Agrupamento de interações não lineares em análise fatorial

Erick da Conceição Amorim

Departamento de Estatística - ICEx - UFMG

Fevereiro de 2020

Agrupamento de interações não lineares em análise fatorial

Erick da Conceição Amorim

Orientador: Vinícius Diniz Mayrink

Tese submetida para avaliação da banca de doutorado no Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como parte dos requisitos necessários à obtenção do Título de Doutor em Estatística.

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

Belo Horizonte, MG - Brasil
Fevereiro de 2020

Ficha catalográfica elaborada pelo bibliotecário Célio Resende Diniz - CRB 6ª Região nº 2.403.

Amorim, Erick da Conceição.

A524a Agrupamento de interações não lineares em análise fatorial / Erick da Conceição Amorim — Belo Horizonte, 2020.
x, 120 f.: il.; 29 cm.

Tese (doutorado) - Universidade Federal de Minas Gerais – Departamento de Estatística.

Orientador: Vinícius Diniz Mayrink.

1. Estatística. – Teses. 2. Análise multivariada – Teses. 3. Análise fatorial – Teses. 4. Mamas - Câncer. – Teses. I. Orientador. II. Título.

CDU 519.2(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

UFMG

ATA DA DEFESA DE TESE DO ALUNO
ERICK DA CONCEIÇÃO AMORIM

Realizou-se, no dia 19 de fevereiro de 2020, às 14:00 horas, 2040 ICEx, da Universidade Federal de Minas Gerais, a 60ª defesa de tese, intitulada *Agrupamento de interações não lineares em análise fatorial*, apresentada por ERICK DA CONCEIÇÃO AMORIM, número de registro 2016675360, graduado no curso de ESTATÍSITCA, como requisito parcial para a obtenção do grau de Doutor em ESTATÍSTICA, à seguinte Comissão Examinadora: Prof(a). Vinícius Diniz Mayrink - Orientador (DEST/UFMG), Prof(a). Rosangela Helena Loschi (DEST/UFMG), Prof(a). Flávio Bambirra Gonçalves (DEST/UFMG), Prof(a). Rafael Izbicki (UFSCAR), Prof(a). Florencia Graciela Leonardi (USP).

A Comissão considerou a tese:

(X) Aprovada

( ) Reprovada

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.
Belo Horizonte, 19 de fevereiro de 2020.

Prof(a). Vinícius Diniz Mayrink (Doutor)

Prof(a). Rosangela Helena Loschi (Doutora)

Prof(a). Flávio Bambirra Gonçalves (Doutor)

Prof(a). Rafael Izbicki (Doutor)

Prof(a). Florencia Graciela Leonardi (Doutora)

Às próximas gerações.

"Our world, our life, our destiny, are dominated by Uncertainty; this is perhaps the only statement we may assert without uncertainty".

de Finetti

Agradecimentos

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

Aos professores da UFPA Marinalva e Héliton Tavares pelo 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 meus Amigos Arthur Lopes e Cássio Iago. E todos os meus colegas que fiz em especial: Guilherme Lopes, Leonardo Brandão, Ana Cláudia (Baldini), Adriana Lima, Douglas Mateus, Patricia Viana (Paty Maionese), Rumenick Pereira, Renato Panaro, Marina Amorim, Frederico Machado e Isabela Severino. Agradeço a eles pelos melhores momentos de minha vida durante estes anos em Belo Horizonte.

E por fim, agradeço a Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio durante o doutorado através de uma bolsa de estudos.

Resumo

A análise fatorial é uma ferramenta poderosa para a redução da dimensão nos estudos de estatística multivariada. Esta tese é dedicada a estender o modelo fatorial com interações não lineares proposto em 2013. A principal contribuição do trabalho é apresentar duas abordagens para agrupar as interações não lineares, e assim desenvolver novos modelos que não são restritos à cenários extremos onde todas as interações não nulas são diferentes ou todas iguais. A primeira estratégia para lidar com os grupos envolve uma mistura finita de componentes degeneradas. A segunda opção é especificada por um processo Dirichlet. Um estudo simulado abrangente é desenvolvido para explorar as propostas e mostrar seus bons desempenhos. Uma análise de sensibilidade é realizada para avaliar as vantagens de estimar o parâmetro de suavização definido na função de covariância do processo Gaussiano que estabelece a não linearidade das interações. Em termos de aplicação, a metodologia é apresentada em análise de expressão de genes relacionados a quatro conjuntos de dados referente ao câncer de mama. Aqui, os genes pertencentes a regiões disjuntas do genoma, com alteração do número de cópias, estão conectados aos fatores principais e suas interações não lineares são estimadas e agrupadas. A investigação conjunta e a comparação desses quatro conjuntos de dados sobre câncer de mama raramente são encontradas na literatura.

Palavras-chave: Mistura, processo Dirichlet, Expressão de Genes, Câncer de mama.

Abstract

Factor analysis is a powerful tool for dimension reduction in a multivariate statistical study. This Thesis is dedicated to extend the factor model with non-linear interactions proposed in 2013. The main contribution of our work is to present two approaches to cluster the non-linear interactions and thus develop new models that are not restricted to the extreme scenarios where all non-null interactions are different or all are the same. The first strategy to handle the clusters involves a finite mixture of degenerated components. The second option is especified via the Dirichlet process. A comprehensive simulation study is developed to explore the proposals and it shows their good performances. A sentitivity analysis is carried out to evaluate advantages of estimating a smoothness parameter defined in a covariance function of the Gaussian process establishing the non-linearity of the interactions. In terms of application, the methodology is illustrated with the analysis of gene expression related to four breast cancer data sets. Here, the genes belonging to disjoint genome regions, with copy number alteration, are connected to the main factors and their non-linear interactions are estimated and clustered. The mutual investigation and comparison of these four breast cancer data sets is rarely found in the literature.

Keywords: Mixture, Dirichlet Process, Gene expression, Breast Cancer.

Sumário

Capítulo 1

Introdução

Nos últimos anos métodos computacionais e técnicas de estatística multivariada têm sido de grande importância para analisar dados com alta dimensão. O modelo fatorial é uma ferramenta estatística poderosa e bastante utilizada para analisar a dependência multivariada verificando padrões e associações nos dados. A principal função da análise fatorial é reduzir ou resumir a informação de 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 e dos métodos iterativos de simulação via cadeias de Markov, vários tipos de modelos fatoriais combinados com a abordagem Bayesiana foram usados para analisar dados 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. Carvalho et al. (2008) também é outra referência que usa distribuições *a priori* esparsas em modelos fatoriais latentes para abordar a redução da dimensão de dados de expressão de genes. Portanto modelos fatoriais Bayesianos têm levado a resultados interessantes para análise de expressão de genes.

A complexa rede de dependência entre genes motivou Mayrink e Lucas (2013) a cons-

truir um modelo fatorial com interações, pois a interação é introduzida para explicar uma parte dessa dependência. Nesta situação, os autores trabalham com uma aplicação prática, em que verificaram ao longo do genoma uma pequena região, pré-definida por Lucas et al. (2010), apresentando um problema conhecido por alteração do número de cópias. Então, foi considerado como grupo 1 os genes daquele trecho. No mesmo genoma também foi identificado uma segunda região (conhecida) que apresentava o mesmo problema de alteração, e os genes localizados nesta segunda parte são usados para formar um grupo 2. Diante dessa situação, um modelo fatorial com interações foi estruturado para que seus fatores estejam relacionados com cada um desses grupos. Neste caso, o fator 1 está associado com os genes do grupo 1 e o fator 2 está associado com os genes do grupo 2. Os efeitos de interação entre os fatores inseridos na modelagem, seriam a representação do efeito combinado das duas regiões com alteração sobre genes localizados em outras partes do genoma.

A forma que a interação entre as regiões com alteração afetam os genes localizados em outras partes do genoma, é estudada por Mayrink e Lucas (2013) como sendo não linear e são introduzidas no modelo por meio de misturas de distribuições. Dois tipos de misturas foram abordadas pelos autores. A primeira atribui às interações uma mistura com duas componentes, uma degenerada em zero e outra sendo um processo Gaussiano com função de covariâncias baseada nos escores dos fatores. Na segunda abordagem, as interações são incorporadas ao modelo a partir de uma mistura com duas componentes degeneradas, uma em zero e outra em um efeito de interação, que será estimado e compartilhado por diversos genes. Os autores também estudaram a suavidade da superfície de interação entre os fatores gerada pelo processo Gaussiano, por meio de uma análise de sensibilidade do parâmetro de comprimento-escala da função de covariâncias Gaussiana. Por isso, uma das contribuições feitas neste presente trabalho será a estimação deste parâmetro que tem uma grande importância para o ajuste do modelo fatorial com interações.

Além do mais, Mayrink e Lucas (2013) consideram casos extremos onde as interações são todas diferentes ou iguais. Pensando nisso, o uso de modelos com propriedades de agrupamentos podem ser bastante utéis para o desenvolvimento de uma nova abordagem, que modifica a forma de como os efeitos de interação são estimados. Neste caso, as espe-

cificações *a priori* para as interações seriam usadas com o interesse voltado à formação de grupos disjuntos de genes afetados pelo mesmo tipo de interação, sendo que estas interações se diferem em cada grupo. Diante disso, modelos com propriedades de agrupamentos, assim como é o caso do processo Dirichlet [Ferguson (1973)], são muito utéis para o desenvolvimento de aplicações Bayesianas não paramétrica. O processo Dirichlet é bastante usado como uma distribuição *a priori* para uma mistura de distribuições desconhecida [McEachern e Muller (1998), Escobar e West (1995) e Neal (2000)]. Por causa da sua propriedade, o processo Dirichlet utilizado em problemas de misturas tem a vantagem de permitir a determinação automática do número de componentes que melhor se ajusta aos dados. A flexibilidade desses modelos para representar diversas configurações de densidades *a priori*, juntamente com o desenvolvimento dos métodos de amostragem via Cadeia de Markov, tem possibilitado sua utilização em diversas aplicações.

No caso do modelo fatorial com interações proposto em 2013, existem duas situações práticas onde ele é aplicado. Uma delas, resaltando novamente, considera uma modelagem com as interações todas diferentes. Enquanto que a outra situação admite uma modelagem com as interações todas iguais. Estes casos acabam sendo uma limitação do modelo apresentado por Mayrink e Lucas (2013), pois deixa a modelagem das interações restrita a duas situações extremas. A ideia proposta aqui é considerar uma situação intermediária para formar grupos com os efeitos de interação. Para fazer os agrupamentos serão cosiderados duas abordagens. Na primeira, será usado uma mistura finita de distribuições com várias componentes degeneradas. Essa abordagem estende uma das modelagens apresentada por Mayrink e Lucas (2013), em que todas as interações não nulas são iguais. A segunda abordagem, utiliza o processo Dirichlet para fazer o agrupamento das interações, pois como esta metodologia não foi abordada pelos autores, ela representa uma outra contribuição deste trabalho.

No contexto de dados de expressão, o agrupamento das interações a partir das abordagens propostas, representariam a atuação conjunta de diversos genes que são influenciados (afetados) por outros localizados nas regiões do genoma pré-especificadas com o problema de alteração. Essa ideia proposta para formação de grupos com as interações é motivada por estudos conhecidos na literatura chamados de redes reguladoras de ge-

nes. Estes estudos procuram relações de dependência entre genes que interagem entre si formando uma complexa rede de associações. Portanto, essa motivação de aplicações em problemas práticos dão um direcionamento sobre a importância do tipo de modelagem que está sendo proposta.

1.1 Organização da Tese

Esta tese de Doutorado está organizado da seguinte maneira:

O Capítulo 2, faz uma descrição resumida sobre a análise de expressão de genes e o pré-tratamento dos dados que serão utilizados neste trabalho. Também será apresentado o problema de alteração do número de cópias, que motivou a aplicação de um modelo fatorial com interações abordado com detalhes no contexto de expressão de genes afetados pela alteração do número de cópias. O Capítulo 2 também apresenta o modelo fatorial com interações proposto em 2013 e mostra um estudo simulado envolvendo a modelagem considerando todos os efeitos de interação estimados sendo diferentes. O objetivo aqui, é fazer a estimação do parâmetro de comprimento-escala da função de covariâncias e verificar o comportamento do modelo fatorial com interações. A estimação do parâmetro de comprimento-escala tem uma grande importância para o ajuste do modelo fatorial. Na modelagem apresentada por Mayrink e Lucas (2013) este parâmetro foi fixado em diversos valores para um estudo de sensibilidade. Por isso, a primeira contribuição para este trabalho será ajustar o modelo fatorial com interações diante de sua estimação e verificar o seu comportamento.

O Capítulo 3 faz uma extensão da modelagem proposta por Mayrink e Lucas (2013) que consideram a situação onde todas as interações estimadas são iguais. Aqui, será desenvolvido uma modelagem fazendo agrupamentos com as interações por meio de misturas de distribuições com ponto de massa. Diversos cenários simulados são avaliados para estudar este tipo de modelagem de agrupamento para as interações, observando como o modelo fatorial se comporta diante da estimação do parâmetro de comprimento-escala. O modelo apresentado neste capítulo, além de estender a modelagem proposta em 2013, acaba sendo um possível concorrente ao modelo em que as interações são agrupadas

via processo Dirichlet.

O Capítulo 4 desenvolve uma abordagem mais flexivel da modelagem vista no Capítulo 3. Será mostrado algumas simulações com o modelo fatorial adotando uma mistura via processo Dirichlet para os efeitos de interação. Os ajustes são feitos considerando a estimação do parâmetro de comprimento-escala. A ideia é usar o processo Dirichlet como um modo alternativo ao modelo de misturas finito, pois suas boas propriedades permitem fazer agrupamentos com as interações, sem especificação prévia da quantidade de grupos. Vários cenários são avaliados e comparados com a modelagem via misturas.

No Capítulo 5, apresenta-se uma aplicação dos modelos com agrupamentos da interações. A aplicação envolve quatro conjuntos de dados reais representando expressões de genes para o câncer de mama. Primeiramente, é feito uma análise com os ajustes das modelagens propostas em apenas uma das bases de dados. Os ajustes são feitos diante de diferentes configurações *a priori* atribuidas aos pesos da mistura. Neste estudo, é verificado que algumas especificações fornecem modelagens mais parcimoniosas que outras. Em seguida, ajustes são feitos para identificação e avaliação de genes afetados por interação que são comuns nos quatro conjuntos de dados.

O Capítulo 6 fará um resumo de tudo aquilo que foi apresentado e discutido neste trabalho destacando as principais conclusões tiradas nos estudos e apresentando algumas propostas de trabalhos futuros.

Capítulo 2

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*. Os *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, será considerado como representação de um gene de interesse. Assim, neste *chip* que contém sequências genéticas é aplicado uma solução com material extraído 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 2.1 apresenta uma imagem ilustrativa desse processo.

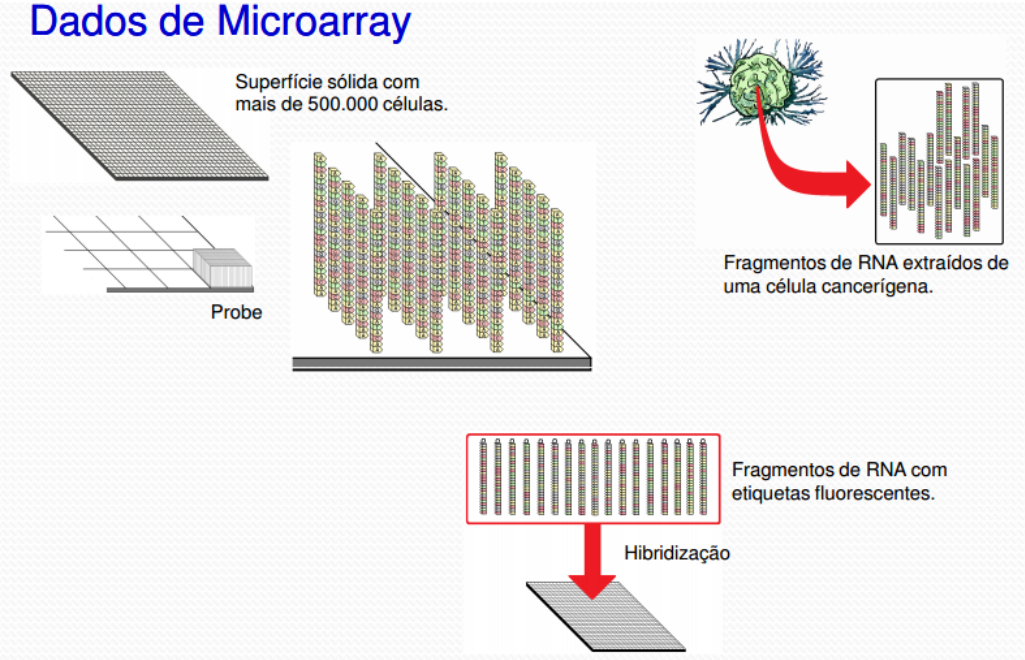

Figura 2.1: *Representação da configuração de um chip usado para a construção de plataformas de microarrays.*

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 2.2 representa a imagem de um *microarray*.

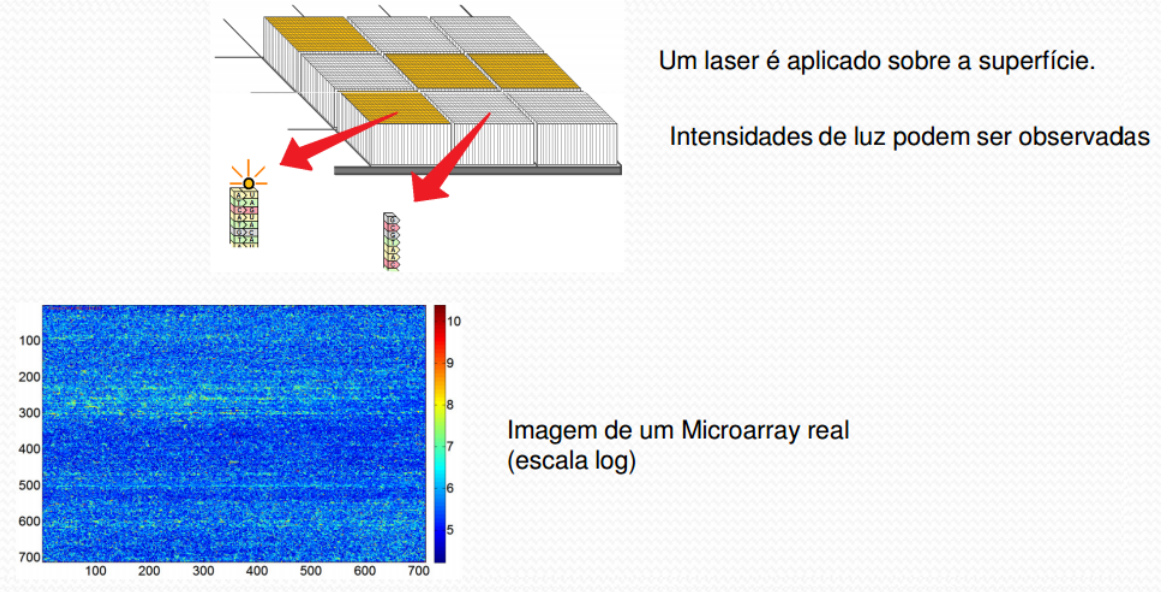

Figura 2.2: *Representação de uma imagem de microarray.*

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 em diversos estudos, como por exemplo, 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 [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 se analisar o nível de expressão. Existem diversos métodos de pré-processamento, entre eles pode-se 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, tem-se 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 2.1 faz uma breve descrição sobre o problema de alteração do número de cópias. Na Seção 2.2 será apresentado o modelo fatorial com interações descrevendo suas principais características para modelagem de dados de expressão. A Seção 2.3 mostra um estudo simulado para avaliar o comportamento e o desempenho do modelo fatorial com interações contexto da alteração do número de cópias. A Seção 2.4 fecha o capítulo com as principais conclusões.

2.1 Alteração no Número de Cópias

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). Neste trabalho estas regiões são pré-definidas e não está sendo proposto nenhum método para encontrá-las. Alguns autores trabalham propondo métodos 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, os autores identificaram 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. Os autores verificaram que um mesmo grupo de genes, localizados em uma região do genoma com CNA para câncer de mama, podem estar ativos e serem classificados como presentes em diferentes tipos de câncer como pulmão e ovário. A classificação como presente ou ausente é feita baseada na significancia ou não do impacto dos fatores que eram encontrados ao longo das amostras. A atenção que é dada para essas regiões do genoma que são conhecidas por exibir CNA, e a identificação dos genes que apresentam este problema é extremamente importante porque estas anormalidades são relevantes para a progressão do câncer.

Neste contexto considere, por exemplo, que o genoma tenha duas regiões com alteração, estas regiões são disjuntas e englobam vários genes (ver Figura 2.3). A ideia é utilizar um modelo fatorial (que será apresentado com mais detalhes adiante) contendo dois fatores, sendo que cada fator estará diretamente relacionado com cada grupo de genes das diferentes regiões com alteração.

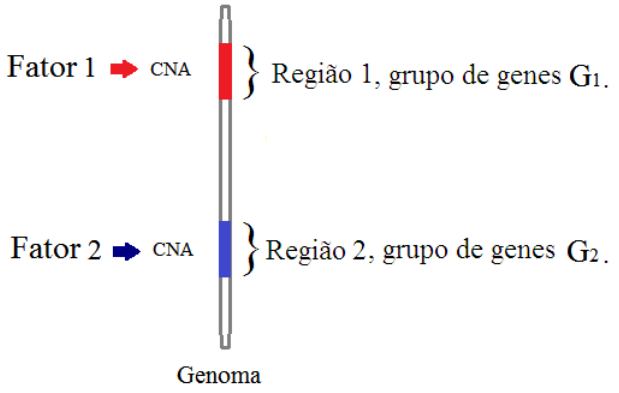

Figura 2.3: *Imagem ilustrativa do problema de CNA no genoma.*

No modelo fatorial apresentado por Mayrink e Lucas (2013) o número de fatores está relacionado ao número de regiões disjuntas no genoma afetadas pelo problema de CNA. Quanto mais regiões com CNA existirem no genoma, maior será o número de fatores adotados na modelagem. O exemplo ilustrado na Figura 2.3 considera uma situação simplificada em que existem apenas dois trechos afetados com CNA. Assim, o efeito principal do modelo seria o primeiro fator fortemente associado com grupo de genes G_1,

localizados em uma região 1, e o segundo fator atuando fortemente no grupo de genes G_2 localizados em uma região 2. Os demais genes localizados fora desses trechos do genoma com CNA, poderiam ter influência de cada fator ou até mesmo de uma interação entre os fatores. Essas situações foram abordadas por Mayrink e Lucas (2013), que construiram um modelo fatorial esparso para investigar a existência do efeito de interação entre os fatores latentes. A incorporação do efeito de interação entre os fatores no modelo, foi motivada pela complexa estrutura de associação e dependência entre os genes que não pode ser estudada em modelos simples, visando responder perguntas como: de que forma a interação entre as regiões do genoma com CNA afetam os genes localizados em outras partes do genoma? Diante deste fato, os autores adotam misturas de distribuições *a priori* esparsas para descrever a incerteza que se tem sobre a atuação conjunta dos grupos G_1 e G_2 em genes de outras regiões do genoma. Neste caso, duas especificações *a priori* são usadas. A primeira atribui às interações dos fatores uma mistura com duas componentes, uma degenerada no vetor nulo e a outra em um processo Gaussiano. Nessa situação, a interação entre os grupos de genes G_1 e G_2, que poderão afetar outros genes, seriam todas diferentes. A segunda especificação *a priori* usada pelos autores seria uma mistura com duas componentes degeneradas, uma no vetor nulo e a outra em um vetor não nulo a ser estimado. Nessa situação o tipo de interação entre os grupos G_1 e G_2, que podem afetar outros genes, seriam as mesmas. Esses tipos de especificações *a priori* tem sido bastante usada para definir a estrutura de modelos que são desenvolvidos, por exemplo, nos trabalhos de West (2003), Lucas et al. (2006) e Carvalho et al. (2008).

