{(t)}, \theta_3^{(t-1)}, ..., \theta_p^{(t-1)}, y)$;

$\vdots$

gerar $\theta_p^{(t)}$ de $p(\theta_p \mid \theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_{p-1}^{(t)}, y)$.

Passo 3 Faça $t = t + 1$ e retorne ao passo 2 até obter a amostra desejada após a convergência das cadeias.

Conforme t cresce, a cadeia aproxima-se de uma condição de equilíbrio. Os valores de $\theta^{(t)}$ obtidos até uma iteração t_0 serão descartados da análise, pois nesta fase a cadeia passa por um período de aquecimento (*burn-in*). A partir da iteração $t_0 + 1$ os valores de $\theta^{(t)}$ serão tratados como amostras da distribuição *a posteriori* $p(\theta \mid y)$. As amostras geradas de $p(\theta \mid y)$ não são totalmente aleatórias, pois há uma dependência entre os valores da cadeia deixando as observações correlacionadas. Uma estratégia utilizada para minimizar esse problema é formar a amostra final com os valores selecionados a cada L iterações, o que chamamos de *lag*.

2.3 Metropolis-Hastings

Quando a distribuição de interesse, por exemplo a condicional completa *a posteriori*, não tem forma conhecida, não será possível gerar amostras diretamente em uma parte do algoritmo *Gibbs Sampling*. Nesse caso, o algoritmo Metropolis-Hastings é uma alternativa que considera o núcleo $h(\theta)$ obtido via regra de Bayes e uma distribuição geradora de candidatos em um teste de aceitação/rejeição. Em outras palavras, se desejamos gerar valores da distribuição *a posteriori* $p(\theta \mid y) \propto h(\theta) \equiv p(y \mid \theta)p(\theta)$, precisamos especificar uma distribuição geradora de candidatos $q(\theta^* \mid \theta^{(t-1)})$ válida no domínio de θ e condicionada em $\theta^{(t-1)}$. Dado o valor inicial $\theta^{(0)}$, o algoritmo procede da seguinte maneira:

Passo 1 gerar θ^* de $q(\theta^* \mid \theta^{(t-1)})$;

Passo 2 calcule a razão $r = \dfrac{h\left(\theta^*\right) q\left(\theta^{(t-1)} \mid \theta^*\right)}{h\left(\theta^{(t-1)}\right) q\left(\theta^* \mid \theta^{(t-1)}\right)}$ e considere $\alpha = \min\{1, r\}$;

Passo 3 gere $u \sim U(0,1)$. Se $u < \alpha$ faça $\theta^{(t)} = \theta^*$, caso contrário $\theta^{(t)} = \theta^{(t-1)}$

Passo 4 Faça $t = t + 1$ e retorne ao Passo 1 até obter a amostra desejada após a convergência das cadeias.

Existe flexibilidade para a escolha da distribuição candidata $q(\theta^* \mid \theta^{(t-1)})$. Uma opção mais usual seria propor uma distribuição simétrica em seus argumentos como a normal, neste caso $q(\theta^* \mid \theta^{(t-1)}) = q(\theta^{(t-1)} \mid \theta^*)$, então a razão no Passo 2 do algoritmo poderia

ser simplificada para $r = \dfrac{h(\theta^*)}{h(\theta^{(t-1)})}$. Entretanto, é importante destacar que a escolha da dispersão de $q(\theta^* \mid \theta^{(t-1)})$ influencia na autocorrelação da cadeia. Valores altos para a variância geram taxas de aceitação muito baixas e isso leva a uma cadeia que permanece constante por várias iterações, entretanto, valores baixos permitem movimentos pequenos no espaço paramétrico e determinam taxas de aceitação altas, porém a cadeia terá forte autocorrelação.

Utilizaremos neste trabalho o Metropolis-Hastings dentro do *Gibbs Sampling* para gerar amostras das distribuições *a posteriori* condicionais completas desconhecidas. A ideia é combinar os algoritmos *Gibbs Sampling* e Metropolis-Hastings construindo um *Gibbs Sampling* que em alguns de seus passos executa uma iteração do Metropolis-Hastings, ou seja, as observações que não podem ser geradas diretamente de suas distribuições condicionais completas, serão geradas através do Metropolis-Hastings (uma única iteração) dentro do ciclo amostral do *Gibbs Sampling*. Isso é conhecido na literatura como *Metropolis within Gibbs*, mais detalhes podem ser vistos em Gamerman e Lopes (2006).

2.4 Conclusões do Capítulo

Este capítulo apresentou alguns conceitos básicos sobre inferência Bayesiana indicando a regra de Bayes como ferramenta para atualizar a distribuição *a priori*. Além disso, vimos que uma escolha adequada da distribuição *a priori* facilita a análise ao obtermos a conjugação. Apresentamos também uma breve descrição dos principais métodos computacionais MCMC utilizados para amostragem indireta da distribuição *a posteriori*: os algoritmos *Gibbs Sampling* e Metropolis-Hastings. O capítulo seguinte apresentará essa abordagem no contexto de análise de expressão de genes, onde um modelo fatorial simples será usado para modelar um padrão de expressão genético.

Capítulo 3

Análise de Expressão de Genes

Recentes tecnologias envolvendo sequências de oligonucleotídeos curtos (fragmentos de DNA ou RNA com 25-30 bases) em pequenos *chips* estão sendo usadas para a construção de plataformas de *microarray*s. O sistema GeneChip, produzido pela Affymetrix (`http://www.affymetrix.com/estore/`), utiliza sequências curtas de oligonucleotídeos depositados em um *chip* configurando em um "grid"com *probes*. As *probes* contêm materiais genéticos compostos por sequências que são projetadas (conhecidas) para combinar com outras sequências genéticas extraídas de uma amostra. Um conjunto composto por 11-20 pares de *probes* formam um *probe set*, que neste trabalho, iremos considerar como representação de um gene de interesse. Assim, neste *chip* que contém sequências genéticas é aplicado uma solução com material de células que se quer analisar. Esse material genético recebe uma marcação fluorescente, e as sequências presentes na solução poderão encontrar seus pares fixos no *chip* e, se isso acontecer, haverá conexão. Esse processo é conhecido como hibridização. A Figura 3.1 apresenta uma imagem ilustrativa desse processo.

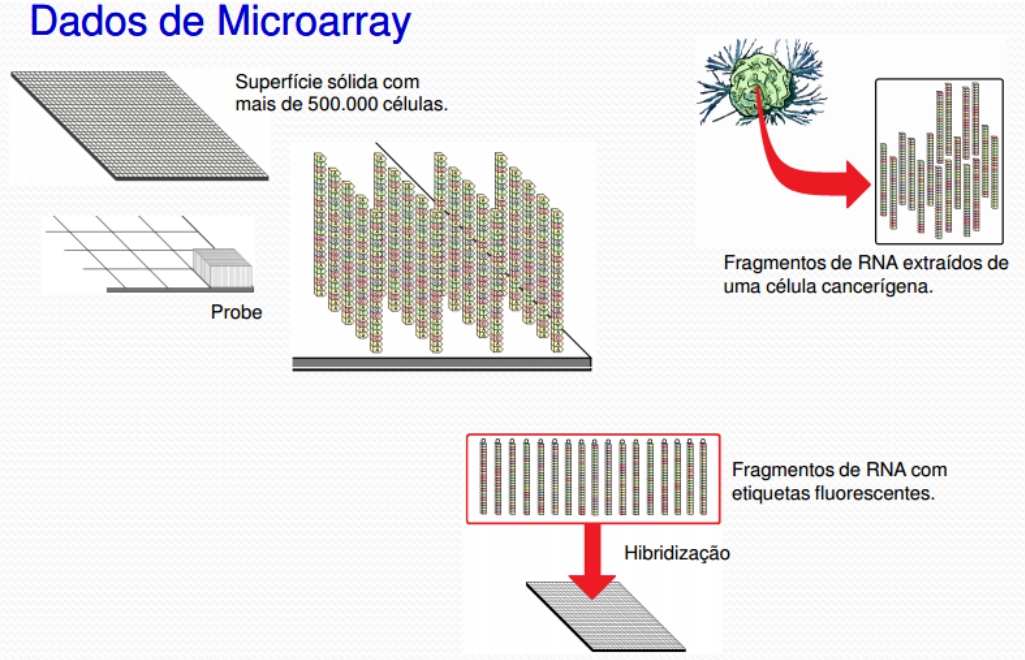

Figura 3.1: *Representação da configuração de um chip usado para a construção de plataformas de microarrays.*

O processo de hibridização por complementariedade dos oligonucleotídeos é avaliado a partir de dois tipos de *probes*: *Perfect Match* (PM) e *Mismatch* (MM). O PM tem uma sequência idêntica a um trecho de um dado gene; neste caso, a ligação entre a sequência do *chip* e da solução é perfeita. Por outro lado, o MM apresenta uma sequência de oligonucleotídeos com a base do meio (13ª posição) alterada; aqui a ligação entre as sequências não ocorre perfeitamente resultando no que se chama de hibridização cruzada. O propósito do *probe* tipo MM é justamente permitir que se investigue a ocorrência da hibridização cruzada.

Após a hibridização, o *chip* é lavado para remover materiais sem conexão, em seguida aplica-se um *laser* para ativar as etiquetas fluorescentes. No final, o *chip* é escaneado medindo-se a luminosidade dos *probes* do *microarray*. Esta luminosidade é representada por um valor real positivo. Onde houver hibridização o valor da luminosidade será alto e onde não houver será baixo. A Figura 3.2 representa a imagem de um *microarray*.

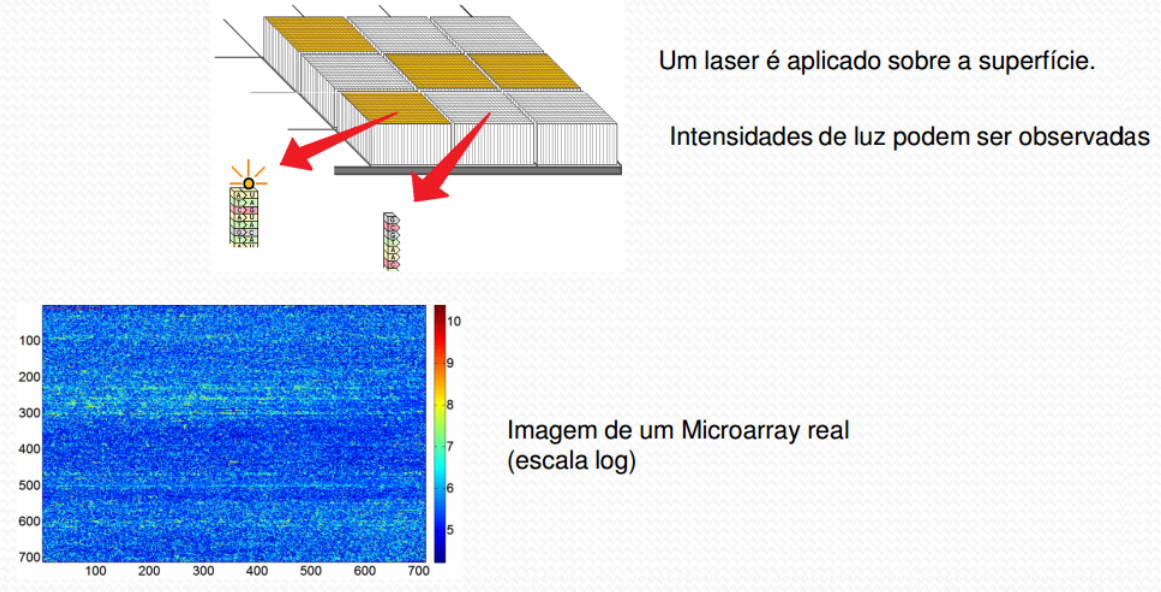

Figura 3.2: *Representação de uma imagem de microarray.*

Esses experimentos produzem um grande conjunto de dados que contêm indícios das atividades de milhares de genes. Os dados que serão utilizados neste trabalho são os valores das luminosidades. Eles representam a intensidade de hibridização e proporcionam uma medida do nível de expressão do gene. Essas medidas podem ser usadas para classificar um gene como presente por meio de um padrão que um *probe set* apresenta em diversos *microarrays*, produzidos para diferentes amostras do mesmo tipo de célula. Podemos classificar o *probe set* (representando um gene) como ausente quando não há um padrão ao longo das amostras; veja Mayrink e Lucas (2015). Alguns problemas como sujeiras, desajuste do *scanner* ou defeitos no *chip* podem causar distorções nas luminosidades, por isso, esses dados passam por um pré-processamento antes de propriamente analisarmos o nível de expressão.

Existem diversos métodos de pré-processamento, entre eles podemos citar o MAS 5.0, RMA e o GCRMA. O MAS 5.0 (*Microarray Suite Version* 5.0) é um *software* desenvolvido pela Affymetrix que incorpora um conjunto de ferramentas para análise de *microarrays*. Entre as ferramentas, temos um método simples de pré-processamento que utiliza ambos os tipos de *probes* PM e MM. Outro método de pré-processamento bastante popular

é conhecido por *Robust Multi-array Average* (RMA). Esse método usa transformações logarítmicas e faz um ajuste linear para corrigir as luminosidades do *chip*, eliminando parte dos erros. Além disso, os dados passam por uma normalização que utiliza projeção de quantis para regular as luminosidades. Finalmente, o conjunto de dados de um *probe set* será sumarizado resultando em uma única luminosidade para cada gene. Para mais detalhes a respeito do MAS 5.0, RMA ou sobre outros métodos de pré-processamento como o CGRMA veja, respectivamente, Affymetrix (2001), Irizarry et al. (2003), Wu et al. (2004).

Este capítulo está organizado da seguinte maneira: A seção 3.1 faz uma breve descrição da modelagem fatorial proposta. Na seção 3.2 apresentamos um estudo simulado para avaliar o comportamento deste modelo. A seção 3.3 fecha o capítulo com as principais conclusões.

3.1 O Modelo Fatorial Latente Simples

Considere X_{ij} a luminosidade pré-processada referente ao *probe* ou gene (*probe set*) i do *microarray* ou amostra j, com $i = 1, 2, ..., m$ e $j = 1, 2, ..., n$. O modelo fatorial tem a seguinte formulação:

$$X = \alpha\lambda + \epsilon, \tag{3.1}$$

sendo α uma matriz de cargas ou *loadings* $(m \times L)$, λ a matriz dos escores dos fatores $(L \times n)$ e ϵ a matriz dos erros $(m \times n)$ com $\epsilon_{ij} \sim N(0, \sigma_i^2)$. Veja que L é o número de fatores adotados no modelo.

As seguintes distribuições *a priori* conjugadas são especificadas:

$$\sigma_i^2 \sim GI(a, b); \tag{3.2}$$

$$\lambda_{\bullet j} \sim N_L(\mathbf{0}, I_L) \text{ com } \lambda_{\bullet j} = (\lambda_{1j}, \lambda_{2j}, ..., \lambda_{Lj})'. \tag{3.3}$$

Considere GI a indicação de uma Gama Inversa e I_L a matriz identidade com dimensão L. A configuração escolhida em (3.3) é utilizada como estratégia padrão para fixar a magnitude de λ e evitar um problema de identificabilidade no produto $\alpha\lambda$. No contexto de análise de expressão de genes, λ é responsável por descrever o padrão da expressão

do *probe* (ou *probe set*) ao longo das amostras. Além disso, no modelo fatorial em (3.1) as cargas atuam como se fossem coeficientes de regressão, podendo ser tanto positivas quanto negativas e determinam a força e a direção de influência dos fatores.

Nesta versão mais simples do modelo fatorial adotamos a seguinte distribuição *a priori*:

$$\alpha'_{i\bullet} \sim N_L(M_\alpha, V_\alpha) \text{ sendo } \alpha'_{i\bullet} = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iL}). \tag{3.4}$$

Através da distribuição *a priori* em (3.4) podemos avaliar a significância das cargas levando em conta o intervalo de credibilidade. Se o zero pertence ao intervalo, α_{il} não seria significativo e, com isso, o fator l não teria influência sobre as observações da linha i na matriz de dados. Uma matriz de cargas com poucos α_{il} significativos indica que o efeito dos fatores no modelo é baixo ou nulo, caracterizando um padrão de expressão fraco ou aleatório em X. Uma matriz com muitos α_{il} significativos, sugere que a influência dos fatores é alta e um padrão de expressão forte pode ser observado na matriz de dados X.

Na Figura 3.3, o painel esquerdo mostra uma matriz de dados que apresenta um padrão de expressão forte; neste caso o modelo indicaria uma matriz α com muitas cargas significativas. Por outro lado, a direita, temos uma matriz de dados com um padrão de expressão fraco ou aleatório, aqui o modelo apresenta uma matriz α com poucas cargas significativas.

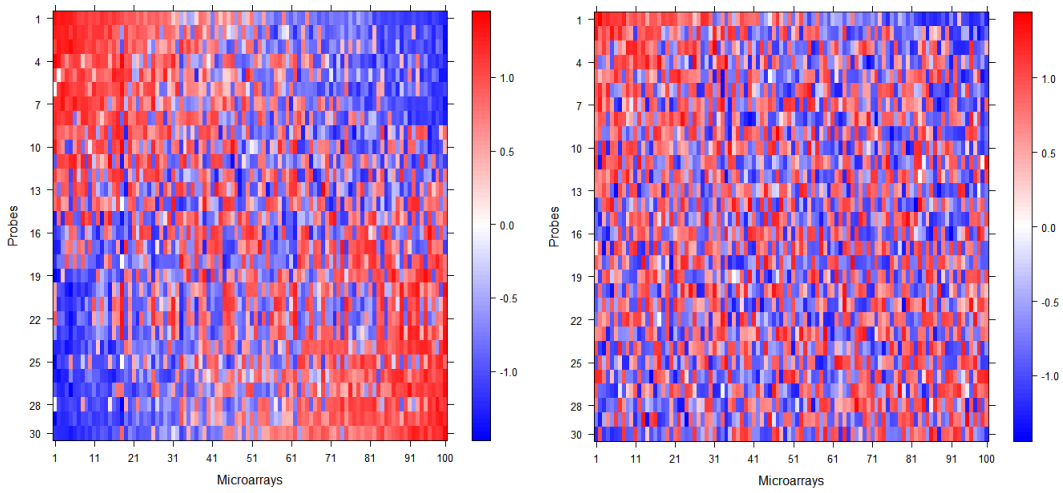

Figura 3.3: *Representação de duas matrizes de dados X, a esquerda observa-se um padrão forte, a direita um padrão fraco ou aleatório.*

Uma das questões a ser levada em conta na estimação dos parâmetros é que α e λ podem trocar de sinais. Este problema de identificabilidade é natural em análise fatorial e no caso Bayesiano fica evidente quando se executa mais de uma vez o algoritmo MCMC, pois em execuções diferentes poderíamos observar que a cadeia converge para o valor correto, porém havendo mudança de sinal entre α e λ. Uma forma de tratar esse problema para realizar inferência é multiplicar os valores da cadeia por -1 quando houver mudança de sinal. Alem da troca de sinal, outro problema de identificação é a troca de colunas dentro de α e linhas dentro de λ. Uma das maneiras de contornar este segundo obstáculo é impor para α a restrição de ser uma matriz triangular superior. Há uma desvantagem neste caso, pois iremos forçar alguns poucos $\alpha_{il} = 0$, sendo que o mesmo não precisa ser necessariamente nulo. Na aplicação que iremos trabalhar, o problema da troca de posições será controlado da seguinte forma: A partir da geração dos dados especificamente da matriz das cargas, que será feito no estudo simulado a seguir com $L = 2$, faremos a primeira coluna ter a maioria das cargas com valores maiores que as da segunda coluna. Desta maneira, no final da execução do MCMC iremos: (i) construir uma matriz com as médias das cadeias de cada carga gerada após o período de *burn-in* (m_{il}), (ii) calcular a diferença absoluta entre as média da primeira coluna (m_{i1}) e os valores reais (α_{il}) gerados para as cargas formando uma matriz com entradas $\text{abs}_{il} = |m_{i1} - \alpha_{il}|$. O que se espera é que a maioria dos abs_{i1} seja menor que abs_{i2}, caso contrario as cadeias geradas para as cargas da primeira coluna serão trocadas de posição com as cadeias da segunda coluna. Consequentemente os escores dos fatores da primeira linha de λ serão trocados de posição com os escores da segunda linha de λ. Em uma aplicação que vamos trabalhar no próximo capítulo, o problema da troca de posições será controlado a partir de especificações *a priori* para alguns elementos de α.

3.2 Estudo Simulado

Para verificar o desempenho do modelo (3.1) em relação a inferência de seus parâmetros, foi simulada uma matriz de dados X com $m = 30$, $n = 100$ e $L = 2$. Este estudo servirá de base para a modelagem mais robusta que será apresentada nos próximos capítulos.

Foi escolhido um modelo com dois fatores, pois no Capítulo 4 faremos um estudo abordando um problema onde cada fator estará relacionado a regiões do genoma afetadas por alteração no DNA. É possível modelar com mais de dois fatores ($L > 2$) porém este não é o foco do presente trabalho. Os seguintes passos foram usados para gerar os dados:

1. Gerar $\alpha'_{i\bullet} \sim N(0, D_\alpha)$, onde $D_\alpha = \text{diag}(0.2, 0.2, ..., 0.2)$;
2. Gerar $\lambda_{\bullet j} \sim N(0, I_L)$;
3. Gerar $\epsilon_{ij} \sim N(0, \sigma_i^2)$, sendo $\sigma_i^2 = 0.2$ $(i = 1, ..., 15)$ e $\sigma_i^2 = 0.1$ $(i = 16, ..., 30)$;
4. Calcular $X = \alpha\lambda + \epsilon$.

A Tabela 3.1 apresenta algumas estatísticas da matriz de dados gerada para este estudo inicial. Nela, pode-se verificar que o valor médio das observações é de 0.0138, além disso, o menor e o maior valor são -3.2607 e 4.3124, respectivamente. O painel (a) na Figura 3.4 apresenta um histograma mostrando a distribuição dos dados em X. No painel (b) podemos observar um padrão na imagem que representa estes dados.

Tabela 3.1: Estatísticas descritivas da matriz de dados.

Estatísticas	Valores
Mínimo	-3.2607
Primeiro Quartil	-0.4131
Mediana	0.0209
Média	0.0138
Terceiro Quartil	0.4390
Máximo	4.3124

(a) (b)

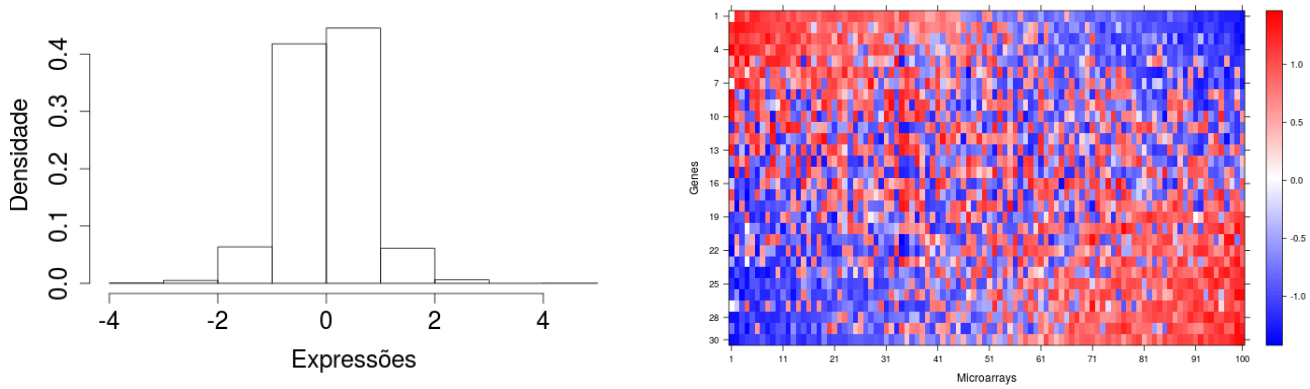

Figura 3.4: *Histograma e imagem da matriz de dados.*

A partir da regra de Bayes obtém-se o núcleo da distribuição *a posteriori* $p(\alpha, \lambda, \sigma^2|X)$ que não é tratável analiticamente, então será aplicado o algoritmo *Gibbs Sampling*. O *Gibbs Sampling* é essencialmente um esquema iterativo de amostragem de uma cadeia de Markov, cujo núcleo de transição é formado pelas distribuições *a posteriori* condicionais completas, que para o modelo em (3.1), apresentam forma fechada. As expressões das distribuições condicionais completas estão apresentadas no Apêndice A. O algoritmo para geração das amostras é como segue:

Passo 1 Escolha os valores iniciais $\alpha_{i\bullet}^{(0)} = (\alpha_{i1}^{(0)}, \alpha_{i2}^{(0)})$, $\lambda_{\bullet j}^{(0)} = (\lambda_{1j}^{(0)}, \lambda_{2j}^{(0)})'$ e $\sigma^{2(0)} = (\sigma_1^{2(0)}, ..., \sigma_m^{2(0)})'$, para $i = 1, ..., m$ e $j = 1, ..., n$. Inicialize o contador de iterações $t = 1$.

Passo 2 Obtenha os novos valores $\alpha_{i\bullet}^{(t)}$, $\lambda_{\bullet j}^{(t)}$ e $\sigma^{2(t)}$, para $i = 1, ..., m$ e $j = 1, ..., n$, a partir das sucessivas gerações abaixo:

Gerar $\sigma_i^{2(t)}$ de $\left[\sigma_i^2 \mid \alpha^{(t-1)}, \lambda^{(t-1)}, \sigma_{-i}^{2(t-1)}, X\right]$.
Gerar $\alpha_{i\bullet}^{(t)}$ de $\left[\alpha_{i\bullet} \mid \alpha_{\{-i\bullet\}}^{(t-1)}, \lambda^{(t-1)}, \sigma^{2(t)}, X\right]$.
Gerar $\lambda_{\bullet j}^{(t)}$ de $\left[\lambda_{\bullet j} \mid \alpha^{(t)}, \lambda_{\{-\bullet j\}}^{(t-1)}, \sigma^{2(t)}, X\right]$.

Passo 3 Faça $t = t + 1$ e retorne ao Passo 2 até obter a amostra desejada após a convergência das cadeias.

O algoritmo utilizado foi implementado na linguagem **R** [R Development Core Team (2016)] utilizando os pacotes Rcpp [Eddelbuettel e Francois (2011), Eddelbuettel (2013)] e RcppArmadillo [Eddelbuettel e Sanderson (2014)] que integra o R e a linguagem C++, além disso, para o cálculo dos intervalos *Highest Posterior Density* (HPD) foi utilizado o pacote coda [Plummer et al. (2006)]. Na simulação, foram assumidas as distribuições *a priori* (3.2), (3.3) e (3.4) com hiperparâmetros $a = 2.1$, $b = 1.1$, $M_\alpha = \mathbf{0}$ e $V_\alpha = \text{diag}(10, ..., 10)$. Veja que a Gama Inversa escolhida *a priori* indica $E(\sigma_i^2) = 1$ e $\text{var}(\sigma_i^2) = 10$. Foi considerado um total de 6000 iterações para execução do MCMC. A convergência de algumas cadeias foi avaliada por meio do critério de Gelman e Rubin (1992).

O diagnóstico de Gelman e Rubin (1992) é baseado na comparação das trajetórias de múltiplas cadeias, com diferentes valores iniciais; elas devem ser muito parecidas após a convergência. A análise consiste em comparar os desvios dentro de cada cadeia e entre as cadeias. Considere m cadeias paralelas e uma função $\psi = t(\theta)$, então tem-se m trajetórias $\{\psi_i^{(1)}, \psi_i^{(2)}, ..., \psi_i^{(n)}\}$ para $i = 1, ..., m$. Com base nisso as variâncias entre as cadeias B e dentro das cadeias W são obtidas como segue:

$$B = \frac{n}{m-1}\sum_{i=1}^{m}(\bar{\psi}_i - \bar{\psi})^2 \quad \text{e} \quad W = \frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}(\psi_i^{(j)} - \bar{\psi}_i)^2$$

sendo $\bar{\psi}_i$ a média das observações da cadeia i e $\bar{\psi}$ a média dessas médias. Sob a hipótese de convergência, todos os mn valores são gerados da distribuição *a posteriori* conjunta e a variância de ψ será estimada de maneira consistente por $\hat{\sigma}_\psi^2 = \left(1 - \frac{1}{n-1}\right) W + \left(\frac{1}{n}\right) B$. Se as cadeias não convergirem até o momento, então $\hat{\sigma}_\psi^2$ vai superestimar σ_ψ^2 e W vai subestimar esta variância. A partir dessa ideia, um indicador de convergência pode ser obtido por meio de um estimador de redução de escala potencial dado por:

$$\hat{R} = \sqrt{\frac{\hat{\sigma}_\psi^2}{W}}.$$

Este estimador é a estatística de Gelman e Rubin (1992) que é sempre maior que 1. Os estimadores $\hat{\sigma}_\psi^2$ e W convergem para σ_ψ^2 a medida que $n \to \infty$, $\hat{R} \to 1$. Se $\hat{R}$ está próximo de 1, então temos evidência de que as M cadeias convergiram para a distribuição de interesse. Os autores sugerem que valores de $\hat{R}$ menores que 1.2 indicam convergência.