2.2 Modelo Fatorial Latente Esparso com Interações

Para este estudo considere X uma matriz com dimensão $m \times n$ construída com os valores das luminosidades pré-processadas. Cada linha de X representa um gene de interesse e cada coluna representa a amostra desse gene. Por exemplo, X_{ij} é a luminosidade pré-processada referente ao gene (*probe set*) i do *microarray* ou amostra j, com $i = 1, 2, ..., m$ e $j = 1, 2, ..., n$. O modelo fatorial com interações tem a seguinte formulação:

$$X = \alpha\lambda + F + \epsilon, \tag{2.1}$$

sendo α uma matriz de cargas ou *loadings* com dimensão $(m \times L)$, λ a matriz dos escores dos fatores com tamanho $(L \times n)$, F a matriz dos efeitos de interação sendo $(m \times n)$, ϵ a matriz dos erros de dimensão $(m \times n)$ com $\epsilon_{ij} \sim N(0, \sigma_i^2)$ e L representa o número de fatores adotados no modelo. Considere aqui $\alpha_{i\bullet}$ um vetor $(1 \times L)$ representando a i-ésima linha da matriz α e $\lambda_{\bullet j}$ um vetor $(L \times 1)$ representando a j-ésima coluna da matriz λ. A representação individual do modelo em (2.1) para cada gene e para cada amostra será:

$$X_{ij} = \alpha_{i\bullet}\lambda_{\bullet j} + F_{ij} + \epsilon_{ij}.$$

Em análise de expressão de genes, λ é responsável por descrever o padrão da expressão dos genes ao longo das amostras. Enquanto que as cargas determinam a força e a direção de influência desses fatores. A Figura 2.4 apresenta uma representação das matrizes do modelo fatorial com interações visto em (2.1) considerando $L = 2$ fatores latentes. Veja que no contexto de genes afetados pela CNA, o modelo mostrado considera que cada fator está representando um grupo de genes (G_1 e G_2) em que esses grupos representam partições (ou submatrizes) bem definidas de X. O grupo G_E é formado pelas linhas de X que representam todos os genes localizados fora das regiões do genoma afetadas pela CNA.

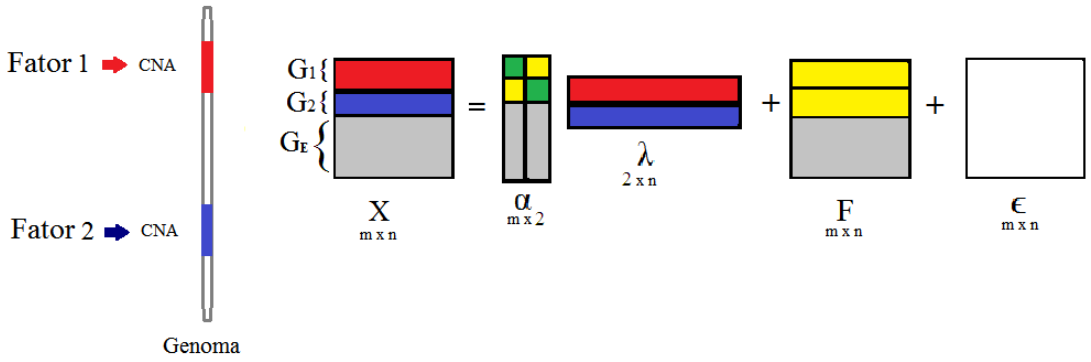

Figura 2.4: *Representação ilustrativa do modelo fatorial com interações relacionado com o problema de CNA no genoma. As regiões em amarelo representam valores nulos. Em verdes estão as cargas diferentes de zero. Em cinza tem-se regiões de incerteza sobre a presença ou ausência do efeito principal dos fatores e/ou da interação entre eles.*

A análise fatorial é aplicada em diversos contextos, não somente em dados de ex-

pressão de genes. Considere por exemplo, que está sendo feito uma pesquisa onde são coletadas informações a respeito dos preços de produtos de supermercados. Neste caso, cada linha da matriz de dados X representa uma variável, ou seja, o preço de cada produto que está sendo medido como os preços do arroz, trigo, banana, laranja e outros. Enquanto que as colunas de X representam supermercados. Neste exemplo, considere uma situação simplificada onde cada supermercado é tratado como independente um do outro e sempre se está medindo as mesmas variáveis. Desta forma, a análise fatorial é usada para se fazer uma redução do conjunto de variáveis originais a partir da construção de novas variáveis às quais são chamadas de fatores latentes. Os fatores são usados para resumir a informação contida dos dados originais. No exemplo com os preços dos produtos de supermercados, o uso do modelo fatorial é feito de maneira que cada fator representaria um grupo de produtos, neste caso, um fator poderia ser usado para representar as variáveis como preço do arroz, trigo, milho, etc, formando um grupo com os cereais. Enquanto que outro fator poderia representar o preço da laranja, maçã, etc, formando o grupo com as frutas. Os fatores seriam novas variáveis representando índices que vão resumir a informação dos grupos de preços. Estes índices teriam um determinado comportamento observado ao longo dos supermercados. Com isso, podem ser observados grupos de alimentos de tipos distintos para os quais cada fator é responsável por descrevê-los.

No modelo apresentado em (2.1), os fatores que irão ser usados para reduzir a informação de X estão em λ. Por exemplo, se forem escolhidos $L = 2$ fatores teremos duas linhas em λ por n amostras (supermercados, *microarrays*, etc). Neste caso, cada linha de λ (fator) poderá estar representando grupos de variáveis em X com características diferentes. Algumas dessas variáveis serão afetadas apenas pelo fator 1, outras variáveis somente pelo fator 2 e pode haver também algumas variáveis que são afetadas pelos dois fatores. Voltando ao exemplo anterior onde são coletados os preços dos produtos de cada supermercado. O modelo fatorial irá resumir o preço do arroz, trigo, laranja, etc, a partir de fatores latentes e dará um coeficiente para cada um desses fatores. O "tamanho" desse coeficiente será a importância que o fator tem sobre a variável. No modelo visto em (2.1) esses coeficientes são representados por α, e podem ser tanto positivos quanto negativos

indicando o impacto ou a força que o fator tem para explicar as variáveis analizadas. Considerando um modelo com $L = 2$ fatores, tem-se que α será uma matriz com duas colunas (reveja a Figura 2.4), neste caso a carga α_{il} é quem será responsável por determinar se o fator l será relevante ou não para explicar a variável i de X. Supondo que α_{il} é significativo, isto é, $\alpha_{il} \neq 0$, então o fator l tem um impacto para explicar a variável i. Se $\alpha_{il} = 0$, então o fator l não tem impacto algum para a variável i. Pode-se ter também que $\alpha_{i1} \neq 0$ e $\alpha_{i2} \neq 0$, neste caso os dois fatores serão importantes tendo um impacto para representar a variável i. Uma matriz de cargas com muitos α_{il} significativos indica que os fatores conseguem explicar muito bem o conjunto de variáveis, enquanto que uma matriz de cargas com poucos α_{il} significativos sugere que a influência dos fatores no conjunto de variáveis é baixa ou nula.

O modelo em (2.1) é todo formado por quantidades aleatórias e problemas de identificação como trocas de colunas de α e, consequentemente, trocas de linhas em λ podem ocorrer. Neste caso, algumas restrições serão feitas a partir de especificações *a priori* estabelecidas para essas quantidades aleatórias. Neste trabalho, a aplicação do modelo fatorial com interações é feito no contexto de expressão de genes afetados pela CNA. Então, para estabelecer a suposição de que cada grupo de genes G_l esteja diretamente relacionado com cada fator l, considere a seguinte especificação *a priori* para as cargas:

$$\begin{aligned} (\alpha_{il} \mid h_{il}, \omega) &\sim (1 - h_{il})\delta_0(\alpha_{il}) + h_{il}N(0, \omega); \\ (h_{il} \mid q_{il}) &\sim \text{Bernoulli}(q_{il}); \\ q_{il} &\sim \text{Beta}(\gamma_1, \gamma_2), \end{aligned} \tag{2.2}$$

sendo $\delta_0(\alpha_{il})$ uma medida de probabilidade represetando a componente da mistura com ponto de massa em zero. Na estrutura adotada em (2.2) q_{il}, $(h_{il} \mid q_{il})$ e $(\alpha_{il} \mid h_{il}, \omega)$ são independentes para o conjunto formado com os pares de índices distintos (i, l). Vale ressaltar que a variável latente binária h_{il} foi introduzida na notação apenas para auxiliar a descrição do modelo. Em cada iteração do algoritmo Monte Carlo via Cadeias de Markov (MCMC) usado para estimação, fazemos $h_{il} = 0$ se $\alpha_{il} = 0$ e $h_{il} = 1$ caso contrário. Desta forma, não há inconsistência entre estes elementos no algoritmo.

A distribuição *a priori* em (2.2) 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 $N(0, \omega)$. A incerteza expressa sobre q_{il} é feita por meio da distribuição Beta e sua estimativa *a posteriori* permite avaliar a significância de α_{il}. Com as escolhas dos hiperparâmetros apropriados para as distribuições Betas, pode ser estabelecido a relação grupo-fator afirmando que cada fator l influenciará cada grupo G_l. Por exemplo, a distribuição Beta(2,1) para q_{il} com $i \in G_1$ e $l = 1$ favorecerá $\alpha_{i1} \neq 0$ em G_1, enquanto que uma distribuição Beta(1,2) para q_{il} com $i \in G_1$ e $l = 2$ favorecerá $\alpha_{i2} = 0$ em G_1. Desta forma, apenas o primeiro fator influenciará G_1. Já a distribuição Beta(1,2) para q_{il} com $i \in G_2$ e $l = 1$ favorecerá $\alpha_{i1} = 0$ em G_2, e uma Beta(2,1) para q_{il} com $i \in G_2$ e $l = 2$ favorecerá $\alpha_{i2} \neq 0$ em G_2. Neste caso, apenas o segundo fator influenciará G_2. Em relação as probabilidades q_{il} envolvendo os elementos de G_E, pode-se assumir uma distribuição Beta(1,1), neste caso o modelo fica livre para definir com base nos dados quais α_{il} serão significativos e quais genes em G_E serão associados a algum fator. Essa especificação *a priori* para q_{il} em (2.2) permite a identificação do modelo ao evitar a troca de colunas dentro de α e, consequentemente, linhas dentro de λ. A Figura 2.5 apresenta uma imagem ilustrativa desse exemplo citado acima.

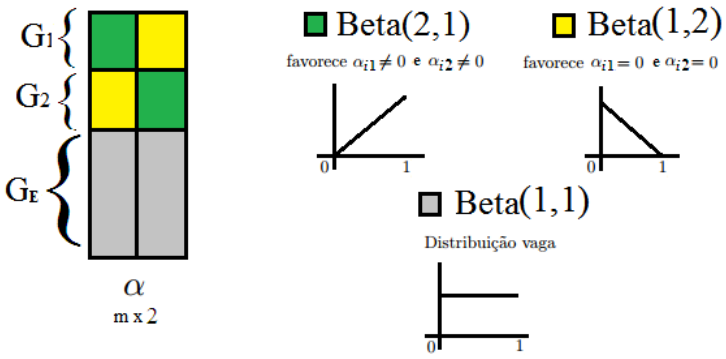

Figura 2.5: *Representação ilustrativa de uma possível escolha para os hiperparâmetros das distribuições Betas atribuidas à q_{il} que estabelecem a relação grupo-fator. As regiões em amarelo representam cargas com valores nulos. Em verde estão as cargas diferentes de zero. Em cinza tem-se uma região de incerteza sobre as cargas relacionadas aos fatores.*

Este estudo abordado aqui, considera o caso mais simples assim como feito por Mayrink e Lucas (2013) em que o número de fatores adotados é $L = 2$. Uma situação que

envolva mais fatores ($L > 2$) também é possível. Contanto que haja mais regiões no genoma com CNA para formação de grupos G_l, que estarão sendo explicados por esses fatores acrecentados. Por exemplo, no caso de haver três grupos um modelo com $L = 3$ fatores seria adotado, então a matriz α teria dimensão $(m \times 3)$. Diante dessa situação a relação grupo-fator seria estabelecida com uma seguinte sugestão de configuração *a priori* para q_{il} apresentada na Tabela 2.1.

Tabela 2.1: Distribuição *a priori* de q_{il} para um modelo com $L = 3$ fatores.

índices	q_{i1}	q_{i2}	q_{i3}
$i \in G_1$	Beta$(2, 1)$	Beta$(1, 2)$	Beta$(1, 2)$
$i \in G_2$	Beta$(1, 2)$	Beta$(2, 1)$	Beta$(1, 2)$
$i \in G_3$	Beta$(1, 2)$	Beta$(1, 2)$	Beta$(2, 1)$
$i \in G_E$	Beta$(1, 1)$	Beta$(1, 1)$	Beta$(1, 1)$

Os diversos aspectos descritos sobre o modelo em (2.1), como em especificar com antecedência o número de fatores e quais grupos de genes com CNA que irão compor a modelagem, são uma característica essencial da análise fatorial confirmatória. Esses aspectos devem estar completamente fundamentados na evidência conhecida que se tem sobre as regiões do genoma afetadas pela CNA. Na prática, o modelo fatorial estruturado em (2.1), também é usado para confirmar quais genes são definidos para esses fatores. Em uma abordagem exploratória, procura-se definir o número de fatores estimando-os, para determinar qual conjunto de genes observados compartilham a característica de serem afetados pela CNA. Embora esse procedimento não seja o foco desta tese, alguns pesquisadores tem contribuído sobre a discursão a respeito da estimação do número de fatores. Lopes e West (2004) desenvolvem e exploram métodos MCMC para modelos fatoriais que tratam o número de fatores como desconhecido. Os autores introduzem o algoritmo *Reversible Jump Markov Chain Monte Carlo* (RJMCMC) [veja Green (1995)] para transitar entre modelos com diferentes números de fatores. Mais detalhes a respeito de modelos fatoriais Bayesianos e sobre como estimar o número de fatores, podem ser encontrados respectivamente, no Capítulo 5 de *Bayesian Inference in the Social Sciences*

de Jeliazkov e Yang (2014) e em Fruhwirth-Schnatter e Lopes (2009).

Além das especificações para as cargas, considere as seguintes distribuições *a priori* para as variâncias e para os escores dos fatores:

$$\sigma_i^2 \sim GI(a, b); \tag{2.3}$$

$$\lambda_{\bullet j} \sim N_L(\mathbf{0}, I_L) \text{ com } \lambda_{\bullet j} = (\lambda_{1j}, \lambda_{2j}, ..., \lambda_{Lj})^\top. \tag{2.4}$$

Considere GI a indicação de uma Gama Inversa e I_L a matriz identidade com dimensão L. A configuração escolhida em (2.4) é utilizada como estratégia padrão para fixar a magnitude de λ e evitar um problema de identificabilidade no produto $\alpha\lambda$.

No modelo fatorial em (2.1) a matriz F é construída como uma interação dos fatores. Para isso, no mínimo dois fatores tem que ser bem definidos. Então, para um caso simplificado, considere o modelo fatorial com interações adotando $L = 2$. Nesta situação, tem-se o efeito principal do primeiro fator controlado por α_{i1}, o efeito principal do segundo fator controlado por α_{i2} e, além disso, se for admitido que os dois fatores atuem de maneira conjunta a partir de interações, esta será representada no modelo pelas linhas de F. Considere como notação $F_{i\bullet}$ representando um vetor $(1 \times n)$ sendo a i-ésima linha da matriz F.

No contexto de expressão de genes a matriz F faz muito sentido, pois como foi mencionado cada linha de X representa um gene e pela suposição de que há muitas interações complexas entre os genes, a estrutura adotada para $F_{i\bullet}$ tenta captar os efeitos de interação dos grupos G_1 e G_2, fortemente relacionados aos dois fatores, no gene i ao longo das amostras. Para os efeitos de interação que podem afetar os genes localizados no grupo G_E, será atribuído a seguinte especificação *a priori*:

$$\begin{aligned}
(F_{i\bullet}^\top \mid \lambda, \phi) &\sim (1 - z_i)\delta_{\mathbf{0}}(F_{i\bullet}) + z_i N_n[\mathbf{0}, K(\lambda, \phi)], \\
(z_i \mid \rho_i) &\sim \text{Bernoulli}(\rho_i), \\
\rho_i &\sim \text{Beta}(\beta_1, \beta_2),
\end{aligned} \tag{2.5}$$

sendo $K(\lambda, \phi)$ uma matriz de covariâncias construída por uma função de covariâncias que será apresentada mais adiante. Considere a medida de probabilidade $\delta_{\mathbf{0}}(F_{i\bullet})$ sendo a componente degenerada no vetor nulo $\mathbf{0}$. Ela representa a i-ésima linha de F igual ao

vetor nulo com probabilidade 1. Aqui é considerado que ρ_i, $(z_i \mid \rho_i)$ e $(F_{i\bullet} \mid z_i, \lambda, \phi)$ são independentes. Além disso, na distribuição *a priori* em (2.5) as interações significativas ao modelo (2.1) são avaliadas a partir da distribuição *a posteriori* de ρ_i. Com os valores dos hiperparâmetros da distribuição Beta atribuída à ρ_i, será assumido que os genes em G_1 e G_2, não são afetados pelas interações. Por exemplo, estabelecendo uma Beta$(1, 2)$ para ρ_i, com $i \in (G_1 \cup G_2)$, permitirá valores baixos para ρ_i favorecendo $z_i = 0$ levando $F_{i\bullet} = \mathbf{0}$. Isso indica que o gene i não é afetado pela interação dos fatores. Estabelecendo uma distribuição Beta$(1, 1)$ para ρ_i, com $i \in G_E$, isso deixa o modelo decidir com base nos dados quais $F_{i\bullet} \in G_E$ serão vetores nulos ou não nulos gerados do processo Gaussiano. A Figura 2.6 apresenta uma imagem ilustrativa desse exemplo.

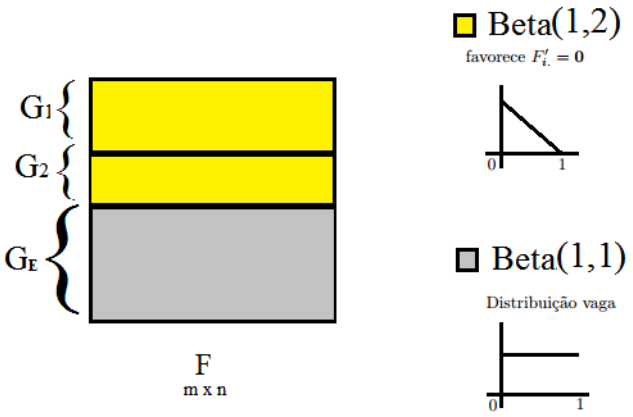

Figura 2.6: *Representação ilustrativa de uma possível escolha para os hiperparâmetros das distribuições Betas atribuídas à ρ_i que estabelecem a relação grupo-fator e ajudam na identificação do modelo. As regiões em amarelo representam efeitos de interações nulos. Em cinza tem-se uma região de incerteza sobre a presença ou não de efeitos de interações.*

Vale ressaltar que as distribuições *a priori* para q_{il} e ρ_i não são únicas. Especificações *a priori* "mais fortes" para q_{il} e ρ_i podem ser usadas para garantir a suposição feita sobre a relação grupo-fator, determinando a identificação do modelo fatorial, principalmente em situações onde a matriz de dados apresenta uma quantidade de genes (linhas) muito grande. Em aplicações envolvendo um conjunto de dados reais, onde há milhares de genes,

as distribuições *a priori* como Beta(2,1) e Beta(1,2), mostradas nos exemplos anteriores seriam facilmente dominadas pelos dados, pois os grupos G_1 e G_2 apresentam menos genes em relação ao grupo G_E. Assim, as distribuições *a priori* usadas neste estudo estão descritas na Tabela 2.2. Com essas configurações *a priori* para q_{il} e ρ_i também está sendo resolvido o problema de trocas de posições nas colunas de α e linhas de λ, garantindo a identificação do modelo fatorial com interações.

Tabela 2.2: 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

Veja que ao considerar $p(q_{i1} = 1) = 1$, $\forall i \in G_1$, tem-se $\alpha_{i1} \neq 0$ em G_1, enquanto que $p(q_{i2} = 0) = 1$ garante $\alpha_{i2} = 0$ para G_1, assim apenas o primeiro fator terá influência em G_1. Já $p(q_{i1} = 0) = 1$ irá fornecer $\alpha_{i1} = 0$ para G_2 enquanto que $p(q_{i2} = 1) = 1$ irá garantir $\alpha_{i2} \neq 0$ para G_2, e neste caso apenas o segundo fator terá influência em G_2. Além disso, ao atribuir $p(\rho_i = 0) = 1$ para ρ_i será considerado que o efeito de interação $F_{i\bullet}$ não influênciará $(G_1 \cup G_2)$. A distribuição Beta$(1, 1)$ atribuída para q_{il} e ρ_i deixa o modelo decidir, com base nos dados, quais α_{il} e $F_{i\bullet}$, em G_E, são significativos. Com isso, o grupo G_E pode ser influenciado pelo efeito principal e/ou pelo efeito de interação dos fatores. Vale ressaltar que outras configurações da distribuição Beta podem ser especificadas para q_{il} e ρ_i referentes a α_{il} e $F_{i\bullet}$ em G_E. Gonçalves et al. (2013) utilizam uma distribuição Beta(γ_1, γ_2), com γ_1 e γ_2 entre 0 e 1, em estudos simulados para modelos da Teoria da Resposta ao Item. O uso desta distribuição Beta seria uma outra opção, pois ela tem um comportamento bimodal, ou seja, formato de "banheira" concentrando a maioria de sua massa probabilística nos extremos do intervalo $(0, 1)$. Entretanto, esta configuração não será abordada aqui.

A forma que a interação dos grupos G_1 e G_2 afetam os genes do grupo G_E é não linear. O termo é dito não linear, porque na estrutura adotada para $F_{i\bullet}$ em (2.5) é atribuído, em uma das componentes da mistura, o processo Gaussiano. Note que o processo Gaussiano depende de λ por meio de uma matriz de covariâncias gerada por uma função $K(\lambda, \phi)_{j_1,j_2}$, e para o modelo usado com $L = 2$ fatores esses efeitos de interação, gerados pelo processo Gaussiano, são representados por superfícies. A suavidade dessa superfície depende do parâmetro ϕ na função de covariâncias a ser utilizada. A superfície não é um plano ou algo linear, e isso dá a noção do que se quer dizer com o termo não linear, pois a interação entre os fatores pode apresentar muitos formatos diferentes.

Muitas funções de covariâncias podem ser utilizadas em (2.5), uma bastante popular na literatura é a função de covariâncias Gaussiana (ou exponencial quadrática) que será usada neste trabalho assim como feito por Mayrink e Lucas (2013). Ela é expressa por:

$$K(\lambda, \phi)_{j_1,j_2} = \upsilon^2 \exp\left\{-\frac{1}{2\phi^2}\left\|\lambda_{\bullet j_1} - \lambda_{\bullet j_2}\right\|^2\right\}, \tag{2.6}$$

sendo υ^2 um parâmetro global que controla a variabilidade, $(j_1, j_2) \in \{1, 2, ..., n\}$ e $\phi > 0$ 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. Em um caso particular, assumindo $\upsilon^2 = 1$ tem-se 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 $\mathbb{R}^L$, maior será a similaridade deles levando a $K(\lambda, \phi)_{j_1,j_2} \approx 1$. Por outro lado, quanto maior a distâncias entre estes vetores, menor será a similaridade entre os mesmos e assim tem-se $K(\lambda, \phi)_{j_1,j_2} \approx 0$.

O parâmetro de comprimento-escala ϕ tem uma grande importância na estimação dos efeitos de interação entre os fatores. Ao aumentar seu valor, será "suavizado" a superfície estimada que representaria a interação entre os fatores. Mayrink e Lucas (2013) estudaram o comportamento de efeitos de interação em que as superficies apresentavam formato de "sela de cavalo". Os autores ajustaram o modelo fatorial com interações fixando o parâmetro ϕ nos valores 0.2, 0.3 e 0.5 e verificaram que ao aumentarem o valor de ϕ os efeitos de interação apresentavam superficies estimadas com um comportamento mais "suavizado" ou plano. Além disso, os autores verificaram a qualidade das estimativas dos

parâmetros a partir de uma medida chamada de distância média absoluta, e mostraram que as estimativas de α, λ, F e σ^2 não eram tão boas quando assumiam o valor $\phi = 0.5$, argumentando que essa característica pode ser interpretada como uma indicação de pior aproximação entre as estimativas *a posteriori* e os verdadeiros valores. A Figura 2.7 apresenta uma superficie estimada que representa um dos efeitos de interação estimado para o modelo onde a superfície real tem forma de sela, veja que o formato da superfície vai se tornando mais plano conforme ϕ aumenta.

$\phi = 0.2$ $\phi = 0.3$ $\phi = 0.5$

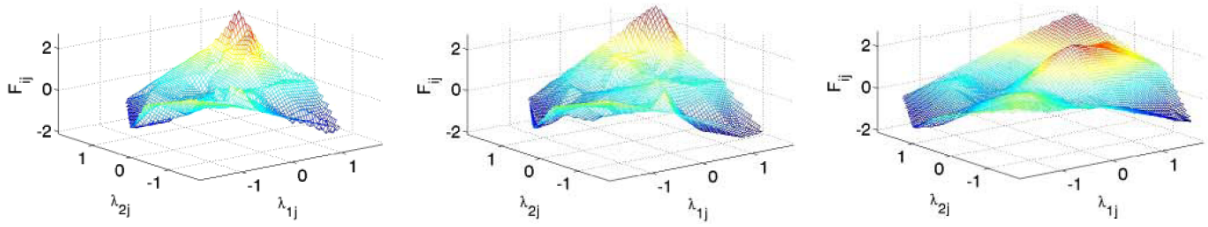

Figura 2.7: *Representação do efeito de interação estimado para o modelo fatorial com interações. Fonte: Mayrink e Lucas (2013).*

A estimação paramétrica de funções de covariâncias em processos Gaussianos é bastante vista em modelos de krigagem [veja Stein (1999), Rasmussen e Williams (2005)]. Fazer a estimação do parâmetro de suavização ϕ da função de covariâncias não é uma tarefa fácil. Neste capítulo, será proposto fazer a estimação do parâmetro de comprimento-escala ϕ, em vez de fixá-lo como feito por Mayrink e Lucas (2013). Assim, tudo o que foi discutido até este ponto do trabalho está abordado no artigo referência de 2013. A partir deste ponto serão feitas algumas contribuições como a estimação do parâmetro ϕ. Para isto, especifica-se a seguinte distribuição *a priori*:

$$\phi \sim U(a, b). \tag{2.7}$$

A escolha de a e b na distribuição em (2.7) é feita lenvando em conta a magnitude (escala) de λ, que neste trabalho está sendo atribuída uma $N(0, 1)$. Aqui, serão utilizados $a = 0.1$ e $b = 0.5$ devido as análises do comportamento dos efeitos de interações (superfícies estimadas) no artigo base de 2013, ou seja, se for atribuído uma distribuição

uniforme com uma amplitude muito "ampla", isso poderia fazer com que as superfícies estimadas (efeitos de interação) apresentassem comportamentos mais planos ou suavizados diante de valores de ϕ muito grandes, enquanto que para valores de ϕ muito pequeno as superficies apresetariam ocilações fortes e mais irregulares, podendo trazer estimativas não tão boas para os demais parâmetros do modelo.

Com a regra de Bayes obteve-se o núcleo da distribuição *a posteriori* $p(\alpha, \lambda, \sigma^2, F, \phi \mid X)$ que não é tratável analiticamente, então foi aplicado o algoritmo *Gibbs Sampling* que é essencialmente um esquema iterativo de amostragem de uma cadeia de Markov, cujo núcleo de transição é formado pelas distribuições *a posteriori* condicionais completas. Em particular foi aplicado o Metropolis-Hastings com passeio aleatório como um passo dentro do *Gibbs Sampling* para gerar $\lambda_{\bullet j}$ e ϕ. Esses parâmetros aparecem em (2.5) e suas distribuições condicionais completas não apresentam forma fechada. Para mais detalhes sobre os algoritmos *Gibbs Sampling* e Metropolis-Hastings veja Gamerman e Lopes (2006). As expressões das distribuições condicionais completas estão apresentadas no Apêndice A. O algoritmo utilizado diante da estimação de ϕ tem os seguintes passos:

Passo 1 Escolher os valores iniciais para $\alpha^{(0)}$, $\lambda^{(0)}$, $\sigma^{2(0)}$, $F^{(0)}$, $\phi^{(0)} = 0.3$, $z^{(0)} = (z_1^{(0)}, ..., z_m^{(0)})^\top$, $h^{(0)} = (h_{1\bullet}^{(0)}, ..., h_{m\bullet}^{(0)})^\top$ 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 $u \leq \rho_i^*$, faça $z_i^{(t)} = 1$ e obtenha $F_{i\bullet}^{(t)}$ de $N_n\left(M_{F_{i\bullet}}, V_{F_{i\bullet}}\right)$. 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 $N\left(M_\alpha, V_\alpha\right)$. Caso contrário, faça $h_{il}^{(t)} = 0$ e $\alpha_{il}^{(t)} = 0$.

Passo 3 Gere ϕ^* de $U\left(\phi^{(t-1)} - 0.01, \phi^{(t-1)} + 0.01\right)$.

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

$$r = \frac{\left\{\prod_{i=1}^{m}\left[N_n(F_{i\bullet}^{\top(t)} \mid M_{F_{i\bullet}^{\top(t)}}(\phi^*), V_{F_{i\bullet}^{\top(t)}}(\phi^*))\right]^{z_i}\right\} p(\phi^{(t-1)} \mid \phi^{(*)})}{\left\{\prod_{i=1}^{m}\left[N_n(F_{i\bullet}^{\top(t)} \mid M_{F_{i\bullet}^{\top(t)}}(\phi^{(t-1)}), V_{F_{i\bullet}^{\top(t)}}(\phi^{(t-1)}))\right]^{z_i}\right\} p(\phi^{(*)} \mid \phi^{(t-1)})},$$

em que $N_n(F_{i\bullet}^{\top(t)} \mid M_{F_{i\bullet}^{\top(t)}}, V_{F_{i\bullet}^{\top(t)}})$ é a densidade da normal multivariada no vetor $F_{i\bullet}^{\top}$; $p(\phi^{(*)} \mid \phi^{(t-1)})$ e $p(\phi^{(t-1)} \mid \phi^{(*)})$ são as densidades das distribuições geradora de candidatos $U(\phi^{(t-1)} - 0.01, \phi^{(t-1)} + 0.01)$ e $U(\phi^* - 0.01, \phi^* + 0.01)$, respectivamente. Essas distribuições mudam quando os candidatos são gerados próximo da borda do espaço paramétrico de ϕ. Por exemplo, se ϕ for gerado próximo de 0.5 teria-se que calcular $p(\phi^{(t-1)} \mid \phi^{(*)})$ usando $U(\phi^* - 0.01, b)$, com $b > 0.5$, mas o que é feito é o cálculo de $p(\phi^{(t-1)} \mid \phi^{(*)})$ com $U(\phi^* - 0.01, 0.5)$. A ideia é semelhante para os valores de ϕ gerados próximos de 0.1.

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

Passo 4 Gere $\lambda_{\bullet j}^*$ de $N_L\left(\lambda_{\bullet j}^{(t-1)}, 0.25 I_L\right)$, para $j = 1, ..., n$.

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

$$r = \frac{N_L(\lambda_{\bullet j}^* \mid M_\lambda, V_\lambda)\left|K(\lambda^*, \phi^{(t)})\right|^{-\frac{1}{2}\sum_{i=1}^{m} z_i^{(t)}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{m} F_{i\bullet}^{(t)} K(\lambda^*, \phi^{(t)})^{-1} F_{i\bullet}^{\top(t)}\right\}}{N_L(\lambda_{\bullet j}^{(t-1)} \mid M_\lambda, V_\lambda)\left|K(\lambda^{(t-1)}, \phi^{(t)})\right|^{-\frac{1}{2}\sum_{i=1}^{m} z_i^{(t)}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{m} F_{i\bullet}^{(t)} K(\lambda^{(t-1)}, \phi^{(t)})^{-1} F_{i\bullet}^{\top(t)}\right\}},$$

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

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

Passo 5 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` [Core Team (2019)] utilizando os pacotes `Rcpp` [Eddelbuettel e Francois (2011), Eddelbuettel (2013)] e `RcppArmadillo` [Eddelbuettel e Sanderson (2014)] que integra o `R` e a linguagem `C++`.

2.3 Estudo Simulado

Para este estudo foram consideradas matrizes de dados com tamanho $m = 200$ e $n = 100$. Utilizou-se um modelo com $L = 2$ fatores e, diferente do que foi feito por Mayrink e Lucas (2013) que simularam apenas um tipo de interação (sela), será simulado diferentes formatos de superfícies (interação). O objetivo com estas simulações é verificar o desempenho e o comportamento do modelo fatorial com interações, diante da estimação do parâmetro de suavização ϕ da função de covariâncias Gaussiana. Admiti-se que cada fator tem uma relação direta com cada grupo de genes G_l, contendo 10 genes, e que esses grupos não serão influenciados pelos efeitos de interação. Com isso, o primeiro fator não terá influência em G_2, assim como o segundo fator não influenciará G_1. Os efeitos de interação e/ou o efeito principal dos fatores podem estar associados com um grupo de 180 genes denominado G_E. Os grupos de genes G_1, G_2 e G_E são partições formando subconjuntos de 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 $u < 0.7$ para $i \in G_E$ e todo l.
   Fixa-se em 0.7 supondo em média 70% 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 $F_{i\bullet} \sim N_n[0, K(\lambda, \phi)]$ se $u < p$ para todo $i \in G_E$. Fixa-se p nos valores 0.10 ou 0.25 ou 0.50 e para cada um desses casos será considerado o parâmetro $\phi = 0.2$ ou 0.4.

4. Gere $\epsilon_{ij} \sim N(0, \sigma_i^2)$, sendo $\sigma_i^2 \sim U(0.2, 0.4)$.

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

Para este estudo foram feitas 50 replicações Monte Carlo ajustando um modelo para cada banco de dados X gerado, considerando três casos específicos para a matriz de interações F. O primeiro considerado no estudo supõe que a matriz F usada para construir

X tenha, em média, 10% de interações não nulas. O segundo, considera que a matriz F usada para gerar os dados tenha, em média, 25% de interações. O terceiro, que F tenha, em média, 50% de interações não nulas. Para cada um desses casos as linhas da matriz de interações F são geradas considerando duas situações: $\phi = 0.2$ ou $\phi = 0.4$.

A Tabela 2.3 apresenta os seguintes cenários construídos para cada caso. Na notação utilizada para os cenários, tem-se em C que o sobrescrito indica o valor real usado para ϕ. Por outro lado, o subscrito indica o tipo de modelagem que está sendo feita. Por exemplo, o Cenário $C_{0.4}^{0.2}$, significa que foi gerado um conjunto de dados com interações $F_{i\bullet}$ obtidas usando o valor real $\phi = 0.2$ na função de covariâncias, e em seguida foi ajustado um modelo considerando o valor de ϕ fixo em 0.4. O Cenário $C_{\hat{\phi}}^{0.2}$ significa que foi gerado um conjunto de dados com $F_{i\bullet}$ obtidas usando o valor real $\phi = 0.2$, em seguida foi ajustado o modelo fatorial diante da estimação do parâmetro ϕ.

Tabela 2.3: Cenários usados no estudo Monte Carlo.

ϕ_{real}	Estratégia ϕ	Cenários
0.2	$\phi_{fixo} = 0.2$	$C_{0.2}^{0.2}$
	$\phi_{fixo} = 0.4$	$C_{0.4}^{0.2}$
	$\hat{\phi}$	$C_{\hat{\phi}}^{0.2}$
0.4	$\phi_{fixo} = 0.2$	$C_{0.2}^{0.4}$
	$\phi_{fixo} = 0.4$	$C_{0.4}^{0.4}$
	$\hat{\phi}$	$C_{\hat{\phi}}^{0.4}$

Na simulação, foi assumido para as distribuições *a priori* em (2.2) e (2.3) os hiperparâmetros $\omega = 10$, $a = 2.1$, $b = 1.1$. Assim, a distribuição Gama Inversa escolhida *a priori* indica $E(\sigma_i^2) = 1$ e $\text{var}(\sigma_i^2) = 10$. Utilizou-se a função de covariâncias em (2.6) com $\upsilon = 1$ e os valores iniciais das cadeias foram $F_{ij}^{(0)} = 0$, $\alpha_{il}^{(0)} = 0$, $\sigma^2 = 1$ e $\lambda_{ij}^{(0)} \sim N(0, 1)$. Para as indicadoras $h_{il}^{(0)}$ e $z_i^{(0)}$ usou-se os valores iniciais de uma distribuiçãoo Bernoulli com probabilidades $q_{il}^{(0)}$ e $\rho_i^{(0)}$ dadas na Tabela 2.2. Foi considerado um total de 4000 iterações para execução do MCMC e as 2500 primeiras amostras foram usadas como *burn in*. A convergência foi observada em todos os parâmetros selecionados para inspeção. Uma análise mais formal envolvendo a convergência das cadeias a partir

de critérios como o de Gelman e Rubin (1992) ou Geweke (1992), podem ser encontrados em Amorim (2016).

Para analisar a qualidade do ajuste e das estimativas dos parâmetros α, λ, σ^2, F e ϕ, serão usadas as medidas *Deviance Information Criterion* (DIC), *Widely Applicabe Information Criterion* (WAIC), *Log Pseudo Marginal Likelihood* (LPML) e Erro Quadrático Médio (EQM). Detalhes sobre os cálculos do DIC, WAIC e LPML são encontrados no Apêndice C. Para o cálculo dos EQMs de matrizes é considerado o seguinte critério: Seja Q a matriz dos parâmetros com m linhas, n colunas e elementos Q_{ij} e considere $\hat{Q}$ a matriz estimada, então o EQM de Q é dado por:

$$EQM(Q) = \frac{1}{nm}\sum_{j=1}^{n}\sum_{i=1}^{m}\left(\hat{Q}_{ij} - Q_{ij}\right)^2.$$

Nesse estudo será multiplicado por -1 os valores do WAIC e LPML para que tenham a mesma interpretação do DIC, isto é, valores baixos para estes critérios indicam melhores ajustes. Mais detalhes sobre o DIC, WAIC e LPML podem ser verificados, respectivamente, em Spiegelhalter et al. (2002), Watanabe (2010) e Gelman et al. (2003). Recomenda-se também o Apêndice C.

Para cada conjunto de dados g gerado na simulação Monte Carlo, será ajustado o modelo fatorial com interações dado em (2.1) e calculado as medidas DIC_g, $WAIC_g$ e $LPML_g$, com $g = 1, ..., 50$, diante de cada cenário. O objetivo neste ponto do trabalho é avaliar o ajuste e o desempenho do modelo fatorial com interações diante da estimação de ϕ, e a medida em que há um aumento na quantidade de interações que afetam os genes no grupo G_E. Para isso as seguintes medidas foram construídas: $\frac{DIC_g^A}{DIC_g^B}$, $\frac{WAIC_g^A}{WAIC_g^B}$, $\frac{LPML_g^A}{LPML_g^B}$, que representam razões de medidas comparativas de modelos calculadas para o conjunto de dados g. Os termos A e B, indicados em sobrescrito, sugerem o tipo de modelo avaliado no numerador e denominador. Estes termos podem ser $\phi_{fixo} = 0.2$, $\phi_{fixo} = 0.4$ ou $\hat{\phi} =$ modelo estimando ϕ. Em todas as análises, será atribuído o modelo gerador dos dados sempre no denominador. Diante disso, o ponto de referência usado como base é o valor 1. Valores da razão acima de 1 indicam pior qualidade de ajuste em relação ao modelo gerador. Será avaliado se as medidas dos modelos em que há a estimação de ϕ são mais próximas de 1 do que aquelas relativas aos modelos em que ϕ é

fixo em 0.4, para os casos onde o verdadeiro valor de ϕ é 0.2. Situações semelhantes são investigadas para os casos onde o verdadeiro valor de ϕ é 0.4.

Os gráficos das medidas construídas com os critérios de comparação podem ser vistos na Figura 2.8. Nela, pode-se observar que as medidas construídas com os Cenários $C_{0.2}^{0.2}$, $C_{0.4}^{0.2}$ e $C_{\hat{\phi}}^{0.2}$ apresentam a maior parte dos *boxplots* acima do 1, indicando que o modelo ajustado fixando ϕ no seu valor real 0.2 apresenta melhor ajuste em relação aos modelos em que se fixa ϕ em 0.4 ou se estima. Além do mais, os *boxplots* das razões dos DICs, WAICs e LMPLs referentes ao modelo em que estima-se ϕ são mais próximos de 1 em comparação ao modelo onde se fixa ϕ em 0.4. Isso ocorre em todos os três casos, ou seja, quando tem-se 10%, 25% e 50% de interações afetando G_E. Veja também que os *boxplots* das razões dos DICs, WAICs e LMPLs diante do Cenário $C_{0.4}^{0.2}$, se distanciam mais do 1 nos casos onde há 50% de interações em G_E, enquanto que as razões dessas medidas diante do Cenário $C_{\hat{\phi}}^{0.2}$ se mantem próximas de 1 para todos os casos.

Para as medidas construídas com os Cenários $C_{0.2}^{0.4}$, $C_{0.4}^{0.4}$ e $C_{\hat{\phi}}^{0.4}$, nota-se que a maior parte dos gráficos estão abaixo do 1. Isso indica que o modelo ajustado considerando o Cenário $C_{0.4}^{0.4}$ apresenta o pior ajuste em relação aos modelos dos Cenários $C_{0.2}^{0.4}$ e $C_{\hat{\phi}}^{0.4}$, ou seja, maiores DICs, WAICs e LPMLs. Uma possível explicação está no fato de que quando o valor de ϕ é grande, a superfície modelada pelo processo Gaussiano (rever Figura 2.7) é suave e mais plana, o que determina um prejuízo na estimação das interações não lineares que de certa forma tem pouca liberdade para variar em relação à vizinhança mais próxima no espaço. Obviamente, isso tem um impacto sobre a estimação dos demais parâmetros do modelo e sobre as medidas de qualidade de ajuste. Veja a partir dos *boxplots*, ainda considerando os Cenários $C_{0.2}^{0.4}$, $C_{0.4}^{0.4}$ e $C_{\hat{\phi}}^{0.4}$, que as medidas construídas diante dos casos $C_{0.2}^{0.4}$ e $C_{\hat{\phi}}^{0.4}$ apresentam medianas bem próximas nas situações com 10% de interações. Isso ocorre para todas as medidas de DICs, WAICs e LPMLs. Note também, que há uma leve melhora no ajuste do modelo para o Cenário $C_{\hat{\phi}}^{0.4}$ quando se está considerando o caso de 50% de interações, pois esse modelo apresenta os *boxplots* com medianas menores em relação aos *boxplots* das medidas diante de $C_{0.2}^{0.4}$. Observa-se ainda, considerando o caso 50%, que há um distanciamento menor do 1 das medidas feitas com os critérios diante dos Cenário $C_{0.2}^{0.4}$. Isso ocorre devido ao fato delas apresentarem DIC, WAIC e LPML

próximos daqueles considerando o Cenário $C_{0.4}^{0.4}$.

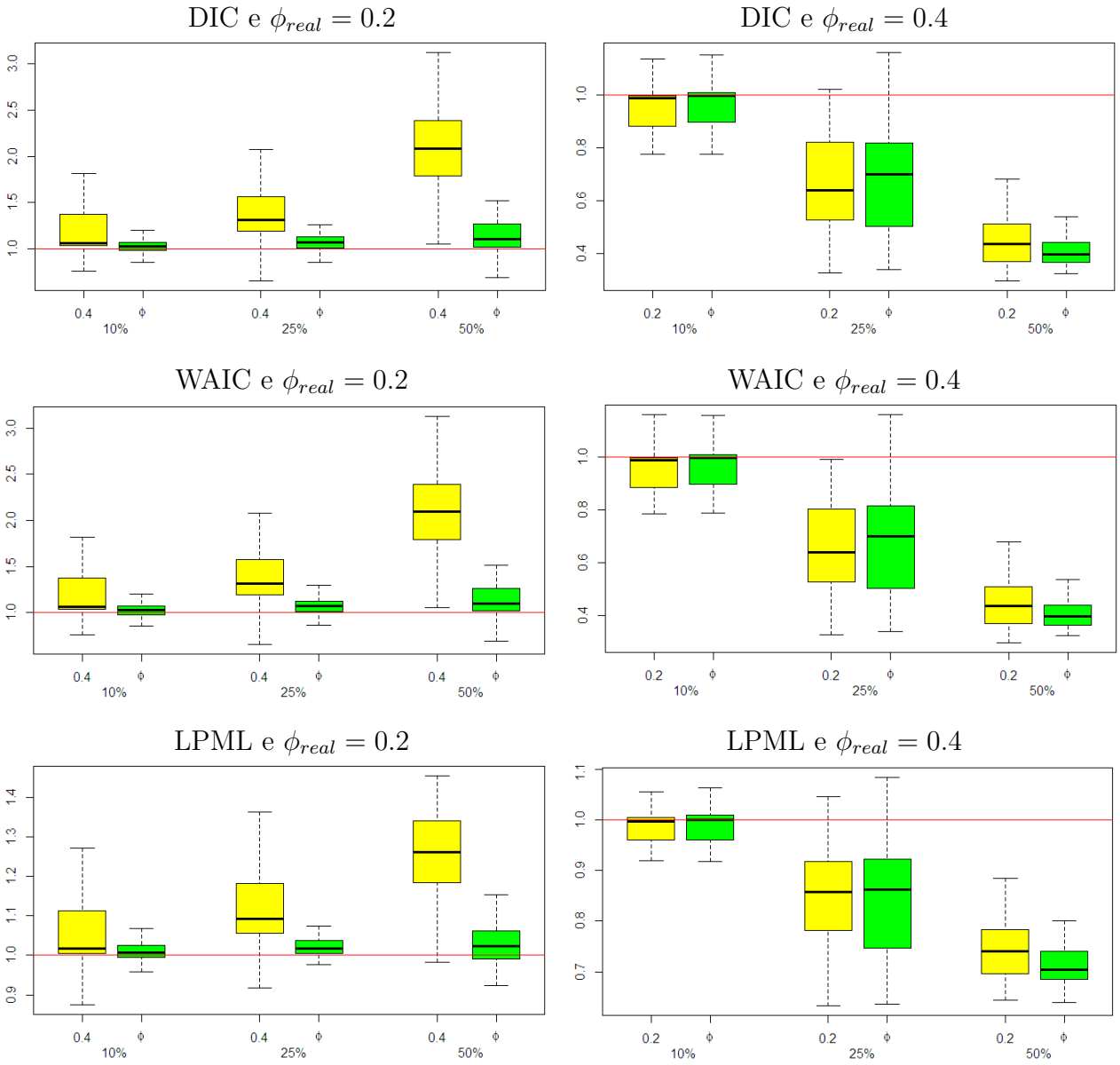

Figura 2.8: *Boxplots das razões dos DIC, WAIC e LPML para os casos onde há 10%, 25% e 50% de interações em F. Nos painéis à esquerda estão os casos onde o valor real de ϕ é 0.2, enquanto que nos painéis à direita estão os casos onde o valor real de ϕ é 0.4. A linha na horizontal representa o valor 1. Os boxplots verdes referem-se as medidas do modelo onde há a estimação de ϕ. Os boxplots amarelos referem-se as medidas do modelo onde ϕ é fixo.*

Na Figura 2.9 pode-se observar os *boxplots* dos EQMs referentes aos seis cenários diante de cada caso (10%, 25% ou 50%) considerado para o grupo G_E na matriz F. Os resultados apresentados consideram a média *a posteriori* dos parâmetros α, λ, F, σ^2 e ϕ. Os EQMs de α foram calculados desconsiderando os $\alpha_{i2} = 0 \quad (\forall \quad i \in G_1)$ e $\alpha_{i1} = 0 \quad (\forall \quad i \in G_2)$, enquanto os EQMs de F foram calculados desconsiderando todos os $F_{i\bullet} = \mathbf{0}$. A retirada dos efeitos de interação nulos no cálculo dos EQMs é feito pois o modelo tem mais facilidade em estimá-los, tornando os valores dos EQMs menores e "mascarando" a qualidade das estimativas de $F_{i\bullet} \neq \mathbf{0}$. Além do mais, a exclusão dos $F_{i\bullet} = \mathbf{0}$ permite fazer comparações de modelos entre os casos 10%, 25% e 50%. Veja na Figura 2.9, que os EQMs de α e λ para o modelo do Cenário $C_{0.2}^{0.2}$ são mais baixos que os dos demais cenários, em todos os casos. Ao considerar o Cenário $C_{\hat{\phi}}^{0.2}$, podem-se observar que os *boxplots* (em verde) dos EQMs de α e λ em todos os casos são mais baixos do que aqueles no Cenário $C_{0.4}^{0.2}$ (*boxplots* em amarelo). Esses resultados refletem em um melhor ajuste para o modelo do Cenário $C_{\hat{\phi}}^{0.2}$ em relação ao do Cenário $C_{0.4}^{0.2}$, confirmando os resultados vistos anteriormente. Já nos casos para os Cenários $C_{0.2}^{0.4}$, $C_{0.4}^{0.4}$ e $C_{\hat{\phi}}^{0.4}$, observa-se a partir dos *boxplots* que os EQMs de α e λ no Cenário $C_{0.4}^{0.4}$ são maiores, determinando um pior ajuste. Ao analisar os *boxplots* dos EQMs de α para os Cenários $C_{\hat{\phi}}^{0.4}$ (*boxplots* em verde) e $C_{0.2}^{0.4}$ (*boxplots* em azul), observa-se que as medianas são próximas em todos casos. Na configuração com 50% de interações, pode ser visto a partir dos *boxplots*, uma menor variabilidade dos EQMs de α e λ para o Cenário $C_{\hat{\phi}}^{0.4}$ em relação aos do Cenário $C_{0.2}^{0.4}$. Essas análises dão uma indicação de que estimar o parâmetro ϕ trazem bons resultados para a estimação das cargas em α.