Neste primeiro estudo, consideremos três cadeias para o mesmo parâmetro partindo dos valores iniciais 1, 2 e 3 para $\sigma^{2(0)}$; 0, -4 e 4 para $\alpha^{(0)}$ e $\lambda^{(0)}$. A Figura 3.5 apresenta os gráficos de algumas cadeias para σ^2, α e λ partindo de diferentes valores inicias, pode-se observar que as trajetórias das cadeias convergem para a distribuição alvo. A estatística de Gelman e Rubin para estas cadeias foi menor que 1.19 confirmando a convergência.

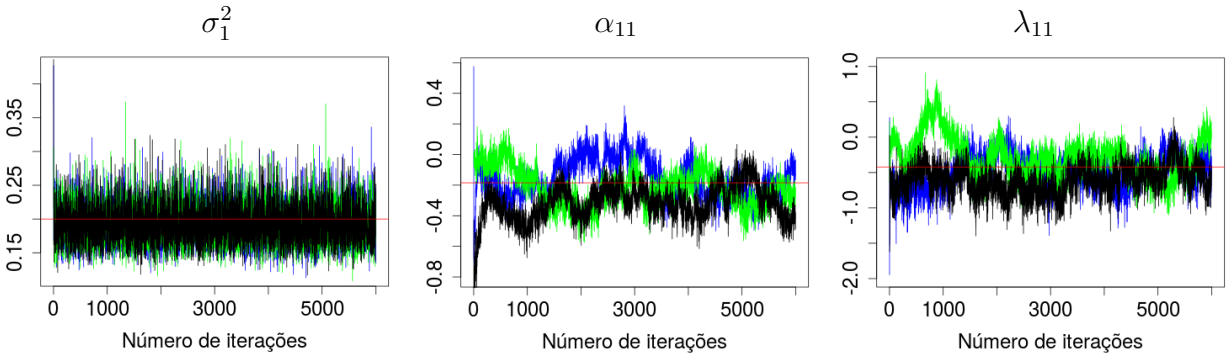

Figura 3.5: *Gráfico de algumas cadeias de α, λ e σ^2. A linha horizontal representa o valor real.*

Outra maneira de avaliar a convergência é usando o critério de Geweke (1992) que sugere o uso de algumas ferramentas baseadas nas cadeia. Após um número suficientemente longo de N iterações, são construídas as médias t_a e t_d, sendo t_a a média das primeiras n_a iterações e t_d a média das últimas n_d iterações. Em seguida é calculada a estatística Z dada pela diferença entre as médias dividida pelo erro padrão assintótico dessa diferença. Geweke (1992) sugere que essas médias só sejam construídas após algumas iterações inicias terem sido descartadas e que sejam usados os 10% inicial da cadeia e os 50% final.

Em nosso estudo, descartamos as 3000 primeiras observações das cadeias que tiveram os valores iniciais $\sigma^{2(0)} = 1$, $\alpha^{(0)} = 0$ e $\lambda^{(0)} = 0$. O pacote "coda"fornece a estatística Z e dois limites para avaliação. Se a maioria das estatísticas Z estiverem dentro dos limites de 95% da Normal, há evidências de convergência da cadeia. A Figura 3.6 apresenta gráficos do teste envolvendo nosso estudo simulado. Os valores da estatística Z (símbolo "×") são investigados para algumas cadeias. Pode-se notar que a maioria dos pontos estão dentro dos limites de 95% sugerindo a convergência.

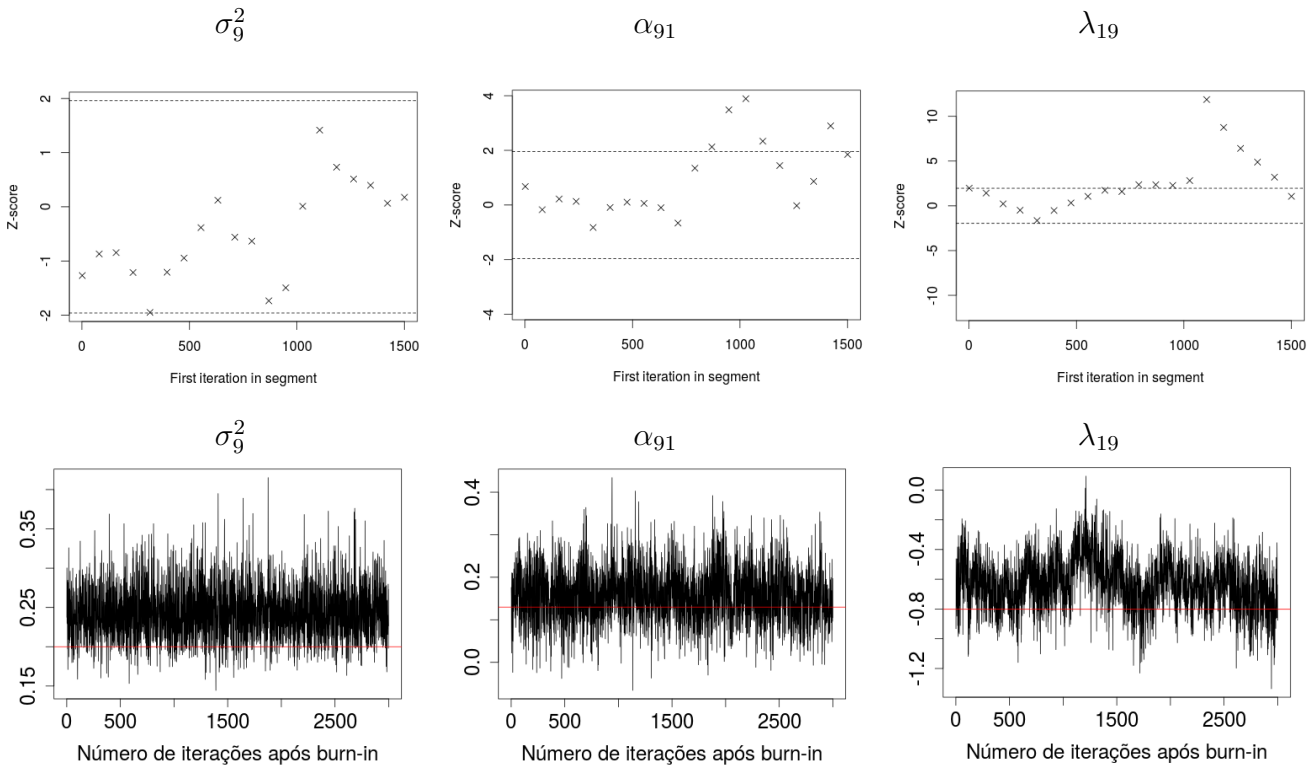

Figura 3.6: *Gráfico com os valores da estatística de Geweke e algumas cadeias de* σ^2, α *e* λ. *O símbolo "×"representa a estatística* Z, *as linhas tracejadas demarcam os valores (-1.96,1.96) correspondendo a região de probabilidade 0.95 na* $N(0,1)$. *As linhas horizontais nos painéis inferiores representam os valores reais.*

A Figura 3.7 apresenta os gráficos de algumas cadeias e correlogramas referentes as observações geradas após o período de *burn-in*. Pode-se observar correlações baixas para σ_5^2 e λ_{15}, porém para as observações referentes a α_{51}, notamos uma autocorrelação razoável para diversos *lag's*. Visualmente podemos observar a convergência da cadeia de α_{51} e poderíamos selecionar algumas observações de forma espaçada para melhorar a inferência.

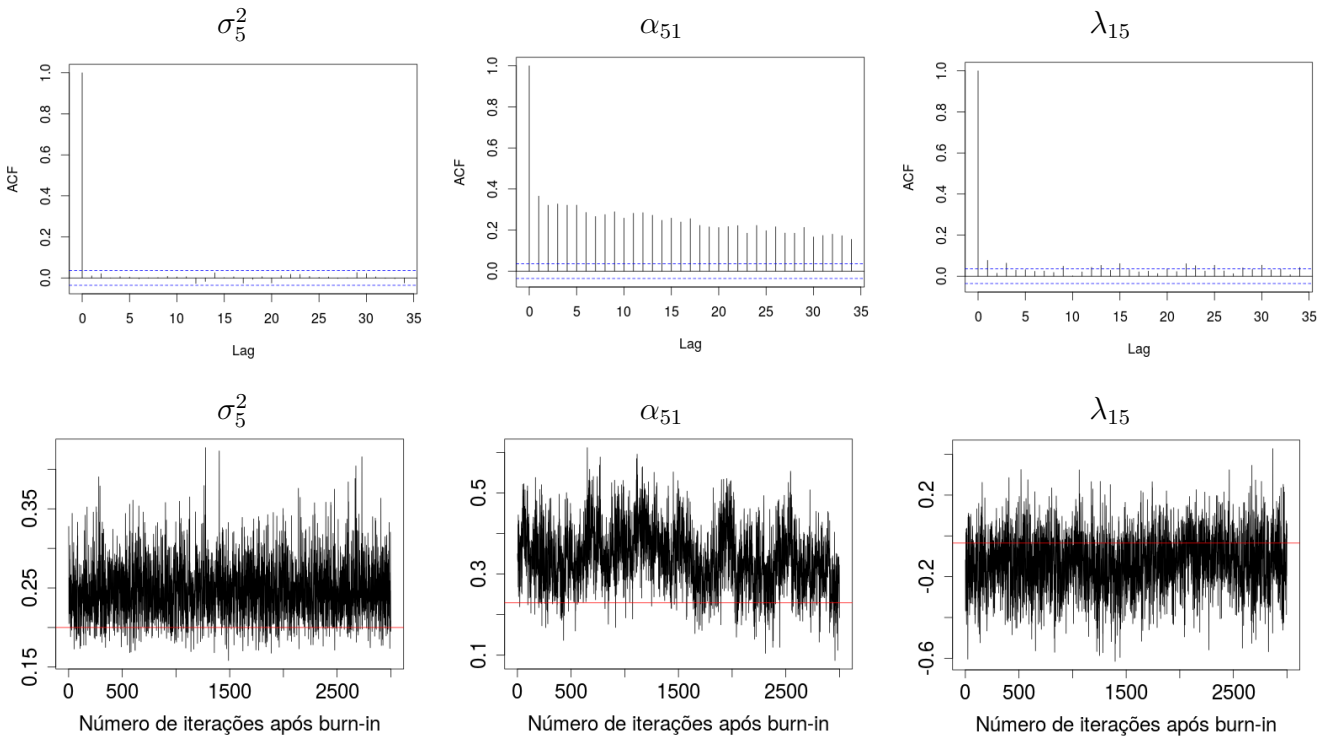

Figura 3.7: *Gráfico de algumas cadeias e correlogramas referentes as observações geradas após execução do MCMC. A linha horizontal nos gráficos inferiores representa o valor real.*

A partir destas primeiras análises devido algumas cadeias apresentarem uma autocorrelação razoável, decidimos utilizar um *lag* de tamanho 15 como estratégia para melhorar a inferência, e isto resultou em uma amostra *a posteriori* de tamanho 200. Na Figura 3.8, pode ser observado os valores reais (asteriscos) comparados com os valores estimados a partir da média *a posteriori* (círculos) além dos intervalos HPD que são representados pelos seguimentos de reta na vertical. Veja que a maioria dos intervalos contém o

verdadeiro valor do parâmetro e apresentam algumas amplitudes curtas, indicando uma pequena variabilidade e boas estimativas. O painel (d) mostra os valores de λ ordenados de forma crescente em relação à média *a posteriori* para uma melhor visualização.

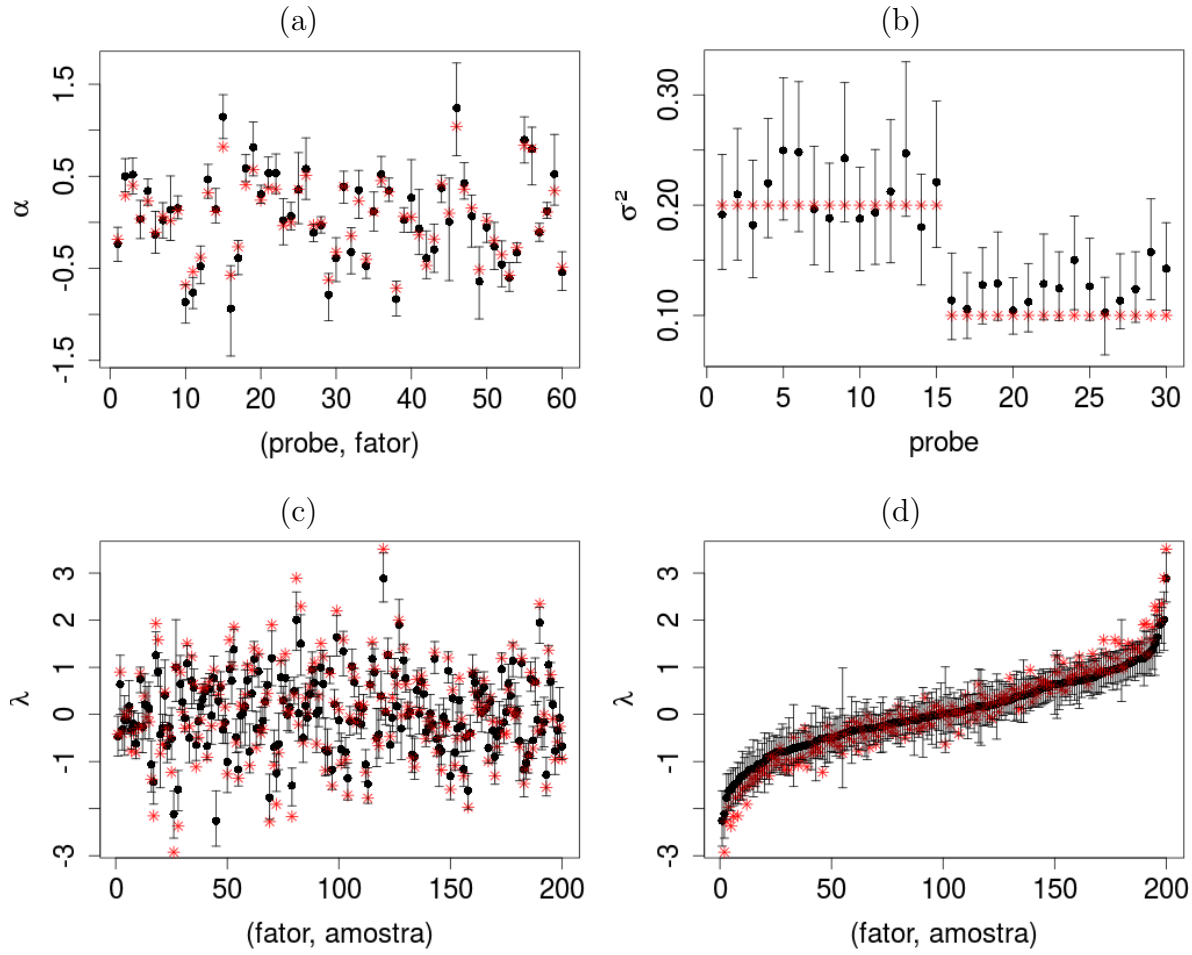

Figura 3.8: *Valores reais (asterisco) comparados com as médias a posteriori (círculo). O intervalo HPD de 95% de credibilidade é representado pelo segmento de reta na vertical. No painel (d) temos os intervalos de λ ordenados pela média a posteriori.*

Ao avaliar a significância das cargas tomando como base a Figura 3.8, observamos 67% dos α_{il} com intervalos HPD sem conter o zero, caracterizando uma matriz de cargas com a maioria significativa. Isso determina um efeito forte dos fatores no modelo indicando a existência de um padrão na matriz X. Lembramos que este padrão controlado pelos escores em λ reflete uma informação subjacente que é compartilhada pelas linhas de X. No contexto de expressão de genes, isto indicaria genes atuando coordenadamente.

3.3 Conclusões do Capítulo

Neste capítulo, fizemos uma breve descrição dos dados de expressão de genes, que motiva este trabalho. Ajustamos um modelo fatorial Bayesiano simples para um conjunto de dados simulados e fizemos uma breve avaliação de alguns critérios de convergência para as cadeias geradas. Os resultados do estudo indicam boas estimativas *a posteriori* e servem para mostrar ao leitor o funcionamento do modelo fatorial Bayesiano em uma aplicação simples. O capítulo a seguir, irá apresentar uma modelagem mais robusta contextualizada para o problema da alteração no número de cópias em seguimentos de DNA definindo regiões do genoma onde a expressão dos genes é mais forte ou mais fraca que o esperado. O modelo fatorial que será apresentado contém uma matriz que representa a interação entre os fatores. A distribuição *a priori* para as cargas será diferente da distribuição utilizada neste capítulo; consideraremos uma mistura para avaliar a significância.

Capítulo 4

Alteração no Número de Cópias e Interação entre Fatores

Diferentes regiões do genoma podem produzir quantidades de mRNA muito acima ou abaixo do esperado, assim medidas feitas em *microarrays* podem ser afetadas por duplicações/eliminação em seguimentos de DNA ocasionando o que chamamos de Alteração no Número de Cópias (Copy Number Alteration - CNA). Alguns métodos são propostos para identificar essas regiões. Lucas et al. (2010) usaram a análise fatorial para determinar regiões do genoma com CNA associadas a resposta de acidose lática e hipóxia em tumores. Eles ajustaram um modelo fatorial latente para a assinatura de genes em um conjunto de 251 amostras referentes ao câncer de mama [Miller et al. (2005)] para gerar 56 fatores latentes. Além disso, o trabalho identificou que a variação no valor da expressão de vários fatores estava altamente associada com a CNA em algumas regiões cromossômicas. A busca por regiões do genoma com CNA, para o câncer de mama, é o foco de outros estudos como: Pollack et al. (2002) e Rueda e Uriarte (2007).

Mayrink e Lucas (2015) utilizam o modelo fatorial latente esparso para avaliar o efeito que a CNA tem no padrão de expressão de genes para diferentes tipos de câncer, incluindo os tumores de mama. Esse artigo utiliza uma distribuição *a priori* esparsa composta de duas componentes Normais centradas em zero e com variância pequena e grande, respectivamente, para a matriz das cargas. Esta especificação *a priori* objetiva avaliar se o fator teria efeito significativo para explicar o padrão na matriz de dados.

Além disso, eles verificaram que um mesmo grupo de genes, localizados em uma região do genoma com CNA para câncer de mama, pode exibir padrões relacionados ao CNA em outros tipos de câncer como pulmão e ovário.

No contexto de dados de CNA vamos considerar, que o genoma tenha duas regiões com alteração; regiões estas que englobam múltiplos genes cada. A ideia é utilizar um modelo com dois fatores sendo que cada fator estará diretamente relacionado com o grupo de genes das diferentes regiões com alteração. Incluiremos também uma influência direta dos fatores ou indireta via interação para grupos de genes fora dessas regões. Assim, o efeito principal do modelo seria o primeiro fator atuando diretamente no grupo de genes G_1 localizados em uma região 1 e o segundo fator atuando diretamente no grupo de genes G_2 localizados em uma região 2. Os demais genes, localizados fora das regiões do genoma com CNA, poderiam ter influência de cada fator ou até mesmo de uma interação entre os fatores. Mayrink e Lucas (2013) utilizam um modelo fatorial e investigam a existência do efeito de interação não linear entre fatores latentes. A incorporação do efeito de interação foi motivada pela existência de uma complexa estrutura de associação entre os genes que não pode ser estudada em modelos simples. Os autores testaram a significância dos termos de interação não linear assumindo uma distribuição *a priori* esparsa sendo uma mistura com uma componente degenerada em zero e um Processo Gaussiano baseado nos escores dos fatores. Esse tipo de especificação *a priori* tem sido usada para definir a estrutura esparsa em modelos que são desenvolvidos, por exemplo, nos trabalhos de West (2003), Lucas et al. (2006) e Carvalho et al. (2008).

Neste capítulo, utilizamos o modelo fatorial latente esparso com interações proposto por Mayrink e Lucas (2013) para modelar a associação entre os genes no contexto da CNA. Mais adiante, no capítulo seguinte, iremos propor uma modificação neste modelo assumindo funções de covariâncias diferentes.

O presente capítulo está organizado como segue: na seção 4.1, apresentamos o modelo com interações que é uma extensão do modelo visto no capítulo anterior, com a utilização de misturas de distribuições *a priori* esparsa para as cargas e interação entre os fatores. Na seção 4.2 mostramos um estudo simulado replicando os resultados expostos em Mayrink e Lucas (2013). Nosso objetivo aqui é confirmar os bons resultados de inferência

para o modelo com função de covariâncias exponencial quadrática. A seção 4.3 finaliza o capítulo com as principais conclusões.

4.1 Modelo Fatorial Latente Esparso com Interações

Seja X_{ij} a luminosidade pré-processada referente ao gene (*probe set*) i na amostra j. Considere o seguinte modelo:

$$X = \alpha\lambda + F + \epsilon, \tag{4.1}$$

sendo α a matriz das cargas $(m \times L)$, λ a matriz dos escores dos fatores $(L \times n)$, F a matriz do efeito de interação entre os fatores $(m \times n)$ e ϵ a matriz dos erros $(m \times n)$ com $\epsilon_{ij} \sim N(0, \sigma_i^2)$.

O modelo em (4.1) tem como objetivo identificar o efeito não linear de interação entre fatores sobre genes no conjunto de dados. Para isso no mínimo dois fatores terão que ser bem definidos para que haja modelagem da interação. Considere que para cada fator $l = 1, 2, ..., L$, há um grupo de genes G_l em X que está diretamente relacionado a esse fator, e um grupo extra G_E em X com genes que podem apresentar uma relação direta com esses fatores e/ou com o efeito de interação entre eles. Além do mais, vamos considerar que esses grupos são dijuntos, isto é, $(G_1 \cup G_2 \cup ... \cup G_L \cup G_E)$ representam as linhas de X. Estabelecemos a relação direta entre grupo-fator envolvendo os elementos de G_l por meio de especificações *a priori* para α e F. Assumimos que os genes em G_l são afetados por um único fator e não serão afetados por interação; a última suposição é induzida por uma distribuição *a priori* que favorece a i-ésima linha de F a ser zero $(F_{i\bullet} = \mathbf{0})$ para cada gene i pertencente a G_l.

Há diferentes versões de modelos fatoriais que podem ser explorados, essas versões se diferenciam por meio de diferentes especificações *a priori* para α_{il} e $F_{i\bullet}$. Neste capítulo, será utilizado uma formulação *a priori* para α_{il} diferente daquela usada no Capítulo 3.

Consideraremos:

$$\begin{aligned} \alpha_{il} &\sim (1 - h_{il})\delta_0(\alpha_{il}) + h_{il}N(0, \omega); \\ h_{il} &\sim \text{Bernoulli}(q_{il}); \\ q_{il} &\sim \text{Beta}(\gamma_1, \gamma_2), \end{aligned} \tag{4.2}$$

sendo $\delta_0(\alpha_{il})$ a componente da mistura com ponto de massa em zero.

Ressaltamos que a variável latente binária h_{il} foi introduzida na notação apenas para auxiliar a descrição do modelo. Em cada iteração do MCMC fazemos $h_{il} = 0$ se $\alpha_{il} = 0$ e $h_{il} = 1$ caso contrário. Desta forma, não há inconsistência entre estes elementos do algoritmo. A distribuição *a priori* em (4.2) é dita esparsa pois avalia, através da probabilidade q_{il}, se α_{il} é um valor nulo proveniente de $\delta_0(\alpha_{il})$ ou não nulo proveniente da $N(0, \omega)$. Expressamos nossa incerteza sobre q_{il} por meio da distribuição Beta e sua estimativa *a posteriori* permite avaliar a significância de α_{il}, e determinar padrões em X. Valores altos de q_{il} favorecerão α_{il} diferente de zero. Por outro lado, valores baixos de q_{il} favorecerão $\alpha_{il} = 0$. Utilizaremos adiante especificações *a priori* para q_{il} visando permitir a identificação do modelo ao evitar a troca de colunas dentro de α e, consequentemente, linhas dentro de λ. Novamente, especificamos *a priori* $\lambda_{\bullet j} \sim N(\mathbf{0}, I_L)$ que é padrão para fixar a magnitude de λ no produto com α.

Assumimos também uma mistura de distribuições *a priori* para o efeito de interação $F_{i\bullet}$, sendo que uma das componentes é a distribuição com ponto de massa em zero e a outra um Processo Gaussiano com função de covariâncias dependendo dos escores dos fatores. Assim a especificação *a priori* do efeito de interação será:

$$F'_{i\bullet} \mid \lambda \sim (1 - z_i)\delta_0(F_{i\bullet}) + z_i N_n[0, K(\lambda)], \tag{4.3}$$

sendo z_i uma variável indicadora com z_i tendo distribuição Bernoulli(ρ_i) e ρ_i com distribuição Beta(β_1, β_2). Valores baixos de ρ_i gerados da distribuição Beta favorecem $z_i = 0$ levando à $F_{i\bullet} = \mathbf{0}$, indicando que o gene i (ou i-ésima linha em X) não é afetado pela interação dos fatores. Entretanto, valores altos de ρ_i favorecem $z_i = 1$ ocasionando $F_{i\bullet} \neq \mathbf{0}$, indicando que o gene i é afetado pela interação dos fatores. No caso em que $F_{i\bullet} \neq \mathbf{0}$ a interação $F_{i\bullet}$ será gerada do Processo Gaussiano com vetor de médias zero e matriz de covariâncias obtida pela função $K(\lambda)$.

Diferentes funções de covariâncias podem ser utilizadas, uma bastante popular na literatura é a função de covariâncias Gaussiana (ou exponencial quadrática) que usaremos neste capítulo para reproduzir os resultados de Mayrink e Lucas (2013). Ela é expressa por:

$$K(\lambda) = \upsilon^2 \exp\left\{-\frac{1}{2l_s^2}\left\|\lambda_{\bullet j_1} - \lambda_{\bullet j_2}\right\|^2\right\}, \tag{4.4}$$

sendo υ^2 um parâmetro global que controla a variabilidade, $(j_1, j_2) \in 1, 2, ..., n$ e l_s o parâmetro de comprimento-escala que controla o quão próximos os escores dos fatores $\lambda_{\bullet j_1}$ e $\lambda_{\bullet j_2}$ devem ser para que sejam considerados associados (considere $\lambda_{\bullet j}$ a j-ésima coluna de λ). Em um caso particular, assumindo $\upsilon^2 = 1$ temos a função de correlação. Veja que a função exponencial quadrática depende da norma euclidiana $\|\lambda_{\bullet j_1} - \lambda_{\bullet j_2}\|$, isto é, quanto mais próximos são os escores $\lambda_{\bullet j_1}$ e $\lambda_{\bullet j_2}$ das amostras j_1 e j_2 no espaço $\Re^L$, maior será a similaridade deles levando a $K(\lambda) \approx 1$. Por outro lado, quanto maior a distâncias entre estes vetores, menor será a similaridade entre os mesmos e assim temos $K(\lambda) \approx 0$.

4.2 Estudo Simulado

Para este estudo vamos considerar uma matriz de dados com tamanho $m = 20$ e $n = 100$, além disso, utilizaremos um modelo com $L = 2$ fatores e definiremos $F_{ij} = \lambda_{1j}\lambda_{2j}$ como o verdadeiro efeito de interação. Nosso objetivo com esta simulação é verificar o desempenho em termos de inferência do modelo fatorial com interações. Admita que cada fator tem uma relação direta com cada grupo de genes G_l, contendo 5 elementos, e que esses grupos não serão influenciados pelo efeito de interação. Com isso, o primeiro fator não terá influência em G_2, assim como o segundo fator não influenciará G_1. O efeito de interação e/ou o efeito principal dos fatores podem estar associados com um grupo de 10 genes denominado G_E. Os grupos de genes G_1, G_2 e G_E são disjuntos e formam as linhas da matriz de dados X que será simulada considerando os seguintes passos:

1. Considere $\alpha_{il} = 0$, para todo $i \in G_1$ sendo $l = 2$, e para todo $i \in G_2$ com $l = 1$.

Gere $\alpha_{il} \sim N(0, 1)$ para $i \in G_1$ com $l = 1$, e $i \in G_2$ com $l = 2$.

Gere $u \sim U(0, 1)$ e obtenha $\alpha_{il} \sim N(0, 0.5)$ se $u < 0.8$ para $i \in G_E$ e todo l.

Fixamos em 0.8 supondo em média 80% de cargas significativas.

2. Gere $\lambda_{lj} \sim N(0,1)$, para $j = 1, 2, ..., 100$ e $l = 1, 2$.

3. Gere a matriz de interações como segue:

$F_{i\bullet} = \mathbf{0}$ se $i \in (G_1 \cup G_2)$. Gere $u \sim U(0,1)$ e faça $F_{ij} = \lambda_{1j}\lambda_{2j}$ se $u < 0.4$ para todo $i \in G_E$.