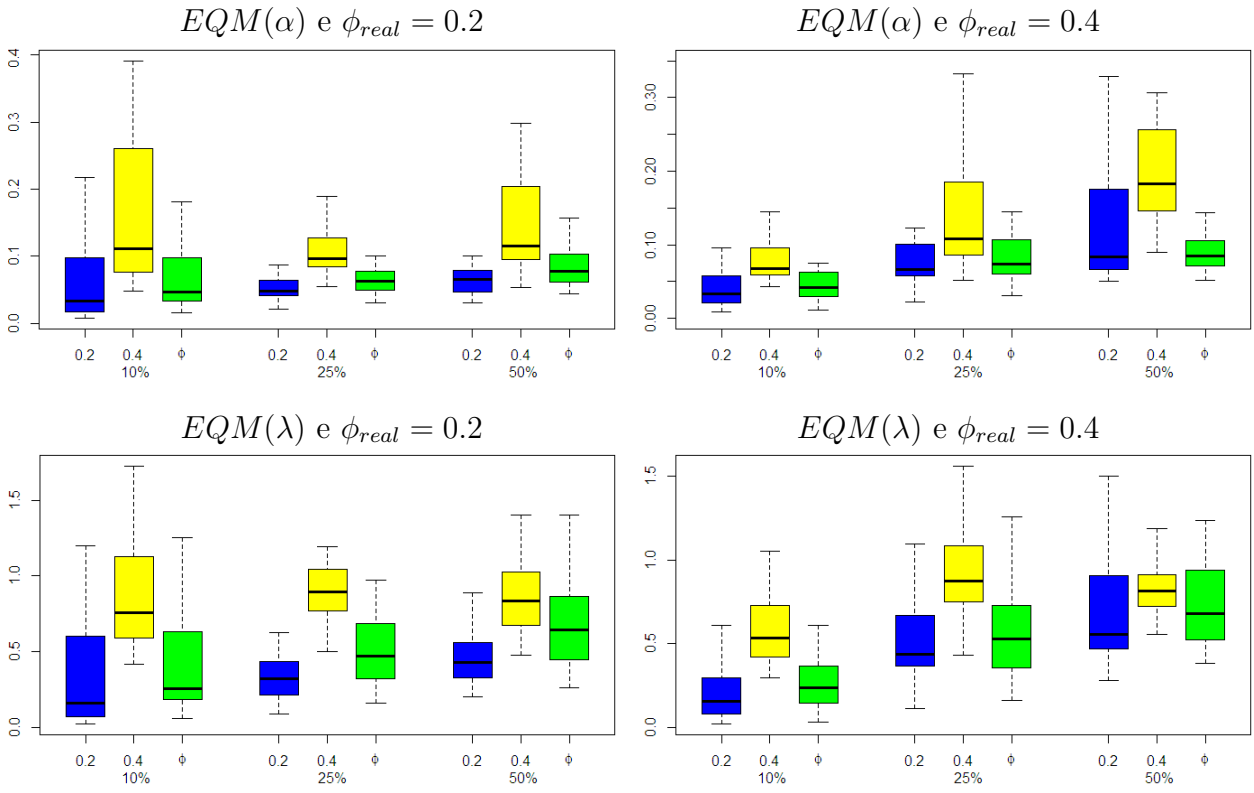

Figura 2.9: *Boxplots dos EQMs dos parâmetros α e λ para os casos onde há 10%, 25% e 50% de interações em F. Nos painéis à esquerda, estão os casos onde o valor real de ϕ é 0.2, enquanto que nos painéis à direita, estão os casos onde o valor real de ϕ é 0.4. Os boxplots azuis referem-se aos EQMs dos parâmetros para os Cenários $C_{0.2}^{0.2}$ e $C_{0.2}^{0.4}$. Os boxplots amarelos referem-se aos EQMs para os Cenários $C_{0.4}^{0.2}$ e $C_{0.4}^{0.4}$. Os boxplots verdes referem-se aos EQMs para os Cenários $C_{\hat{\phi}}^{0.2}$ e $C_{\hat{\phi}}^{0.4}$.*

Na Figura 2.10, observa-se os *boxplots* dos EQMs referentes as variâncias e as interações. Note que para o Cenário $C_{\hat{\phi}}^{0.2}$, os *boxplots* (em verde) dos EQMs de σ^2 são mais baixos em relação aos *boxplots* (em amarelo) dos EQMs no modelo do Cenário $C_{0.4}^{0.2}$, além disso, a variabilidade dos *boxplots* para o Cenário $C_{\hat{\phi}}^{0.2}$ são menores em relação as do Cenário $C_{0.4}^{0.2}$. No gráfico acima e a direita na Figura 2.10, é mostrado apenas os *boxplots* dos EQMs para os Cenários $C_{0.2}^{0.2}$ e $C_{\hat{\phi}}^{0.2}$. Veja que, devido a escala do gráfico ser muito pequena, a qualidade das estimativas das variâncias para o modelo do Cenário $C_{\hat{\phi}}^{0.2}$ são tão boas quanto as do modelo no Cenário $C_{0.2}^{0.2}$.

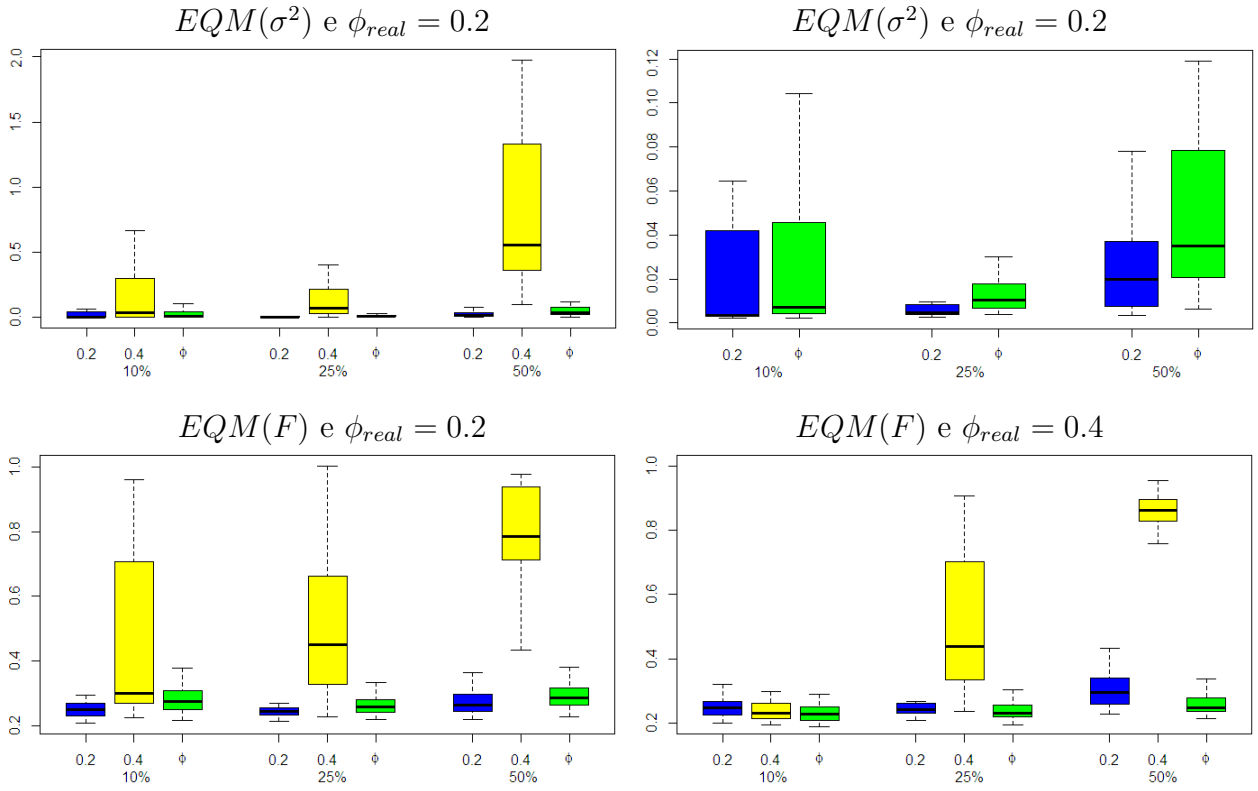

Figura 2.10: *Boxplots dos EQMs dos parâmetros σ^2 e F para os casos onde há 10%, 25% e 50% de interações em F. Nos painéis à esquerda estão os casos onde o valor real de ϕ é 0.2. Os boxplots azuis referem-se aos EQMs dos parâmetros para os Cenários $C_{0.2}^{0.2}$ e $C_{0.2}^{0.4}$. Os boxplots amarelos referem-se aos EQMs para os Cenários $C_{0.4}^{0.2}$ e $C_{0.4}^{0.4}$. Os boxplots verdes referem-se aos EQMs para os Cenários $C_{\hat{\phi}}^{0.2}$ e $C_{\hat{\phi}}^{0.4}$. Para facilitar a visualização devido a pequena escala, o painel acima e à direita mostra os boxplots dos EQMs de σ^2 dos Cenários $C_{0.2}^{0.2}$ (em azul) e $C_{\hat{\phi}}^{0.2}$ (em verde). O último painel à direita mostra os EQMs de F para os casos onde o valor real de ϕ é 0.4.*

Ainda na Figura 2.10 pode ser observado, a partir dos *boxplots*, que os EQMs das interações diante do Cenário $C_{\hat{\phi}}^{0.2}$ são mais baixos do que os EQMs das interações do modelo diante do Cenário $C_{0.4}^{0.2}$. Além disso, os *boxplots* (em verde) dos EQMs para as interações de todos os casos considerando o Cenário $C_{\hat{\phi}}^{0.2}$ apresentam menor variabilidade quando comparado com os casos do Cenário $C_{0.4}^{0.2}$. Veja também, que ao aumentar o número de interações em G_E (10%, 25% e 50%), tem-se que os EQMs de F no modelo

do Cenário $C_{0.4}^{0.2}$ também aumentam, enquanto que os EQMs de F para o modelo do Cenário $C_{\hat{\phi}}^{0.2}$ se mantém estáveis. Para os Cenários $C_{0.2}^{0.4}$, $C_{0.4}^{0.4}$ $C_{\hat{\phi}}^{0.4}$, pode-se observar que os EQMs de F são semelhantes para o caso 10%, pois apresentam *boxplots* com medianas próximas. Já para os demais casos, onde tem-se 25% e 50% de interações afetando G_E, nota-se um aumento dos EQMs de F para o Cenário $C_{0.4}^{0.4}$, enquanto que os EQMs para o modelo do Cenário $C_{\hat{\phi}}^{0.4}$ se mantém estáveis e são levemente menores do que os EQMs de F no caso $C_{0.2}^{0.4}$ considerando o caso de 50% de interações.

A Figura 2.11 apresenta os gráficos dos EQMs referentes as variâncias dos modelos para os Cenários $C_{0.2}^{0.4}$, $C_{0.4}^{0.4}$ e $C_{\hat{\phi}}^{0.4}$. O último painel à direita mostra os *boxplots* referentes aos casos de 25% e 50% juntos. Veja que, no caso onde há 10% de interações afetando G_E, os *boxplots* apresentam medianas próximas além de serem pequenas indicando boas estimativas para σ^2. Nos outros casos, quando há 25% e 50% de interações em G_E, nota-se um aumento dos EQMs de σ^2 diante do Cenário $C_{0.4}^{0.4}$, enquanto que para o modelo considerando os Cenários $C_{\hat{\phi}}^{0.4}$ e $C_{0.2}^{0.4}$, os EQMs parecem se manter estáveis. Entretanto, ao observar o último gráfico à direita, é possível notar que medida que há um aumento das interações em F de 25% para 50%, a variabilidade para o Cenário $C_{0.2}^{0.4}$ aumenta, enquanto que para o modelo do Cenário $C_{\hat{\phi}}^{0.4}$ ela se mantém estável.

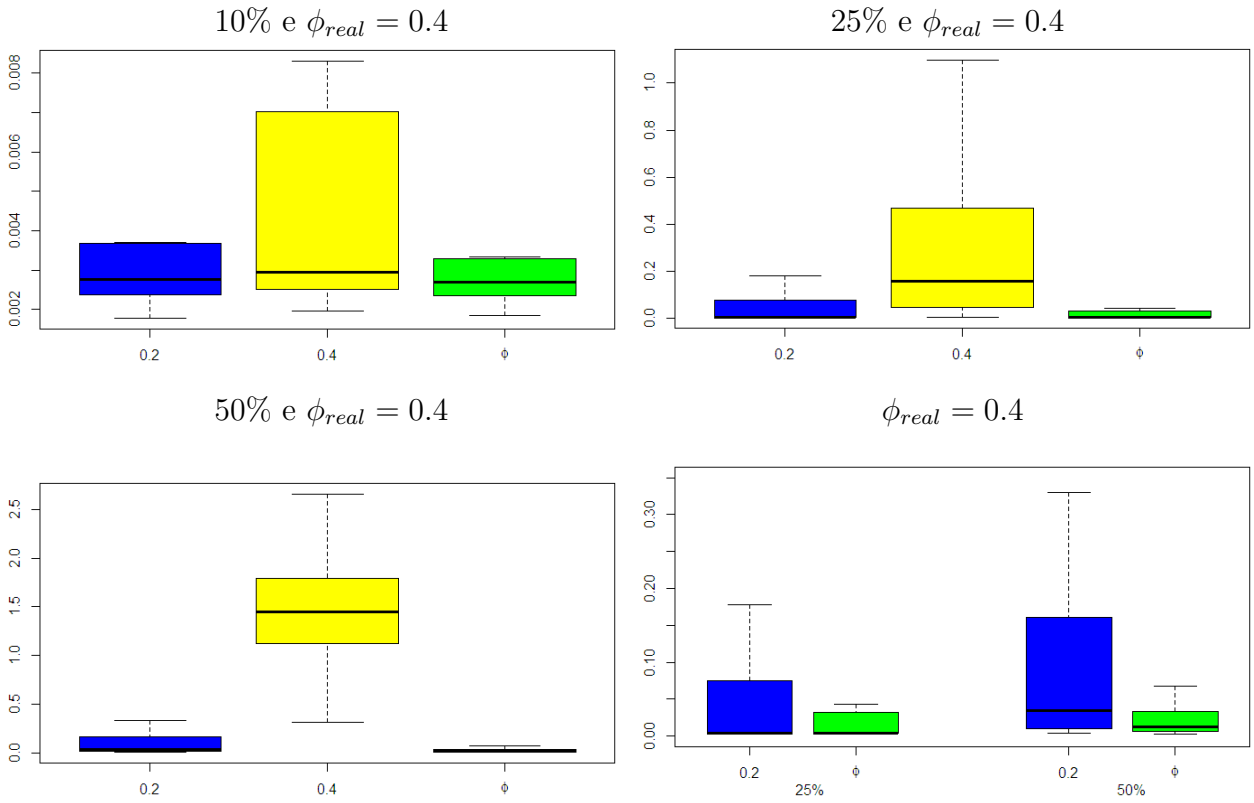

Figura 2.11: *Boxplots dos EQMs dos parâmetros σ^2 para os casos onde há 10%, 25% e 50% de interações em F. Nos painéis estão os casos onde o valor real de ϕ é 0.4, eles são mostrados separadamente para uma melhor visualização. Os boxplots azuis referem-se aos EQMs dos parâmetros para o Cenário $C_{0.2}^{0.4}$. Os boxplots amarelos referem-se aos EQMs para o Cenário $C_{0.4}^{0.4}$. Os boxplots verdes referem-se aos EQMs para o Cenário $C_{\hat{\phi}}^{0.4}$. O último painel à direita mostra os boxplots dos EQMs de σ^2 dos Cenários $C_{0.2}^{0.4}$ (em azul) e $C_{\hat{\phi}}^{0.4}$ (em verde) para os casos onde há 25% e 50% de interações em F.*

A Figura 2.12 mostra os *boxplots* das médias *a posteriori* de ϕ, observa-se que as medianas estão próximas uma das outras, concentrando seus valores em torno de 0.3 e este comportamento se reflete tanto na situação em que o verdadeiro ϕ é 0.2, como na situação onde seu valor verdadeiro é 0.4. Este comportamento está presente em todos os casos considerados (10%, 25% e 50%) da matriz F. Análises similares são feitas em relação aos EQMs deste parâmetro.

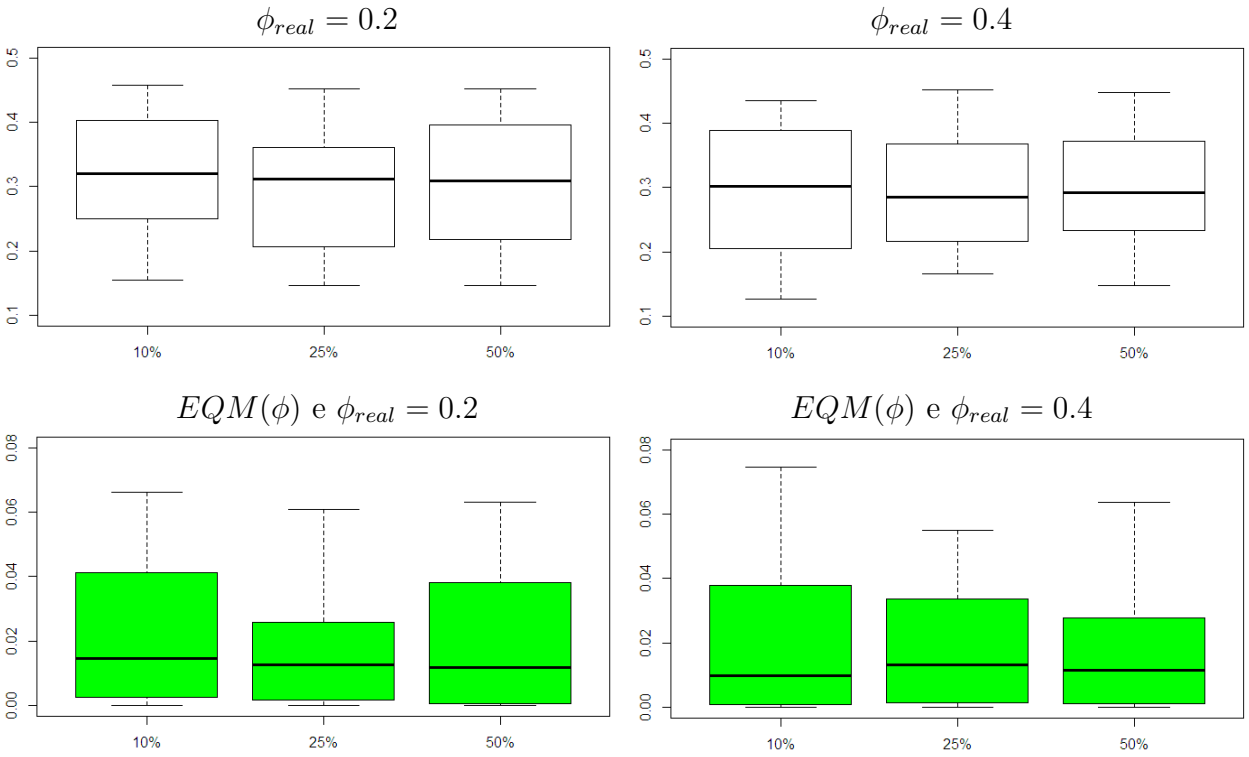

Figura 2.12: *Boxplots das médias e EQMs de ϕ para os casos onde há 10%, 25% e 50% de interações. Nos painéis a esquerda estão os casos onde o valor real de ϕ é 0.2, enquanto que a direita estão os casos onde o valor real de ϕ é 0.4. Os boxplots nos dois últimos painéis (em verde) referem-se aos EQMs de ϕ.*

2.4 Conclusões do Capítulo

Este capítulo fez uma breve descrição dos dados de expressão de genes, além do problema de CNA envolvendo regiões disjuntas do genoma onde grupos distintos de genes estão localizados, para os quais foram definidos os fatores do modelo. Em seguida, foi apresentado o modelo fatorial 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 em sua estrutura a função de covariâncias Gaussiana.

Foi estudado o ajuste e a qualidade das estimativas dos parâmetros do modelo fatorial com interações. Diferente do que foi visto por Mayrink e Lucas (2013), este estudo abor-

dou o comportamento do modelo considerando diferentes formatos de interações diante da estimatição do parâmetro de suavização ϕ. A partir das simulações, verificou-se que há uma vantagem na modelagem quando há estimação de ϕ. Pois os resultados sugerem que o ajuste do modelo fatorial com interações diante da estimação desse parâmetro, onde o valor de ϕ é desconhecido, pode proporcionar um melhor ajuste e trazer melhores estimativas para a maioria dos demais parâmetros e principalmente para os efeitos de interação. Portanto, esta etapa do trabalho acaba contribuindo e confirmando alguns resultados dos trabalhos desenvolvidos por Mayrink e Lucas (2013), dando suporte para futuros trabalhos que serão desenvolvidos usando o modelo fatorial com interações. O capítulo a seguir irá apresentar uma modelagem diferente para os efeitos de interação e o que será proposto é uma modelagem onde se pode fazer o agrupamentos delas.

Capítulo 3

Abordagem de Agrupamento das Interações

Estudos conhecidos por Rede Regulatória de Genes (Gene Regulatory Network - GRN) procuram relações entre genes que interagem um com o outro não somente de maneira individual, mas em conjunto com outros genes formando uma complexa rede de associações e interações. Uma rede de regulação genética descreve a relação entre pares de genes com base na sua dependência mútua de expressão. Redes reguladoras de genes são inferidas a partir de experimentos de *microarrays* de larga escala, frequentemente compostos de múltiplas condições observacionais ou experimentais [veja Emmert-Streib et al. (2014)]. Pesquisas que focam no desenvolvimento de métodos para estudar redes de genes a partir de *microarrays* são uma parte importante da bioinformática.

Visando investigar genes que interagem entre si, formando esta complexa associação, Mayrink e Lucas (2013) propuseram o modelo fatorial com interações, visto no capítulo anterior, para tentar explicar uma parte dessas associações, que seriam imperceptíveis atráves de modelos mais simples. Os autores modelaram os efeitos de interação a partir da distribuição *a priori* vista em (2.5) assumindo todos os $F_{i\bullet}$ significativos como diferentes entre si. Nessa situação, cada $F_{i\bullet}$ representaria o efeito de interação não linear de grupos de genes, afetados pela CNA, influenciando um determinado gene. Uma outra maneira de estimar as interações foi considerada pelos autores ao usarem a seguinte distribuição

a priori:

$$
\begin{aligned}
(F_{i\bullet}^{\top} \mid F^*) &\sim (1 - z_i)\delta_{\mathbf{0}}(F_{i\bullet}) + z_i\delta_{F^*}(F_{i\bullet}); \\
(F^* \mid \lambda, \phi) &\sim N_n[\mathbf{0}, K(\lambda, \phi)]; \\
(z_i \mid \rho_i) &\sim \text{Bernoulli}(\rho_i); \\
\rho_i &\sim \text{Beta}(\beta_1, \beta_2).
\end{aligned} \tag{3.1}
$$

Veja que foi introduzido, na primeira linha, uma componente degenerada no vetor F^* sendo $\delta_{F^*}(F_{i\bullet})$ que representa $p(F_{i\bullet} = F^*) = 1$. Em outras palavras, este caso explora uma situação onde a interação não linear é não significativa ou ela é a mesma afetando diversos genes. Conforme comentado por Mayrink e Lucas (2013), esta abordagem pode não ser realística. Uma segunda estratégia afim de explorar o conhecimento *a priori* da variável indicadora z_i, também foi abordada pelos autores, assumindo a seguinte configuração:

$$
\begin{aligned}
(z_i \mid \rho) &\sim \text{Bernoulli}(\rho); \\
\rho &\sim \text{Beta}(\beta_1, \beta_2).
\end{aligned}
$$

Nesta segunda estratégia é assumido uma probabilidade global para todas as interações representando a "força" da interação que afeta os genes. Uma vez que a distribuição *a posteriori* condicional completa de ρ é atualizada, ela leva em conta todos os z_i. Esta estratégia é diferente da configuração abordada em (3.1) para z_i, que considera na atualização da distribuição condicional completa de ρ_i apenas o z_i correspondente. Além do mais, os autores esperam poucas linhas de F contendo interações não nulas e argumentam que ρ tende a ser muito pequeno quando o número de genes m é muito grande. Estes casos favorecem $z_i = 0$ fazendo com que a esparsidade da matriz F seja maior do que o esperado.

A especificação *a priori* em (3.1) estabelece os efeitos de interação $F_{i\bullet}$ que não são zeros, como sendo iguais a F^*. Desta forma, o efeito de interação entre fatores é o mesmo para mais de um gene. Em resumo, o modelo fatorial com interações proposto em 2013, consegue identificar situações extremas em que a interação dos fatores para cada gene i em G_E são todas diferentes ou todas iguais. Isso induziu a pensar em uma

situação intermediária, onde seria possível identificar ou formar grupos de genes tal que as interações são as mesmas dentro de cada grupo, mas são diferentes entre os grupos. A proposta neste capítulo é formar grupos de genes para os quais o efeito de interação não linear $F_{i\bullet}$ seria o mesmo, representando a atuação conjunta de diversos genes, afetados pela CNA, que estão influenciando outros em localizações diferentes no genoma. Essa ideia é motivada pelos estudos de GRN que procuram essas relações complexas entre genes.

A organização deste capítulo é como segue: A Seção 3.1 apresenta uma proposta de modelagem para os efeitos de interação que estende o modelo apresentado por Mayrink e Lucas (2013), que usam a distribuição em (3.1) para as interações. A Seção 3.2 mostra um estudo de simulações feito com a análise fatorial diante da proposta indicada na Seção 3.1, apresentando os principais resultados obtidos com e sem a estimação do parâmetro ϕ no modelo. Na Seção 3.3 é exibido um estudo simulado extra comparando os modelos que são propostos aqui com a modelagem do capítulo anterior e o modelo fatorial sem interações. A Seção 3.4 encerra o capítulo mostrando suas principais conclusões.

3.1 Agrupamento das Interações via Mistura

Nesta etapa do trabalho, propõe-se estender o modelo fatorial com interações não lineares, visto em (2.1), através de uma nova abordagem para identificação das interações, ou seja, será feito uma modificação no modo de como as interações são estimadas. Uma maneira de fazer isso é formar grupos com as interações a partir de modelos de misturas finito. Para isso, considere a seguinte especificação *a priori* que é uma extensão de (3.1):

$$\begin{aligned}(F_{i\bullet}^{\top} \mid F_1^*, \cdots, F_R^*) &\sim z_{i0}\delta_{\mathbf{0}}(F_{i\bullet}) + z_{i1}\delta_{F_1^*}(F_{i\bullet}) + z_{i2}\delta_{F_2^*}(F_{i\bullet}) + \cdots + z_{iR}\delta_{F_R^*}(F_{i\bullet}); \quad (3.2)\\ (F_r^* \mid \lambda, \phi) &\sim N_n[\mathbf{0}, K(\lambda, \phi)];\ \text{com } r = 1, 2, \cdots, R.\end{aligned}$$

Sendo cada F_r^* representando um vetor linha $(1 \times n)$ e $(F_r^* \mid \lambda, \phi)$ são independentes. As interações F_r^*'s são distintas, mas com a mesma distribuição $N_n[\mathbf{0}, K(\lambda, \phi)]$. Veja em (3.2) que é de interesse classificar as interações segundo $R + 1$ componentes. Para cada efeito de interação $F_{i\bullet}$ na matriz F, será associado um vetor aleatório $z_i = (z_{i0}, z_{i1}, \cdots, z_{iR})$

indicando qual componente a interação $F_{i\bullet}$ pertence. Assuma:

$$z_{ir} = \begin{cases} 1, & \text{se } F_{i\bullet} \text{ pertence à componente } r; \\ 0, & \text{caso contrário.} \end{cases}$$

Tem-se que $\sum_{r=0}^{R} z_{ir} = 1$. Considere $\rho_{ir} = p(z_{ir} = 1)$ representando a probabilidade de que a interação não linear $F_{i\bullet}$ pertença à r-ésima componente da mistura. Através deste modelo, está sendo admitido que podem haver $R+1$ grupos de interações não lineares. Um destes grupos é formado pela interação nula. Os demais diferem entre si, mas cada um deles pode estar associado a diversos genes. Assim, ρ_{ir} seria o "peso de associação" da interação $F_{i\bullet}$ à r-ésima componente da mistura de distribuições. No contexto de expressão de genes, podemos dizer que ρ_{ir} é a probabilidade do gene i pertencer ao grupo r que é afetado pela interação F_r^*. Portanto, pode-se ter $R+1$ grupos cada um formado por m_r genes, onde cada um dos genes são impactados pela interação F_r^*.

A especificação *a priori* feita para os efeitos de interação pode ser dada pela seguinte representação hierárquica:

$$\begin{aligned} (F_{i\bullet}^{\top} \mid F_r^*, z_{ir} = 1) &\sim \delta_{F_r^*}(F_{i\bullet}); \text{ com } r = 0, \cdots, R; \\ (F_r^* \mid \lambda, \phi) &\sim N_n[\mathbf{0}, K(\lambda, \phi)]; \\ (z_i \mid \rho_i) &\sim \text{Mult}(1, \rho_i); \\ \rho_i &\sim \text{Dir}(\nu). \end{aligned} \tag{3.3}$$

Considere $\delta_{F_0^*} = \delta_{\mathbf{0}}$, $\nu = (\nu_0, \nu_1, \cdots, \nu_R)$ e ρ_i, $(z_i \mid \rho_i)$, $(F_{i\bullet}^{\top} \mid F_r^*, z_{ir} = 1)$ independentes. Além disso, assuma que Mult e Dir denotam as distribuições Multinomial e Dirichlet respectivamente. Uma segunda configuração também será adotada para z_i, assim como foi abordado por Mayrink e Lucas (2013) assuma: $(z_i \mid \rho) \sim \text{Mult}(1, \rho)$ e $\rho \sim \text{Dir}(\nu)$.

Problemas de identificabilidade envolvendo os efeitos de interação em (3.3) podem ocorrer. Esses problemas, conhecidos como *label switching*, são bastante vistos na literatura quando se trabalha com modelos de misturas [Stephens (2000) e Jasra et al. (2005)]. Tal problema prejudica a inferência dos parâmetros. Especificamente no modelo proposto, o *label switching* está ligado aos termos F_r^*'s. O problema de identificação refere-se ao fato de que não há nada na distribuição conjunta dos $F_{i\bullet}$ que possa distinguir cada componente r da mistura em (3.3). Uma estratégia usada para resolver esse

problema seria impor alguma restrição na estrutura da mistura. Uma alternativa comum é ordenar as médias das componentes. Entretanto, as compontentes utilizadas são distribuições degeneradas em vetores. Pode-se então ordenar os vetores F_r^* a partir de alguma medida como, por exemplo, a distância Euclidiana. Outra estratégia, que será adotada neste trabalho para contornar o problema, é ordenar os pesos $\rho_i = (\rho_{i0}, \rho_{i1}, \cdots, \rho_{iR})$. Para isso foi observado em cada iteração do algoritmo MCMC a matriz z formada pelas linhas $z_i = (z_{i0}, z_{i1}, \cdots, z_{iR})$. Então foram calculadas as quantidades $\frac{m_0}{m}, \frac{m_1}{m}, \cdots, \frac{m_R}{m}$ tal que $m_r = \sum_{i=1}^{m} z_{ir}$ com $r = 0, 1, \cdots, R$. Sendo que m é o número de genes (linhas) em X. A medida m_r representa o número de interações $F_{i\bullet}$ iguais a F_r^* na iteração do algoritmo. Assim, no contexto de expressão de genes, a quantidade $\frac{m_r}{m}$ representa a proporção de genes que são afetados pela interação F_r^*. Diante disso, as proporções serão ordenadas de forma decrescente $\frac{m_0}{m} > \frac{m_1}{m} > \cdots > \frac{m_R}{m}$, assim como os pesos ρ_i, os z_i's e os F_r^*'s correspondentes. Os resultados das simulações que serão exibidos a frente mostram que o problema de *label switching* foi solucionado para esta modelagem. Outros detalhes que abordam problemas de identificação em modelos de mistura podem ser vistos nos trabalhos de Fruhwirth-Schnatter (2001) e Celeux et al. (1999).

3.2 Estudo Simulado

Assim como feito no capítulo anterior, a matriz de dados X usada neste estudo terá tamanho $m = 200$ e $n = 100$. Será usado um modelo com $L = 2$ fatores assumindo que cada fator l tenha uma relação direta com cada grupo de genes em G_l ($l = 1$ ou 2). Os grupos G_1 e G_2 contém 10 genes cada, enquanto que em G_E há 180 genes. Para este estudo também será usada a função de covariâncias Gaussiana com o valor real de $\upsilon = 1$ e $\phi = 0.2$. O objetivo neste capítulo é estudar o modelo de agrupamento via misturas para as interações observando como o modelo fatorial se comporta ao estimar as interações F_r^*. Serão realizados ajustes de modelos fixando e estimando o valor de ϕ. Partindo da motivação criada pelas GRN, é razoável pensar no agrupamento das interações. Por isso, a matriz de dados X é gerada usando a seguinte configuração:

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 $u < 0.7$ para $i \in G_E$ e todo l.

Supõe-se em média 70% de cargas significativas.

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

3. Para a matriz de interações, foram considerados 4 casos. Aqui é mostrado um deles correspondendo a $R = 2$ grupos não nulos ($R > 2$ será comentado adiante).

Gere $F_{i\bullet} = \mathbf{0}$ para todo $i \in (G_1 \cup G_2)$. Gere F_1^* e F_2^* da $N_n[\mathbf{0}, K(\lambda, 0.2)]$. Das 180 linhas em G_E seleciona-se aleatoriamente 90 delas, para as quais 45 representam F_1^* e 45 representam F_2^*. Desta forma, a matriz F tem 90 linhas diferentes de zero sendo elas de dois tipos.

4. Gere $\epsilon_{ij} \sim N(0, \sigma_i^2)$, sendo $\sigma_i^2 \sim U(0.2, 0.4)$.

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

Foram gerados 4 tipos de matrizes de dados X. Elas se diferenciam pela forma como foi estruturada a matriz F. O estudo feito com 50% de interações em F (90 linhas) permite fazer uma análise mais simplificada para o presente trabalho. Os Demais casos envolvendo 10% ou 25% de interações em F, como feitos no Capítulo 2, são deixados para trabalhos futuros.

A atribuição de 45 linhas do tipo F_1^* e 45 do tipo F_2^* em F (totalizando 50% de interações) é conveniente visto que a princípio não se quer avaliar as quantidades de interações de cada tipo e sim se F_1^* e F_2^* foram bem estimadas. Se fosse colocado uma quantidade maior de F_1^* do que F_2^*, espera-se diferentes qualidades de estimação destes vetores.

A Tabela 3.1 apresenta os cenários feitos para 4 tipos de conjuntos de dados gerados e cada valor de ϕ usado no ajuste do modelo. Na notação utilizada para o Cenário $C_{0.2}^{2F^*}$ o sobrescrito $2F^*$ indica o tipo de dado gerado assumindo apenas duas interações não nulas, e o subscrito 0.2 indica o tipo de modelagem utilizada (fixando $\phi = 0.2$). Em todos os cenários o ajuste do modelo fatorial é feito usando a especificação *a priori* em (3.3) com $R = 5$ grupos não nulos. A segunda estratégia usada para o vetor z_i também é avaliada nas análises, mas apenas para os casos onde é feito a estimação de ϕ. O

Cenário $C^{2F^*}_{\hat{\phi}\rho_i}$ indica que os dados foram gerados com duas interações não nulas (F_1^* e F_2^*) e o modelo ajustado considera a estimação de ϕ com a configuração usada para z_i sendo ρ_i. É importante ressaltar que para todos os ajustes estará sendo utilizado um modelo assumindo uma mistura com 5 interações não nulas (denotamos esta informação por $5F^*$). Veja que na prática não se sabe a quantidade de interações, então a estratégia seria ajustar um modelo com R grande e avaliar seu desempenho para situações com poucas ou muitas interações.

Os conjuntos de dados representados por $3F^*$ apresentam uma matriz F contendo 90 linhas selecionadas aleatoriamente das quais 30 são F_1^*, 30 linhas F_2^* e 30 são do tipo F_3^*. Nos dados representados por $4F^*$ a matriz F foi gerada com 90 linhas selecionadas aleatoriamente, das quais 23 são F_1^*, 23 F_2^*, 22 são F_3^* e 22 são F_4^*. Para os dados gerados no caso $5F^*$, a matriz F contém 90 linhas de interações com 18 linhas para cada um dos cinco tipos de interação.

Tabela 3.1: Tipos de dados simulados e estratégias para ϕ e z_i usados em cada ajuste do modelo.

Tipos de dados	Estratégia	Modelos
$2F^*$	$\phi_{fixo} = 0.2$	$C_{0.2}^{2F^*}$
	$\phi_{fixo} = 0.3$	$C_{0.3}^{2F^*}$
	$\phi_{fixo} = 0.4$	$C_{0.4}^{2F^*}$
	$(\hat{\phi},\ \rho_i)$	$C_{\hat{\phi}}^{2F^*}$
	$(\hat{\phi},\ \rho)$	$C_{\hat{\phi}\rho}^{2F^*}$
$3F^*$	$\phi_{fixo} = 0.2$	$C_{0.2}^{3F^*}$
	$\phi_{fixo} = 0.3$	$C_{0.3}^{3F^*}$
	$\phi_{fixo} = 0.4$	$C_{0.4}^{3F^*}$
	$(\hat{\phi},\ \rho_i)$	$C_{\hat{\phi}\rho_i}^{3F^*}$
	$(\hat{\phi},\ \rho)$	$C_{\hat{\phi}\rho}^{3F^*}$
$4F^*$	$\phi_{fixo} = 0.2$	$C_{0.2}^{4F^*}$
	$\phi_{fixo} = 0.3$	$C_{0.3}^{4F^*}$
	$\phi_{fixo} = 0.4$	$C_{0.4}^{4F^*}$
	$(\hat{\phi},\ \rho_i)$	$C_{\hat{\phi}\rho_i}^{4F^*}$
	$(\hat{\phi},\ \rho)$	$C_{\hat{\phi}\rho}^{4F^*}$
$5F^*$	$\phi_{fixo} = 0.2$	$C_{0.2}^{5F^*}$
	$\phi_{fixo} = 0.3$	$C_{0.3}^{5F^*}$
	$\phi_{fixo} = 0.4$	$C_{0.4}^{5F^*}$
	$(\hat{\phi},\ \rho_i)$	$C_{\hat{\phi}\rho_i}^{5F^*}$
	$(\hat{\phi},\ \rho)$	$C_{\hat{\phi}\rho}^{5F^*}$

Para cada um dos 4 tipos de conjuntos de dados foram feitas 50 replicações Monte Carlo ajustando o modelo fatorial considerando situações em que o parâmetro ϕ é fixado ou estimado. Será avaliado o comportamento do modelo fatorial diante da estimação de ϕ a medida que há um aumento no número de tipos de interações que afetam G_E, ou seja, estará sendo observado o desempenho do modelo quando há um aumento dos grupos formados pelas interações F_r^*. As avaliações da qualidade das estimativas dos

parâmetros α, λ, σ^2 e F, são feitas por meio dos EQM's construídos usando a média *a posteriori*. Para a qualidade dos ajustes foram usadas razões construídas com os critérios DIC, WAIC e LPML (assim como foi feito no capítulo anterior). Também será avaliado a proporção de acerto que os modelos tem para identificar corretamente cada tipo de interação F_r^* em G_E, diante das situações onde o valor de ϕ é fixado ou estimado, além disso, serão utilizadas as duas estratégias abordadas para o vetor z_i.

As distribuições *a priori* que foram usadas para α, σ^2, λ e ϕ estão em (2.2), (2.3), (2.4) e (2.7), respectivamente, e são as mesmas vistas no capítulo anterior, assim como a distribuição *a priori* de q_{il} que está na Tabela 2.2. Para ρ_i considere a configuração *a priori* apresentada na Tabela 3.2 sendo $(1, 0, \cdots, 0)$ e $(1, 1, \cdots, 1)$ vetores linhas de dimensão $[1 \times (R+1)]$. A escolha do vetor $(1, 0, \cdots, 0)$ é feita para que a componente $\delta_{\mathbf{0}}(F_{i\bullet})$ tenha peso $\rho_{i0} = 1$ enquanto que as demais componentes $\delta_{F_r^*}(F_{i\bullet})$ tenham pesos $\rho_{ir} = 0$ para $r = 1, \cdots, R$. Relembrando o que foi dito nas Seções 2.1 e 2.2 no Cápitulo 2. O que se quer é a toral identificação dos grupos com seus fatores. Por isso, as especificações *a priori* exibidas na Tabela 3.2 determinam que nos grupos G_1 e G_2, não haja interação dos fatores, e que os grupos sejam explicados ou identificados apenas por cada um dos fatores adotados na modelagem, isto é, o fator 1 apenas represetará os genes em G_1 e o fator 2 representará os genes em G_2. Veja que a distribuição Dirichlet adotada corresponde a distribuição Beta usada em casos mais simples, quando há duas componentes, como em (3.1). Outra alternativa, seria atribuir uma distribuição $\text{Dir}(\tau_0, \tau_1, \cdots, \tau_R)$ com $\tau_r < 1$. Essa especificação faria o modelo tomar uma decisão mais direta, não deixando dúvidas a respeito de classificar as interações estimadas para uma das componentes (grupos), porém esta configuração mencionada não será adotada aqui.

Tabela 3.2: Especificação *a priori* utilizada para ρ_i e ρ.

índices	ρ_i
$i \in G_1$	$p[\rho_i = (1, 0, \cdots, 0)] = 1$
$i \in G_2$	$p[\rho_i = (1, 0, \cdots, 0)] = 1$
$i \in G_E$	$\text{Dir}(1, 1, \cdots, 1)$

A partir da regra de Bayes foi obtido o núcleo da distribuição *a posteriori* que por não ser analiticamente tratável, requer o uso do algoritmo *Gibbs Sampling* com passos usando o Metropolis-Hastings para gerar $\lambda_{\bullet j}$ e ϕ. As distribuições condicionais completas estão apresentadas no Apêndice B. Neste estudo, foi considerado um total de 4000 iterações na execução do MCMC onde as 2500 primeiras amostras foram usadas como *burn in*. A convergência foi observada em todos os parâmetros selecionados para inspeção.

Na Figura 3.1 observa-se os *boxplots* das medidas com as razões dos DICs, WAICs e LPMLs dos modelos onde os valores de ϕ são fixados e estimado. Veja que a maior parte dos gráficos está acima do 1, indicando que no denominador dessas razões, onde tem-se o modelo do Cenários $\phi_{fixo} = 0.2$, estão os menores DIC, WAIC e LPML. Nota-se, pelos *boxplots* das medidas, que o modelo ajustado se comporta relativamente bem diante dos tipos de dados. Pode-se observar que os *boxplots* apresentam medianas muito próximas do 1 indicando que o DIC, WAIC e LPML dos modelos $\phi_{fixo} = 0.3$, $\phi_{fixo} = 0.4$ e ϕ estimado, são próximos do DIC, WAIC e LPML do modelo do Cenário $\phi_{fixo} = 0.2$ (que é o modelo gerador dos dados). Os gráficos mostrados nos painéis à direita são os mesmos que estão nos painéis da esquerda com exceção do caso $5F^*$, eles são apresentados para melhor visualização. A partir dos *boxplots*, nota-se também que de certa forma há um distanciamento do 1, mas apesar de visualmente apresentarem essas diferenças, elas não são tão grandes devido a escala dos gráficos no eixo vertical serem pequenas. Note que no caso dos conjuntos de dados $5F^*$, os *boxplots* dos modelos ajustados também apresentam medianas próximas de 1, entretanto pode-se observar que os *boxplots*(em verde) para o modelo ajustado diante da estimação de ϕ com ρ_i, apresentam uma variabilidade menor do que nos demais. Esses breves resultados indicam que pode haver alguma vantagem da modelagem de agrupamento diante da estimação de ϕ.