Fixamos em 0.4 supondo em média 40% de interações significativas

4. Gere $\epsilon_{ij} \sim N(0, \sigma_i^2)$, sendo $\sigma_i^2 = 0.2$ $(i = 1, ..., 10)$ e $\sigma_i^2 = 0.1$ $(i = 11, ..., 20)$.

5. Calcule $X = \alpha\lambda + F + \epsilon$.

Com a regra de Bayes obtemos o núcleo da distribuição *a posteriori* $p(\alpha, \lambda, F, \sigma^2 \mid X)$ que não é tratável analiticamente, por isso utilizamos o algoritmo *Gibbs Sampling* para amostrar dessa distribuição; em particular será aplicado o Metropolis-Hastings com passeio aleatório como um passo dentro do *Gibbs Sampling* para gerar $\lambda_{\bullet j}$. Esse parâmetro alvo aparece em (4.3) e sua distribuição condicional completa não apresenta forma fechada. As distribuições condicionais completas são apresentadas no Apêndice B. O algoritmo utilizado tem os seguintes passos:

Passo 1 Escolher os valores iniciais para $\alpha^{(0)}$, $\lambda^{(0)}$, $\sigma^{2(0)}$, $F^{(0)}$, $z^{(0)} = (z_1^{(0)}, ..., z_m^{(0)})'$, $h^{(0)} = (h_{1\bullet}^{(0)}, ..., h_{m\bullet}^{(0)})'$ com $h_{i\bullet}^{(0)} = (h_{i1}^{(0)}, h_{i2}^{(0)})$. Inicialize o contador de iterações $t = 1$.

Passo 2 Obtenha os novos valores $\alpha^{(t)}$, $\lambda^{(t)}$, $\sigma^{2(t)}$, $F^{(t)}$, $z_i^{(t)}$, $h_{il}^{(t)}$ $\rho_i^{(t)}$, para $i = 1, ..., m$ e $l = 1, 2$, a partir das sucessivas gerações abaixo:

2.1 Gere $\sigma_i^{2(t)}$ de $\left[\sigma_i^2 \mid \alpha^{(t-1)}, \lambda^{(t-1)}, F^{(t-1)}, \sigma_{-i}^{2(t-1)}, X\right]$.

2.2 Gere $\rho_i^{(t)}$ de $\left[\rho_i \mid z_i^{(t-1)}\right]$.

2.3 Calcule ρ_i^* condicional a $\rho_i^{(t)}$, $\lambda^{(t-1)}$, $F^{(t-1)}$, $\alpha^{(t-1)}$, $\sigma^{2(t)}$, X.

2.4 Gere $u \sim U(0,1)$. Se $u \leq \rho^*$, faça $z_i^{(t)} = 1$ e obtenha $F_{i\bullet}^{(t)}$ de $N_n(M_F, V_F)$. Caso contrário, faça $z_i^{(t)} = 0$ e $F_{i\bullet}^{(t)} = \mathbf{0}$.

2.5 Gere $q_{il}^{(t)}$ de $\left[q_{il} \mid h_{il}^{(t-1)}\right]$.

2.6 Calcule q_{il}^* condicional a $q_{il}^{(t)}$, $\sigma^{2(t)}$, $F^{(t)}$, $\lambda^{(t-1)}$, $\alpha^{(t-1)}$, X.

2.7 Gere $u \sim U(0,1)$. Se $u \leq q_{il}^*$, faça $h_{il}^{(t)} = 1$ e obtenha $\alpha_{il}^{(t)}$ de $N(M_\alpha, V_\alpha)$. Caso contrário, faça $h_{il}^{(t)} = 0$ e $\alpha_{il}^{(t)} = 0$.

Passo 3 Gere $\lambda^*_{\bullet j}$ de $N_L\left(\lambda^{(t-1)}_{\bullet j}, \nu I_L\right)$, para $j = 1, ..., n$.

3.1 Calcule a razão r e faça $\eta = \min\{1, r\}$ sendo,

$$r = \frac{N_L(\lambda^*_{\bullet j} \mid M_\lambda, V_\lambda)\,|K(\lambda^*)|^{-\frac{1}{2}\sum_{i=1}^m z_i^{(t)}} \exp\left\{-\frac{1}{2}\sum_{i=1}^m F^{(t)}_{i\bullet} K(\lambda^*)^{-1} F'^{(t)}_{i\bullet}\right\}}{N_L(\lambda^{(t-1)}_{\bullet j} \mid M_\lambda, V_\lambda)\,|K(\lambda^{(t-1)})|^{-\frac{1}{2}\sum_{i=1}^m z_i^{(t)}} \exp\left\{-\frac{1}{2}\sum_{i=1}^m F^{(t)}_{i\bullet} K(\lambda^{(t-1)})^{-1} F'^{(t)}_{i\bullet}\right\}}.$$

Onde $N_L(\lambda_{\bullet j} \mid M_\lambda, V_\lambda)$ é a densidade da normal multivariada no vetor $\lambda_{\bullet j}$.

3.2 Gere $u \sim U(0, 1)$. Se $u < \eta$, faça $\lambda^{(t)}_{\bullet j} = \lambda^*_{\bullet j}$, caso contrário $\lambda^{(t)}_{\bullet j} = \lambda^{(t-1)}_{\bullet j}$.

Passo 4 Faça $t = t + 1$ e retorne ao Passo 2 até obter a amostra desejada após a convergência das cadeias.

As especificações *a priori* em (4.2) e (4.3) podem ser usadas para tratar de alguns problemas de identificabilidade do modelo, como por exemplo, se considerarmos a i-ésima linha de $\alpha\lambda + F$ (no caso $\alpha_{i\bullet}\lambda + F_{i\bullet}$) poderia-se ter um problema como $\alpha_{i\bullet}\lambda + F_{i\bullet} = C\alpha_{i\bullet}\lambda + F^*_{i\bullet}$, onde $F^*_{i\bullet} = (1-C)\alpha_{i\bullet}\lambda$, sendo C um número real. Além disso, podem ocorrer trocas de posições nas colunas de α e nas linhas de λ. Estes problemas são resolvidos a partir das especificações *a priori* exibidas na Tabela 4.1.

Tabela 4.1: Distribuições *a priori* utilizadas para as probabilidades q_{il} e ρ_i.

Parâmetros	Distribuição *a priori*	Valores Iniciais	Índices
q_{il}	Beta(2, 1)	0.9999	$i \in G_1, l = 1$ ou $i \in G_2, l = 2$
	Beta(1, 2)	0.0001	$i \in G_1, l = 2$ ou $i \in G_2, l = 1$
	Beta(1, 1)	0.5	$i \in G_E, l = \{1, 2\}$
ρ_i	Beta(1, 2)	0.0001	$i \in (G_1 \cup G_2)$
	Beta(1, 1)	0.5	$i \in G_E$

A Tabela 4.1 apresenta as configurações *a priori* para as probabilidades q_{il} e ρ_i, que serão atualizadas em cada iteração do algoritmo MCMC. Com a escolha destas distribuições Betas estamos afirmando que cada fator l influenciará cada grupo G_l, pois a distribuição Beta(2,1) favorecerá $\alpha_{i1} \neq 0$ em G_1, enquanto que uma distribuição Beta(1,2)

favorecerá $\alpha_{i2} = 0$ em G_1. Desta forma, apenas o primeiro fator influenciará G_1. Já a distribuição Beta(1,2) favorecerá $\alpha_{i1} = 0$ em G_2 e $\alpha_{i2} \neq 0$ em G_2. Neste caso, apenas o segundo fator influenciará G_2. Além disso, ao atribuir a distribuição Beta(1,2) para ρ_i estamos considerando que o efeito de interação $F_{i\bullet}$ não influenciará o grupo $(G_1 \cup G_2)$, mas ao assumir a Beta(1,1) tanto para q_{il} quanto para ρ_i estamos deixando o modelo livre para definir com base nos dados quais α_{il} e $F_{i\bullet}$, em G_E, são significativos. O grupo G_E pode ser influenciado pelo efeito principal e/ou pelo efeito de interação.

Outras distribuições Beta poderiam ser especificadas para q_{il} e ρ_i referentes a α_{il} e $F_{i\bullet}$ no grupo G_E; como por exemplo, a Beta(γ_1, γ_2), com γ_1 e γ_2 entre 0 e 1. Gonçalves (2006) ao realizar estudos simulados para modelos com detecção de DIF (Differential Item Functioning) na Teoria da Resposta ao Item, sugere esta distribuição ao fazer simulações com o objetivo de analisar e comparar dois algoritmos para gerar amostras da distribuição *a posteriori* conjunta. A utilização dessa distribuição *a priori* Beta seria uma opção interessante, pois ela tem o formato de "banheira", concentrando sua massa nos extremos do intervalo (0,1).

Considere novamente a distribuição *a priori* em (3.3) para $\lambda_{\bullet j}$, faça $\omega = 10$ para a componente da mistura em (4.2) e assuma $GI(2.1, 1.1)$ para σ_i^2 (média 1 e variância 10). Para a função de covariâncias em (4.4) consideramos $\upsilon^2 = 1$ e $l_s = 0.3$. Em termos dos valores iniciais da cadeia foram indicados $F_{ij}^{(0)} = 0$, $\alpha_{il}^{(0)} = 0$, $\sigma^{2(0)} = 1$ e $\lambda_{lj}^{(0)}$ gerado da $N(0, 0.3)$. Para as indicadoras $h_{il}^{(0)}$ e $z_i^{(0)}$ foram considerados os valores iniciais de uma distribuição Bernoulli conforme estabelecido na Tabela 4.1. Além disso, se o valor das estimativas *a posteriori*, observadas nas cadeias, converge para o valor real do parâmetro mas com o sinal trocado, basta multiplicar a cadeia por -1 para corrigir. Na geração dos dados, a matriz de interação F foi simulada contendo as linhas 12, 13, 14, 15 e 17 diferentes de zero, sendo estas consideradas como efeitos de interação real. Para a construção do algoritmo MCMC consideramos um total de 4000 iterações. A partir da Figura 4.1 pode-se observar visualmente a convergência de algumas cadeias de α, λ, σ^2 e F. Realizamos o teste de Geweke (1992) que confirmou a convergência destas cadeias; veja o Apêndice C.

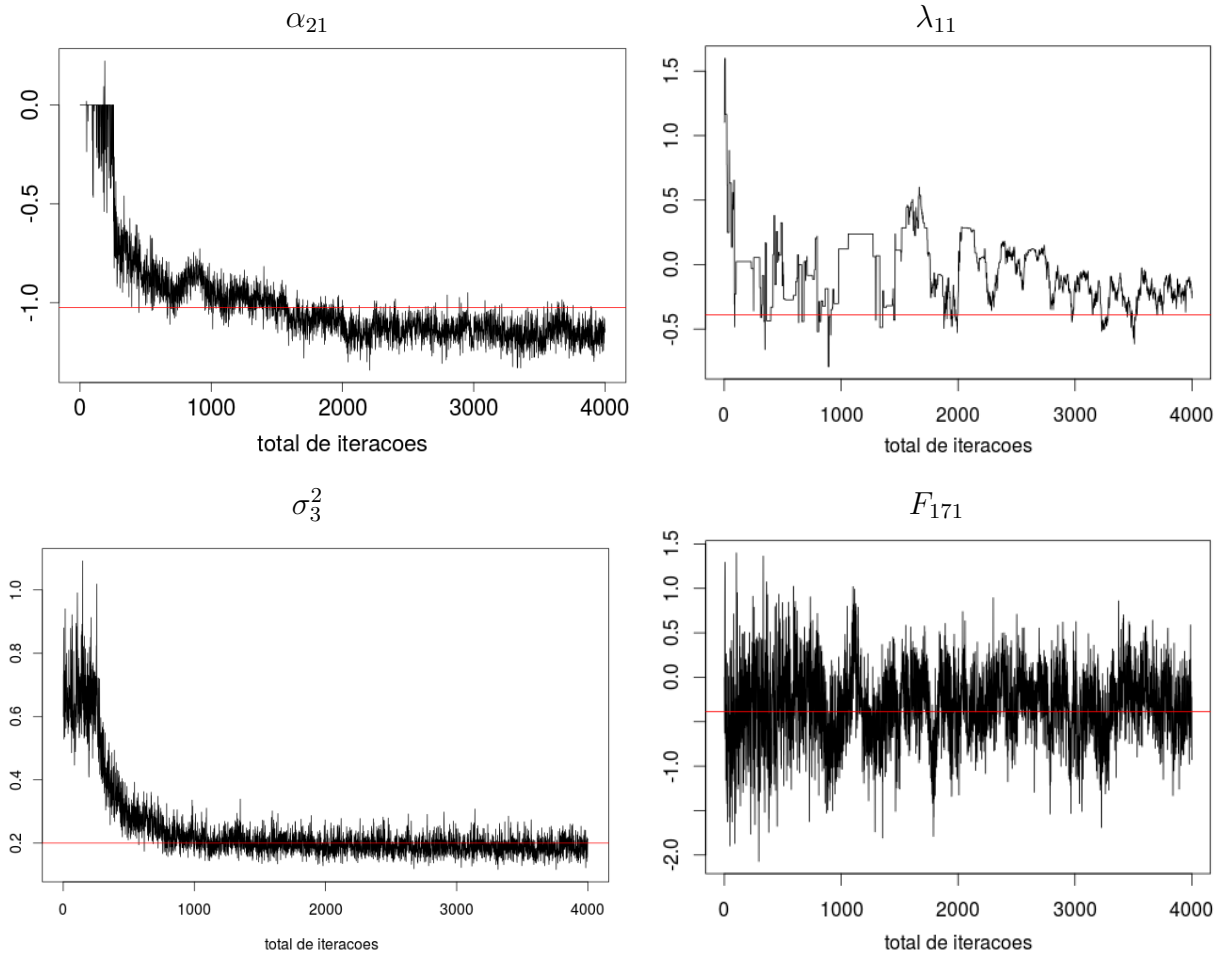

Figura 4.1: *Cadeias de algumas observações geradas de α, λ, σ^2 e F. A linha horizontal representa o valor real. Podemos visualizar a convergência dessas cadeias.*

As primeiras 3000 observações foram definidas como período de *burn-in* e removidas da análise. Não foi utilizado *lag* (observações espaçadas) e a taxa de aceitação média do Metropolis-Hastings, feita para os vetores $\lambda_{\bullet j}$ foi de 41%. A Figura 4.2 apresenta as estimativas dos parâmetros α, σ^2, λ e F. Nela, pode-se observar que a maioria dos intervalos com 95% de credibilidade contém o valor real do parâmetro dando indícios de bom desempenho. No painel (a) observamos que as cargas α_{i1} em G_1 e α_{i2} em G_2 são diferentes de zero, enquanto que α_{i2} em G_1 e α_{i1} em G_2 apresentam valores próximos a zero. Isso é correspondente ao esperado conforme a geração dos dados.

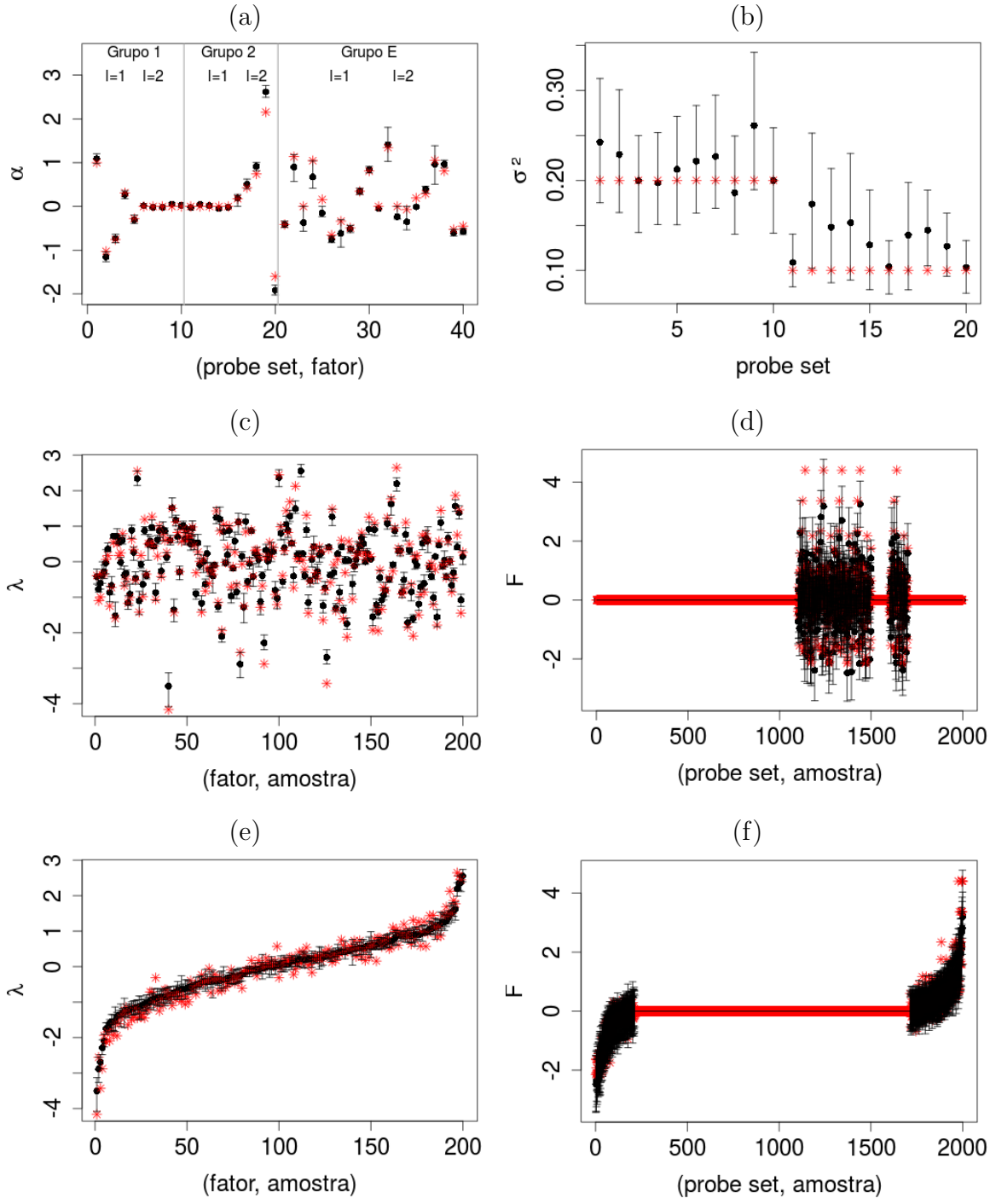

Figura 4.2: *Valores reais (asterisco) e médias a posteriori (círculo). O intervalo HPD de 95% de credibilidade é representado pelo segmento de reta na vertical. Os dois últimos painéis exibem os intervalos ordenados de acordo com a média a posteriori.*

A Figura 4.3 mostra gráficos de superfície representando os efeitos de interação assumindo $l_s = 0.3$. A superfície real tem o formato de "sela de cavalo"e é exibida no primeiro painel. Os demais painéis mostram os $F_{i\bullet}$ estimados. Pode-se observar que o formato das superfícies estimadas reflete bastante a configuração da real. Isto é uma indicação de que o modelo consegue capturar bem a interação não linear alvo. Algumas irregularidades são notadas na superfície estimada, porém isto é natural visto que estamos estimando e há incerteza *a posteriori*.

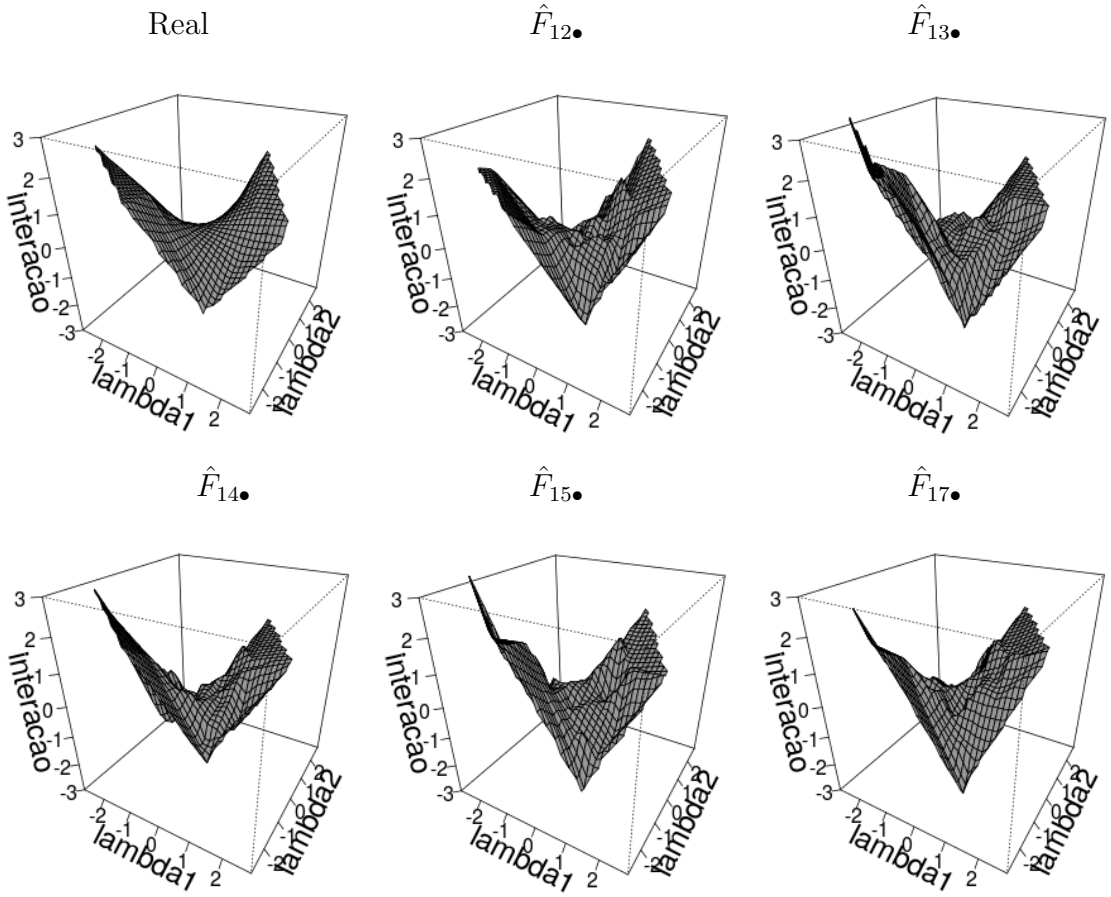

Figura 4.3: *Gráficos de superfície representando o efeito de interação real e estimado. O parâmetro de comprimento-escala é igual a 0.3.*

A partir da diferença entre o valor real e o estimado podemos verificar a qualidade do ajuste. Na Figura 4.4 temos os gráficos das diferenças entre o efeito real e o estimado. Uma situação ideal seria a visualização de um plano centrado na origem, entretanto, há algumas irregularidades cuja magnitude nos dá ideia sobre o erro de estimação. Perceba que não há registros de picos e vales extremos. As superfícies são suaves e sugerem boa estimação *a posteriori*.

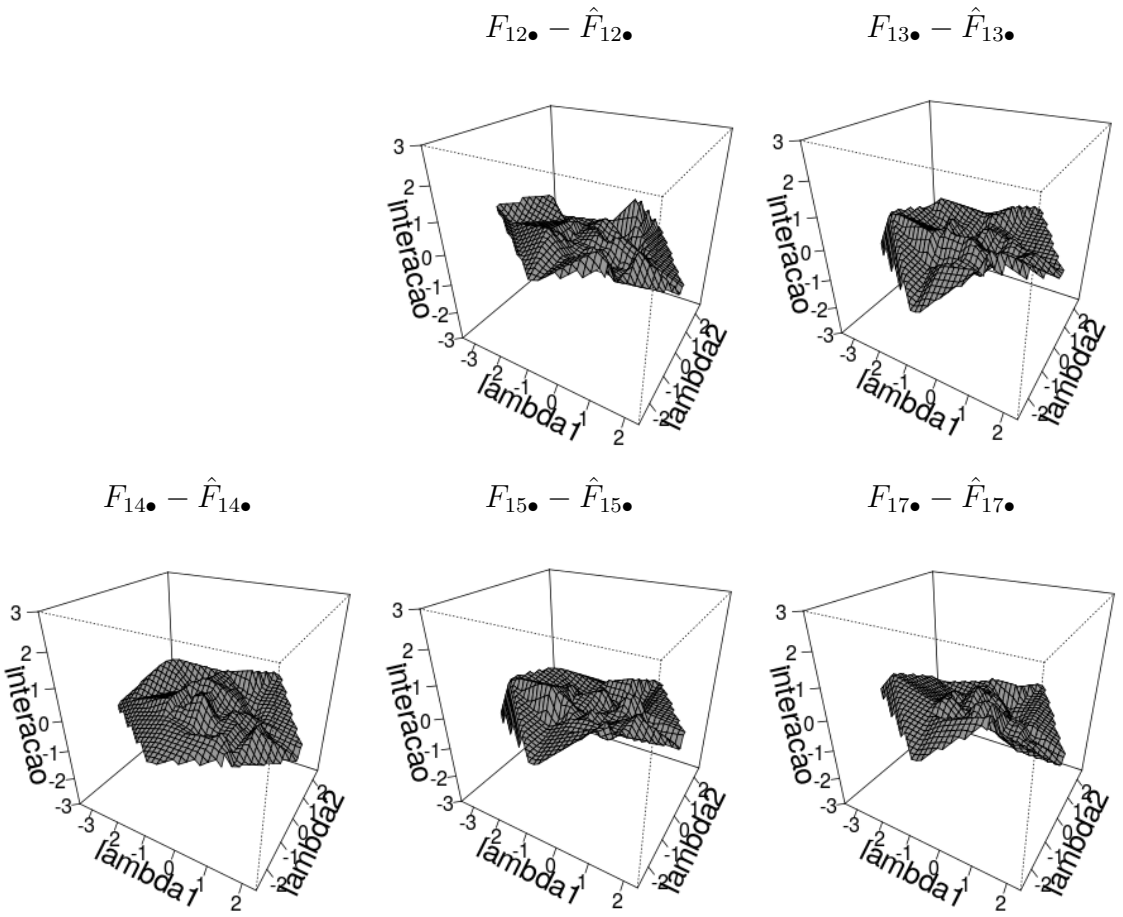

Figura 4.4: *Gráfico de superfície representando a diferença entre o efeito de interação real e o estimado.*

As Figuras 4.5 e 4.6 apresentam gráficos de superfície do efeito de interação estimado considerando l_s igual a 0.1 e 0.2, respectivamente. Comparando os resultados das Figuras 4.3, 4.5 e 4.6 podemos observar algumas diferenças entre as formas de superfície. Veja que quando l_s aumenta, as irregularidades da superfícies $\hat{F}_{12\bullet}$ parecem diminuir deixando-as

um pouco mais suave.

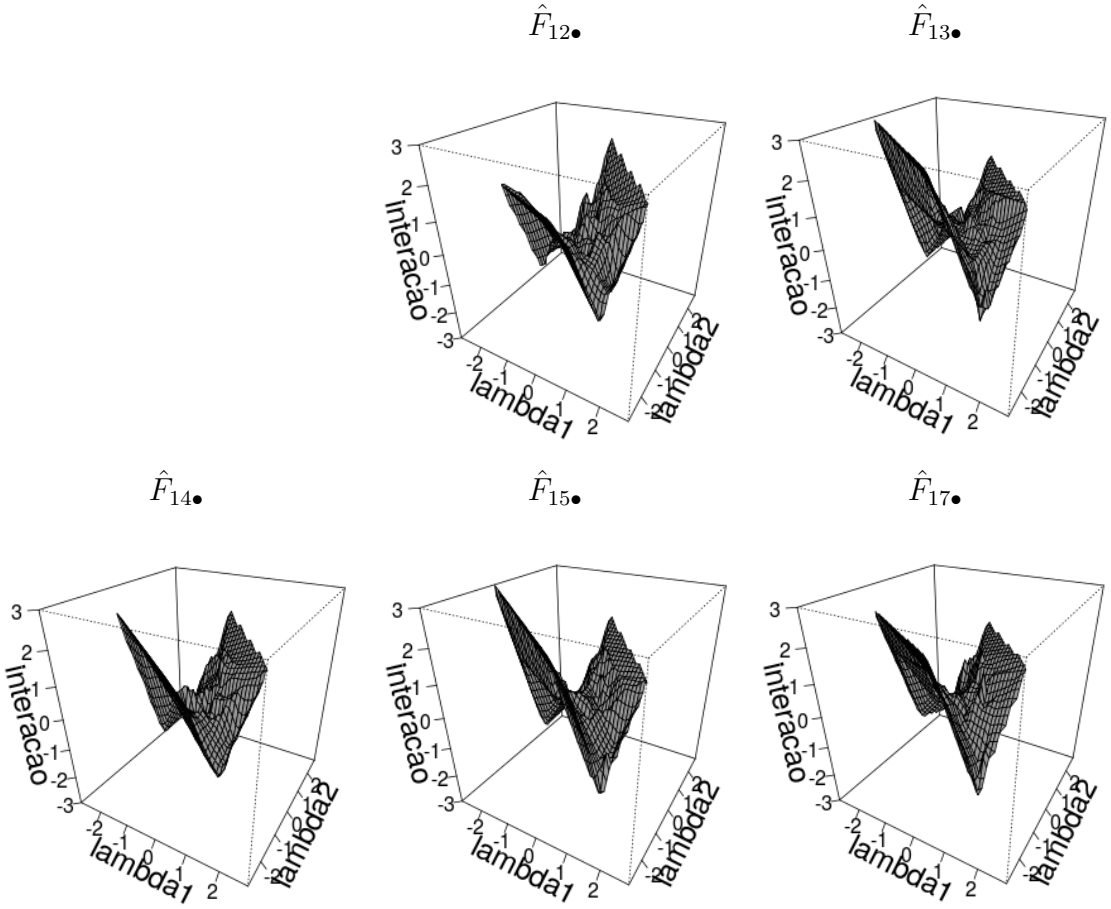

Figura 4.5: *Gráfico de superfície do efeito de interação estimado considerando* $l_s = 0.1$.