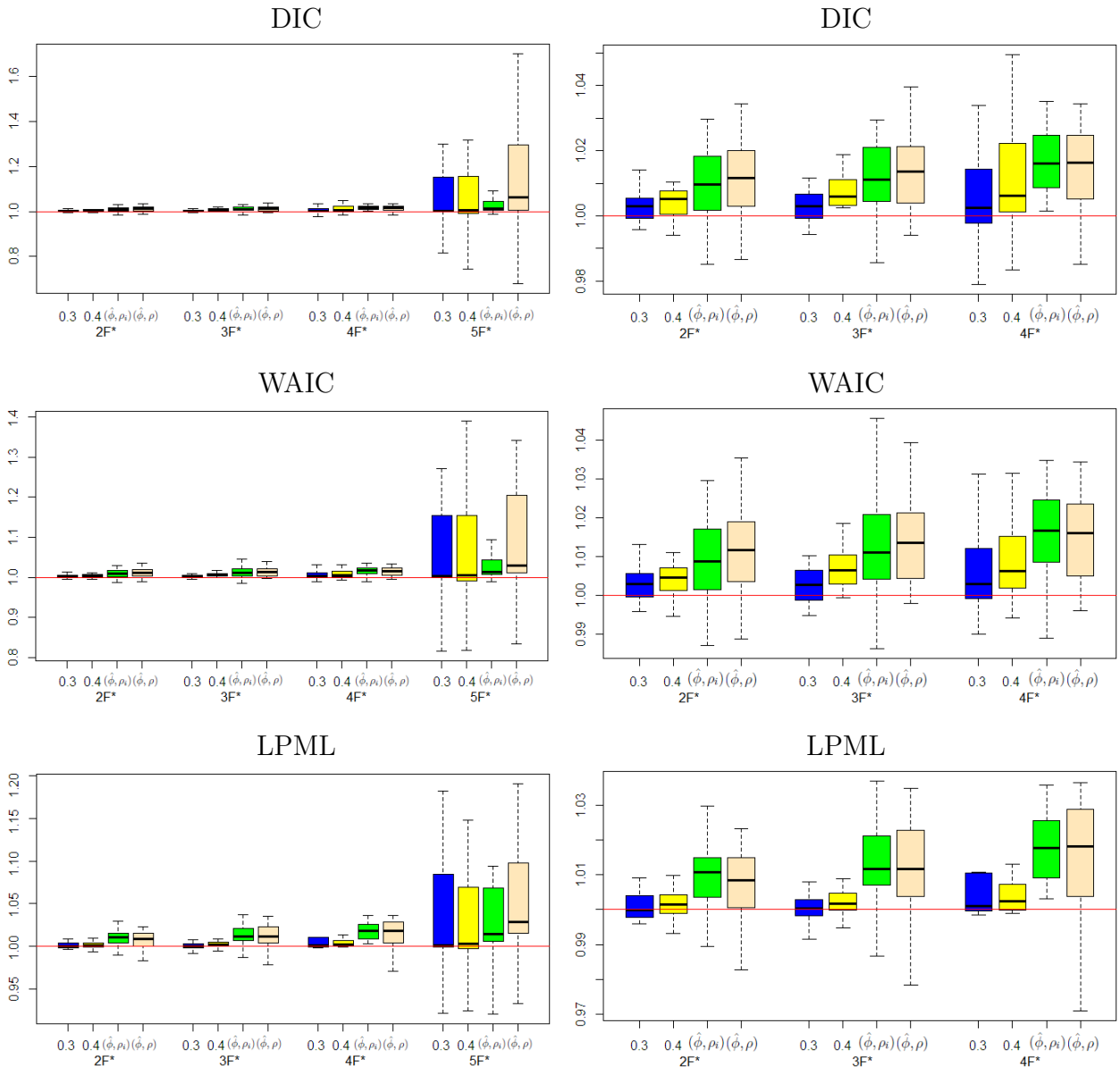

Figura 3.1: *Boxplots das razões dos DIC, WAIC e LPML para os tipos de dados* $2F^*$, $3F^*$, $4F^*$ *e* $5F^*$ *considerando* ϕ *fixo e estimado. Nos painéis à direita estão os tipos de dados* $2F^*$, $3F^*$ *e* $4F^*$ *para melhor visualização. Os boxplots verdes e rosas referem-se as medidas do modelo onde há a estimação de* ϕ *diante das estratégias 1 e 2 consideradas para* z_i *respectivamente. Os boxplots amarelos referem-se as medidas do modelo onde* $\phi_{fixo} = 0.4$. *Os boxplots azuis referem-se as medidas do modelo onde* $\phi_{fixo} = 0.3$. *A linha na horizontal representa o valor 1.*

A Figura 3.2 apresenta os *boxplots* dos EQMs de α, λ e F do modelo para cada cenário. Observa-se a partir dos *boxplots* (em cinza), que a maioria dos EQMs dos parâmetros α, λ e F são mais baixos nos Cenários $C_{0.2}^{2F^*}$, $C_{0.2}^{3F^*}$, $C_{0.2}^{4F^*}$ e $C_{0.2}^{5F^*}$. Esses resultados eram esperados, já que estamos fixando o parâmetro ϕ no seu valor real. Entretanto, ao se analisar os demais casos é possível notar pelas medianas dos *boxplots* (em verde), que os EQMs de α e λ, nos Cenários $C_{\hat{\phi}\rho_i}^{2F^*}$, $C_{\hat{\phi}\rho_i}^{3F^*}$, $C_{\hat{\phi}\rho_i}^{4F^*}$ e $C_{\hat{\phi}\rho_i}^{5F^*}$ são menores em relação aos demais quando ϕ é fixo em 0.3 ou 0.4. Ao observar os *boxplots* dos EQMs das interações, nota-se que o comportamento deles são mais estáveis, apresentando medianas bem próximas em quase todos os casos. Veja que os EQMs de F no *boxplot* do Cenário $C_{\hat{\phi}\rho_i}^{5F^*}$ apresentam uma variabilidade menor em relação aos Cenários $C_{0.3}^{5F^*}$ e $C_{0.4}^{5F^*}$, além do mais, é possível observar no gráfico que as medianas nos Cenários $C_{\hat{\phi}\rho_i}^{5F^*}$, $C_{\hat{\phi}\rho}^{5F^*}$ e $C_{0.2}^{5F^*}$ são bem próximas indicando boas qualidades das estimativas para as interações.

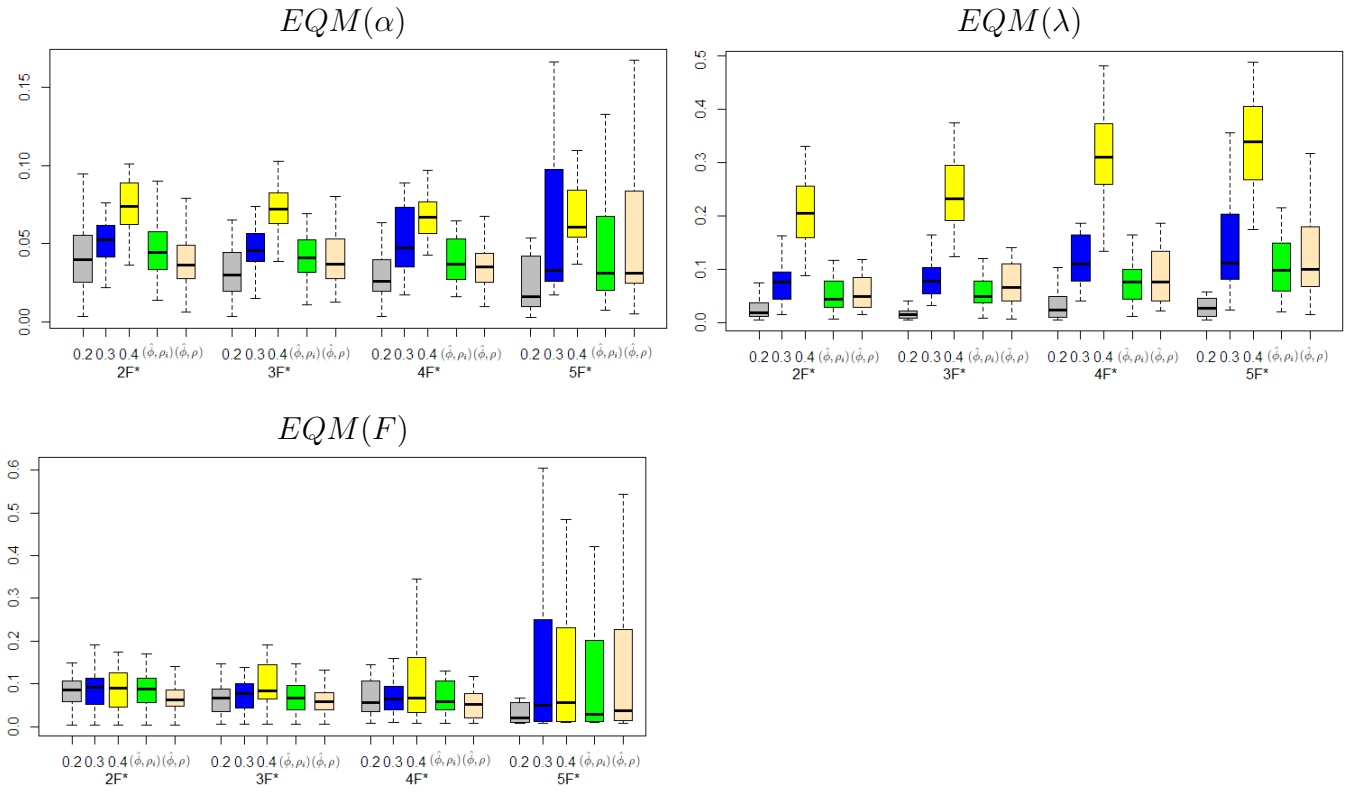

Figura 3.2: *Boxplots dos EQMs dos parâmetros α, λ e F em todos os cenários. Os boxplots em cinza mostram os EQMs dos parâmetros modelando com $\phi_{fixo} = 0.2$. Os boxplots azuis mostram os EQMs dos parâmetros modelando com $\phi_{fixo} = 0.3$. Os boxplots amarelos mostram os EQMs dos parâmetros modelando com $\phi_{fixo} = 0.4$. Os boxplots verdes e rosas mostram os EQMs dos parâmetros no modelo estimando ϕ diante das estratégias 1 e 2 consideradas para z_i respectivamente.*

Na Figura 3.3, pode-se observar os *boxplots* referentes aos EQMs de σ^2 para cada tipo de dado. Veja que a posição dos *boxplots* está em uma região próxima de zero e que as escalas dos gráficos são bem pequenas indicando boas estimativas para as variâncias. Observa-se que as medianas dos *boxplots* em todos os cenários estão bem próximas. Nota-se também que, a pesar da escala dos gráficos serem pequenas, a variabilidade no *boxplot* do Cenário $C^{5F^*}_{\hat{\phi}\rho_i}$ é menor em relação aos demais cenários.

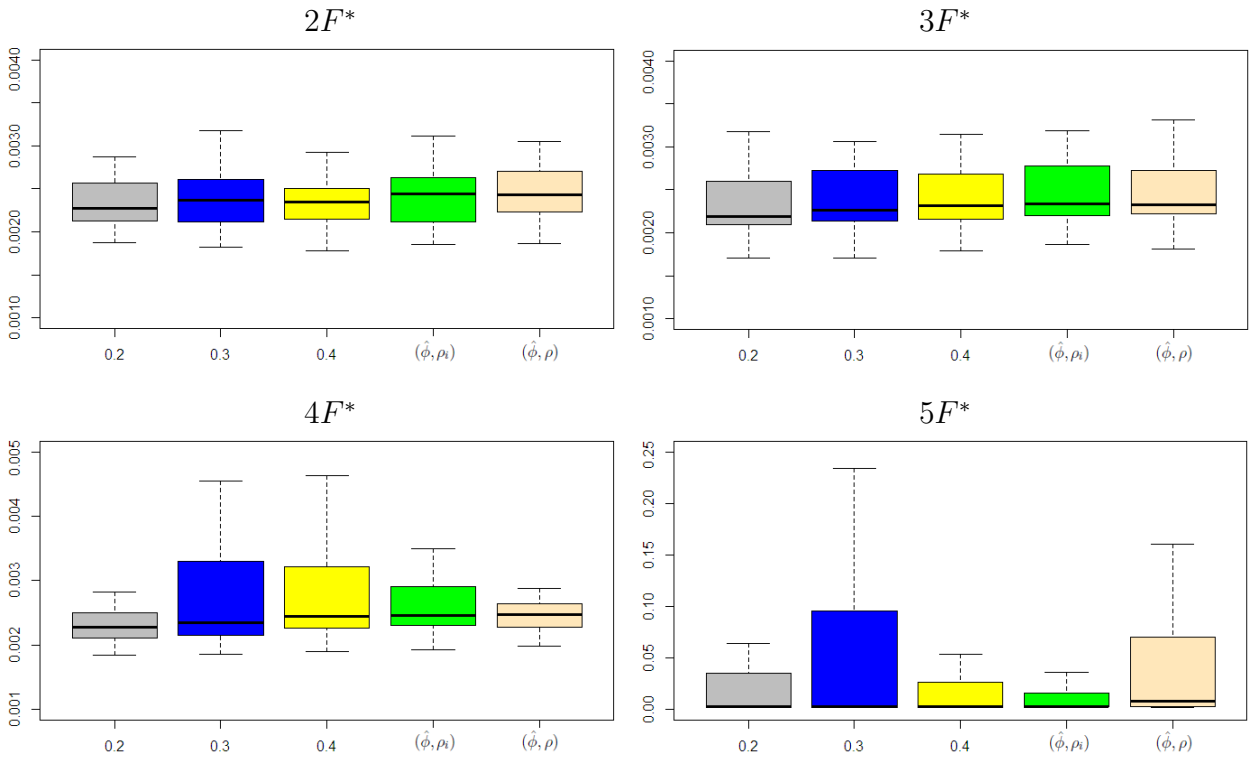

Figura 3.3: *Boxplots dos EQMs de σ^2 para cada tipo de conjunto de dados. Os boxplots em cinza mostram os EQMs do parâmetros modelando com $\phi_{fixo} = 0.2$. Os boxplots azuis mostram os EQMs dos parâmetros modelando com $\phi_{fixo} = 0.3$. Os boxplots amarelos mostram os EQMs dos parâmetros modelando com $\phi_{fixo} = 0.4$. Os boxplots verdes e rosas mostram os EQMs dos parâmetros no modelo estimando ϕ diante das estratégias 1 e 2 consideradas para z_i respectivamente.*

A Figura 3.4 apresenta os *boxplots* com as médias *a posteriori* e os EQMs de ϕ referentes ao modelo fatorial com interações ajustado para cada tipo de conjunto de dados. Pode-se observar, no painel à esquerda, que as médias apresentam uma certa estabilidade em torno do valor 0.3 e este comportamento se reflete para cada um dos tipos de dados analizados. Esse resultado acaba conduzindo a pensar que seria razoável ajustar o modelo fatorial com interações considerando o valor de ϕ fixado em 0.3. Entretanto, todos os resultados vistos até aqui, indicam que há uma vantagem em ajustar o modelo fatorial diante da estimação do parâmetro de comprimento-escala da função de covariâncias Gaussiana. Pois, foi mostrado que a qualidade das estimativas dos parâmetros α, λ, σ^2 e F, para os modelos onde há a estimação de ϕ, são em várias situações melhores do que as estimativas dos parâmetros dos modelos com $\phi_{fixo} = 0.3$ ou $\phi_{fixo} = 0.4$. Estas análises revelam o bom desempenho do modelo fatorial com agrupamento das interações, sugerindo que, em situações onde não se conhece o valor real de ϕ, ajustar o modelo fatorial com a estimação de ϕ pode ser a melhor opção.

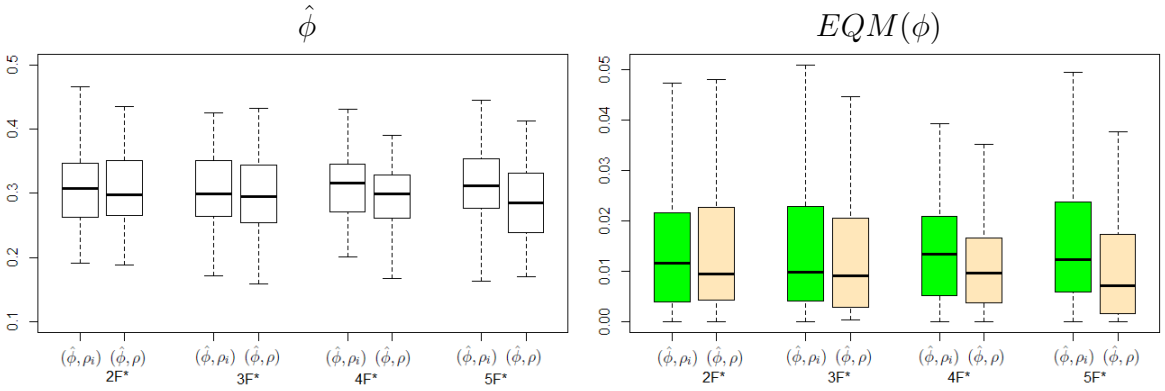

Figura 3.4: *Boxplots das médias e EQMs de ϕ dos modelos ajustados para cada tipo de conjunto de dados. O painél à esquerda exibe os boxplots com as médias a posteriori enquanto que o painél à direita apresenta os boxplots com os EQMs de ϕ diante das estratégias 1 e 2 consideradas para z_i.*

Para avaliar o nível de acerto que o modelo tem ao identificar corretamente cada interação F_r^* $(r = 0, \cdots, R)$ foi calculado a proporção de acerto obtida com as linhas da matriz F. Em cada amostra Monte Carlo é avaliado a matriz estimada $\hat{F}$ juntamente

com a matriz real F. Identificam-se em G_E as linhas de $\hat{F}$ correspondentes às linhas de F em que as interações F_r^* foram estimadas corretamente. Neste caso, os estudos indicam que foram recuperados muito bem os valores dos F_r^*, pois os $\hat{F}_r^*$ são semelhantes aos F_r^* ($\hat{F}_{i\bullet} = \hat{F}_r^* \approx F_r^* = F_{i\bullet}$), porém em algumas linhas esse resultado não vai estar correto. Por exemplo, suponha que na primeira amostra Monte Carlo, considerando o tipo de dado $2F^*$, observa-se na matriz F a linha $F_{25\bullet} = F_1^*$ e ao verificar a matriz $\hat{F}$ foi observado a linha $\hat{F}_{25\bullet} = \hat{F}_1^*$, isso será considerado acerto. Porém, se fosse observado em $\hat{F}$ a linha $\hat{F}_{25\bullet} = \mathbf{0}$ ou $\hat{F}_{25\bullet} = \hat{F}_2^*$ ou $\hat{F}_{25\bullet} = \hat{F}_3^*$ ou $\hat{F}_{25\bullet} = \hat{F}_4^*$ ou $\hat{F}_{25\bullet} = \hat{F}_5^*$ seria computado um erro. O que se quer dizer quando é mensionado que a linha $\hat{F}_{i\bullet} = \hat{F}_r^*$ é semelhante ou próxima da linha $F_{i\bullet} = F_r^*$, é que quando foi analizado o gráfico com as médias *a posteriori* de $\hat{F}_i = \hat{F}_r^*$, observou-se que 80% ou mais dos $\hat{F}_r^*$ continham o verdadeiro valor F_r^*, diante desta situação diz-se que a estimação de $\hat{F}_r^*$ é boa. A Figura 3.5 exibe dois gráficos para entendimento visual do que foi dito. Veja que o painel à esquerda mostra mais de 80% dos intervalos *Highest Posterior Density* (HPD) contendo o verdadeiro valor de F_1^*, observe que os comprimentos dos intervalos não são longos e as médias *a posteriori* de $\hat{F}_1^*$ estão próximas de seu valor real. No painel à direita observa-se uma estimação ruim de F_1^* exibindo a maioria dos intervalos HPD não contendo o verdadeiro valor de F_1^*.

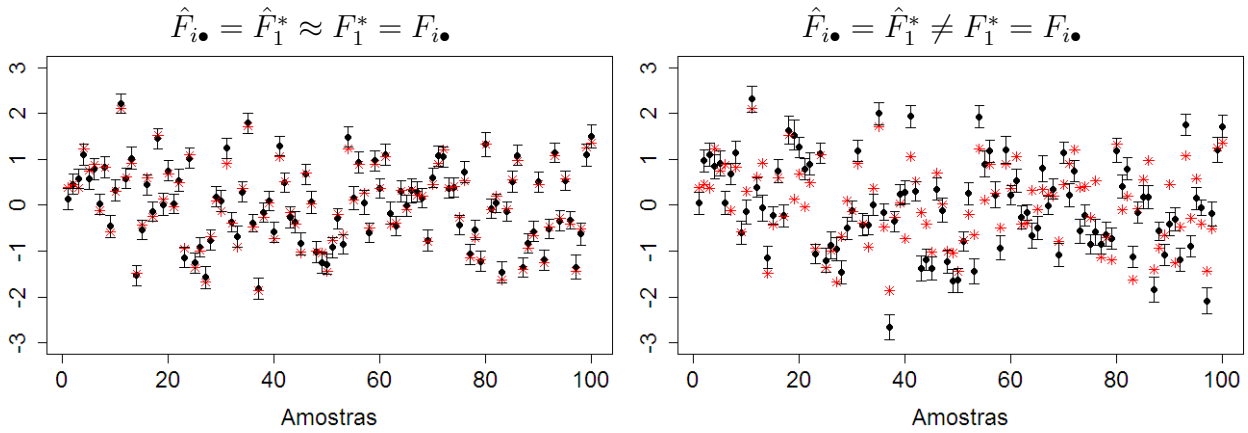

Figura 3.5: *Gráfico com valores reais (asterísco) e médias a posteriori (círculo) de* F_1^*. *O intervalo HPD é representado pelo segmento de reta na vertical. Estes gráficos representam estimativas para um único ajuste da modelagem feita usando o estudo Monte Carlo.*

Neste caso está sendo considerando como acerto o par $(F_{i\bullet} = F_r^*,\ \hat{F}_{i\bullet} = \hat{F}_r^*)$ com $\hat{F}_r^* \approx F_r^*, \forall\ i \in G_E$ e $r = 0, 1, \cdots, R$. Assuma que $F_0^* = \mathbf{0}$. Portanto, será contado o número de acertos em seguida esse resultado será divido pelo total de linhas de G_E. Desta forma tem-se a proporção de interações identificadas corretamente na primeira amostra Monte Carlo.

Na Figura 3.6 são apresentados os *boxplots* das proporções Monte Carlo para cada cenário. Podem ser observadas proporções de acerto altas (acima de 70%) para todos os tipos de dados mostrando o bom desempenho do modelo fatorial ajustado com 5 tipos de interações. Isso indica que não é preciso ajustar o modelo tendo uma noção de quantos tipos de interações F_r^* existem, pois nas simulações foi observado que para os tipos de dados $2F^*$, $3F^*$ e $4F^*$, as componentes extras da mistura adotada no ajuste apresentavam pesos muito baixos ou nulos. Conforme o esperado, as proporções de acertos são maiores nos cenários onde ϕ é fixo no valor real. Observe que as proporções de acerto nos cenários onde há a estimação ϕ, tanto usando a estratégia 1 quanto a 2 para z_i, são maiores que as dos demais cenários com ϕ fixo em 0.3 ou 0.4. Isso mostra que o modelo estimando ϕ tem um desempenho melhor para identificar corretamente as interações estimadas em relação aos modelos com ϕ fixo em 0.3 ou 0.4. Veja que todas as proporções aumentam

a medida em que há um aumento no número de tipos de interações contidas nos dados. Este comportamento reflete o fato de que há uma melhora na estimação ao aproximar o número de interações não nulas ao número de tipos adotados na modelagem (sempre $5F^*$). Note que no Cenário $C_{0.2}^{5F^*}$ a proporção de acerto é quase 1, indicando que o modelo ajustado consegue identificar corretamente quase todas as interações referentes ao tipo de dado $5F^*$.

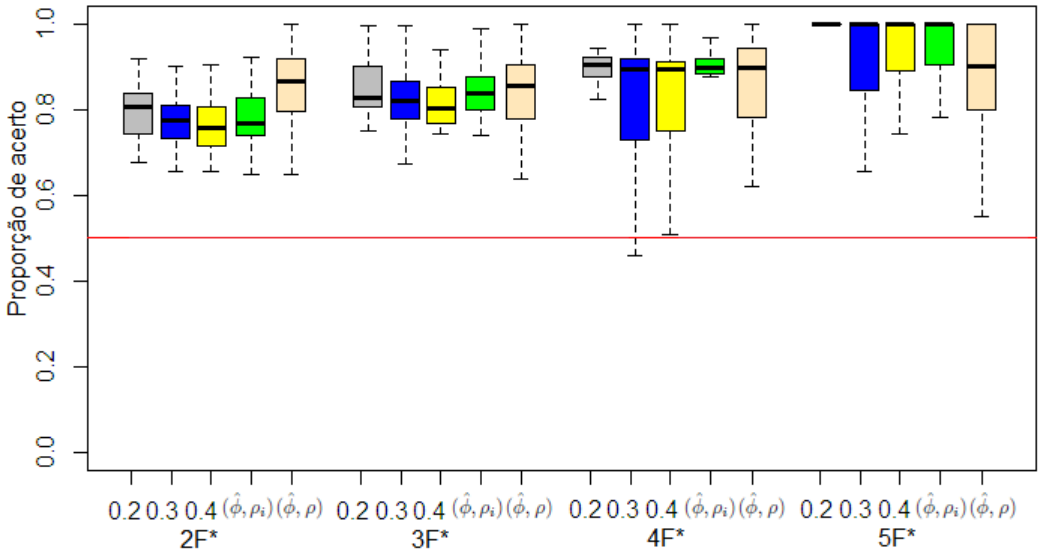

Figura 3.6: *Gráfico das proporções de acerto na identificação correta das interações em cada cenário. Os boxplots em cinza mostram as proporções de acerto para o modelo onde* $\phi_{fixo} = 0.2$. *Os boxplots em azul mostram as proporções de acerto para o modelo onde* $\phi_{fixo} = 0.3$. *Os boxplots em amarelo mostram as proporções de acerto para o modelo onde* $\phi_{fixo} = 0.4$. *Os boxplots verdes e rosas mostram as proporções de acerto para o modelo estimando* ϕ *diante das estratégias 1 e 2 consideradas para* z_i *respectivamente. A linha na horizontal representa o valor 0.5.*

3.3 Estudo Simulado Extra

Nesta Seção será avaliado como o modelo fatorial usado no Capítulo 2 (denotado por F^{2013}) e o modelo sem interação (denotado por SF) se comportam ao serem ajustados

nos tipos de dados exibidos na Seção 3.2. A Figura 3.7 exibe os *boxplots* das medidas com as razões dos DICs, WAICs e LPMLs dos modelos: com o agrupamento das interações (*boxplots* verde e rosa), o modelo fatorial com interações F^{2013} (*boxplot* vermelho) e o modelo sem interação SF (*boxplot* branco). Observa-se claramente que a modelagem proposta é superior aos modelos do Capítulo 2 e o sem interação, pois apresentam *boxplots* mais próximos do valor de referência 1. Isso indica que o tipo de modelagem proposta seja utilizada diante da suspeita de haver algum agrupamento das interações entre genes. É claro que está-se partindo da principal motivação do trabalho que seriam os estudos de GRN. Mas em uma situação onde não há suspeitas de interação entre genes, que seria algo menos realístico, o ajuste de um modelo fatorial sem interações seria a melhor opção. Na Figura 3.8 apresentam-se os *boxplots* com os EQMs dos parâmetros α, λ, σ^2 e F referentes a estes modelos. Veja que ao ajustar o modelo sem interações foram obtidas as piores estimativas para os demais parâmetros, pois apresentam *boxplots* mais altos, refletindo no pior ajuste conforme visto na Figura 3.7. Observe que ao ajustar o modelo considerado no Capítulo 2, que considera todas as interações não nulas sendo diferentes, foram também obtidas estimativas ruins para os parâmetros.

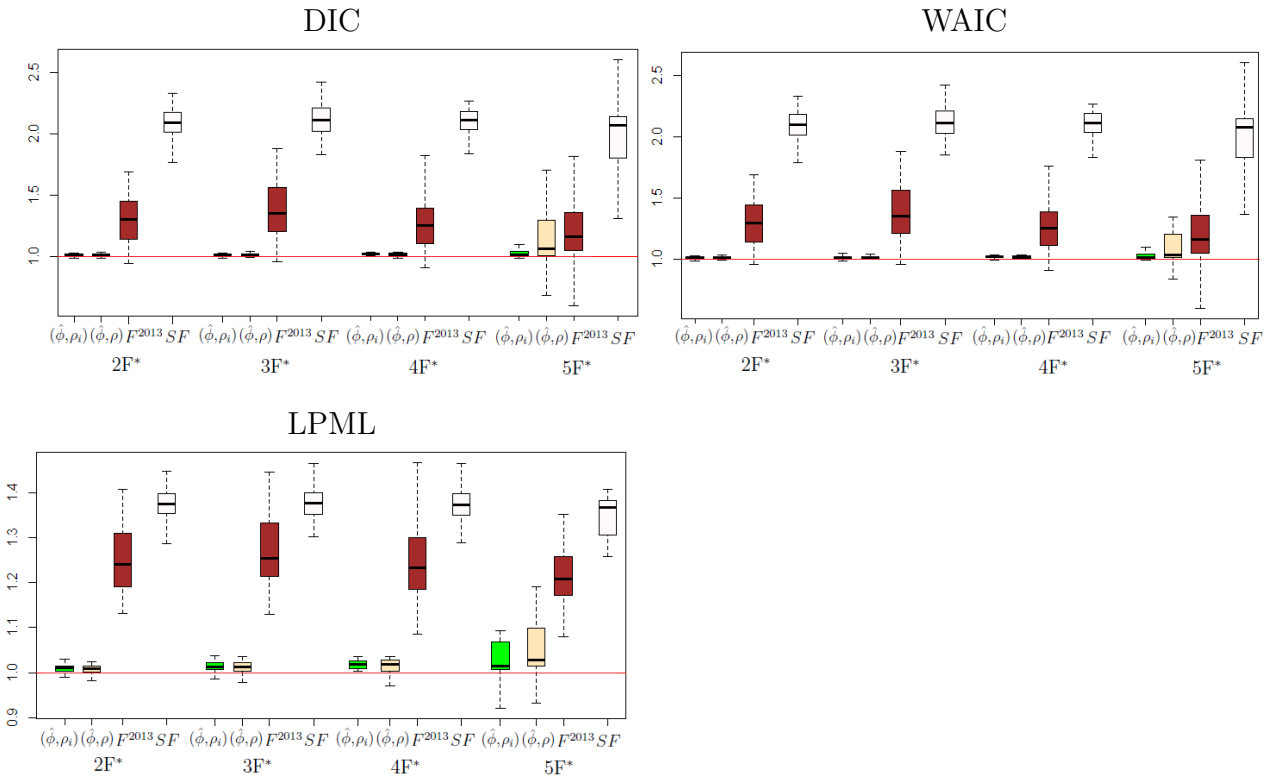

Figura 3.7: *Boxplots das razões dos DIC, WAIC e LPML referentes ao modelo proposto com a estimação de ϕ, juntamente com o modelo fatorial com interações abordado no Capítulo 2, e o modelo fatorial sem interações. A linha na horizontal representa o valor 1.*

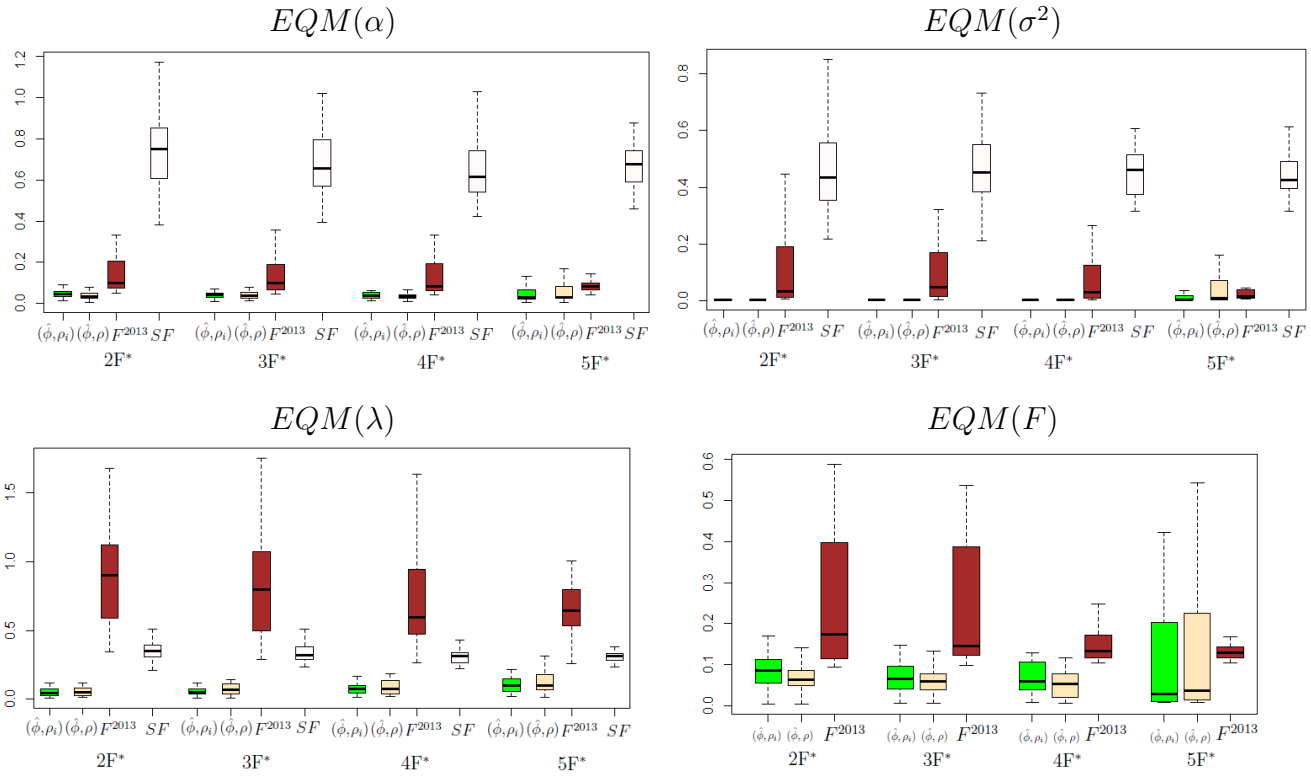

Figura 3.8: *Boxplots dos EQMs dos parâmetros α, λ, σ^2 e F em todos os cenários. Os boxplots verdes e rosas mostram os EQMs dos parâmetros no modelo estimando ϕ diante das estratégias 1 e 2 consideradas para z_i respectivamente. Os boxplots em vermelho mostram os EQMs dos parâmetros do modelo fatorial do capítulo anterior estimando ϕ enquato que os boxplots brancos mostram os EQMs dos parâmetros do modelo fatorial sem interações.*

A Tabela 3.3 exibe os tempos computacionais de um ajuste dos modelos. Veja que a modelagem feita sem interação (SF) apresenta o menor tempo por ser um modelo mais parcimonioso, enquanto que o ajuste do modelo fatorial com interações (F^{2013}) necessita de um tempo maior do que os demais para ser ajustado, pois esta modelagem considera que todos os efeitos de interação estimados são diferentes um do outro. Note que apesar da quantidade de efeitos de interação ser a mesma em cada tipo de dado, o tempo de um ajuste do modelo F^{2013} fica maior quando há um aumento na quantidade de tipos de interação.

Tabela 3.3: Tipos de dados simulados e tempos computacionais (em minutos) de um ajuste do modelo em um conjunto de dados com dimensão de 200 linhas e 100 colunas. A simulação é feita usando a mesma máquina com processador Intel Core $i7$ com 16GB de mémoria RAN e sistema operacional Ubunto 16. A execução do código foi feita com exclusividade sem nenhum outro programa executado em paralelo.

Tipos de dados	Modelos	Tempos
2F*	SF	4
	$(\hat{\phi},\ \rho_i)$	9
	$(\hat{\phi},\ \rho)$	9
	F^{2013}	36
3F*	SF	4
	$(\hat{\phi},\ \rho_i)$	9
	$(\hat{\phi},\ \rho)$	9
	F^{2013}	55
4F*	SF	4
	$(\hat{\phi},\ \rho_i)$	9
	$(\hat{\phi},\ \rho)$	9
	F^{2013}	56
5F*	SF	4
	$(\hat{\phi},\ \rho_i)$	9
	$(\hat{\phi},\ \rho)$	9
	F^{2013}	56

3.4 Conclusões do Capítulo

Neste capítulo foi apresentado uma proposta de modelagem para os efeitos de interação que estendeu o modelo fatorial abordado por Mayrink e Lucas (2013), o qual consideram todas as interações estimadas sendo iguais. Além do mais, o modelo apre-

sentado em 2013 se limita a situações extremas onde as interações estimadas são todas diferentes ou iguais. Na modelagem proposta para os efeitos de interação foi abordado uma situação intermediária na qual foi feito agrupamentos com as interações via misturas de distribuições com ponto de massa. As interações, neste caso, foram estimadas por meio de um processo Gaussiano usando a função de covariâncias exponencial quadrática.

A partir de simulações Monte Carlo feitas com a modelagem proposta para as interações, foi verificado por meio de vários cenários a qualidade das estimativas dos parâmetros e o ajuste do modelo através dos critérios EQM, DIC, WAIC e LPML. Os resultados indicaram o bom desempenho do modelo diante da estimação do parâmetro ϕ na função de covariâncias. Além disso, a modelagem realizada apresentou uma boa proporção de acerto ao identificar corretamente as interações reais, mostrando que não é preciso ter uma noção de quantos tipos interações F_r^* existem, e que na prática basta ajustar um modelo com uma quantidade de grupos grande. Foi verificado ao ajustar um modelo com 5 tipos de interações não nulas, em dados gerados com menos, que o modelo determina alta proporção de acerto. O fato da modelagem proposta apresentar uma alta proporção de acertos das interações, além de ter um bom desempenho diante da estimação de ϕ, mostra que em situações práticas (ϕ é desconhecido) o melhor a se fazer é ajustar o modelo diante de sua estimação.

Capítulo 4

Abordagem de Agrupamento via Processo Dirichlet

4.1 Definição do Processo

O Processo Dirichlet (PD) é um processo estocástico no qual as realizações são distribuições de probabilidades. Esse processo é muito usado na modelagem Bayesiana não paramétrica, principalmente em modelos de misturas por processo Dirichlet, também conhecido como modelos de misturas infinito. Introduzido por Ferguson (1973), o PD é descrito por uma medida de probabilidade aleatória G definida sobre um espaço mensurável $(\Omega, \Im)$, onde Ω é não enumeravel podendo ser $(\mathbb{R}, \mathbb{R}^T, [0,1], ...)$ e $\Im$ é a σ-álgebra de subconjuntos de Ω. Essa medida G é vista como um parâmetro em inferência Bayesiana não paramétrica e assume valores em um conjunto de todas as medidas de probabilidade definidas sobre $(\Omega, \Im)$ [Antoniak (1974)].

Por definição, diz-se que G é distribuída segundo um PD se para qualquer partição $(B_1, B_2, \cdots, B_R)$ de Ω, o vetor de probabilidades $[G(B_1), G(B_2), \cdots, G(B_R)]$ tem distribuição Dirichlet com vetor de parâmetros $[\tau G_0(B_1), \tau G_0(B_2), \cdots, \tau G_0(B_R)]$. Como notação considere:

$$G \sim PD(\tau, G_0).$$

A distribuição base G_0 pode ser interpretada como o valor esperado do processo, ou seja, o

PD gera distribuições em torno de G_0. De uma forma intuitiva seria como gerar valores da distribuição normal em torno de sua média. O parâmetro de concentração τ é um número real positivo e controla o quão distintas ou dispersas serão essas realizações (distribuições) do processo. Portanto o PD *a priori* é centrado em um modelo paramétrico especificado por G_0, enquanto que τ permite controlar a incerteza dessa escolha. Para mais detalhes a respeito da prova da existência do PD e das suas propriedades veja Ferguson (1973), Ferguson (1974), Antoniak (1974), Ghosh e Ramamoorthi (2003).

4.2 Representação via *Stick-Breaking* e Modelos de Misturas

Ao atribuir misturas de distribuições aos efeitos de interações do modelo fatorial, foi apresentado uma abordagem que estendeu a modelagem introduzida por Mayrink e Lucas (2013). Isso proporcionou o agrupamento das interações estimadas, além de uma interpretação para os genes afetados por estes grupos. A nova abordagem que será feita aqui, introduz uma incerteza sobre o modelo de mistura (medida misturadora) que será descrito usando um processo Dirichlet. Essa ideia traz mais flexibilidade para a modelagem de agrupamento do capítulo anterior, pois não requer a especificação prévia do número de componentes na mistura. Considere:

$$\begin{aligned}(F_{1\bullet}^{\top}, ..., F_{m\bullet}^{\top} \mid G) &\sim G; \\ G &\sim PD(\tau, G_0).\end{aligned}$$

Onde os efeitos de interação $F_{i\bullet}^{\top}$'s são independentes e estão sendo gerados de um processo Dirichlet com parâmetro de concentração τ e distribuição base $G_0 \equiv N_n[\mathbf{0}, K(\lambda, \phi)]$. Veja que o processo Dirichlet está sendo definido como uma distribuição *a priori* para a medida misturadora G (modelo de mistura) que é desconhecida e portanto aleatória. Neste caso o PD definido acima será usado para descrever a incerteza que se tem sobre G.

A representação de um processo Dirichlet pode ser feita de diversas formas como, o processo de restaurante Chinês, veja por exemplo Blei et al. (2010), um esquema de urna

de Polya de Blackwell e MacQueen (1973) ou por uma soma ponderada de massas pontuais, conhecida na literatura como construção via *stick-breaking* [Sethuraman (1994)]. Explorando a representação de Sethuraman (1994) para gerar e agrupar os efeitos de interação estimados pode-se escrever:

$$G = \rho_0\delta_{F_0^*}(F_{i\bullet}) + \rho_1\delta_{F_1^*}(F_{i\bullet}) + \rho_2\delta_{F_2^*}(F_{i\bullet}) + \cdots = \sum_{r=0}^{\infty}\rho_r\delta_{F_r^*}(F_{i\bullet}), \tag{4.1}$$

com os F_r^*'s independentes e gerados da distribuição base $N_n[\mathbf{0}, K(\lambda,\phi)]$. Note que a representação do processo Dirichlet via construção *stick-breaking* gera uma mistura enumerável de pontos de massa $\delta_{F_r^*}(\bullet)$. Nessa definição, tem-se que o modelo em (4.1) é semelhante ao modelo de mistura finito em (3.2), a principal diferença é que o número de componentes é um conjunto infinito e os pesos são construídos como segue:

$$V_r \sim \text{Beta}(1,\tau), \quad \text{com } r = 0,1,2,...,$$
$$\rho_0 = V_0, \quad \text{e para } r \geq 1 \quad \rho_r = V_r\prod_{s=0}^{r-1}(1-V_s).$$

Sendo cada V_r gerado independente dos F_r^*'s e $\sum_{r=0}^{\infty}\rho_r = 1$. Para descrição do processo *stick-breaking* considere inicialmente, por exemplo, uma barra de comprimento 1 representando a probabilidade total de todas as observações amostrais $F_{i\bullet}$'s serem alocadas em Ω. Então, quebra-se um pedaço desta barra de tamanho V_0 gerado da distribuição Beta$(1,\tau)$ e atribui-se a probabilidade $\rho_0 = V_0$, como peso da primeira observação amostral $F_{\mathbf{0}}^*$. O comprimento restante $(1-V_0)$ é atribuído a toda as outras observações amostrais. Em seguida, quebra-se novamente um comprimento V_1 gerado da Beta$(1,\tau)$ representando uma porcentagem de $(1-V_0)$ e atribui-se a probablidade $\rho_1 = V_1(1-V_0)$, para a segunda observação amostral $F_{1\bullet}$ gerada de G_0. Esse procedimento se repete e em cada etapa o comprimento da barra vai se encurtando cada vez mais até que os pesos calculados tenderão a zero. Nesta definição do PD, os pesos ρ_r decaem estocasticamente a uma taxa que depende do parâmetro de concentração τ, pois $E[V_r] = \dfrac{1}{1+\tau}$. Esse parâmetro tem um papel extremamente importante para o agrupamento das observações $F_{i\bullet}$'s relacionadas aos genes em G_E. Elas representaram as interações do modelo fatorial e serão discutidas com detalhes adiante. Devido a esta construção, tem-se que para um R suficientemente grande $\sum_{r=R+1}^{\infty}\rho_r = 0$, isto é, $\rho_r \approx 0$ para todo $r = R+1, R+2, \cdots$.

A questão que surge é qual R escolhido é suficientemente grande? Na prática 25 ou 50 são comumente utilizados como padrão na literatura. Autores como Gelman et al. (2003) (Cápitulo 23, pág. 552) argumentam que raramente há uma necessidade de se ter mais do que 10 ou 15 grupos para ajustar com precisão um modelo desconhecido.

O que está implícito na definição construtiva do *stick-breaking* é o fato de que as realizações do processo Dirichlet são discretas (quase certamente). Em muitas aplicações essa discretização parece estranha. Por exemplo, em problemas de estimação de densidade de uma variável aleatória $y_i \sim G$, $i = 1, ..., n$, seria inapropriado assumir $G \sim PD$ se já é sabido que a distribuição dos y_i's é contínua. Mas uma simples extensão do processo Dirichlet conserta o que se parece estranho. Tal procedimento consiste em utilizar uma mistura entre uma distribuição contínua e uma medida de probabilidade gerada de um PD. Considere que $y_i \sim H$ e

$$H(y) = \int h(y \mid \theta) dG(\theta), \quad \text{com } G \sim PD(\tau, G_0). \tag{4.2}$$

Veja que a distribuição H é escrita como uma mistura com respeito a uma medida *a priori* G sendo um PD. Aqui, o núcleo $h(y \mid \theta)$ é algum modelo indexado por θ e representa uma família de distribuições. Ao explorar a representação *stick-breaking* pode-se escrever (4.2) como:

$$H(y) = \sum_{r=0}^{\infty} \rho_r h(y \mid \theta_r^*). \tag{4.3}$$

Note que o modelo não paramétrico em (4.3) é semelhante a um modelo de misturas finito. A principal diferença é que o número de componentes é um conjunto inifinito e os pesos são construídos via representação *stick-breaking*. Nestes casos não paramétricos, em vez de se ter um pequeno número de parâmetros que caracteriza a família de distribuições dos dados, conceitualmente, existe um número inifinito de parâmetros. Na prática o que se faz é truncar esses modelos para ter uma configuração grande e finita do número de parâmetros. Nos casos de estimações de densidades, trabalhar com um número infinito de componentes pode ser uma forma mais atrativa de garantir que o modelo de mistura usando o PD tenha um suporte sobre a ampla classe de distribuições.

O procedimento de atribuir um PD como distribuição *a priori* para o conjunto de todas as distribições de misturas é chamado na literatura de modelo de mistura por

processo Dirichlet. Uma representação hierárquica de (4.2) é:

$$\begin{aligned}(Y_i \mid \theta_i) &\sim h(y \mid \theta_i), i = 1, ..., n;\\ \theta_1, ..., \theta_n \mid G &\sim G;\\ G &\sim PD(\tau, G_0).\end{aligned}$$

Com os $(Y_i \mid \theta_i)$ independentes e os θ_i's independente e identicamente distribuídos, o modelo acima é uma mistura de densidades que dependem de θ_i. A medida misturadora G é desconhecida e portanto o PD é utilizado para descrever a incerteza *a priori* do comportamento de $h(y \mid \theta)$ por meio de θ. Ao considerar variáveis latentes z_i's para dizer que y_i pertence à componente r, com probabilidade ρ_r, tem-se um modelo hierárquico equivalente:

$$\begin{aligned}(y_i \mid z_i = r) &\sim h(y \mid \theta_r^*), i = 1, ..., n;\\ (\theta_r^* \mid G_0) &\sim G_0;\\ (z_i \mid \rho) &\sim \text{Mult}(1, \rho);\\ \rho &\sim stick(\tau).\end{aligned} \tag{4.4}$$

Sendo $\rho \sim stick(\tau)$ uma notação abreviada para denotar que os pesos são gerados de um PD usando a representação via *stick-breaking* com parâmetro de concentração τ. Os θ_r^* são gerados de forma independente. Ao modelar as observações $(y_1, ..., y_n)$ usando um conjunto de parâmetros latentes $(\theta_1, ..., \theta_n)$, cada θ_i é amostrado de forma independente e identicamente distribuída de $G = \sum_{r=1}^{\infty} \rho_r \delta_{\theta_r^*}$. Como G é discreta, vários θ_i's terão valores iguais. Desta forma, no modelo de mistura em (4.4) os y_i's caracterizados com $\theta_i = \theta_r^*$ pertencerão ao mesmo grupo.

Representação proposta nesta tese

Para o caso que será considerado neste trabalho, a distribuição com ponto de massa representada por $\delta_{F_r^*}(F_{i\bullet})$ atuará como a densidade $h(y \mid \theta_r^*)$. Nesta situação considere a

seguinte configuração *a priori* para os efeitos de interação:

$$
\begin{aligned}
(F_{i\bullet}^{\top} \mid F_r^*, z_{ir} = 1) &\sim \delta_{F_r^*}(F_{i\bullet}); \text{ com } r = 0, \cdots, R. \\
(F_r^* \mid \lambda, \phi) &\sim N_n[\mathbf{0}, K(\lambda, \phi)]; \\
(z_i \mid \rho_i) &\sim \text{Mult}(1, \rho_i); \\
\rho_i &\sim stick(\tau).
\end{aligned} \tag{4.5}
$$

Reforçando mais uma vez que esta representação em (4.5) é parecida com o modelo de mistura finito proposto no capítulo anterior. Entretanto, na representação anterior em (3.3) ao estabelecer a distribuição *a priori* para a coleção $(\{\rho_{ir}\}, \{F_r^*\})$, estava-se colocando uma específica distribuição *a priori* para G. Enquanto que na configuração em (4.5) tem-se uma vantagem, pois o PD estabelece mais flexibilidade na modelagem das interações, permitindo um suporte completo sobre todas as distribuições de misturas. Além disso, ao modelar as interações utilizando (4.5) pode-se determinar automaticamente o número de grupos que melhor se ajusta aos dados. Uma segunda estratégia que assume uma probabilidade global para todas as interações também será utilizada. Ela consiste em adotar a seguinte configuração para o vetor indicador z_i em (4.5): $(z_i \mid \rho) \sim \text{Mult}(1, \rho)$ e $\rho \sim stick(\tau)$. Lembrando que esta estratégia leva em conta todos os z_i na atualização da distribuição condicional completa *a posteriori* de ρ.