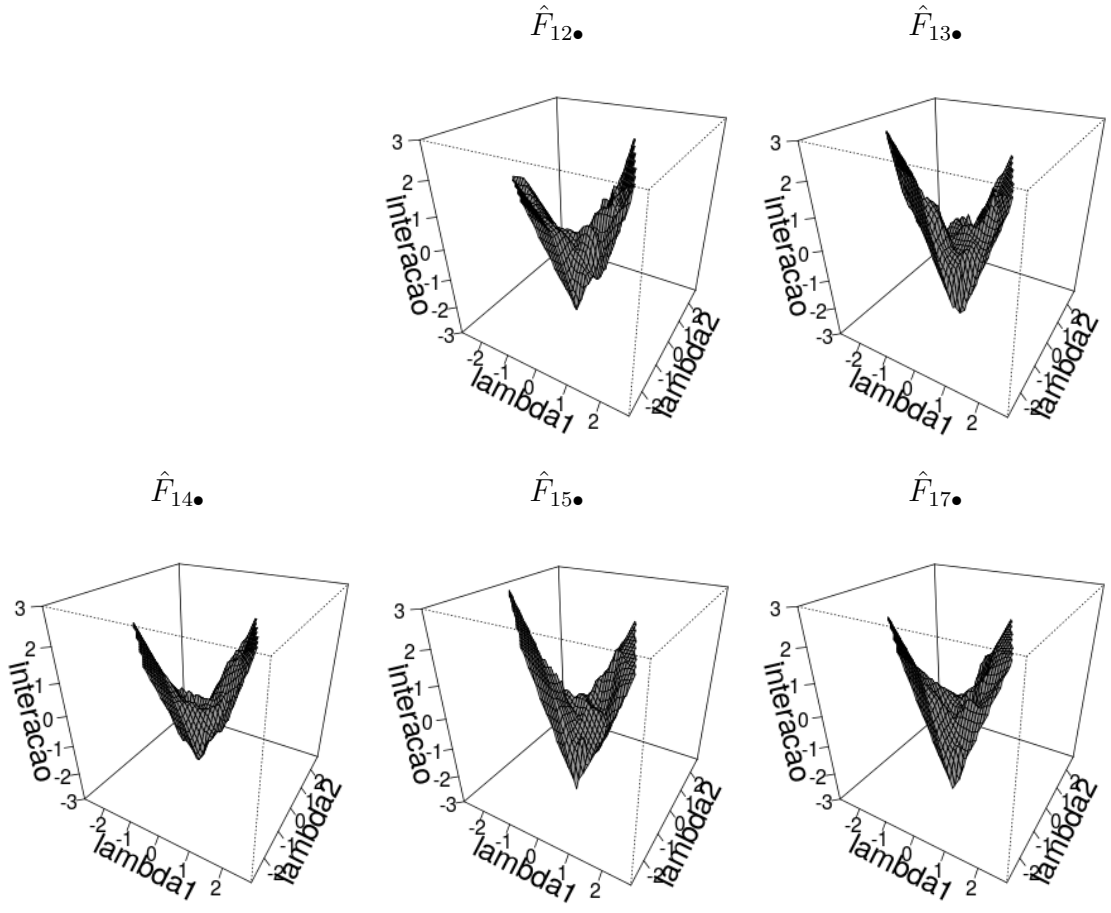

Figura 4.6: *Gráfico de superfície do efeito de interação estimado considerando* $l_s = 0.2$.

Para verificar a significância das cargas e analisar a influência dos fatores em cada grupo G_l foi utilizado a distribuição *a posteriori* condicional completa $q^*_{il} = p(h_{il} = 1 \mid \alpha, \lambda, \sigma^2, F, q_{il}, X)$. Essa medida *a posteriori* é responsável por descrever a probabilidade de cada α_{il} ser diferente de zero, ou seja, significativo. Além disso, a influência ou não do efeito de interação entre os fatores nos genes foi verificado a partir da distribuição *a posteriori* condicional completa $\rho^*_i = p(z_i = 1 \mid \alpha, \lambda, \sigma^2, F, \rho_i, X)$ que avalia $F_{i\bullet} \neq 0$. A Figura 4.7 apresenta as estimativas médias destas probabilidades *a posteriori*. Nela, podemos observar que as médias *a posteriori* de q^*_{i1} referentes ao grupo G_1 são maiores que 0.5 sugerindo $\alpha_{i1} \neq 0$ enquanto que valores de q^*_{i2} referentes ao grupo G_1 estão abaixo de 0.5 indicando $\alpha_{i2} = 0$, conforme o esperado. Além do mais, ao analisar as médias *a posteriori* de q^*_{i1} referentes à G_2, tem-se valores abaixo de 0.5, indicando $\alpha_{i1} = 0$. Já

as médias *a posteriori* de q^*_{i2} referentes a G_2 estão acima de 0.5 indicando $\alpha_{i2} \neq 0$. No grupo G_E pode ser visto que algumas estimativas apresentam intervalos de credibilidade bastantes amplos de maneira que englobam o valor 0.5 e isso aumenta a incerteza *a posteriori* a respeito da significância de algumas cargas deste grupo. Na geração dos dados, as cargas $\alpha_{11,2}$, $\alpha_{13,1}$ e $\alpha_{13,2}$ são iguais a zero, desta forma apenas o primeiro fator influencia o gene 11, enquanto que o gene 13 não sofre influência de nenhum fator. Já as cargas $\alpha_{14,2}$, $\alpha_{15,1}$ e $\alpha_{15,2}$ apresentaram valores próximos de zero. Voltando à Figura 4.7, pode-se observar que a média *a posteriori* da probabilidade $q^*_{13,1}$, referente a significância da carga $\alpha_{13,1}$, além das médias *a posteriori* das probabilidades referentes as cargas $\alpha_{13,2}$, $\alpha_{14,2}$, $\alpha_{15,1}$ e $\alpha_{15,2}$, estão abaixo de 0.5 sugerindo que essas cargas não são significativas, apesar destas estimativas de probabilidades apresentarem intervalos de credibilidade amplo. Com isso pode-se dizer que, em geral, o modelo dá indícios de bom comportamento em relação a significância das cargas.

Quando analisamos a média de ρ^*_i, verificamos que as interações $F_{i\bullet}$ são diferentes de zero no grupo G_E somente para $i = 12, 13, 14, 15$ e 17. Essa configuração estimada corresponde ao real. Ainda na Figura 4.7, note que as médias de ρ^*_i, referentes a $F_{12\bullet}$,$F_{13\bullet}$,$F_{14\bullet}$,$F_{15\bullet}$ e $F_{17\bullet}$, apresentam valores acima de 0.5 indicando que o efeito destas interações é significativo. Observe também que nos grupos G_1 e G_1, as médias de ρ^*_i estão abaixo de 0.5 indicando $F_{i\bullet} = \mathbf{0}$, este resultado está conforme o esperado devido a escolha da distribuição *a priori* para ρ_i referentes a estes grupos.

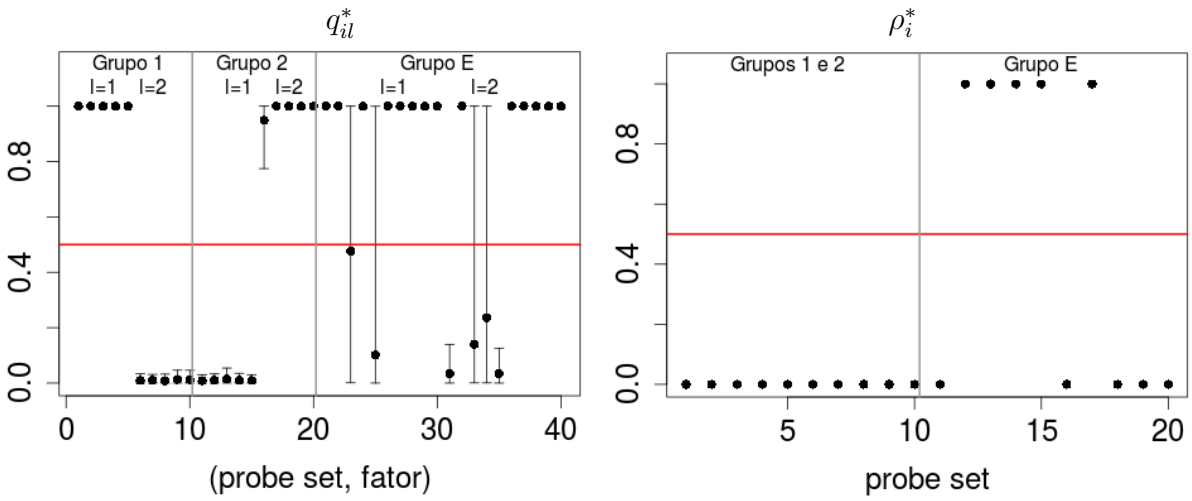

Figura 4.7: *Médias a posteriori (círculo) de q^*_{il} e ρ^*_i. O intervalo HPD de 95% de credibilidade é representado pelo segmento de reta na vertical.*

4.3 Conclusões do Capítulo

Primeiramente foi feita uma breve descrição sobre o problema onde temos regiões do genoma com CNA e grupos de genes localizados nessas regiões para os quais definimos os fatores do modelo. Em seguida, foi estudado o modelo fatorial latente esparso com interações assumindo distribuições *a priori* na forma de mistura para as cargas e para o efeito de interação, sendo que o Processo Gaussiano usado na componente de mistura do efeito de interação apresenta a função de covariâncias exponencial quadrática, conforme estudado em Mayrink e Lucas (2013). Para verificar o desempenho do modelo foi realizado um estudo simulado replicando os resultados do artigo. Neste estudo de simulação, observamos boas estimativas para os parâmetros do modelo. No próximo capítulo, serão apresentados alguns resultados de um estudo feito usando funções de covariâncias diferentes. A função de covariâncias exponencial quadrática permite tratar a suavidade do processo somente por meio do parâmetro l_s. As novas opções que iremos investigar apresentam um parâmetro extra que permite maior controle sobre a suavização.

Capítulo 5

Aplicações Usando Outras Funções de Covariâncias

Neste capitulo, será ajustado um modelo fatorial latente esparso com interações para modelar os tipos de associações entre genes que apresentam CNA assim como feito no capítulo anterior. Utilizaremos o modelo em (4.1) juntamente com as distribuições *a priori* em (4.2) e (4.3), sendo que o Processo Gaussiano na componente de mistura do efeito de interação será definido usando outras funções de covariâncias, diferente da exponencial quadrática. Iniciamos o estudo com a função exponencial potência na próxima seção. Além de apresentar as funções de covariâncias, este capítulo também desenvolve um estudo simulado de avaliação.

5.1 Função Exponencial Potência

Primeiramente considere $t = \|\lambda_{\bullet j1} - \lambda_{\bullet j2}\|$, então a função de covariâncias exponencial potência é definida por:

$$K(t) = v^2 \exp\left\{-\left|\frac{t}{l_s\sqrt{2}}\right|^{\kappa}\right\} \text{ para } t > 0 \text{ e } 0 < \kappa \leq 2. \tag{5.1}$$

Diferente da função de covariâncias Gaussiana, esta opção apresenta em sua composição um parâmetro extra que também será responsável pela suavização. Ao considerarmos $\kappa = 1$, tem-se a função de covariâncias exponencial, para $\kappa = 2$ temos a função de

covariâncias Gaussiana que foi estudada no capítulo 4 e abordada por Mayrink e Lucas (2013). O parâmetro l_s é difícil de estimar, por isso optamos por fixá-lo junto com κ. A Figura 5.1 apresenta três tipos de gráficos da função de covariâncias exponencial potência assumindo: $v^2 = 1$, l_s com valores 0.1, 0.2, 0.3 e κ sendo 0.5, 1.0, 1.5 e 2.0. Observe que a medida que os valores de κ aumentam ou valores de l_s diminuem, essa função apresenta um decaimento mais intenso. Isso ocorre porque ao usarmos a função com $\kappa = 0.5$ ela considera as maiores distâncias entre os escores dos fatores $\lambda_{\bullet j_1}$ e $\lambda_{\bullet j_2}$, enquanto que ao utilizarmos $\kappa = 1.0$ a função está considerando distâncias curtas entre os escores dos fatores, observe por exemplo, que no terceiro painel da Figura 5.1 a função decai com mais intensidade quando aumentamos os valores de κ. Da mesma forma ao aumentarmos os valores de l_s esta função também considera distâncias maiores entre os escores dos fatores, veja por exemplo na Figura 5.1 que ao fixarmos $\kappa = 0.5$ e aumentarmos l_s, notamos a curva se distanciando das coordenadas do gráfico. Este tipo de comportamento poderia trazer mais informação na estimação das interações do modelo fatorial e suavizar mais a superfície estimada.

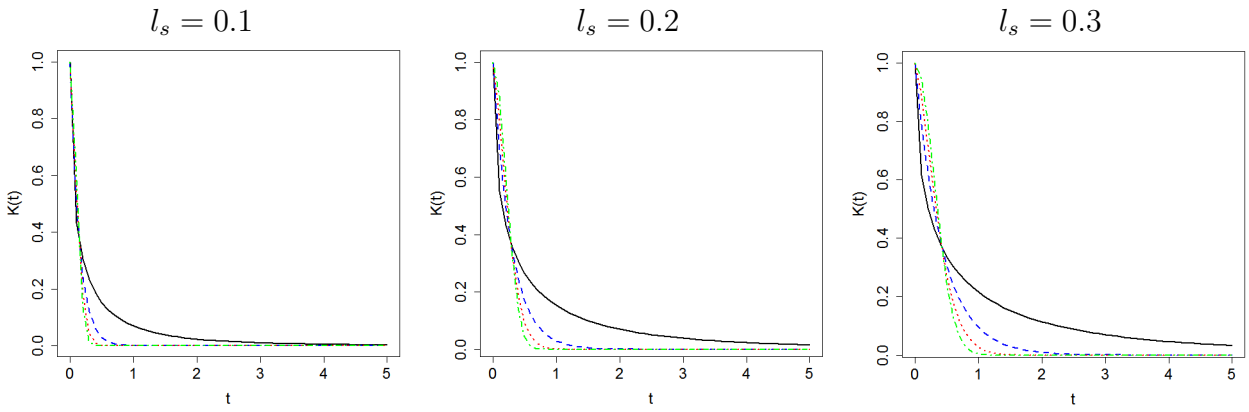

Figura 5.1: Exemplos de funções de covariâncias exponencial potência com: $v^2 = 1$ e $l_s = 0.1$, 0.2 e 0.3, respectivamente, e com $\kappa = 0.5$ (linha sólida); $\kappa = 1$ (linha tracejada), $\kappa = 1.5$ (linha pontilhada) e $\kappa = 2.0$ (linha pontilhada e tracejada).

5.2 Classe Matérn de Funções de Covariâncias

A classe de funções de covariâncias Matérn é uma outra opção a ser utilizada no modelo fatorial em (4.1). Ela é expressa por:

$$K(t) = \frac{\upsilon^2 2^{1-\kappa}}{\Gamma(\kappa)} \left(\frac{2t\sqrt{2\kappa}}{l_s} \right)^{\kappa} B_\kappa \left(\frac{2t\sqrt{2\kappa}}{l_s} \right) \text{ para } t > 0 \text{ e } \kappa > 0, \tag{5.2}$$

sendo κ um parâmetro de suavização e B_κ a função de Bessel modificada (veja Abramowitz e Stegun (1965)). Essa função é obtida pela solução da seguinte equação diferencial:

$$x^2 \frac{d^2}{dx} y + x \frac{d}{dx} y + (x^2 - \kappa^2) y = 0,$$

que fornecerá as funções de Bessel de primeiro e segundo tipo. Ao fazer uma alteração no tipo 2 tem-se a chamada função de Bessel modificada utilizada em (5.2).

Essa classe apresenta características interessantes ao modificarmos o parâmetro de suavização. No caso em que $\kappa \to \infty$, tem-se a função de covariâncias exponencial quadrática. Para $\kappa = \frac{1}{2}$, tem-se a função exponencial. Se utilizarmos a função de covariâncias Matérn com o parâmetro de suavização $\kappa = \frac{3}{2}$ ou $\kappa = \frac{5}{2}$, teremos as seguintes formas respectivamente:

$$K(t) = \upsilon^2 \left(1 + t\frac{\sqrt{3}}{l_s} \right) \exp\left\{ -t\frac{\sqrt{3}}{l_s} \right\} \text{ para } t > 0,$$

$$K(t) = \upsilon^2 \left(1 + t\frac{\sqrt{5}}{l_s} + \frac{5t^2}{3l_s^2} \right) \exp\left\{ -t\frac{\sqrt{5}}{l_s} \right\} \text{ para } t > 0.$$

Nesta classe de funções de covariâncias iremos utilizar estes casos especiais citados, conforme especificação de κ, em um estudo de simulação adiante para avaliar o modelo fatorial. Para mais detalhes a respeito dessa classe de funções e de outras veja Benerjee et al. (2004).

5.3 Critérios de Comparação

Nesta seção, apresentamos alguns critérios de comparação que iremos utilizar nas aplicações adiante, para avaliar o desempenho dos modelos com diferentes configurações

da função de covariâncias no Processo Gaussiano. O Erro Quadrático Médio (EQM) é uma medida que pode ser usada para nosso proposito de comparação. Neste trabalho, ele é calculado como segue:

$$EQM(\alpha) = \frac{1}{mL}\sum_{l=1}^{L}\sum_{i=1}^{m}\left(\hat{\alpha}_{il} - \alpha_{il}\right)^2;\ EQM(\lambda) = \frac{1}{Ln}\sum_{l=1}^{L}\sum_{j=1}^{n}\left(\hat{\lambda}_{lj} - \lambda_{lj}\right)^2;$$

$$EQM(\sigma^2) = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{\sigma^2}_i - \sigma_i^2\right)^2 \quad \text{e} \quad EQM(F_{i\bullet}) = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{F}_{ij} - F_{ij}\right)^2.$$

Outro critério a ser utilizado é o *Deviance Information Criterion* (DIC). O DIC conforme Spiegelhalter et al. (2002) é baseado em duas componentes, uma que mede a qualidade do ajuste e outra que penaliza o modelo levando em conta a complexidade medida pela estimativa do número efetivo de parâmetros. A deviance tem um papel fundamental no cálculo do DIC, ela é dada por:

$$D(x,\theta) = -2\log p(x \mid \theta), \text{ sendo } x = (x_1, ..., x_n) \text{ os dados.}$$

A discrepância entre os dados e o modelo depende tanto de θ quanto de x. Para resumir essa dependência apenas de x, pode-se definir:

$$D_{\hat{\theta}}(x) = D(x, \hat{\theta}(x)), \tag{5.3}$$

que usa algum estimador pontual para θ como, por exemplo, a média *a posteriori*. Do ponto de vista Bayesiano, talvez seja mais atrativo usar a média da deviance sobre a distribuição *a posteriori*, dada por:

$$D_{avg}(x) = E\left[D(x,\theta) \mid x\right],$$

que pode ser estimada usando as observações $\theta^{(s)}$ geradas nas simulações a partir do estimador:

$$\hat{D}_{avg}(x) = \frac{1}{S}\sum_{s=1}^{S} D(x, \theta^{(s)}). \tag{5.4}$$

Para Gelman et al. (2003) a média em (5.4) é um melhor resumo do erro do modelo que a discrepância da estimativa pontual em (5.3). A estimativa pontual usada faz com que o modelo se ajuste bem, enquanto que a média $\hat{D}_{avg}$ usa uma variedade de valores possíveis do parâmetro.

A partir dessas informações o DIC pode ser calculado por:

$$DIC = 2\hat{D}_{avg}(x) - D_{\hat{\theta}}(x),$$

com $\hat{D}_{avg}$ e $D_{\hat{\theta}}$ definidos em (5.4) e (5.3), respectivamente. Valores baixos do DIC indicam melhor ajuste. Para mais detalhes veja Gelman et al. (2003).

O *Widely Applicabe Information Criterion* (WAIC), introduzido por Watanabe (2010), utiliza a verossimilhança para calcular duas componentes. Uma delas é a componente baseada na densidade preditiva para a qualidade do ajuste, que pode ser calculada pelo seguinte estimador Monte Carlo:

$$\widehat{\text{lpd}} = \sum_{i=1}^{n} \log \left[\frac{1}{S} \sum_{s=1}^{S} p(x_i \mid \theta^{(s)}) \right], \tag{5.5}$$

sendo S o número de observações geradas no MCMC da distribuição *a posteriori*. A segunda é a estimativa para o número efetivo de parâmetros, que é calculada usando a variância *a posteriori* da log densidade preditiva para cada dado x_i, descrito por:

$$\hat{p}_{WAIC} = \sum_{i=1}^{m} V_{s=1}^{S} \left(\log p(x_i \mid \theta^{(s)}) \right), \tag{5.6}$$

sendo $V_{s=1}^{S}(a_s) = \frac{1}{S-1} \sum_{s=1}^{S} (a_s - \bar{a})^2$, uma variância amostral.

Assim, a partir das equações (5.5) e (5.6) pode-se calcular o WAIC (valores altos indicam melhor ajuste do modelo) como segue:

$$\widehat{WAIC} = \widehat{\text{lpd}} - \hat{p}_{WAIC}.$$

Uma quarta abordagem para avaliação e seleção de modelos é usar a distribuição preditiva para obter uma medida chamada *Conditional Predictive Ordinate* (CPO). Os CPO's são densidades de validação cruzadas que sugerem quais valores de observações x_i são prováveis quando o modelo é ajustado a todas as observações exceto a i-ésima. O CPO gera uma medida para cada observação individualmente e quando somamos o logaritmo de cada um, isso nos proporciona a medida *Log Pseudo Marginal Likelihood* (LPML). É possível calcular um CPO para cada x_i usando apenas um MCMC explorando as amostras da distribuição *a posteriori* e a verossimilhança. Um estimador Monte Carlo para calcular o CPO é dado por:

$$\widehat{CPO}_i = \left[\frac{1}{S} \sum_{s=1}^{S} \frac{1}{p(x_i \mid \theta^{(s)})} \right]^{-1}, \tag{5.7}$$

sendo S o número de iterações do algoritmo MCMC utilizado para calcular a média harmônica em (5.7). Valores altos de $\widehat{CPO}_i$ indicam melhores ajustes. Finalmente, o LPML pode ser calculado da seguinte maneira:

$$LPML = \sum_{i=1}^{n} \log \widehat{CPO}_i.$$

5.4 Estudo Simulado Envolvendo as Funções Exponencial Potência e Matérn

Neste estudo, ajustamos o modelo fatorial considerando o mesmo conjunto de dados simulados usado no capítulo anterior. As especificações *a priori* utilizadas para α_{il}, $F_{i\bullet}$ e $\lambda_{\bullet j}$ estão em (4.2), (4.3) e (3.3), respectivamente. Assumimos uma $GI(2.1, 1.1)$ para σ_i^2, $\omega = 10$ e usamos a função exponencial potência com $\upsilon = 1$, l_s assumindo os valores 0.1, 0.2, 0.3 e κ considerando os valores 0.5, 1.0, 1.5 e 2.0. Ao avaliarmos a função Matérn utilizamos κ sendo $\frac{3}{2}$ e $\frac{5}{2}$. Inicialmente exploramos as distribuições *a priori* para q_{il} e ρ_i vistas na Tabela 4.1, entretanto na execução do MCMC estas especificações não estavam garantindo a suposição estabelecida de que cada fator teria influência em cada grupo individualmente. Por isso decidimos atribuir distribuições *a priori* mais "fortes". A Tabela 5.1 apresenta as distribuições *a priori* usadas neste estudo.

Tabela 5.1: Distribuições *a priori* e valores iniciais.

Índices	q_{i1}	q_{i2}	ρ_i
$i \in G_1$	Beta$(3,1)$	Beta$(1,3)$	Beta$(1,3)$
$i \in G_2$	Beta$(1,3)$	Beta$(3,1)$	Beta$(1,3)$
$i \in G_E$	Beta$(1,1)$	Beta$(1,1)$	Beta$(1,1)$
Valores iniciais	0.9999	0.0001	0.0001
	0.0001	0.9999	0.0001
	0.5	0.5	0.5

Nesta simulação, consideramos um total de 4000 iterações sendo as 3000 primeiras formando o período de *burn-in*. Não foram utilizados *lag's* e os valores iniciais das cadeias são: $\alpha_{il}^{(0)} = 0$, $F_{ij}^{(0)} = 0$, $\sigma_i^{2(0)} = 1$ e $\lambda_{lj} \sim N(0, 0.3)$. Para as probabilidades $q_{il}^{(0)}$ e $\rho_i^{(0)}$ consideramos os valores iniciais mostrados na Tabela 5.1. Calculamos o EQM das estimativas *a posteriori*, além do DIC, WAIC e LPML para comparar os modelos com diferentes configurações de (κ, l_s). A Tabela 5.2 mostra o EQM para alguns parâmetros do modelo. Podemos notar que os melhores resultados (menores EQM's) para α e λ ocorrem quando assumimos $\kappa = 0.5$ e $l_s = 0.2$ na função exponencial potência. A opção $\kappa = 2.0$, também fornece baixos EQM's para α e λ superando as configurações restantes da função potência. Avaliando com mais atenção os resultados com $\kappa = 2.0$ na função potência, observamos que neste caso o modelo captura melhor os efeitos de interação com EQM's baixos para $F_{13\bullet}$, $F_{14\bullet}$ e $F_{15\bullet}$. Observe que ao utilizarmos a configuração $(\kappa = 1.5, l_s = 0.3)$ obtivemos o pior cenário em termos de ajuste na análise dos EQM's, e isso também pode ser visto graficamente no Apêndice C.

Assim como no modelo com função exponencial potência tendo $\kappa = 2.0$, o modelo fatorial ajustado com a função Matérn usando $\kappa = \dfrac{3}{2}$ também apresenta bom desempenho. Observe que os EQM's referentes a α, λ e alguns F quando usarmos a função Matérn com $\left(\kappa = \dfrac{3}{2}, l_s = 0.1\right)$ são próximos dos EQM's ao utilizarmos a função potência com $(\kappa = 2.0, l_s = 0.3)$ e melhores em relação a algumas configurações como $\kappa = 1.0$ e 1.5 na

função exponencial potência e $\frac{5}{2}$ na classe Matérn.

Tabela 5.2: Erro quadrático médio para os parâmetros α, λ, σ^2 e F no modelo fatorial com interações assumindo as funções de covariâncias exponencial potência (EP) e Matérn (M).

Função	κ	l_s	α	λ	σ^2	$F_{12\bullet}$	$F_{13\bullet}$	$F_{14\bullet}$	$F_{15\bullet}$	$F_{17\bullet}$
EP	0.5	0.1	0.0480	0.0789	0.0020*	0.2825	0.1314	0.2611	0.1536	0.2291
		0.2	**0.0380**	**0.0572**	0.0014	0.2969	0.0988	0.3034	0.1473	0.2760
		0.3	0.0648	0.0797	0.0009	0.3053	0.1069	0.3790	0.1577	0.3252
EP	1.0	0.1	0.1209	0.1299	0.0015	**0.2639**	0.1384	0.2460	0.1544	**0.2194**
		0.2	0.1269	0.1407	0.0009	0.3115	0.1684	0.4310	0.1475	0.2627
		0.3	0.1440	0.1314	0.0007	0.3290	0.1577	0.4342	0.1848*	0.3120
EP	1.5	0.1	0.1105	0.1373	0.0014	0.3072	0.1378	0.3322	0.1582	0.2470
		0.2	0.3161*	0.2477*	0.0005	0.2846	0.1775	0.3388	0.1487	0.3033
		0.3	0.2002	0.1889	**0.0004**	0.3346*	0.2508*	0.4522*	0.1602	0.3391*
EP	2.0	0.1	0.0958	0.1573	0.0014	0.3155	**0.0881**	0.2795	0.1488	0.2234
		0.2	0.0536	0.0905	0.0010	0.2990	0.0991	**0.2068**	**0.1309**	0.2382
		0.3	0.0424	0.0735	0.0019	0.2650	0.1796	0.2845	0.1389	0.2275
M	$\frac{3}{2}$	0.1	**0.0648**	**0.0918**	0.0016*	**0.2724**	**0.0932**	**0.2152**	0.1480	**0.1492**
		0.2	0.1975	0.1912	0.0007	0.2905	0.2056*	0.2658	**0.1469**	0.2874
		0.3	0.2380	0.2177*	**0.0005**	0.3128	0.1029	0.2468	0.1513	0.2853
M	$\frac{5}{2}$	0.1	0.0698	0.1118	0.0012	0.2897	0.1504	0.2876	0.1476	0.2144
		0.2	0.1399	0.1549	0.0008	0.3278*	0.1488	0.2884	0.1480	0.2768
		0.3	0.2452*	0.2136	0.0006	0.2847	0.1419	0.3029*	0.1536*	0.3387*

Em negrito e com marcação * estão os menores e maiores EQM, respectivamente, considerando a função EP e M.

Ao considerarmos o DIC, WAIC e LPML, que são medidas globais para avaliação de um modelo, notamos que estes critérios apontam direções diferentes em relação ao EQM quando consideramos cenários diferentes para κ. Observamos na Tabela 5.3 que os valores do WAIC e LPML são menores com $\kappa = 0.5$ e $l_s = 0.1$ na função potência. E ao utilizar $\kappa = 1.5$ e $l_s = 0.3$ obtivemos o maior LPML. Um fato interessante é que neste cenário obtivemos uma boa estimação para σ^2 (cenário com menor EQM). Para o cenário ($\kappa = 2.0$, $l_s = 0.3$) foi obtido o menor DIC e o maior WAIC confirmando o bom ajuste global do modelo com esta configuração na função de covariâncias exponencial potência. Veja também que os valores do LPML para este cenário com $\kappa = 2.0$ e $l_s = 0.3$ são próximos dos melhores valores obtidos com $\kappa = 1.5$.

Observe também na Tabela 5.3, que todos os critérios globais apontam para o modelo com configuração $\left(\kappa = \frac{3}{2}, l_s = 0.3\right)$ na função Matérn como sendo aquele com melhor ajuste. Entretanto, o modelo fatorial não apresenta um bom ajuste ao usarmos a função com esta configuração. Os gráficos com as médias *a posteriori* e os intervalos HPD para o cenário $\left(\kappa = \frac{3}{2}, l_s = 0.3\right)$ são apresentados no Apêndice C; observe que eles não sugerem um bom ajuste.

Concluindo, esta análise acaba dando suporte para os bons resultados obtidos em Mayrink e Lucas (2013), confirmando que o modelo fatorial ao utilizar a função exponencial quadrática (ou potência usando $\kappa = 2.0$) com parâmetro $l_s = 0.3$ apresenta um bom ajuste.

Tabela 5.3: Critérios para avaliação de modelos considerando diferentes valores para κ e l_s nas funções exponencial potência (EP) e Matérn (M).

Função	κ	l_s	DIC	WAIC	LPML
EP	0.5	0.1	4505.75	-2320.79^*	-1287.23^*
		0.2	4303.72	-2060.50	-1260.63
		0.3	3877.10	-1904.78	-1188.64
EP	1.0	0.1	4943.86^*	-1953.68	-1246.45
		0.2	4074.99	-1903.11	-1215.49
		0.3	3810.92	-1772.37	-1165.71
EP	1.5	0.1	3981.01	-1967.08	-1227.22
		0.2	3613.02	-1715.78	-1146.19
		0.3	3285.47	-1580.55	**-1141.32**
EP	2.0	0.1	4586.53	-2017.58	-1231.63
		0.2	3264.24	-1640.85	-1158.03
		0.3	**2962.04**	**-1378.28**	-1155.67
M	$\frac{3}{2}$	0.1	4592.53^*	-2162.36^*	-1258.07^*
		0.2	3791.98	-1767.84	-1161.08
		0.3	**3636.92**	**-1579.32**	**-1146.74**
M	$\frac{5}{2}$	0.1	4523.13	-2051.89	-1238.05
		0.2	4375.75	-1762.60	-1174.21
		0.3	4022.41	-1675.58	-1157.40

Em negrito e com marcação * estão os melhores e piores valores, respectivamente, nas funções EP e M.

A Figura 5.2 apresenta os gráficos com as estimativas de α, σ^2, λ e F, considerando o cenário ($\kappa = 2.0, l_s = 0.3$) que mostrou bons resultados na análise anterior. Podemos observar que a maioria dos intervalos com 95% de credibilidade contêm o verdadeiro valor do parâmetro confirmando o bom desempenho do modelo. Os gráficos do cenário ($\kappa = 0.5, l_s = 0.2$) são apresentados no Apêndice C; eles também sugerem bom ajuste.

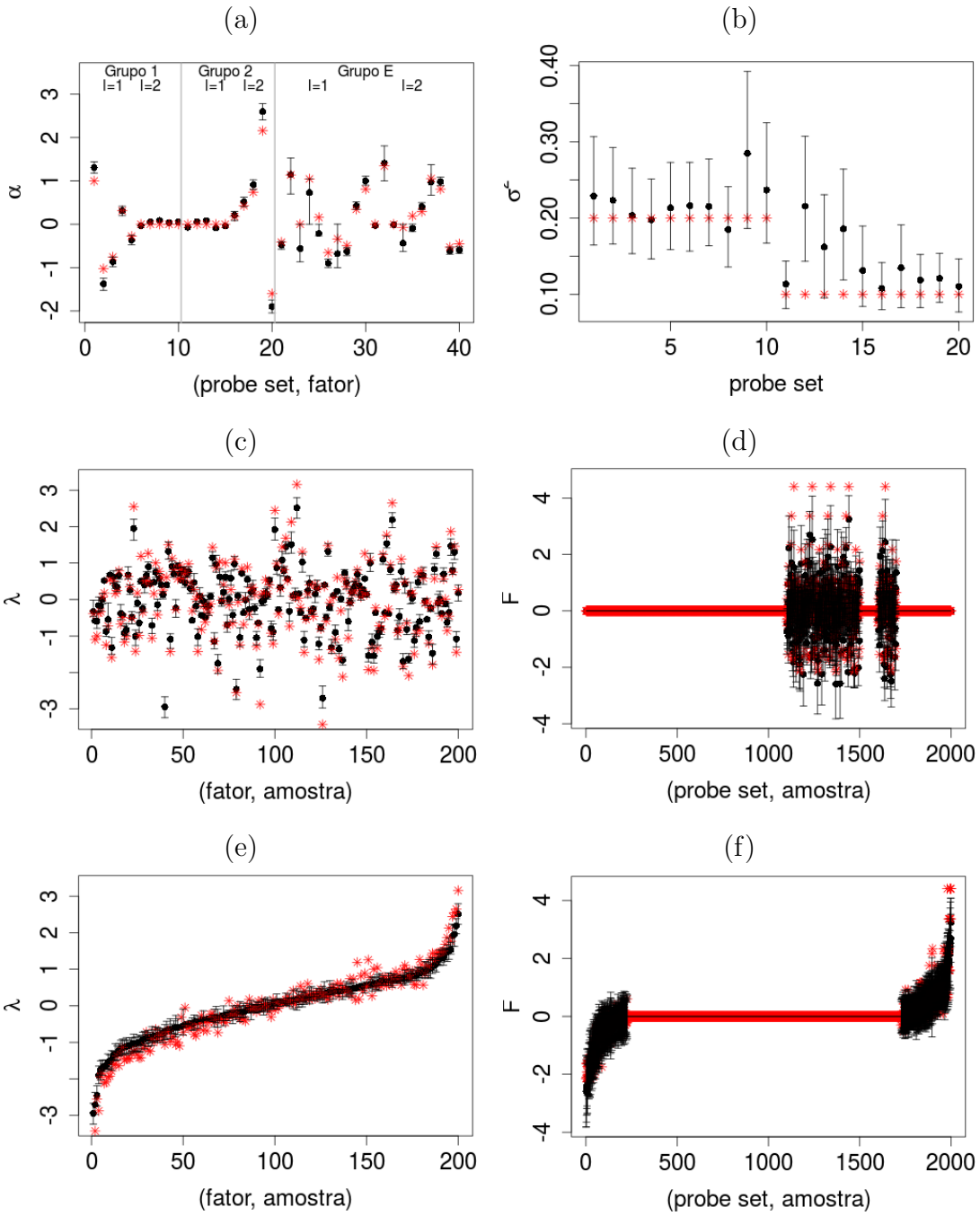

Figura 5.2: *Valores reais (asterisco) e médias a posteriori (círculo). O intervalo HPD de 95% de credibilidade é representado pelo segmento de reta na vertical. Os dois últimos painéis exibem os intervalos ordenados de acordo com a média a posteriori. Cenário com* $\kappa = 2.0$ *e* $l_s = 0.3$ *na função exponencial potência.*

As Figuras 5.3 e 5.4 apresentam os gráficos de superfície real e das estimativas $\hat{F}_{i\bullet}$ considerando os cenários ($\kappa = 2.0, l_s = 0.3$) e ($\kappa = 0.5, l_s = 0.2$), respectivamente. Note que o modelo consegue capturar bem o formato de sela. Observe no segundo painel na Figura 5.3 que a estimativa $\hat{F}_{12\bullet}$ apresenta uma leve melhora da suavidade na superfície em relação a estimativa $\hat{F}_{12\bullet}$ na Figura 5.4 quando considerarmos ($\kappa = 0.5, l_s = 0.2$) na função de covariâncias exponencial potência. Uma maneira de entender o motivo disso é pelo fato de l_s ser maior, pois se aumentarmos o valor de l_s a função estará considerando as maiores distâncias entre os escores do fatores estimados, tornando o raio de influência entre os $\hat{\lambda}_{\bullet j}$ maior, e isso acaba trazendo mais informação na estimação de $F_{i\bullet}$ suavizando a superfície estimada. Veja novamente a Figura 5.1 para visualizar esta relação que explica a maior suavidade. Os gráficos de superfície para o cenário ($\kappa = 1.5, l_s = 0.3$) são mostrados no Apêndice C; veja que o formato real de sela também é bem estimado neste caso.

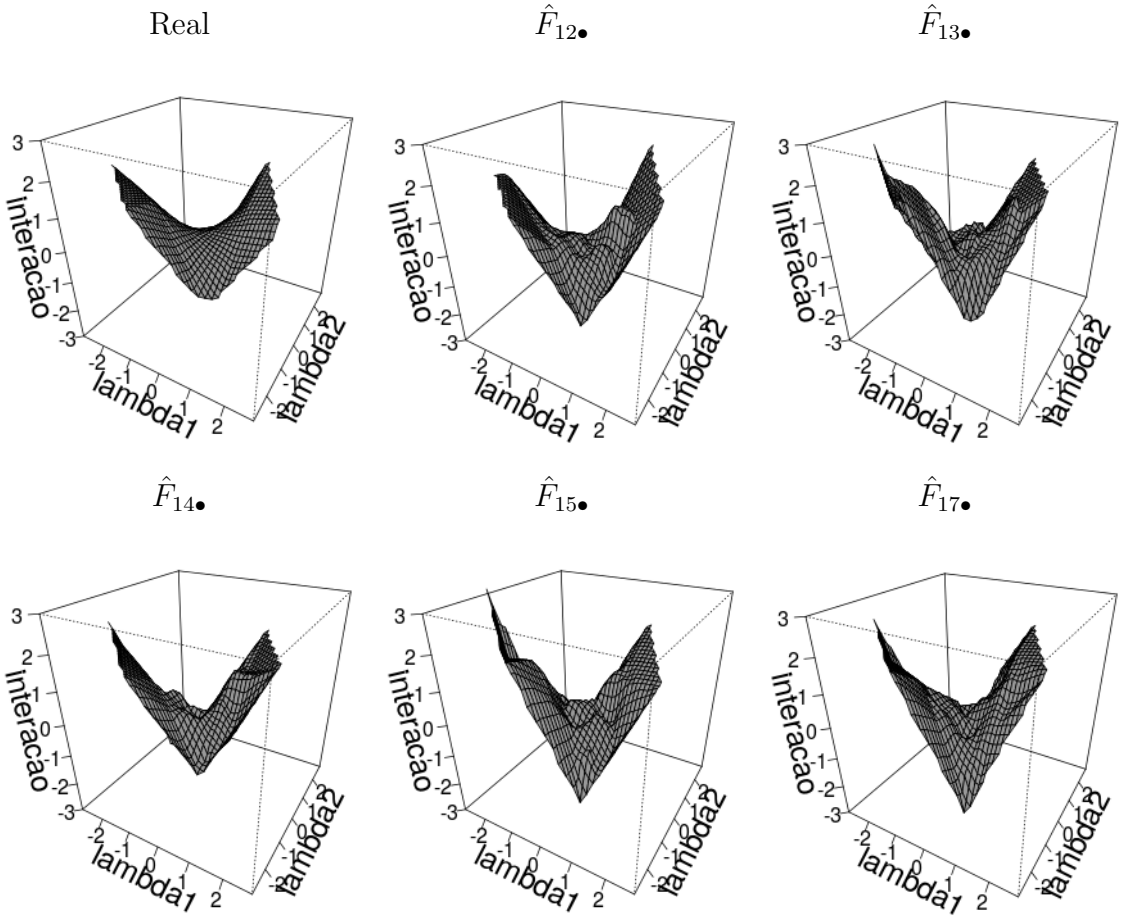

Figura 5.3: *Gráfico de superfície do efeito de interação real e estimado considerando os parâmetro* $\kappa = 2.0$ *e* $l_s = 0.3$ *na função exponencial potência.*

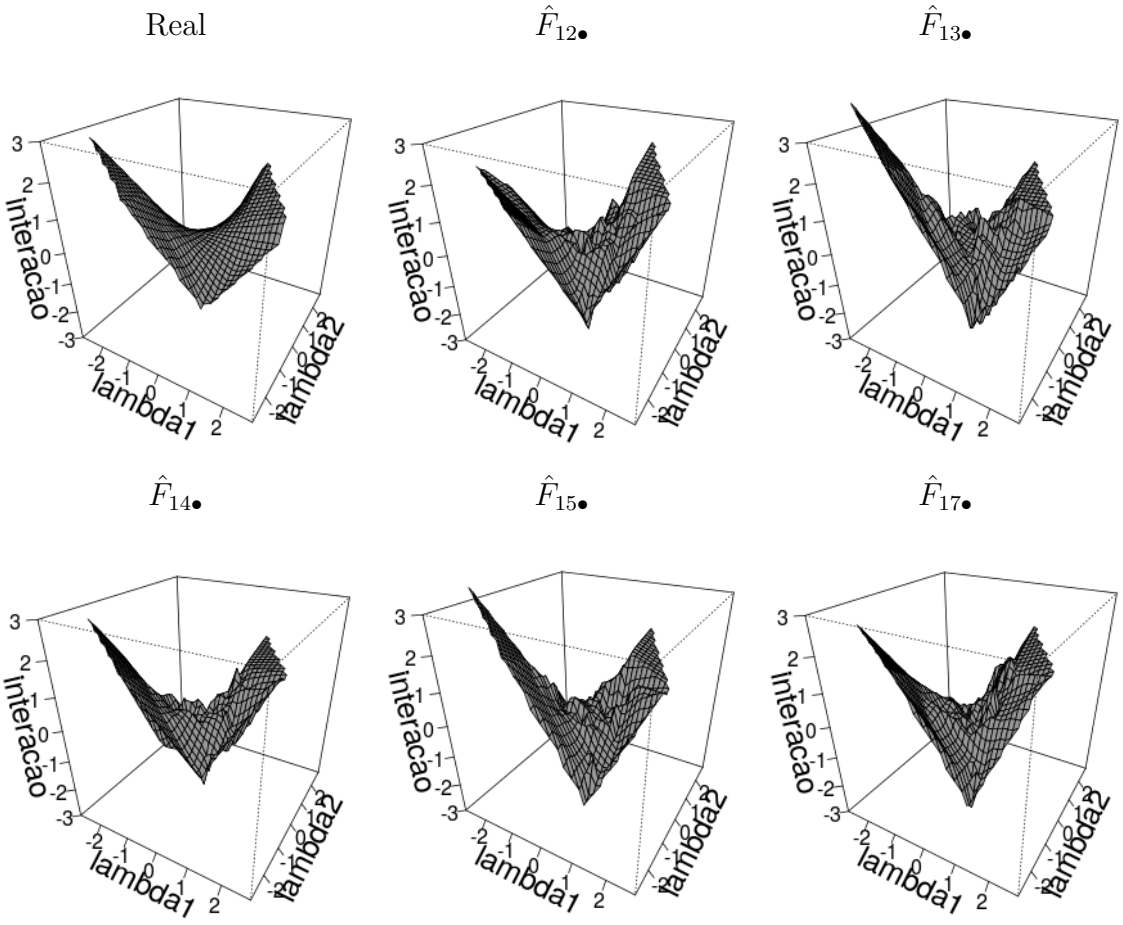

Figura 5.4: *Gráfico de superfície do efeito de interação real e estimado considerando os parâmetro* $\kappa = 0.5$ *e* $l_s = 0.2$ *na função exponencial potência.*

Para avaliar a qualidade das estimativas do efeito de interação, além dos EQM's apresentados na Tabela 5.4, as Figura 5.5 e 5.6 mostram os gráficos da diferença entre o efeito real de interação e o estimado, nos cenários ($\kappa = 2.0, l_s = 0.3$) e ($\kappa = 0.5, l_s = 0.2$), respectivamente. Novamente, ressaltamos que neste caso, em uma situação ideal seria esperado um plano centrado na origem. Podemos observar algumas irregularidades inerentes ao processo de estimação, que indicam o quão distantes as estimativas estariam do valor real. Em uma análise visual fica difícil comparar as superfícies correspondentes nestas duas figuras. Ressaltamos apenas que estes resultados são visualmente mais planos (melhores) que os demais cenários; veja as Figuras C.4 e C.5 no Apêndice C.

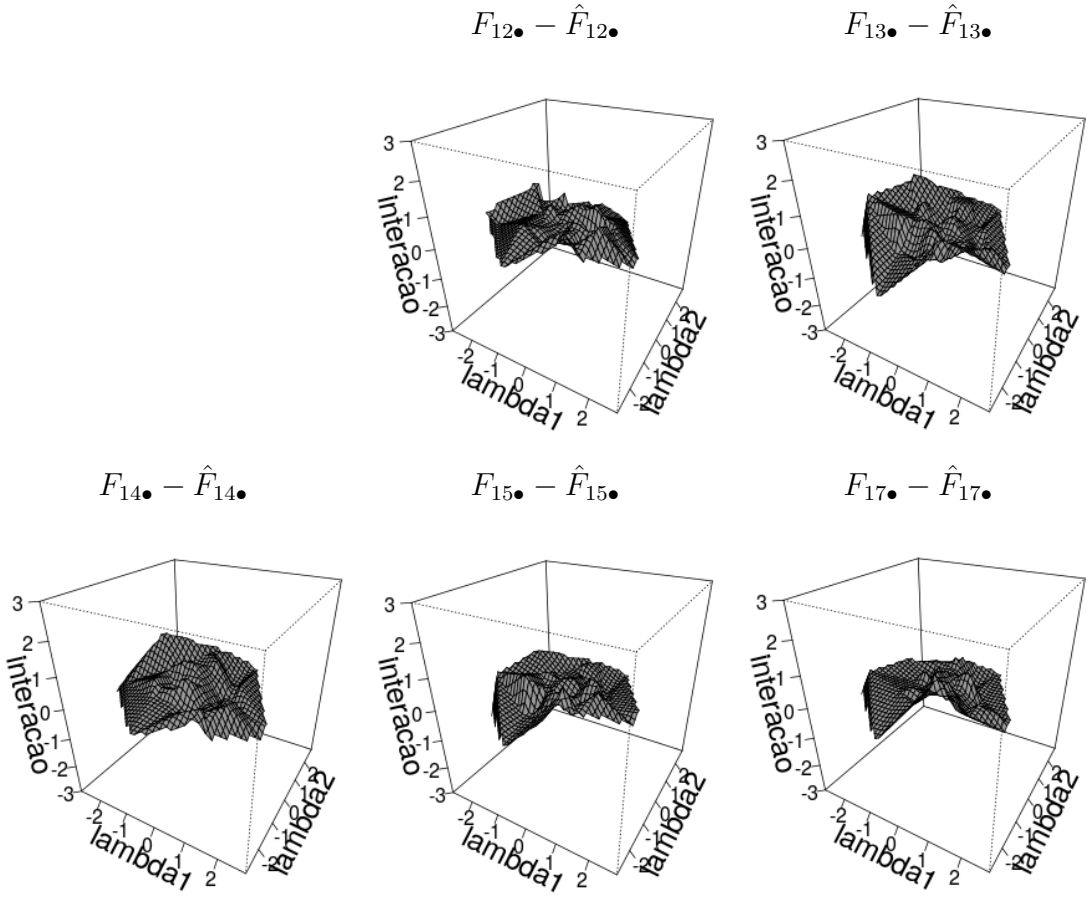

Figura 5.5: *Gráfico de superfície da diferença entre o efeito de interação real e estimado considerando* $\kappa = 2.0$ *e* $l_s = 0.3$.

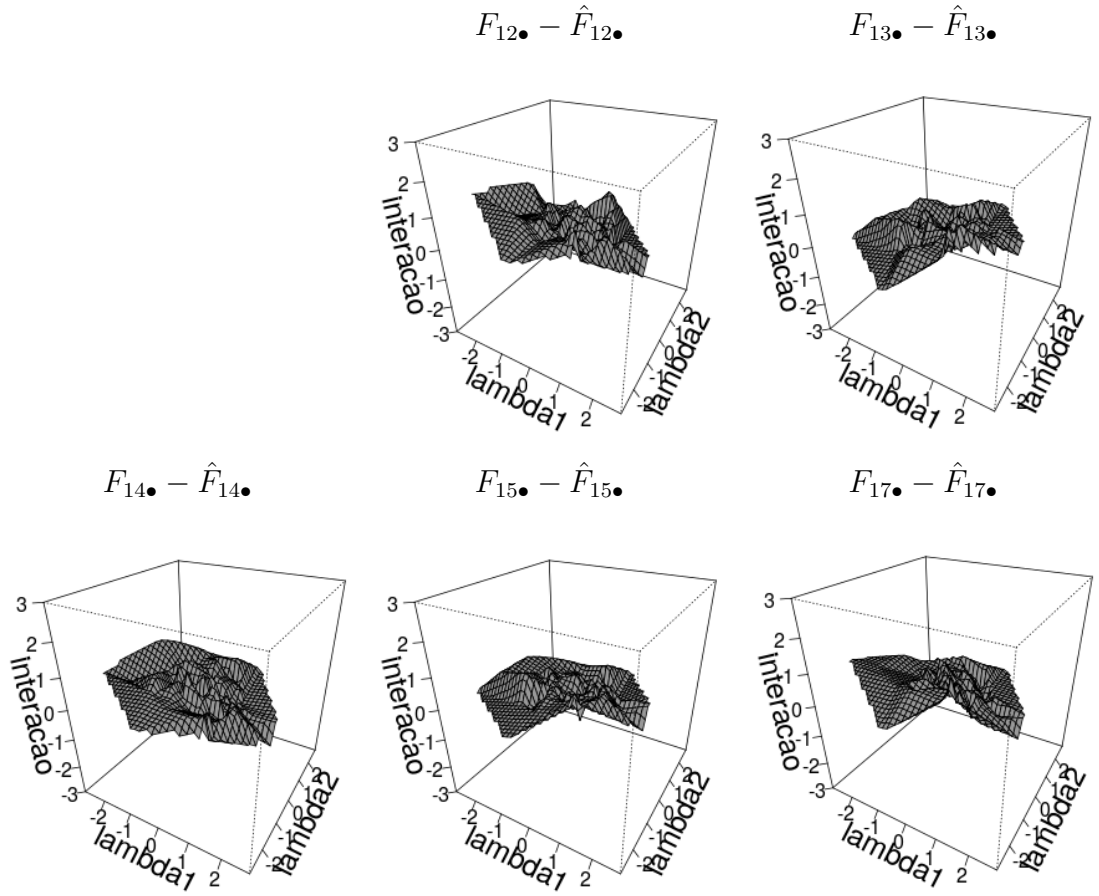

Figura 5.6: *Gráfico de superfície da diferença entre o efeito de interação real e estimado considerando* $\kappa = 0.5$ *e* $l_s = 0.2$.

Na Figura 5.7, pode-se observar o bom ajuste do modelo fatorial a partir dos gráficos incluindo as médias *a posteriori* e intervalos HPD utilizando a função Matérn com $\kappa = \dfrac{3}{2}$ e $l_s = 0.1$. Este foi considerado um dos melhores cenários baseado nas análises dos EQM's, pois estima bem α e λ como quando usamos a função potência com $\kappa = 0.5$, e consegui capturar e estimar bem as interações como na função Gaussiana (potência com $\kappa = 2.0$). Note que as estimativas estão próximas do verdadeiro valor do parâmetro. Veja também que os intervalos de credibilidade englobam em sua maioria os valores reais.

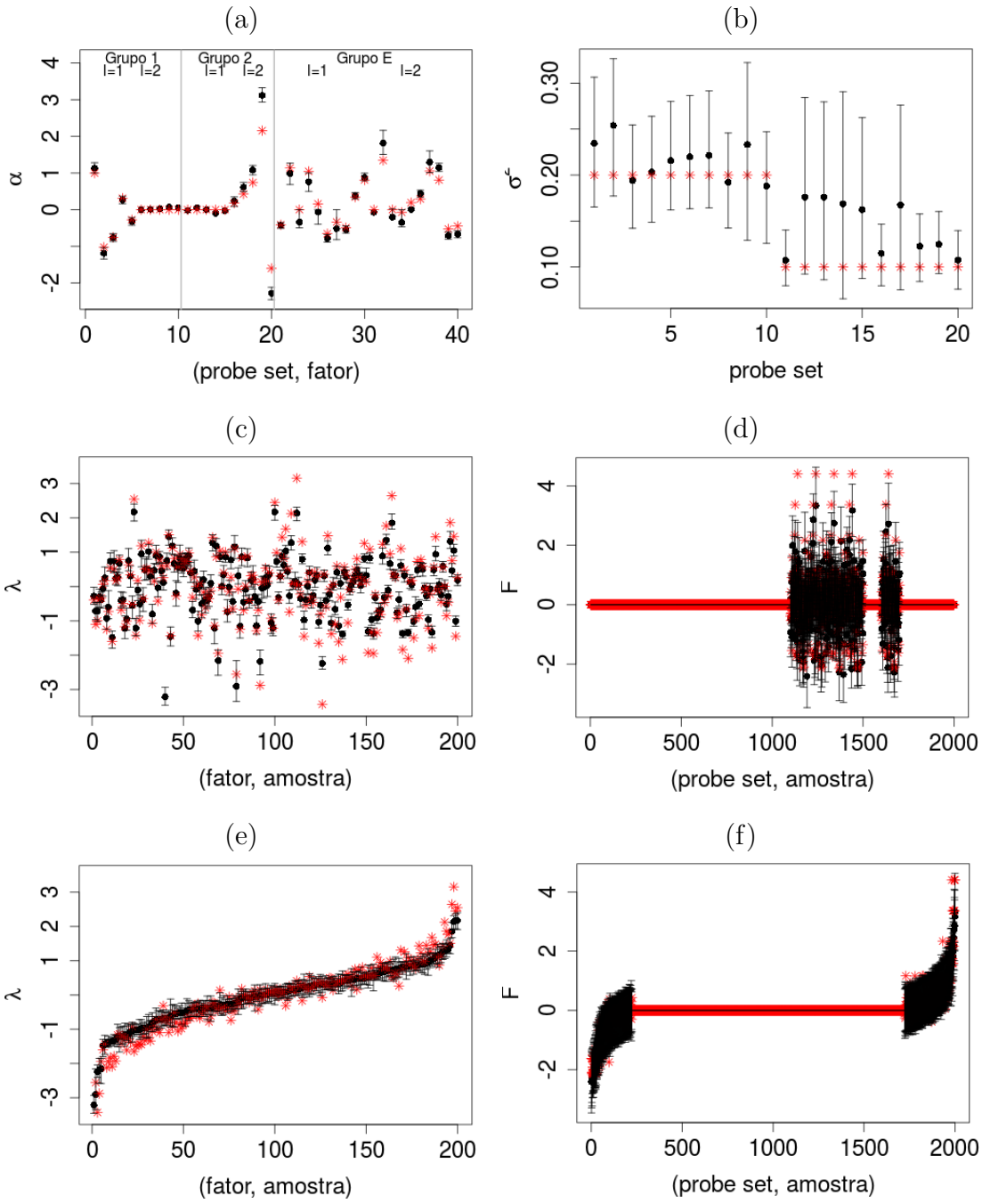

Figura 5.7: *Valores reais (asterisco) e médias a posteriori (círculo). O intervalo HPD de 95% de credibilidade é representado pelo segmento de reta na vertical. Os dois últimos painéis exibem os intervalos ordenados de acordo com a média a posteriori. Cenário considerado* $\kappa = \frac{3}{2}$ *e* $l_s = 0.1$ *na função Matérn.*

As Figuras 5.8 e 5.9 apresentam os gráficos das diferenças entre o efeito de interação real e os estimados considerando a função Mátern com $\left(\kappa = \frac{3}{2}, l_s = 0.1\right)$ e $\left(\kappa = \frac{3}{2}, l_s = 0.3\right)$, respectivamente. Podemos observar que as variações ou irregularidades nas superfícies, que representam o quão ruins são as estimativas dos parâmetros, são menores quando temos $\left(\kappa = \frac{3}{2}, l_s = 0.1\right)$ em relação a $\left(\kappa = \frac{3}{2}, l_s = 0.3\right)$. Observamos também, que no cenário $\left(\kappa = \frac{3}{2}, l_s = 0.1\right)$ temos visualmente estimativas de superfícies mais planas (melhores) que ao usarmos a configuração $\left(\kappa = \frac{5}{2}, l_s = 0.3\right)$ ou ao utilizar o modelo com a função potência e $(\kappa = 1.5, l_s = 0.3)$. Os gráficos de superfície da diferença entre o efeito real e o estimado para o cenário $\left(\kappa = \frac{5}{2}, l_s = 0.3\right)$ na função Matérn são mostrados na Figura C.8 no Apêndice C.

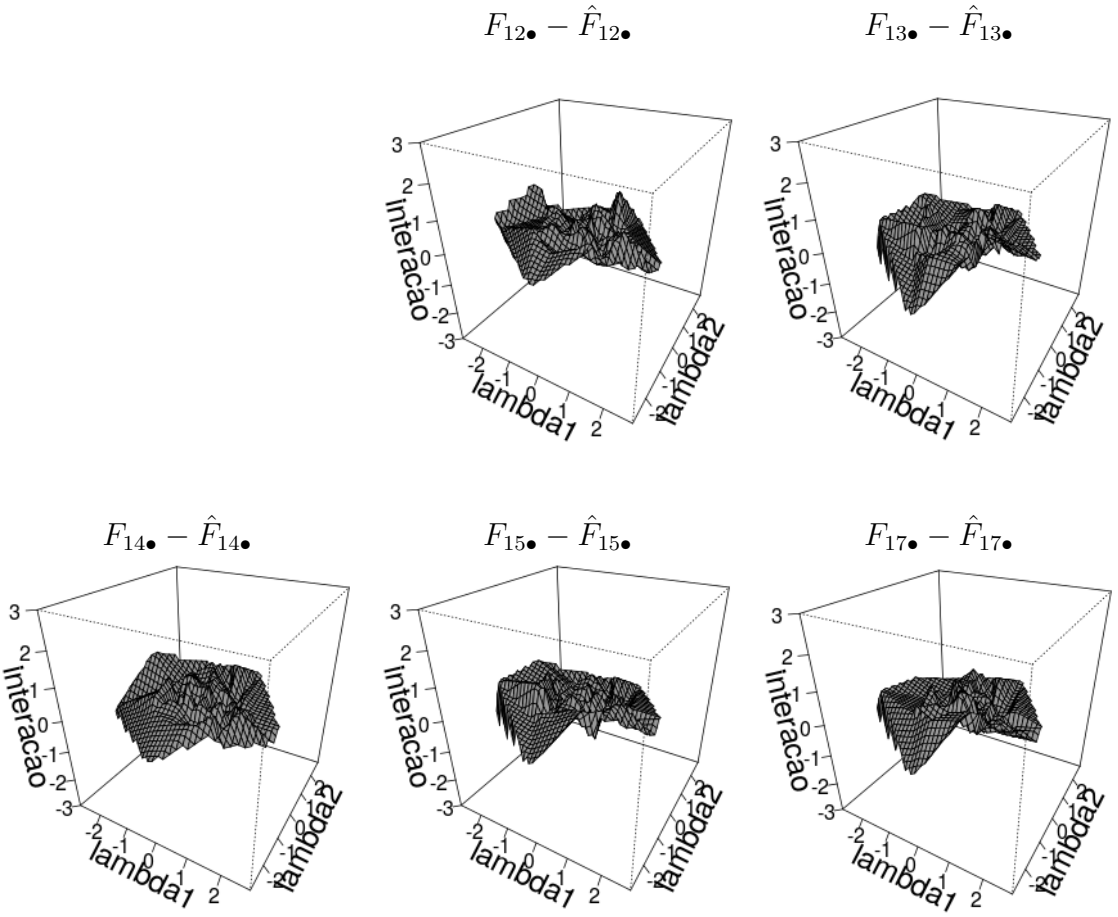

Figura 5.8: *Gráfico de superfície da diferença entre o efeito de interação real e estimado. Considerando* $\kappa = \dfrac{3}{2}$ *e* $l_s = 0.1$ *na função Matérn.*

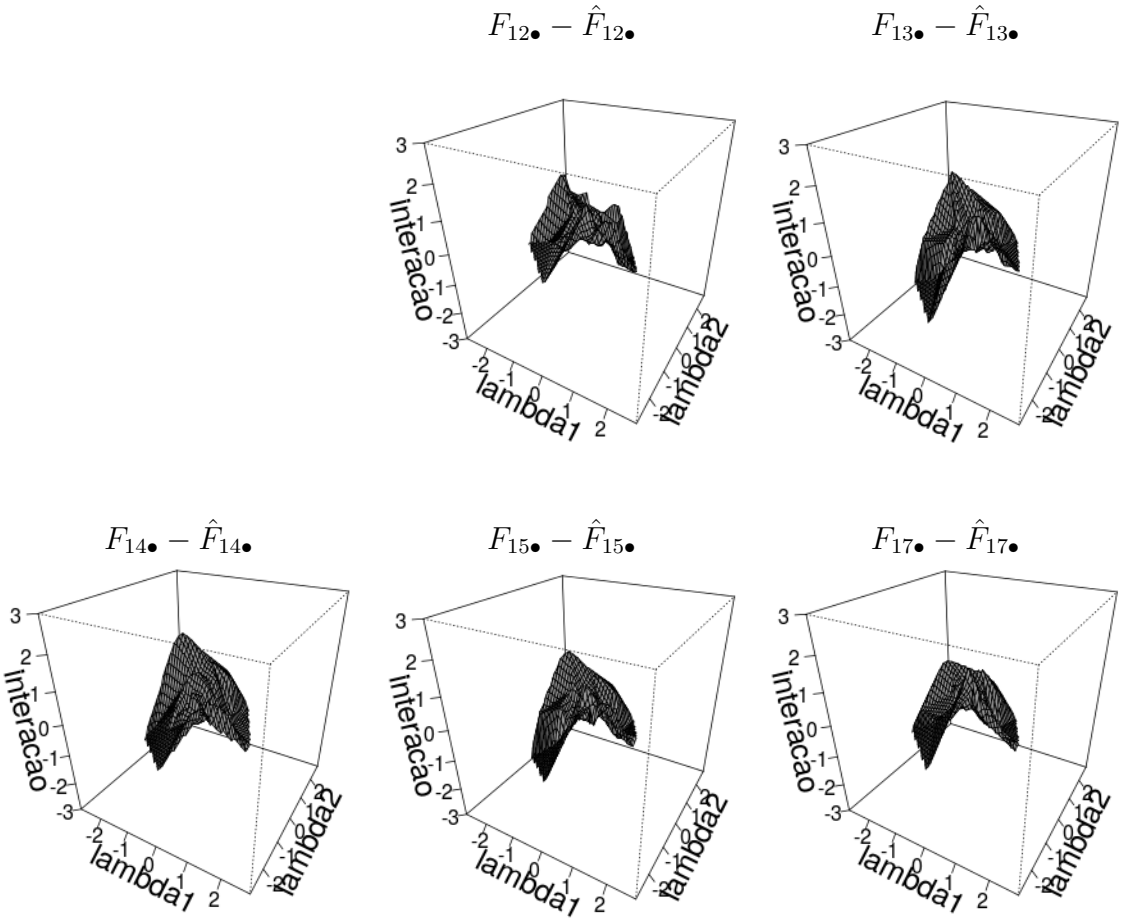

Figura 5.9: *Gráfico de superfície da diferença entre o efeito de interação real e estimado. Considerando* $\kappa = \dfrac{3}{2}$ *e* $l_s = 0.3$ *na função Matérn.*

5.5 Conclusões do Capítulo

Nesta etapa do trabalho, foi verificado o bom ajuste do modelo fatorial latente com interações ao utilizar outras funções de covariâncias como, a exponencial potência e as da classe Matérn. Nos estudos feitos, foi observado que ao utilizar estas funções obtivemos estimativas para os parâmetros tão boas quanto ao utilizar a função de covariâncias exponencial quadrática abordada por Mayrink e Lucas (2013). Nos gráficos de superfícies representando efeito de interação estimado, observamos que as estimativas conseguem capturar bem o formato de sela, que representa o efeito de interação real. Além disso, verificamos a partir dos gráficos da diferença entre a interação real e a estimada, o quão distantes as estimativas estariam do efeito de interação real. O erro quadrático médio assim como o DIC, WAIC e LPML foram calculados e usados como critério de comparação. Foi observado que para algumas especificações de parâmetros nas funções de covariâncias, como $(\kappa = 2.0, l_s = 0.3)$ na exponencial potência e $\left(\kappa = \frac{3}{2}, l_s = 0.1\right)$ na Matérn, o modelo fatorial apresentou bom ajuste. No capítulo seguinte, iremos fazer uma aplicação a dados reais utilizando o modelo fatorial com função Matérn.

Capítulo 6

Aplicação a Dados Reais

Neste capítulo, desenvolvemos uma análise de dados reais tomando como base as expressões de genes registradas em 118 *microarrays* relativos ao câncer de mama e avaliado em Chin et al. (2006). Este foi um dos conjuntos de dados utilizado por Mayrink e Lucas (2013), que investigaram resultados para dois grupos de genes relacionados a regiões do genoma com CNA. A primeira, localizada na posição 35152961 do cromossomo 22 (a qual denotamos como G_1) e a segunda região foi localizada na posição 68771985 do cromossomo 16 (que denotaremos por G_2). Os grupos G_1 e G_2 apresentam 50 e 42 genes, respectivamente. A seleção desses genes é baseada em um intervalo ao redor da posição localizada no genoma. Os *microarrays* selecionados para a aplicação representam 22283 genes replicados em 118 amostras. Para diminuir o custo computacional foi realizado um procedimento de limpeza descrito com detalhes na seção E do material suplementar de Mayrink e Lucas (2013). Esse procedimento reduz o tamanho da matriz de dados X com 22283 linhas, selecionando os principais genes para aplicação. O conjunto de dados que iremos investigar aqui tem G_1 com 22 genes, G_2 com 18 e G_E com 3704 genes. Ressaltamos também que estes dados foram pré-processados via RMA. A Figura 6.1 apresenta a matriz X que será utilizada neste estudo.

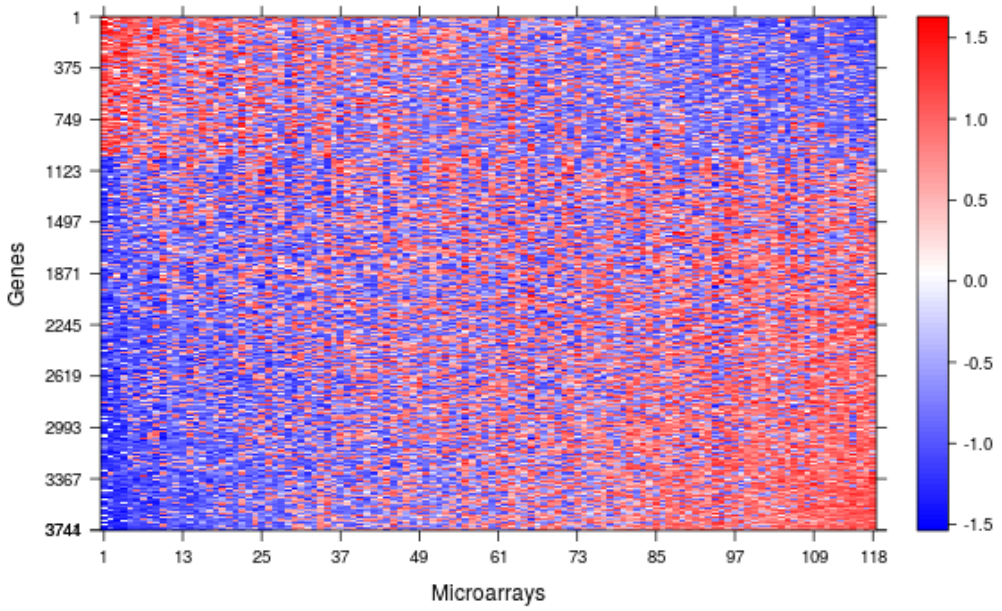

Figura 6.1: *Conjunto de dados de câncer de mama utilizados em Mayrink e Lucas (2013) e Chin et al. (2006).*

O modelo em (4.1) com dois fatores será utilizado em nossa análise e cada fator é responsável por descrever o padrão de expressão a partir das amostras para cada região onde a CNA foi detectada. Usaremos nesta aplicação a função de covariâncias Matérn com $\kappa = \dfrac{3}{2}$ e $l_s = 0.1$, pois ela apresentou resultados mais interessantes em termos do EQM no capítulo anterior. As especificações *a priori* para α_{il} e $F_{i\bullet}$ estão descritas em (4.2) e (4.3), respectivamente. Utilizamos $\lambda_{\bullet j} \sim N_L(\mathbf{0}, I_n)$, $\sigma_i^2 \sim GI(2.1, 1.1)$ e $\omega = 10$. Estas são as mesmas distribuições *a priori* vistas no capítulo anterior e foram usadas por Mayrink e Lucas (2013) quando utilizaram a função de covariâncias Gaussiana. Ressaltamos que estamos mudando apenas a função de covariâncias.

Como a matriz de dados apresenta muitos genes (linhas), iremos utilizar distribuições *a priori* mais "fortes"para q_{il} e ρ_i apresentadas na Tabela 6.1. Esta estratégia, também usada por Mayrink e Lucas (2013), é importante e garante a suposição feita sobre a relação grupo-fator determinando a identificação do modelo. Note que o grupo G_E é muito maior que $(G_1 \cup G_2)$ e, desta forma, as especificações *a priori* anteriores seriam facilmente dominadas pelos dados. Veja que não assumiremos efeito de interação para os

genes em $(G_1 \cup G_2)$, supomos que eles serão influenciados por cada fator individualmente.

Tabela 6.1: Distribuições *a priori* e valores iniciais utilizados para q_{il} e ρ_i.

Índices	q_{i1}	q_{i2}	ρ_i
$i \in G_1$	$p(q_{i1} = 1) = 1$	$p(q_{i2} = 0) = 1$	$p(\rho_i = 0) = 1$
$i \in G_2$	$p(q_{i1} = 0) = 1$	$p(q_{i2} = 1) = 1$	$p(\rho_i = 0) = 1$
$i \in G_E$	Beta$(1, 1)$	Beta$(1, 1)$	Beta$(1, 1)$
	1	0	0
Valores iniciais	0	1	0
	0.1	0.1	0.5

Os valores iniciais utilizados para as cadeias foram: $\alpha_{i1}^{(0)} \sim N(0, 1)$ para $i \in G_1$, $\alpha_{i2}^{(0)} \sim N(0, 1)$ para $i \in G_2$ e $\alpha_{il}^{(0)} = 0$ para os demais i, l; $\lambda_{lj}^{(0)} \sim N(0, 1)$, $\sigma_i^{2(0)} = 1$ e $F_{ij}^{(0)} = 0$. Para $q_{il}^{(0)}$ e $\rho_i^{(0)}$, usados nas variáveis latentes binária $h_{il}^{(0)}$ e $z_i^{(0)}$, utilizamos os valores iniciais mostrados na Tabela 6.1. Consideramos um total de 5000 iterações com um *burn-in* de 3000 e não foram utilizados *lag's*. Gráficos mostrando a convergência de algumas cadeias de α, σ^2 e λ ; podem ser visualizadas na Figura D.1 no Apêndice D, para estas cadeias foi realizado o teste de Geweke (1992) que confirmam a convergência visual.

6.1 Resultados da Aplicação

A Figura 6.2 apresenta as médias *a posteriori* das cargas referentes aos grupos G_1 e G_2 e seus intervalos HPD. Podemos observar que a maioria das cargas α_{i1} em G_1 são negativas enquanto que as cargas α_{i2} em G_2 são positivas mostrando que a direção do efeito de cada fator é oposta sendo eles bem definidos. Estes resultados são semelhantes aos encontrados por Mayrink e Lucas (2013), lembrando que estes autores utilizam o modelo fatorial com a função exponencial quadrática.

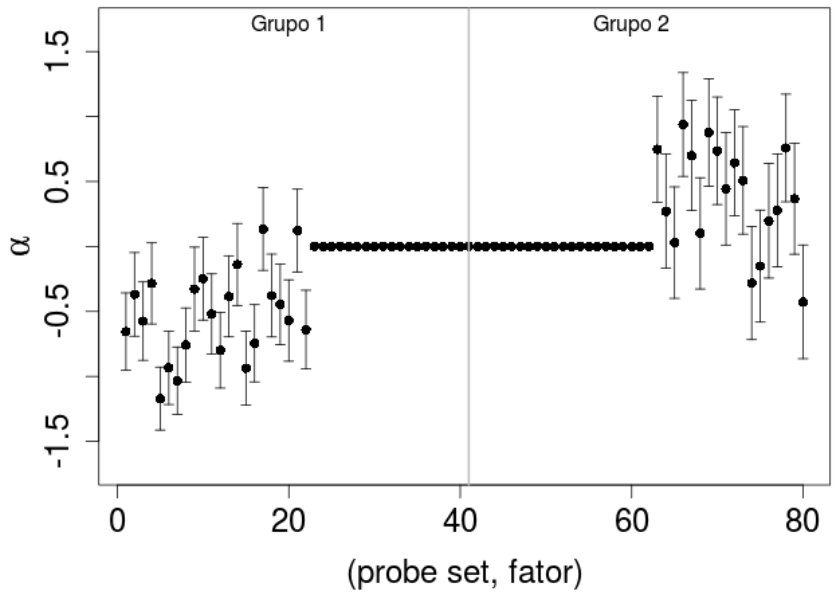

Figura 6.2: *Gráfico com as médias a posteriori (círculo) e intervalos HPD de 95% de credibilidade sendo representado pelo segmento de reta na vertical.*

Ao analisarmos as médias *a posteriori* das probabilidades ρ_i^* e seus intervalos HPD, observamos que 280 genes apresentavam intervalo para ρ_i^* com o limite inferior acima de 0.5. Isto indica que as linhas de F são significativas, isto é, 280 genes em G_E estariam sendo afetados pelo efeito de interação. Ao avaliarmos esta significância das interações pela média *a posteriori* de ρ_i^*, temos 453 médias acima de 0.5 sugerindo 453 linhas de F significantes. Notamos que o número de linhas em X afetadas por F tem relação com a escolha *a priori* da distribuição Beta atribuída a ρ_i. Em um breve estudo que fizemos, considerando $\rho_i \sim$Beta$(1, 10)$ (esta Beta indica que supomos *a priori* que há poucas interações afetando G_E), obtivemos 142 genes com intervalos para ρ_i^* completamente acima de 0.5. Ao avaliarmos a significância por meio da média *a posteriori*, obtivemos 273 casos acima de 0.5. Para o mesmo banco de dados Mayrink e Lucas (2013) identificaram 275 interações afetando os genes ao utilizar a Beta$(1, 1)$ para ρ_i.

A Figura 6.3 mostra a matriz F estimada. O painel (a) exibe a matriz F completa, é possível perceber a presença de algumas linhas na horizontal que representam as interações significativas ($F_{i\bullet} \neq 0$). Já o painel (b), exibe somente as 280 linhas de F que

foram identificadas com interação significativa. Podemos observar também que o painel (b) apresenta no topo e na base padrões distintos representando os diferentes efeitos de interação.

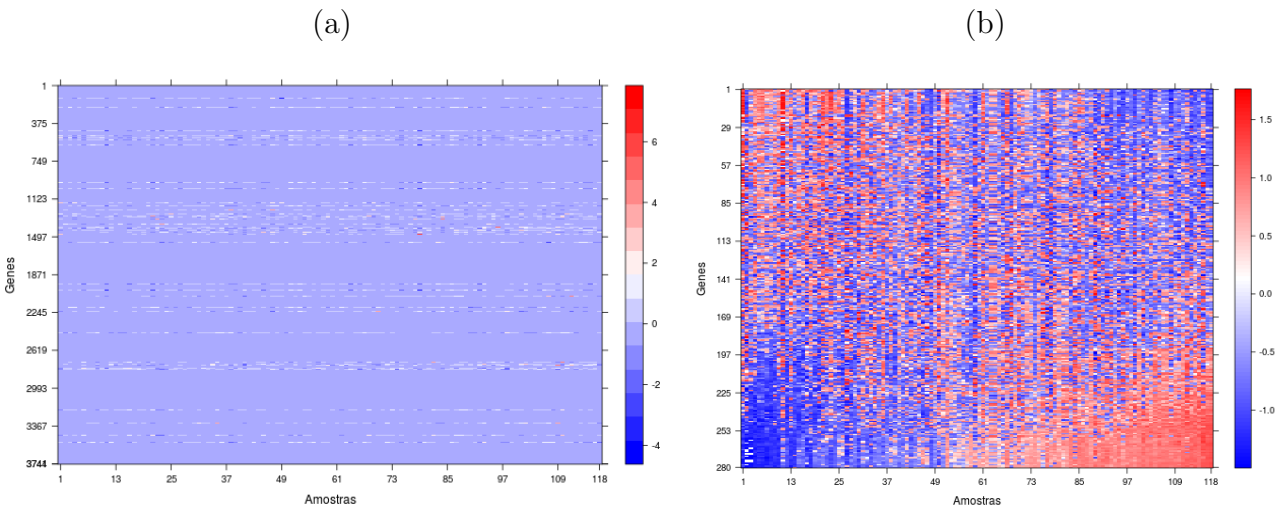

Figura 6.3: *Imagens da matriz F. O painel (a) mostra a matriz completa e o painel (b) exibe os casos onde $F_{i\bullet} \neq 0$.*

A Figura 6.4 mostra alguns gráficos de superfícies, médias *a posteriori* e intervalos HPD para alguns dos efeitos de interação. As superfícies apresentam formatos irregulares diferentes sugerindo interações distintas afetando cada gene. Nos painéis a direita, observamos as médias *a posteriori* utilizadas na construção da superfície e seus intervalos de credibilidade indicando nossa incerteza *a posteriori* relacionada a estimação.

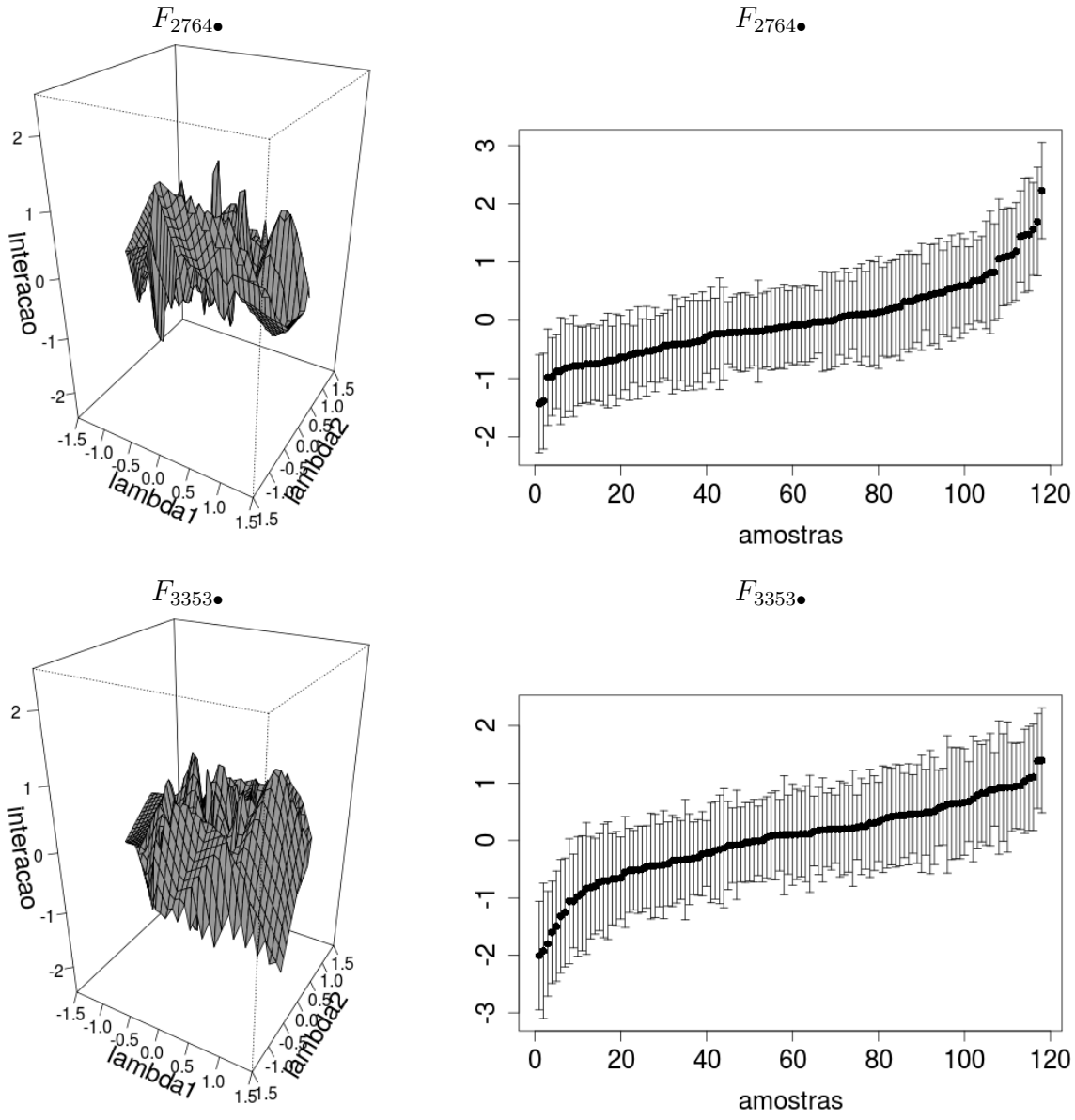

Figura 6.4: *Gráficos de superfícies e médias a posteriori (círculo) dos efeitos de interação* $F_{2764\bullet}$ *e* $F_{3353\bullet}$*. Os intervalos HPD de 95% de credibilidade são representados pelos segmentos de retas na vertical. Os painéis a direita estão organizados de forma crescente em relação a média.*

6.2 Conclusões do Capítulo

Neste capítulo fizemos uma aplicação mostrando o uso do modelo fatorial com interações utilizando a função de covariâncias Matérn com $\kappa = \frac{3}{2}$ e $l_s = 0.1$. Investigamos a existência dos efeitos de interação envolvendo os fatores latentes por meio da modelagem de misturas. Testamos a existência de interações a partir das probabilidades ρ_i^*. Notamos que número de interações significativas tem uma relação com a escolha da distribuição Beta atribuída *a priori* para ρ_i. Dependendo da informação *a priori* sobre a quantidade de interações teremos a identificação de muitas ou poucas iterações ($F_{i\bullet} \neq 0$) afetando os genes localizados fora das regiões com CNA. Distintos formatos de interações (superfícies) são obtidos para cada gene individualmente. Mostramos que o modelo fatorial utilizando a função de covariâncias Matérn identificou 280 interações. Este resultado está próximo ao encontrado por Mayrink e Lucas (2013).

Capítulo 7

Conclusões

Neste trabalho, consideramos o modelo fatorial latente esparso com interações proposto por Mayrink e Lucas (2013) em um estudo que explora outras opções de funções de covariâncias. Estudamos o modelo diante das funções exponencial potência e da classe Matérn com diferentes configurações para (κ, l_s), comparando os resultados com a exponencial quadrática (usada em Mayrink e Lucas (2013) e sendo um caso particular da função exponencial potência).

Regiões do genoma englobando grupos de genes afetados pelo problema da CNA juntamente com uma complexa associação entre genes, motivou o uso de um modelo com 2 fatores, sendo cada um deles associados a diferentes regiões com CNA. No modelo, utilizamos misturas para modelar as cargas e as interações. Além disso, as especificações *a priori* Beta para as probabilidades (q_{il} e ρ_i) que avaliam a significância de α_{il} e $F_{i\bullet}$ ajudam a resolver problemas de identificação do modelo.

Os estudos feitos nesta dissertação, explorou inicialmente dados simulados e versões mais simples da modelagem fatorial. Em todas as análises, obtivemos bons resultados de inferência que comprovam o funcionamento dos algoritmos implementados e dos modelos explorados.

As análises do EQM e dos critérios de comparação DIC, WAIC e LPML forneceram indicações diferentes sobre o melhor modelo. O EQM é um critério que faz uma avaliação para cada parâmetro individualmente, enquanto que os demais são medidas globais da qualidade do ajuste. Conforme as análises do EQM, DIC e WAIC, o modelo

fatorial usando a função de covariâncias exponencial potência com ($\kappa = 2.0, l_s = 0.3$), captura melhor os efeitos de interação e apresenta um bom ajuste, confirmando o bom desempenho do estudo em Mayrink e Lucas (2013). Pelo LPML, observamos que os valores referentes as configurações ($\kappa = 2.0, l_s = 0.3$) e ($\kappa = 1.5, l_s = 0.3$) são bem próximos apesar do LPML indicar a configuração ($\kappa = 1.5, l_s = 0.3$) como tendo o melhor ajuste. Ao avaliar o modelo com a função Matérn, a configuração $\left(\kappa = \frac{3}{2}, l_s = 0.1\right)$ forneceu os melhores resultados em termos do EQM. Entretanto, os critérios DIC, WAIC e LPML apontam para o modelo com configuração $\left(\kappa = \frac{3}{2}, l_s = 0.3\right)$. A partir das análises gráficas notamos que esta opção não se ajusta bem apesar de apresentar uma boa estimação de σ^2 (menor EQM entre os analisados).

O uso de ferramentas computacionais do C++ através do Rcpp no R, teve grande importância no desenvolvimento deste trabalho para que pudéssemos analisar o banco de dados real. Os dados de câncer de mama explorado aqui, contêm 3744 genes replicados em 118 amostras. Desta forma, ajustar o modelo fatorial com interações aqui é um desafio computacional.

Na aplicação real, utilizamos as mesmas distribuições *a priori* que Mayrink e Lucas (2013) tendo como diferença a função de covariâncias Matérn, pois ela apresentou melhores resultados nas análises dos EQM's. Os resultados da estimação de α e do número de interações significativas se mostraram bem similares aos de Mayrink e Lucas (2013). No artigo, os autores identificaram 275 genes afetados pelos efeitos de interação. Em nosso estudo identificamos 280 genes afetados quando avaliamos os casos em que ρ_i^* apresentam intervalos HPD completamente acima de 0.5 (453 genes quando avaliamos as médias *a posteriori* de ρ_i acima de 0.5).

7.1 Trabalhos futuros

Para trabalhos futuros poderíamos utilizar *a priori* uma distribuição Beta "banheira"com parâmetros menores que 1, conforme sugere Gonçalves (2006), em uma análise de sensibilidade para o modelo. Outra proposta interessante seria estimar os parâmetros da função de covariâncias atribuindo alguma distribuição *a priori* ou até mesmo construir uma função de covariâncias válida (conforme mostra Benerjee et al. (2004)) que venham a contribuir para melhorar o desempenho do modelo fatorial.

Apêndice A: Verossimilhança e condicionais completas no modelo fatorial simples.

O cálculo das distribuições *a posteriori* é feito utilizando a forma mais apropriada da função de verossimilhança que pode ser escrita de duas maneiras:

Verossimilhança 1: Primeiramente denote $X_{\bullet j}$ como a j-ésima coluna da matriz de dados X, então temos $(X_{\bullet j}|\alpha, \lambda, \sigma^2) \sim N_m(\alpha\lambda_{\bullet j}, D)$, onde $D = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_m^2)$. Assumindo independência condicional entre as colunas de X pode-se escrever:

$$\begin{aligned} p(X \mid \alpha, \lambda, \sigma^2) &= \prod_{j=1}^{n} p(X_{\bullet j} \mid \alpha, \lambda, \sigma^2) \\ &= \prod_{j=1}^{n} (2\pi)^{-\frac{m}{2}} \mid D \mid^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \left[X_{\bullet j} - \alpha\lambda_{\bullet j}\right]' D^{-1} \left[X_{\bullet j} - \alpha\lambda_{\bullet j}\right] \right\}. \end{aligned}$$

Verossimilhança 2: Agora denote por $X_{i\bullet}$ a i-ésima linha da matriz X, então $(X'_{i\bullet}|\alpha, \lambda, F, \sigma^2) \sim N_n(\lambda'\alpha'_{i\bullet}, \sigma_i^2 I_n)$ e assumindo independência condicional entre as linhas de X temos:

$$\begin{aligned} p(X \mid \alpha, \lambda, \sigma^2) &= \prod_{i=1}^{m} p(X_{i\bullet} \mid \alpha, \lambda, \sigma^2) \\ &= \prod_{i=1}^{m} (2\pi)^{-\frac{n}{2}} \mid \sigma_i^2 I_n \mid^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \left[X'_{i\bullet} - \lambda'\alpha'_{i\bullet}\right]' (\sigma_i^2 I_n)^{-1} \left[X'_{i\bullet} - \lambda'\alpha'_{i\bullet}\right] \right\}. \end{aligned}$$

As funções de verossimilhanças descritas acima são equivalentes e serão utilizadas, conforme conveniência, no cálculo das distribuições *a posteriori* condicionais completas de cada parâmetro. Considere as seguintes notações: $\alpha_{-\{i\bullet\}}$ é o conjunto de elementos da matriz α com exceção da linha $\alpha_{i\bullet}$; $\lambda_{-\{\bullet j\}}$ é o conjunto dos elementos da matriz λ com exceção da coluna $\lambda_{\bullet j}$; e $\sigma^2_{-i} = \left\{\sigma_1^2, \sigma_2^2, ..., \sigma_{i-1}^2, \sigma_{i+1}^2, ..., \sigma_m^2\right\}$ o vetor de variâncias sem a i-ésima componente.

Utilizando a regra de Bayes e a verossimilhança 2 a distribuição *a posteriori* de $\alpha_{i\bullet}$ será:

$$p(\alpha'_{i\bullet} \mid \alpha_{-\{i\bullet\}}, \lambda, \sigma^2, X) \propto \left[\prod_{i=1}^{m} p(X \mid \alpha, \lambda, \sigma^2)\right] p(\alpha'_{i\bullet})$$
$$\propto \exp\left\{-\frac{1}{2}\left[\alpha_{i\bullet}\left(\frac{\lambda\lambda'}{\sigma^2} + V_\alpha^{-1}\right)\alpha'_{i\bullet} - 2\alpha_{i\bullet}\left(\frac{\lambda X'_{i\bullet}}{\sigma_i^2} + V_\alpha^{-1}M_\alpha\right)\right]\right\}.$$

Este é o núcleo da distribuição $N_L(M_\alpha^*, V_\alpha^*)$ com $V_\alpha^* = \left(\frac{\lambda\lambda'}{\sigma^2} + V_\alpha^{-1}\right)^{-1}$ e $M_\alpha^* = V_\alpha^*\left(\frac{\lambda X'_{i\bullet}}{\sigma_i^2} + V_\alpha^{-1}M_\alpha\right)$.

A distribuição *a posteriori* condicional completa de $\lambda_{\bullet j}$, é obtida utilizando a função de verossimilhança 1, então via regra de Bayes, temos:

$$p(\lambda_{\bullet j} \mid \alpha, \lambda_{-\{\bullet j\}}, \sigma^2, X) \propto \left[\prod_{j=1}^{n} p(X \mid \alpha, \lambda, \sigma^2)\right] p(\lambda_{\bullet j})$$
$$\propto \exp\left\{-\frac{1}{2}\left[\lambda'_{\bullet j}\left(\alpha' D^{-1}\alpha + I_L\right)\lambda_{\bullet j} - 2\lambda_{\bullet j}\alpha' D^{-1}X_{\bullet j}\right]\right\},$$

que é o núcleo de uma $N_L(M_\lambda^*, V_\lambda^*)$ com $V_\lambda^* = \left(\alpha' D^{-1}\alpha + I_L\right)^{-1}$ e $M_\lambda = V_\lambda^*\left(\alpha' D^{-1}X_{\cdot j}\right)$.

Utilizando a verossimilhança 1 chega-se a distribuição condicional completa *a posteriori* de σ_i^2:

$$p(\sigma_i^2 \mid \alpha, \lambda, \sigma_{-i}^2, X) \propto \left[\prod_{i=1}^{m} p(X \mid \alpha, \lambda, \sigma^2)\right] p(\sigma_i^2)$$
$$\propto (\sigma_i^2)^{-\frac{n}{2}-a-1} \exp\left\{-\frac{1}{\sigma_i^2}\left[b + \frac{1}{2}\left(X_{i\bullet}X'_{i\bullet} - 2\alpha_{i\bullet}\lambda X'_{i\bullet} + \alpha_{i\bullet}\lambda\lambda'\alpha'_{i\bullet}\right)\right]\right\}.$$

Podemos observar que esta é a distribuição $GI(a^*, b^*)$, com $a^* = a + \frac{n}{2}$ e $b^* = b + \frac{1}{2}\left(X_i X'_{i\bullet} - 2\alpha_{i\bullet}\lambda X'_{i\bullet} + \alpha_{i\bullet}\lambda\lambda'\alpha'_{i\bullet}\right)$.

Apêndice B: Verossimilhança e condicionais completas no modelo fatorial com interações.

Assumiremos que as observações X_{ij} são condicionalmente independentes dado os parâmetros. Novamente, a função de verossimilhança será escrita de duas maneiras para facilitar as contas. Para isso considere que $F_{i\bullet}$ e $F_{\bullet j}$ são vetores que representam a i-ésima linha e a j-ésima coluna de F, respectivamente. As funções de verossimilhança são como segue:

Verossimilhança 1: $(X_{\bullet j}|\alpha,\lambda,F,\sigma^2) \sim N_m(\alpha\lambda_{\bullet j}+F_{\bullet j}, D)$, sendo $D = \text{diag}(\sigma_1^2,\sigma_2^2,...,\sigma_m^2)$.

$$p(X \mid \alpha,\lambda,F,\sigma^2) = \prod_{j=1}^{n} p(X_{\bullet j} \mid \alpha,\lambda,F,\sigma^2).$$

Verossimilhança 2: $(X'_{i\bullet}|\alpha,\lambda,F,\sigma^2) \sim N_n(\lambda'\alpha'_{i\bullet} + F'_{i\bullet}, \sigma^2 I_n)$ e

$$p(X \mid \alpha,\lambda,F,\sigma^2) = \prod_{i=1}^{m} p(X_{i\bullet} \mid \alpha,\lambda,F,\sigma^2).$$

A função de verossimilhança mais apropriada será utilizada para calcular as distribuições *a posteriori* condicionais completas. A partir da regra de Bayes obtemos os seguintes resultados:

- $(\sigma^2 \mid \alpha,\lambda,F,\sigma^2_{-i},X) \sim GI(A,B)$, sendo $A = a + \frac{n}{2}$ e

$B = \frac{1}{2}\left[X_{i\bullet}X'_{i\bullet} - 2\alpha_{i\bullet}\lambda(X'_{i\bullet} - F'_{i\bullet}) - 2F_{i\bullet}X'_{i\bullet} + F_{i\bullet}F'_{i\bullet} + \alpha_{i\bullet}\lambda\lambda\alpha'_{i\bullet}\right] + b.$

- Se $h_{il} = 0$, a distribuição *a posteriori* condicional completa de α_{il} será $\delta_0(\alpha)$.
- Se $h_{il} = 1$, a condicional completa de α_{il} será $N(M_\alpha, V_\alpha)$ com $V_\alpha = \left[\frac{1}{w} + \frac{1}{\sigma_i^2}\sum_{j=1}^{n}\lambda_{lj}^2\right]^{-1}$ e $M_\alpha = V_\alpha\left[\frac{1}{\sigma_i^2}\sum_{j=1}^{n}\lambda_{lj}\left(X_{ij} - F_{ij} - \sum_{l^*\neq l}\alpha_{il^*}\lambda_{l^*j}\right)\right]$.
- Para avaliar a significância das cargas α_{il}, calculamos a seguinte probabilidade:

$$q^*_{il} = p(h_{il} = 1 \mid \alpha,\lambda,F,\sigma^2,q_{il},X) = \frac{q_{il}}{q_{il} + (1-q_{il})\dfrac{N(0|M_\alpha,V_\alpha)}{N(0|0,\omega)}}; \quad \text{e}$$

$$(q_{il} \mid h_{il}) \sim \text{Beta}(\gamma_1 + h_{il}, \gamma_2 + 1 - h_{il}).$$

- Se $z_i = 0$, a distribuição *a posteriori* condicional completa de $F'_{i\bullet}$ será $\delta_0(F_{i\bullet})$.
- Se $z_i = 1$, a condicional completa de $F'_{i\bullet}$ será a $N_n(M_F, V_F)$ com $V_F = \left[\frac{1}{\sigma_i^2}I_n + K(\lambda)^{-1}\right]^{-1}$ e $M_F = V_F\left[\frac{1}{\sigma_i^2}\left(X'_{i\bullet} - \lambda'\alpha'_{i\bullet}\right)\right]$.

- Para avaliar a significância de $F_{i\bullet}$, calculamos a seguinte probabilidade:

$$\rho_i^* = p(z_i = 1 \mid \alpha, \lambda, F, \sigma^2, \rho_i, X) = \frac{\rho_i}{\rho_i + (1-\rho_i)\dfrac{N(0|M_F, V_F)}{N(0|0, K(\lambda))}}; \text{ e}$$

$$(\rho_i \mid z_i) \sim \text{Beta}(\beta_1 + z_i, \beta_2 + 1 - z_i).$$

- Para $\lambda_{\bullet j}$ temos a condicional completa:

$$p(\lambda_{\bullet j} \mid \alpha, \lambda_{-\{\bullet j\}}, F, \sigma_i^2, X) \propto p(X \mid \alpha, \lambda, F, \sigma^2)\ p(F \mid \lambda, z)\ p(\lambda_{\bullet j})$$
$$\propto N_L(\lambda_{\bullet j}|M_\lambda, V_\lambda)\,|K(\lambda)|^{-\sum_{i=1}^m \frac{z_i}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^m F_{i\bullet}K(\lambda)^{-1}F_{i\bullet}'\right\},$$

sendo $V_\lambda = [\alpha' D^{-1}\alpha + I_L]^{-1}$ e $M_\lambda = V_\lambda\,[\alpha' D^{-1}(X_{\bullet j} - F_{\bullet j})]$.

Este núcleo não permite reconhecer uma distribuição de probabilidade, então será necessário um método para amostragem indireta desta condicional completa. Consideramos o algoritmo Metropolis-Hastings com passeio aleatório para gerar candidatos.

Apêndice C: Gráficos extras dos estudos simulados.

A Figura C.1 apresenta gráficos do teste envolvido em nosso estudo simulado. Os valores da estatística Z (símbolo "$\times$") são investigados para as cadeias de α_{21}, λ_{11}, σ_3^2 e F_{171}. Pode-se notar que a maioria dos pontos estão dentro dos limites de 95% indicando a convergência.

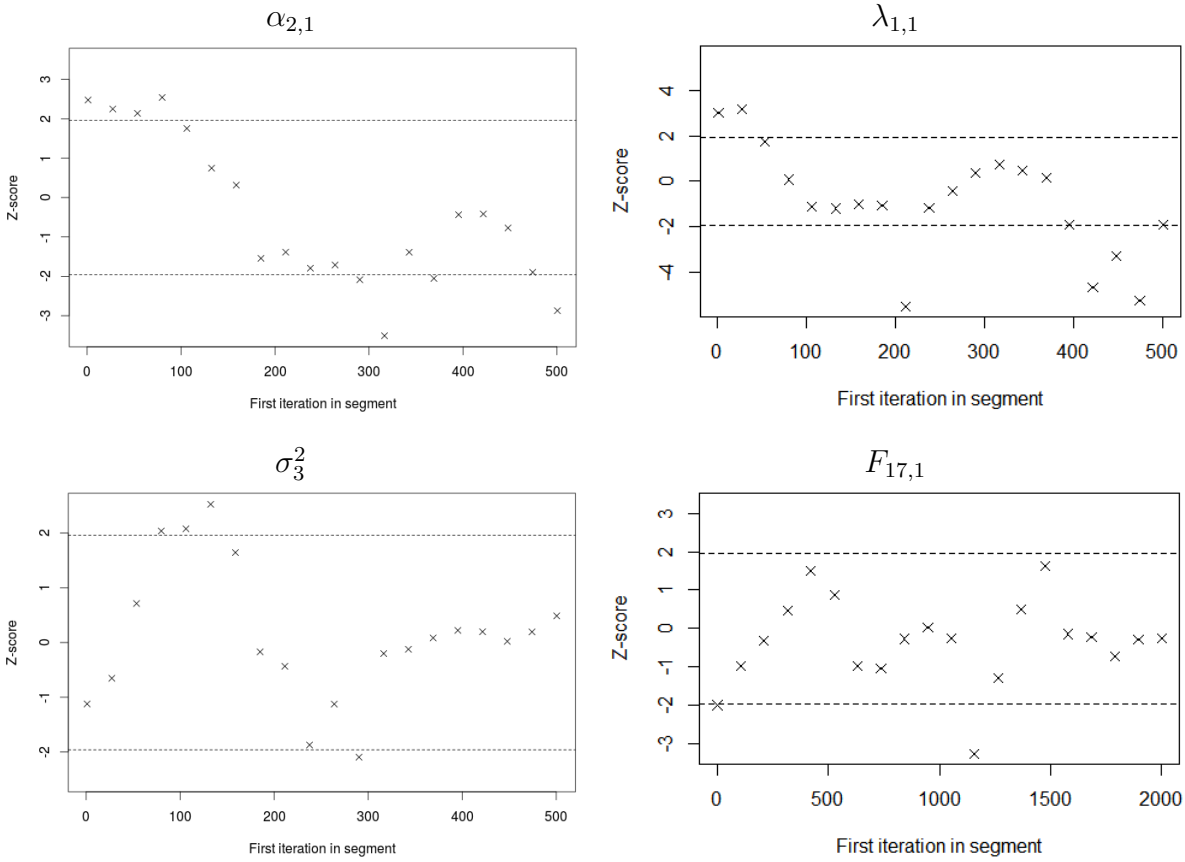

Figura C.1: *Gráfico com os valores da estatística de Geweke e algumas cadeias de* α, σ^2 *e* λ. *O símbolo "*$\times$*"representa a estatística* Z, *as linhas tracejadas demarcam os valores (-1.96,1.96) correspondendo a região de probabilidade 0.95 na* $N(0,1)$.

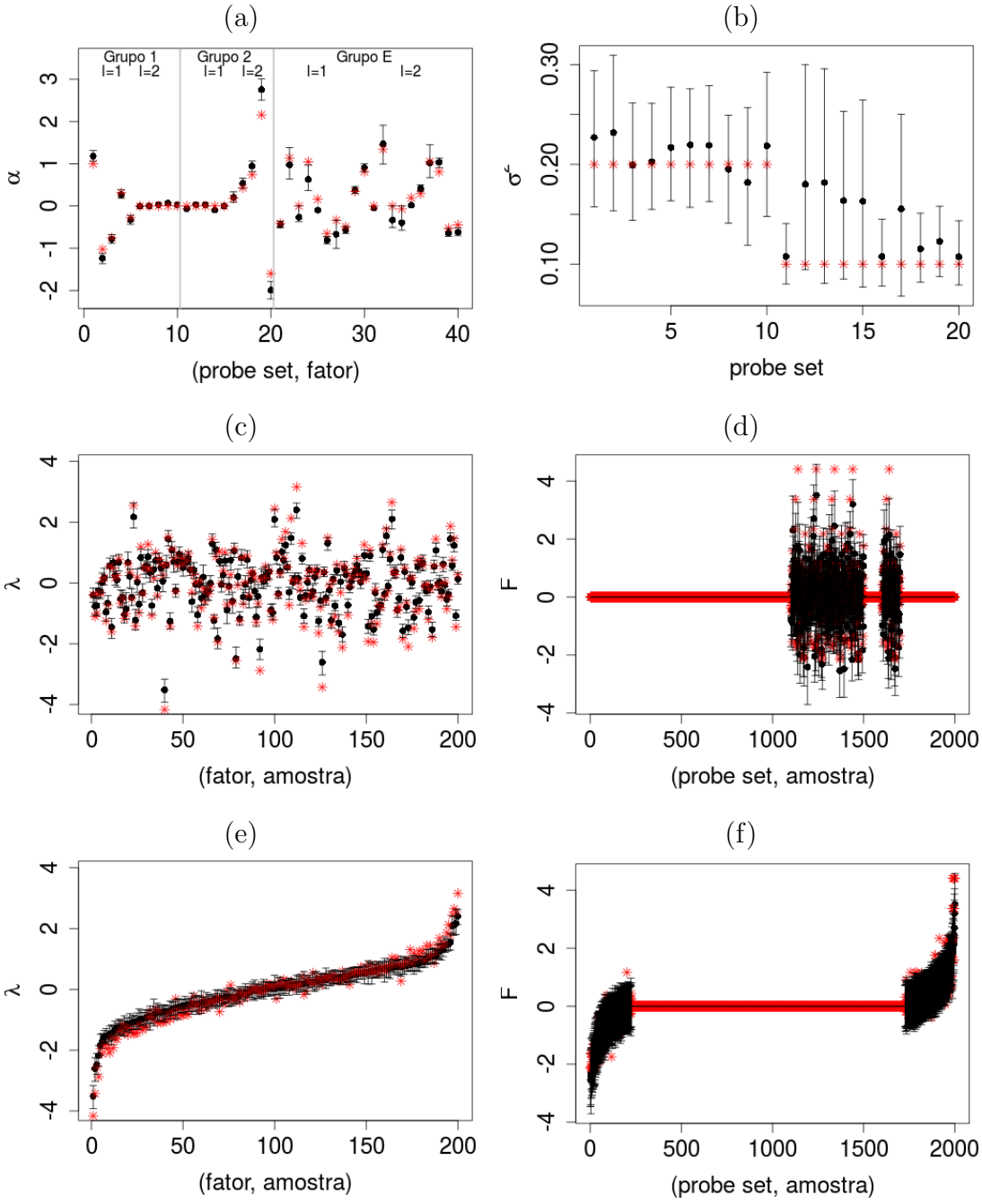

Figura C.2: *Valores reais (asterisco) e médias a posteriori (círculo). O intervalo HPD de 95% de credibilidade é representado pelo segmento de reta na vertical. Os dois últimos painéis exibem os intervalos ordenados de acordo com a média a posteriori. Cenário considerado* $\kappa = 0.5$ *e* $l_s = 0.2$ *na função exponencial potência.*

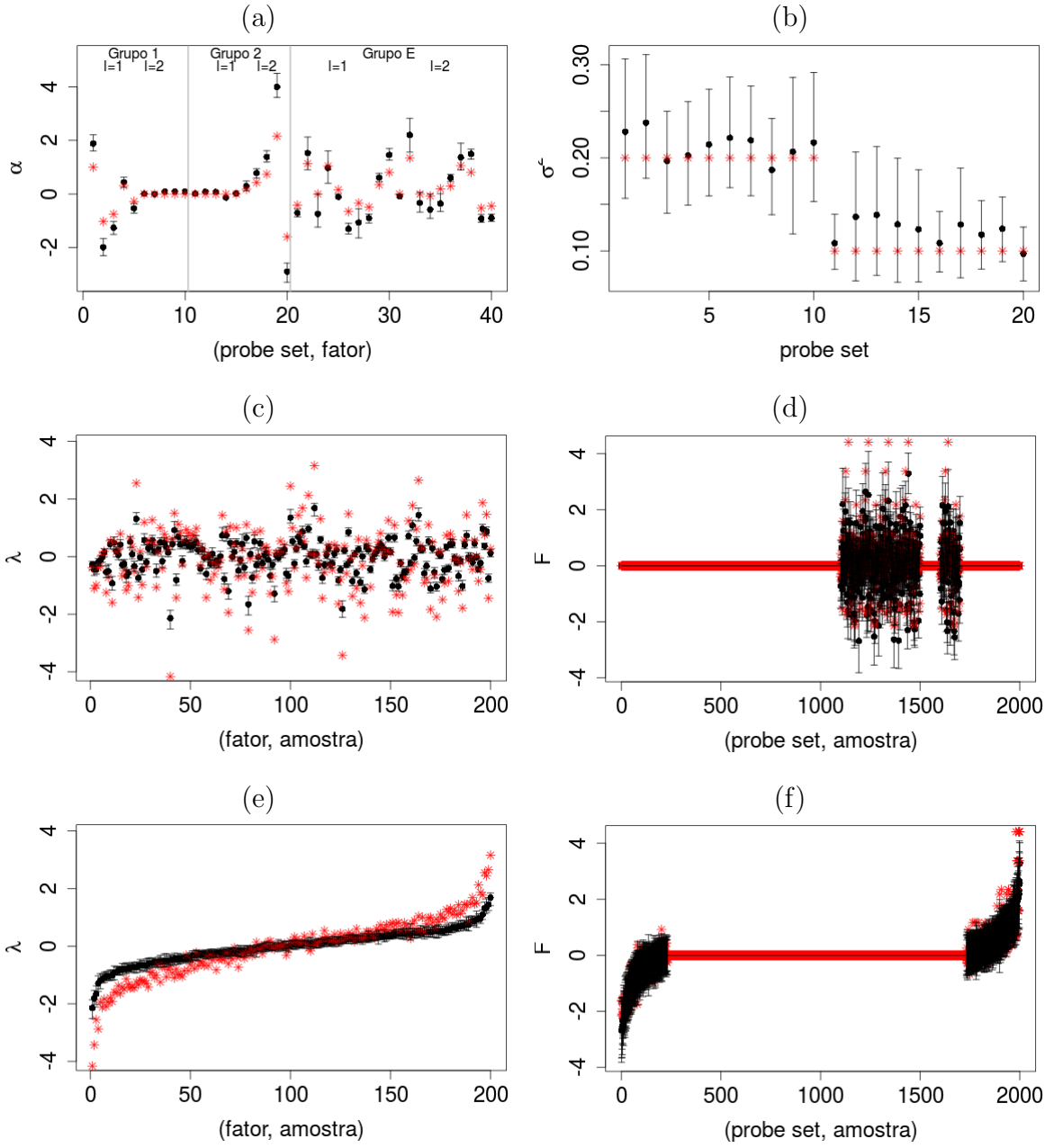

Figura C.3: *Valores reais (asterisco) e médias a posteriori (círculo). O intervalo HPD de 95% de credibilidade é representado pelo segmento de reta na vertical. Os dois últimos painéis exibem os intervalos ordenados de acordo com a média a posteriori. Cenário considerando* $\kappa = 1.5$ *e* $l_s = 0.3$ *na função exponencial potência. Este foi o pior cenário na análise dos EQM's.*

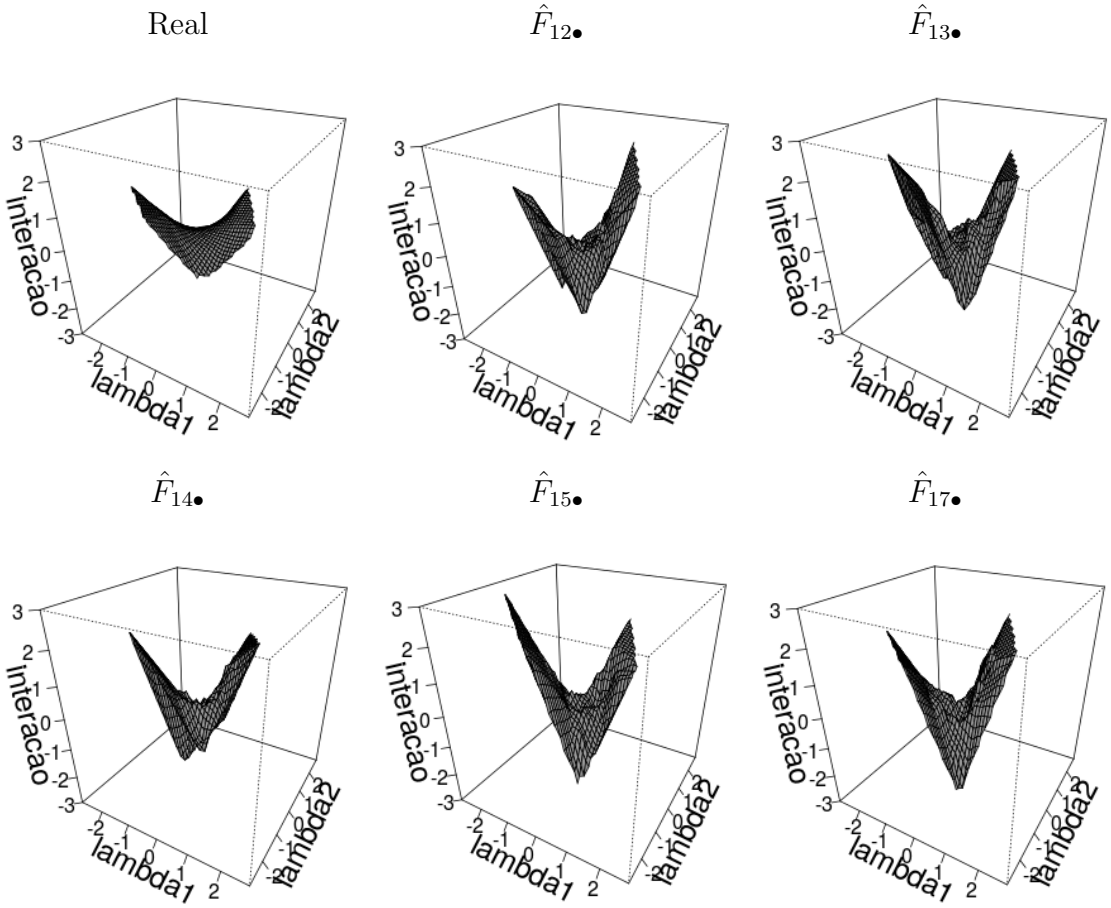

Figura C.4: *Gráfico de superfície representando o efeito de interação real e estimado considerando os parâmetro* $\kappa = 1.5$ *e* $l_s = 0.3$ *na função de covariâncias exponencial potência. Este é o pior cenário na análise dos EQM's.*

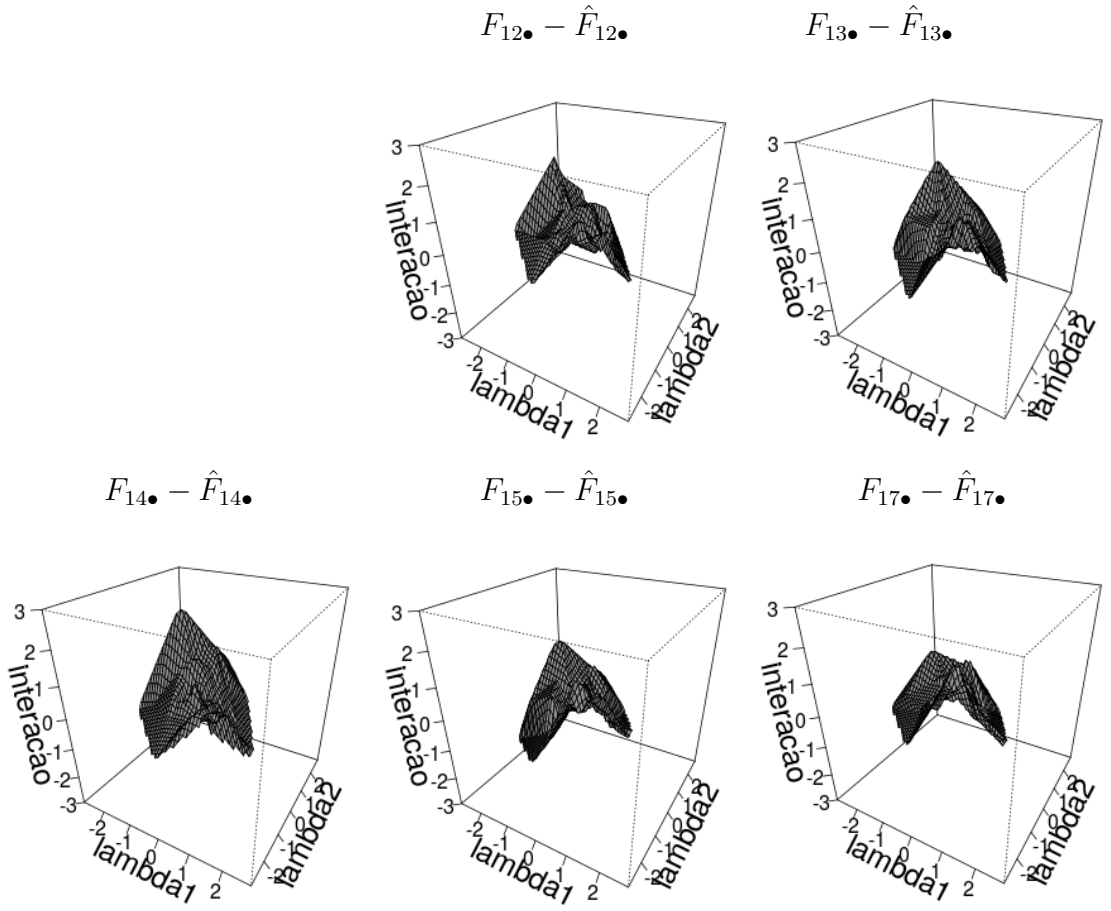

Figura C.5: *Gráfico de superfície da diferença entre o efeito de interação real e estimado. Considerando* $\kappa = 1.5$ *e* $l_s = 0.3$ *na função exponencial potência. Este é o pior cenário na análise dos EQM's.*

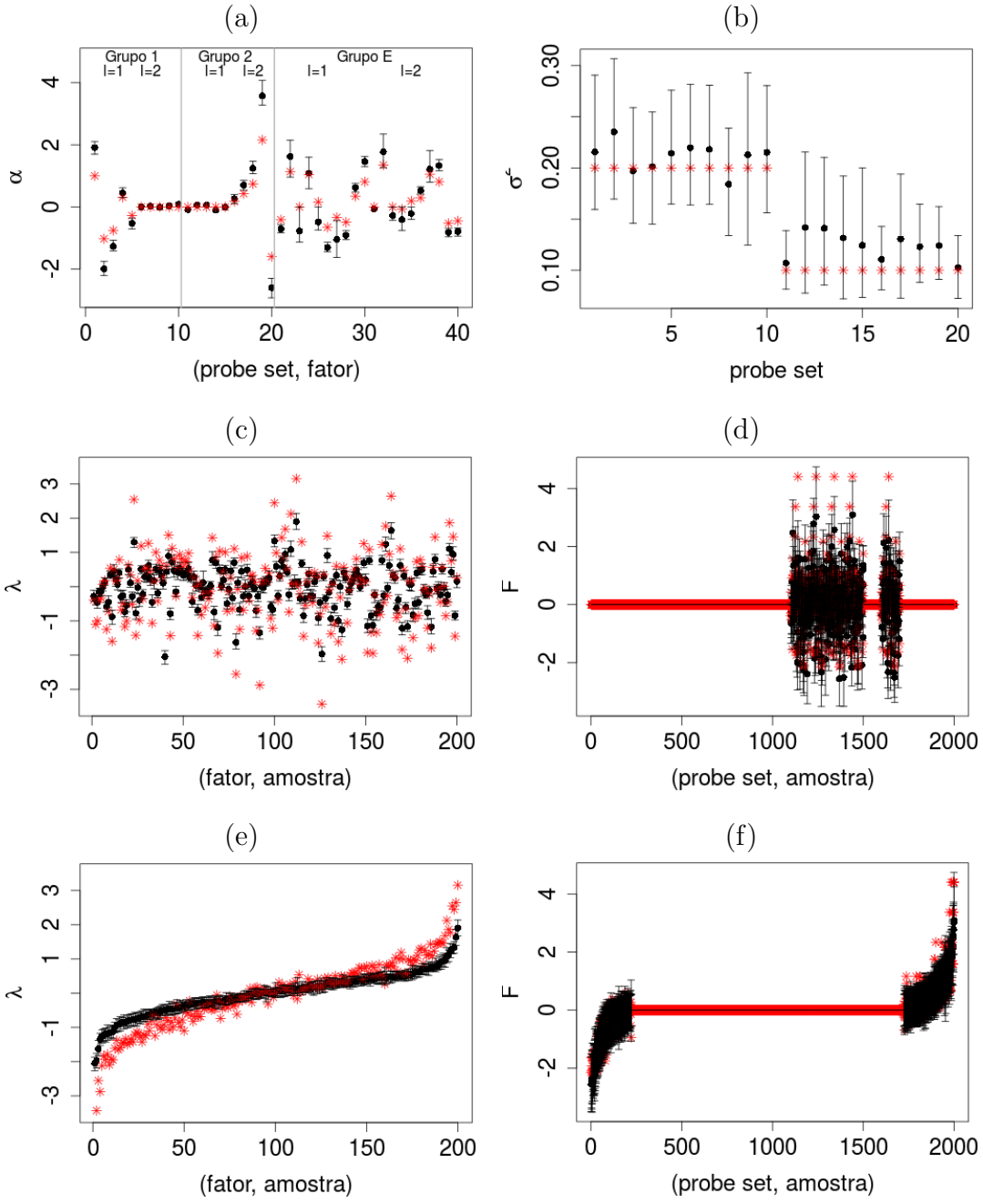

Figura C.6: *Valores reais (asterisco) e médias a posteriori (círculo). O intervalo HPD de 95% de credibilidade é representado pelo segmento de reta na vertical. Os dois últimos painéis exibem os intervalos ordenados de acordo com a média a posteriori. Cenário considerado* $\kappa = \frac{3}{2}$ *e* $l_s = 0.3$ *na função de covariâncias Matérn.*

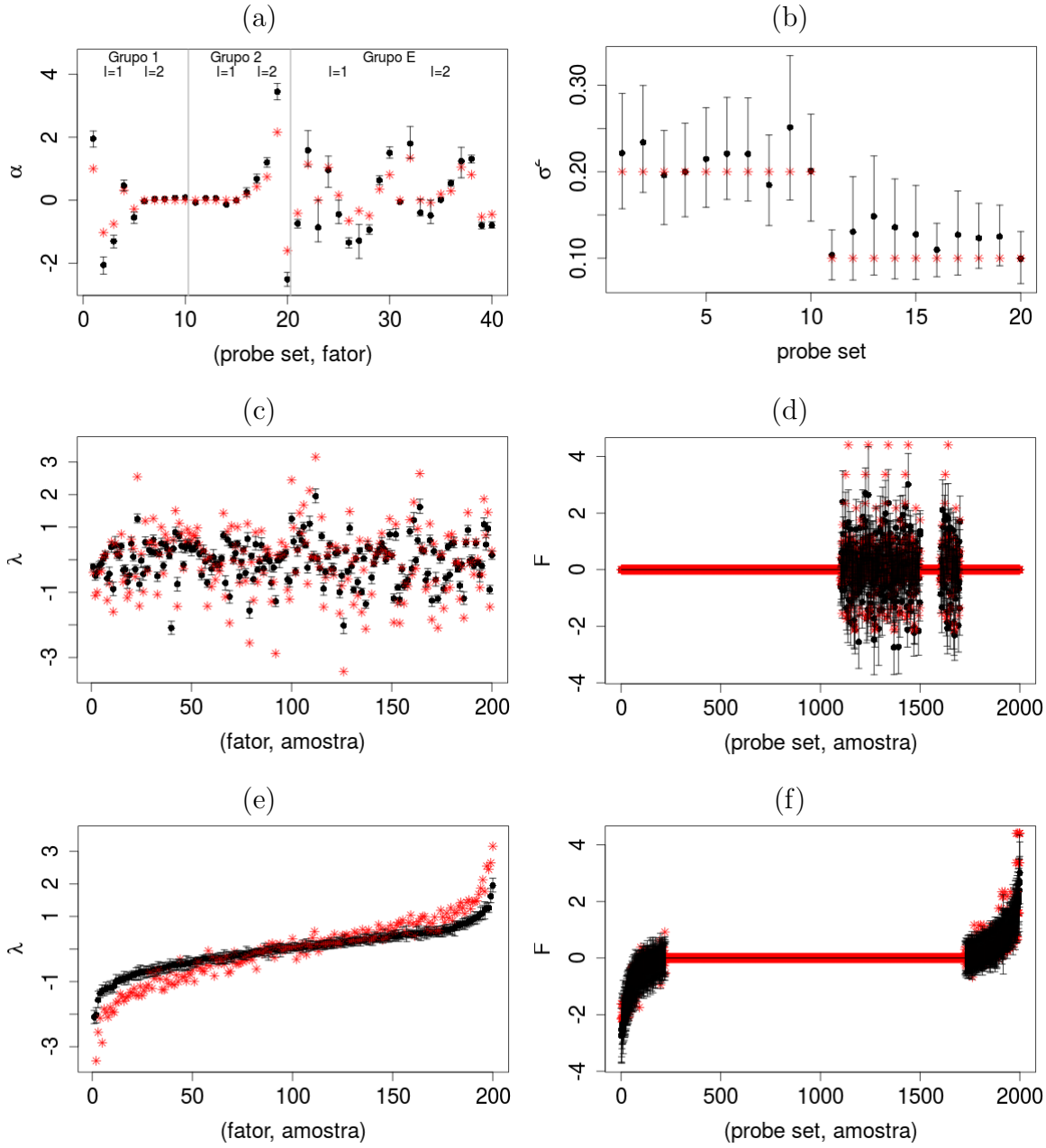

Figura C.7: *Valores reais (asterisco) e médias a posteriori (círculo). O intervalo HPD de 95% de credibilidade é representado pelo segmento de reta na vertical. Os dois últimos painéis exibem os intervalos ordenados de acordo com a média a posteriori. Cenário considerado* $\kappa = \frac{5}{2}$ *e* $l_s = 0.3$ *na função de covariâncias Matérn.*

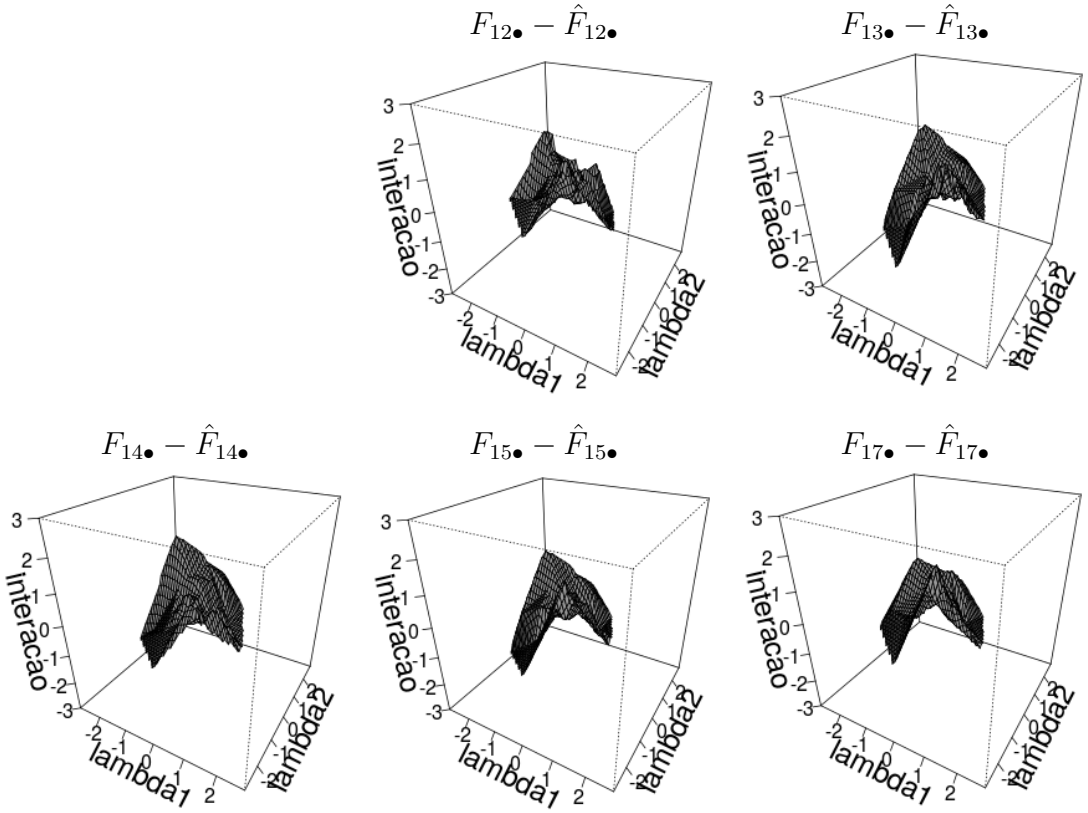

Figura C.8: *Gráfico de superfície da diferença entre o efeito de interação real e estimado. Considerando* $\kappa = \dfrac{5}{2}$ *e* $l_s = 0.3$ *na função de covariâncias Matérn. Este é o pior cenário na análise dos EQM's.*

Apêndice D: Gráficos da aplicação real.

As taxas de aceitação dos λ apresentados ficam em torno de 21% a 35%.

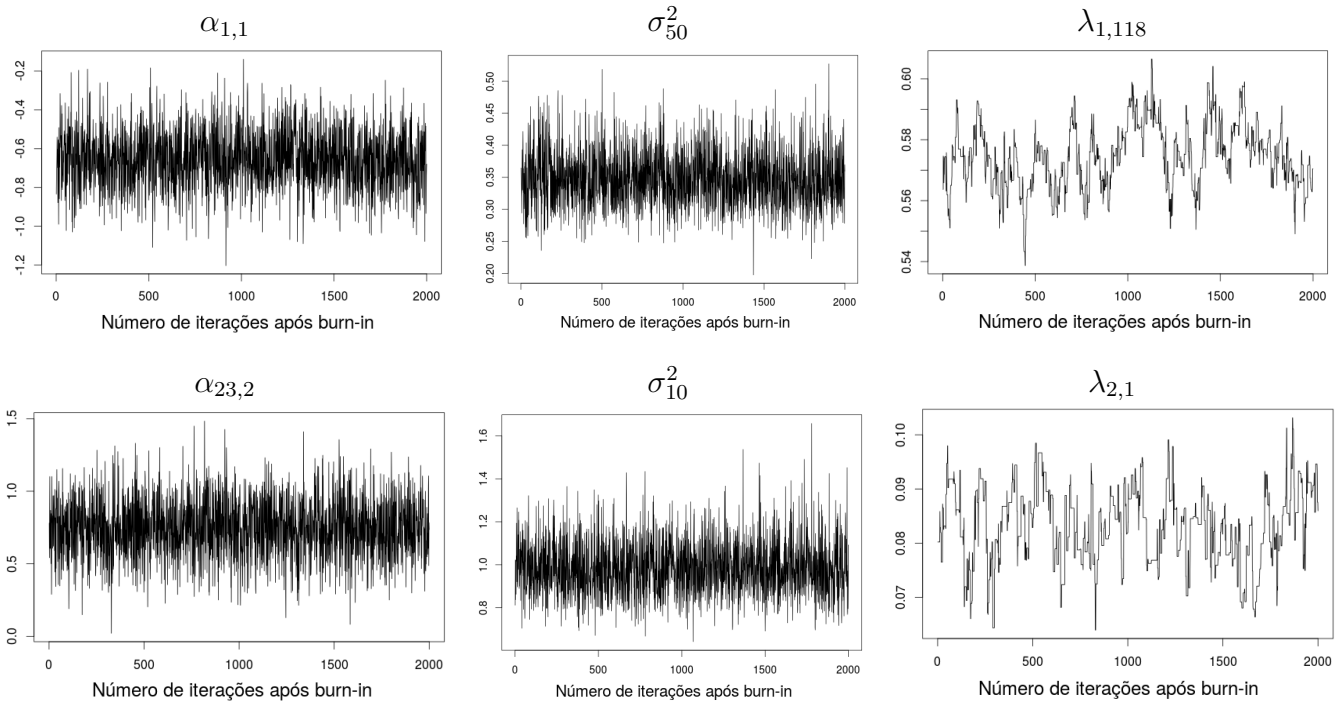

Figura D.1: *Algumas cadeias de* α, σ^2 *e* λ.

A Figura D.2 apresenta gráficos do teste envolvido em nossa aplicação real. Os valores da estatística Z (símbolo "$\times$") são investigados para algumas cadeias de α, σ^2 e λ. Pode-se notar que a maioria dos pontos estão dentro dos limites de 95% indicando a convergência.

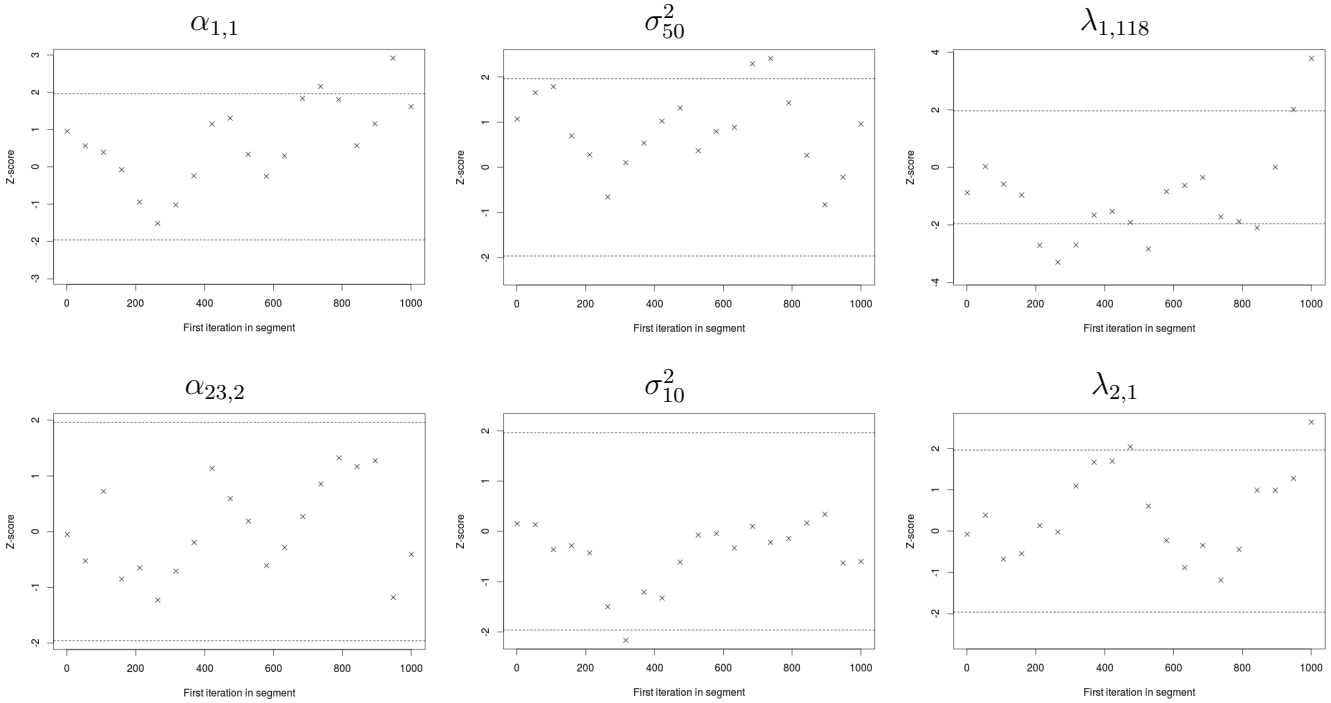

Figura D.2: *Gráfico com os valores da estatística de Geweke e algumas cadeias de α, σ^2 e λ. O símbolo "$\times$"representa a estatística Z, as linhas tracejadas demarcam os valores (-1.96,1.96) correspondendo a região de probabilidade 0.95 na $N(0,1)$.*

Referências Bibliográficas

Abramowitz, M. e Stegun, I. A. (1965), *Handbook of Mathematical Functions*, Dover, New York.

Affymetrix (2001), *Statistical algorithms reference guide*, Affymetrix Technical Report.

Benerjee, S., Carlin, B. P., e Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall/CRC.

Carvalho, M. C., Chang, J., Lucas, J. E., Wang, J. R. N. Q., e West, M. (2008), "High-dimensional sparse factor modelling: Applications in gene expression genomics," *Journal of the American Statistical Association*, 103, 1438–1456, MR2655722.

Chin, K., DeVriers, S., J, Fridlyand, Spellman, P. T., roydasgupta, R., Kuo, W. L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., Chen, F., Feiler, H., Tokuyasu, T., Esserman, L., Albertson, D. G., Waldman, F. M., e Gray, J. W. (2006), "Genomic and transcriptional aberrations linked to breast cancer pathophysiologies," *Cancer Cell*, 10, 529–541.

Eddelbuettel, D. (2013), *Seamless R and C++ Interations with Rcpp*, Springer, ISBN 978-1-4614-6867-7.

Eddelbuettel, D. e Francois, R. (2011), "Rcpp: Seamless R and C++ Integration," *Journal of Statistical Software*, 40, 1–18.

Eddelbuettel, D. e Sanderson, C. (2014), "RcppArmadRcpp: Accelerating R with high-performance C++ linear algebra," *Computational Statistics and Data Analysis*, 71, 1054–1063.

Gamerman, D. e Lopes, H. F. (2006), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, vol. 68, Chapman and Hall/CRC, London, 2 edn.

Gelfand, A. E. e Smith, A. F. M. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A. e Rubin, D. B. (1992), "Inference from iterative simulation using multiple sequences," *Statistical Science*, 7, 457–472.

Gelman, A., Carlin, J. B., Stern, H. S., e Rubin, D. B. (2003), *Bayesian Data Analysis*, Chapman and Hall/CRC, second edn.

Geman, S. e Geman, D. (1984), "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of imagens," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Geweke, J. (1992), "Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments (with discussion)." *In Bayesian Statistics 4, J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds.)*, pp. 169–193, Oxford University Press, Oxford.

Gilks, W. R. e Wild, P. (1992), "Adaptative rejection sampling for Gibbs sampling," *Applied Statistics*, 41, 337–348.

Gonçalves, F. B. (2006), "Análise Bayesiana da Teoria da Resposta ao Item: Uma Abordagem Generalizada," Dissertação de Mestrado, IM-UFRJ.

Hastings, W. K. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, 57, 97–109.

Henao, R. e Winther, O. (2011), "Sparse linear identifiable multivariate modeling," *Journal of Machine Learning Research*, 12, 663–705.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., e Speed, T. P. (2003), "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, 4, 249–264.

Lawrence, N. D. (2004), "Gaussian process models for visualisation of high dimensional data," *Advances in Neural Information Processing Systems*, 16, 329–336, eds. Thrun, S., Saul, L. and Scholkopf B., Cambridge, MA, MIT Press.

Lawrence, N. D. (2005), "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *Journal of Machine Learning Research*, 6, 1783–1816.

Lucas, J. E., Carvalho, C., Wang, Q., Bild, A., Nevins, J. R., e West, M. (2006), "Sparse statistical modelling in gene expression genomics," *In Bayesian Inference for Gene Expression and Proteomics (P. Muller, K. Do and M. Vannucci, eds.)*, Cambridge University Press.

Lucas, J. E., Kung, H. N., e Chin, J. T. (2010), "Cross-Study Projections of genomics biomarkes: an evaluation in cancer genomics," *PLoS Computational biology*, 6, e1000920.

Mayrink, V. D. e Lucas, J. E. (2013), "Sparse Latent Factor Model with Interations: Analysis of Gene Expression," *The Annals of Applied Statistics*, 7, 799–822.

Mayrink, V. D. e Lucas, J. E. (2015), "Bayesian factor model for the detection of coherent patterns in gene expression data," *Brazilian Journal of Probability and Statistics*, 29, 1–33.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., e Teller, E. (1953), "Equations of state calculations by fast computing machines," *Journal of Chemical Physics*, 21, 1087–1092.

Miller, D. L., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T., e Bergh, J. (2005), "An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival," *PNAS - Proceedings of the National Academy of Science of the United States of America*, 112, 13550–13555.

Neal, R. M. (2003), "Slice sampling (with discussion)," *Annals of Statistics*, 31, 705–767.

Plummer, M., Best, N., Cowles, K., e Vines, K. (2006), "CODA: Convergence Diagnosis and Output Analysis for MCMC," *R News*, 6, 7–11.

Pollack, J. R., Sorlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Botstein, R. T. D., Dale, A. L. B., e Brown, P. O. (2002), "Microarrays analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors," *Proceedings of the National Academy of Sciences of the United States of America*, 99, 12963–12968.

R Development Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rueda, O. M. e Uriarte, R. D. (2007), "Flexible and accurate detection of genomic copy number changes from aCGH," *PLoS Computational Biology*, 3, e122.

Spiegelhalter, D. J., Best, N. G., e van der Linde, B. P. C. A. (2002), "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society*, pp. 583–639, Série B.

Watanabe, S. (2010), "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory," *Journal of Machine Learning Research*, 11, 3571–3594.

West, M. (2003), "Bayesian factor regression models in the large p, small n paradigm," *Bayesian Statistics*, 7, 723–732, eds. Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A. and West, M., Oxford University Press.

Wu, Z., Irizarry, R. A., Gentleman, R., Murillo, F. M., e Spencer, F. (2004), "A model based background adjustment for oligonucleotide expression arrays," *Journal of the American Statistical Association*, 99, 909–917.