Vale ressaltar, que outras formas de deixar o modelo de mistura finito do capítulo anterior mais flexível são: *i*) Fazer a estimação do número de componentes adotando alguma distribuição *a priori* para R. *ii*) Ajustar vários modelos com diferentes valores de R e selecionar aquele com melhor ajuste baseado em algum critério (LPML, DIC, WAIC e outros). Entretanto, tais procedimentos não serão avaliados neste trabalho, pois métodos como estes provavelmente implicariam em um custo compoutacional maior, além de dificuldades no processo de estimação porque teria-se que usar algoritimos do tipo RJMCMC. Lembrando mais uma vez, que a representação *stick-breaking* utilizada em (4.5) estabelece uma ordem das componentes da mistura via PD, pois os pesos decaem estocasticamente conforme r aumenta, ou seja, $\rho_{ir} = 0$ para todo r suficientemente grande. Isso implica que um número infinito de grupos (componentes) não serão ocupados por interações não nulas.

O parâmetro de concentração τ pode ser fixado ou estimado. No caso de estimação, por ser um valor não negativo é muito comum atribuir uma distribuição Gamma(u_1, u_2). Isso irá permitir que os dados informem sobre o valor mais apropriado para τ. Outro fato importante é que o parâmetro de concentração influencia a quantidade de componentes do modelo em (4.5). Valores de τ muito pequenos fazem com que haja poucos grupos criados na modelagem, pois na construção via *stick-breaking* os pesos ρ_{ir} das primeiras componentes tenderão a ser maiores, enquanto que os demais tenderão a zero mais rápido $\left(\sum_{r=R+1}^{\infty} \rho_{ir} \to 0\right)$. Neste caso, muitos tipos de efeitos de interação serão nulos, porque a primeira componente em (4.5) é $\delta_{\mathbf{0}}(F_{i\bullet})$. Valores de τ muito grandes permitirão que mais componentes sejam criadas e muitas delas com pesos ρ_{ir} próximos a zero. Isso implicará em um modelo com muitos tipos de efeitos de interação não nulos. Além do mais, o número de grupos $R+1$ pode ser interpretado como um limite superior para o número de componentes que podem ser ocupadas pelas interações. Algumas dessas compontentes não serão criadas, outras poderão ser ocupadas por uma única observação $F_{i\bullet}$.

Um questionamento que pode surgir é se a modelagem usando o PD é comparável ao modelo de mistura finito utilizado no capítulo anterior. A resposta para isso é sim, pois anteriormente a representação adotada para $F_{i\bullet}$ em (3.3) tinham os pesos $\rho_i = (\rho_{i0}, \rho_{i1}, \cdots, \rho_{iR}) \sim \text{Dir}\,(\nu_0, \nu_1, \cdots, \nu_R)$. Os tamanhos relativos dos parâmetros da distribuição Dirichlet descrevem a média da distribuição *a priori* para os pesos ρ_i. Se for considerado que $\dfrac{1}{R+1} = \nu_r,\ \forall\ r = 0, 1, \cdots, R$, então $\tau = \sum_{r=0}^{R} \nu_r$ representará uma medida da força da distribuição *a priori*. Neste casos, diz-se que ρ_i segue um PD com distribuição base $\dfrac{1}{R+1}$ e parâmetro de concentração τ. É muito comum em modelos de mistura adotar $\nu = (1, 1, \cdots, 1)$, pois essa configuração padrão tende a alocar mais observações (em nosso caso interações) para diferentes componentes de forma uniforme, já que a Dir$(1, 1, \cdots, 1)$ é uma generalização da distribuição Beta$(1, 1)$ que por sua vez é equivalente a U$(0, 1)$.

4.3 Estudo Simulado

Nesta etapa do trabalho será apresentada algumas simulações feitas com o modelo fatorial adotando a especificação *a priori* em (4.5) para os efeitos de interação. Os ajustes do modelo fatorial serão realizados nos mesmos dados do Capítulo 3 que foram gerados considerando o valor real $\phi = 0.2$, onde foi ajustado o modelo fatorial com misturas finitas para agrupamento das interações. Todos os ajustes feitos consideram a estimação do parâmetro ϕ. A escolha do parâmetro de concentração $\tau = 6$ adotado aqui, é feito com o propósito de comparar a modelagem via PD com o modelo de misturas finito abordado no capítulo anterior. Para analisar o efeito que τ tem na quantidade máxima de componentes $R+1$, foi ajustado o modelo com a representação *stick-breaking* truncando $R = 5$ e $R = 10$ (6 e 11 componentes, respectivamente). Para estes casos adimitiu-se $\tau = 6$ e $\tau = 11$ valorizando a possibilidade de mais interações e grupos que podem ser criados. A Tabela 4.1 apresenta os cenários com os ajustes dos modelos usados nos 4 tipos de conjuntos de dados vistos no Capítulo 3. Aqui, também foram feitas 50 replicações Monte Carlo ajustando o modelo fatorial diante da estimação do parâmetro de suavização ϕ. O Cenário M_{ρ_i} indica que o modelo fatorial é ajustado adotando a mistura finita em (3.3) para $F_{i\bullet}$ e usando a primeira estratégia para z_i. O Cenário $PD_{\rho_i,6}$ representa o ajuste do modelo com a especificação *a priori* em (4.5) para $F_{i\bullet}$ e com $\tau = 6$. As demais distribuições *a priori* usadas para α, σ^2, λ e ϕ são as mesmas do capítulo anterior e encontram-se em (2.2), (2.3), (2.4) e (2.7), respectivamente. Para os pesos q_{il} e ρ_i, atribuem-se as mesmas configurações *a priori* que estão nas Tabelas 2.2 e 3.2, respectivamente, com exceção dos pesos ρ_i e ρ em G_E referentes aos modelos em que usa-se o PD, pois estes são construídos via representação *stick-breaking*.

A avaliação da qualidade dos ajustes é feita a partir das medidas construídas com as razões dos critérios DIC, WAIC e LPML. Para avaliar a qualidade das estimativas dos parâmetros α, λ, σ^2, F e ϕ foram calculados os EQMs assim como feito no capítulo anterior. Além disso, foram calculadas as proporções de acerto Monte Carlo em que o modelo consegue identificar corretamente cada tipo de efeito de interação F_r^* em G_E, considerando as duas estratégias adotadas para z_i.

Tabela 4.1: Cenários com os ajustes de modelos.

Valores de τ	Modelos
-	M_{ρ_i}
-	M_{ρ}
6	$PD_{\rho_i,6}$
6	$PD_{\rho,6}$
6	$PD^*_{\rho_i,6}$
11	$PD^*_{\rho_i,11}$
11	$PD^*_{\rho,11}$

* modelo com representação de Sethuraman (1994) truncado em 11 componentes.

Na Figura 4.1 observa-se os *boxplots* das medidas feitas com as razões dos DICs, WAICs e LPMLs obtidas via estudo Monte Carlo. Lembrando que no denominador dessas razões estão os valores dos critérios do modelo gerador dos dados. Veja que a maioria dos gráficos apresentam *boxplots* com medianas bem próximas e estão acima do 1, indicando que os modelos ajustados se comportam relativamente bem em relação aos 4 tipos de dados. Note que no tipo de dado $4F^*$ os *boxplots* (azul e laranja) referentes aos Cenários $PD_{\rho_i,6}$ e $PD_{\rho,6}$ são maiores e mostram maior variabilidade em comparação aos demais cenários de $4F^*$. Além disso, pode-se observar que para o tipo de dado $5F^*$, os *boxplots* (azul e laranja) referentes a $PD_{\rho_i,6}$ e $PD_{\rho,6}$ apresentam um distanciamento do valor 1, isso pode sugerir que o modelo nestes cenários não se adequem bem quando há um aumento nos tipos de interações (grupos). Entretanto, a modelagem para estes casos é feita usando o truncamento em 6 componentes ($R = 5$) na construção via *stick-breaking*. Note que quando aumenta-se o número de componentes para 11 ($R = 10$), na representação de Sethuraman (1994), pode-se observar no *boxplot* (marron) que o modelo no Cenário $PD^*_{\rho_i,6}$ apresenta uma variabilidade menor aproximando-se de 1. Isso reflete em um impacto positivo na estimação de F e σ^2 para os tipos de dados $4F^*$ e $5F^*$, que podem ser vistos pelos EQMs a seguir. Em resumo, a modelagem via misturas para os $F_{i\bullet}$'s representada nos Cenários M_{ρ_i} e M_{ρ}, exibem gráficos (verde e rosa) mais estáveis e próximos de 1 em todos os tipos de dados avaliados. Isso indica um melhor ajuste

em relação à modelagem via PD exibida nos Cenários $PD_{\rho_i,6}$ e $PD_{\rho,6}$, que apresentam resultados mais irregulares.

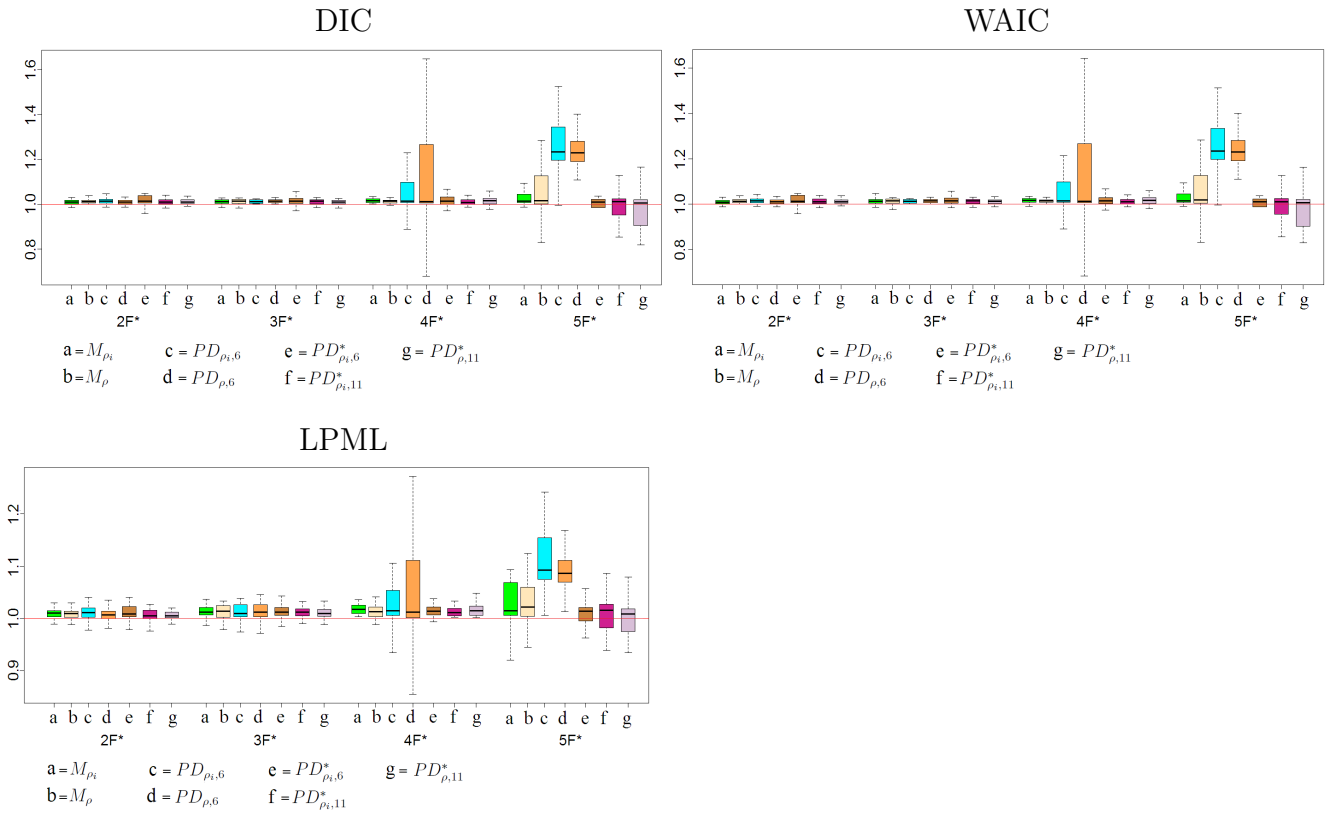

Figura 4.1: *Boxplots das razões dos DIC, WAIC e LPML obtidos via Monte Carlo e referentes ao modelo proposto com a estimação de ϕ. Os boxplots em verde e rosa referem-se as medidas do modelo exibido no Capítulo 3 diante das estratégias 1 e 2 para z_i, respectivamente. Os boxplots azul e laraja referem-se as medidas para os modelos dos Cenários $PD_{\rho_i,6}$ e $PD_{\rho,6}$ respectivamente. Os boxplots marron, roxo e lilás, referem-se as medidas para os modelos do Cenários $PD^*_{\rho_i,6}$, $PD^*_{\rho_i,11}$ e $PD^*_{\rho,11}$, respectivamente. A linha na horizontal representa o valor 1.*

A Figura 4.2 apresenta os *boxplots* dos EQMs de α, λ e F para os modelos em cada cenário. Veja que nos tipos de dados $2F^*$ e $3F^*$ a maioria dos *boxplots* se mantém com uma certa estabilidade. Enquanto que para o tipo de dado $4F^*$ os Cenários $PD_{\rho_i,6}$ e $PD_{\rho,6}$ apresentam *boxplots* (azul e laranja) maiores mostrando uma variabilidade alta

em relação aos demais cenários de $4F^*$. Observe que no painel referente aos EQMs de F, os gráficos dos Cenários $PD^*_{\rho_i,6}$, $PD^*_{\rho_i,11}$ e $PD^*_{\rho,11}$ se mantém estáveis quando há um aumento na quantidade de tipos de interação (grupos). Nestes cenários também observam-se *boxplots* com medianas próximas, além serem mais baixos em relação aos Cenários $PD_{\rho_i,6}$ e $PD_{\rho,6}$ em $5F^*$.

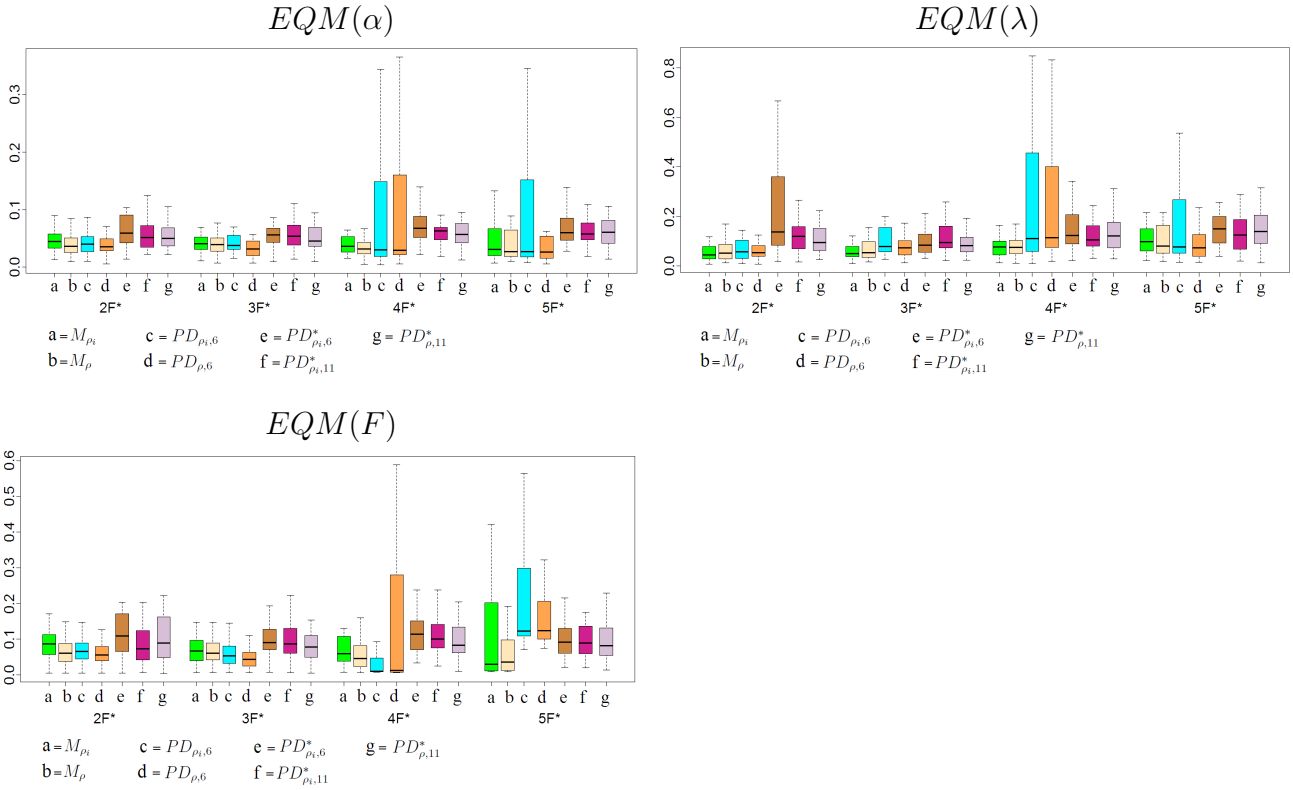

Figura 4.2: *Gráficos dos EQMs obtidos via Monte Carlo para os parâmetros α, λ e F em todos os cenários. Os boxplots em verde e rosa referem-se aos EQMs para os modelos exibidos no Capítulo 3 diante das estratégias 1 e 2 para z_i, respectivamente. Os boxplots azul e laraja mostram os EQMs para os modelos dos Cenários $PD_{\rho_i,6}$ e $PD_{\rho,6}$ respectivamente. Os boxplots marron, roxo e lilás, exibem os EQMs para os modelos do Cenários $PD^*_{\rho_i,6}$, $PD^*_{\rho_i,11}$ e $PD^*_{\rho,11}$, respectivamente.*

A Figura 4.3 exibe os gráficos com os EQMs de σ^2. Cada painel apresenta os *boxplots* dos cenários avaliados diante de cada quantidade de tipos de interação. Observe que nos painéis $2F^*$ e $3F^*$ todos os cenários mostram *boxplots* baixos e com medianas próximas. Veja que o Cenário $PD^*_{\rho_i,6}$ em $2F^*$ apresenta um *boxplot* maior, mas com mediana ainda próxima dos demais. Observa-se também em $2F^*$ que quando há um aumento no parâmetro de concentração τ de 6 para 11, os EQMs de σ^2 exibem menor variabilidade melhorando a estimação de σ^2, compare os Cenários $PD^*_{\rho_i,6}$ e $PD^*_{\rho_i,11}$. No painel com tipo de dado $4F^*$ observam-se que a maioria dos gráficos estão bem próximos de zero. Note que os Cenários $PD_{\rho_i,6}$ e $PD_{\rho,6}$ exibem *boxplots* maiores indicando uma variabilidade maior do que os demais. Entretanto suas medianas ainda são próximas a zero como nos demais cenários, isso indica uma boa estimação de σ^2. No painel em $5F^*$, pode-se observar que os modelos dos Cenários $PD_{\rho_i,6}$ e $PD_{\rho,6}$ mostram *boxplots* mais acima que os outros, sugerindo estimativas não tão boas para σ^2. Porém, ao ajustar o modelo do Cenário $PD_{\rho_i,6}$, usando a representação via *stick-breaking* truncando em $R = 10$ (11 componentes), é possível notar que há uma melhora na estimação de σ^2. Compare o gráfico do Cenário $PD_{\rho_i,6}$ com o *boxplot* do Cenário $PD^*_{\rho_i,6}$. Além disso, note mais uma vez que o modelo via misturas no Cenário M_{ρ_i} não apresentou grandes variações como a modelagem via PD.

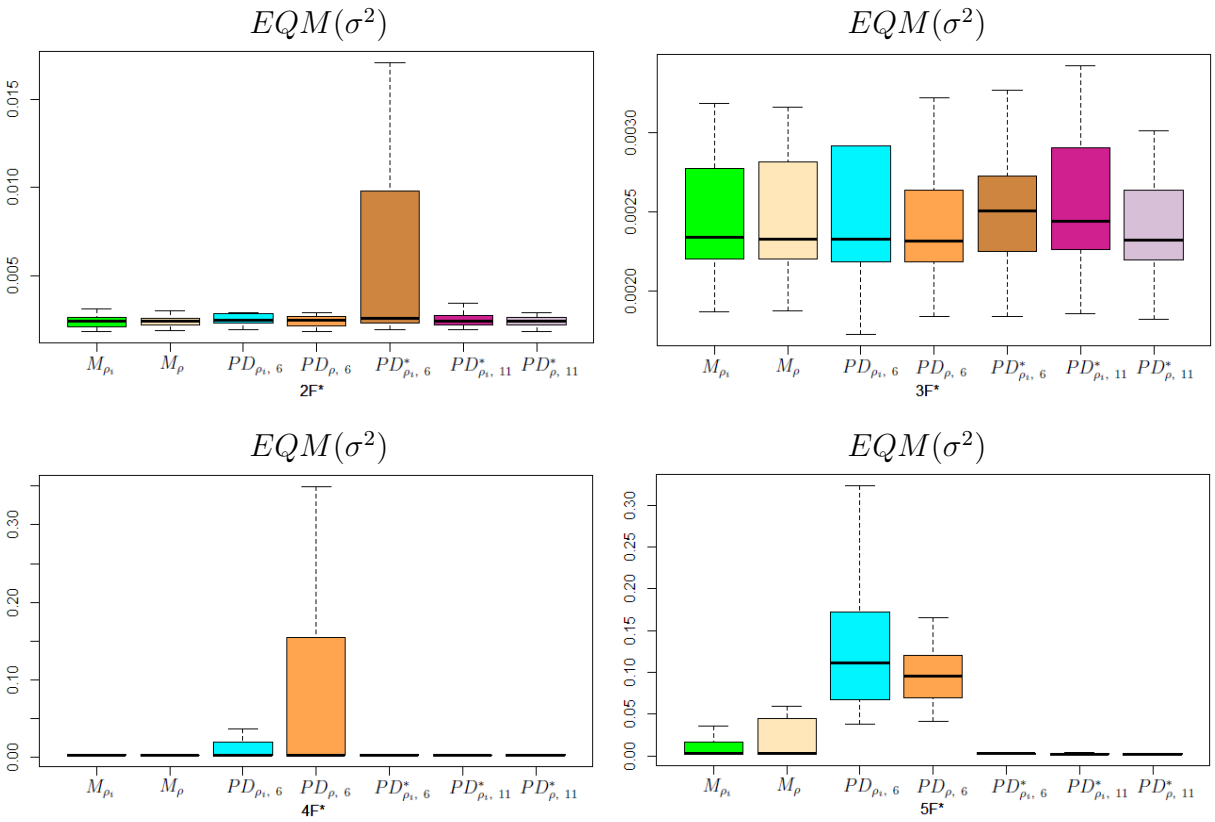

Figura 4.3: *Gráficos dos EQMs de σ^2 para cada tipo de conjunto de dados. Os boxplots verdes e rosas referem-se ao EQMs para os modelos exibido no Capítulo 3 diante das estratégias 1 e 2 para z_i respectivamente. Os boxplots azul e laraja mostram os EQMs para os modelos do Cenários $PD_{\rho_i,6}$ e $PD_{\rho,6}$ respectivamente. Os boxplots marron, roxo e lilás, exibem os EQMs para os modelos do Cenários $PD^*_{\rho_i,6}$, $PD^*_{\rho_i,11}$ e $PD^*_{\rho,11}$, respectivamente.*

Na Figura 4.4 observa-se os *boxplots* com as médias *a posteriori* e os EQMs de ϕ referentes aos modelos ajustados para cada tipo de conjunto de dados. Veja que no painel à esquerda, as médias apresentam uma certa estabilidade em torno do valor 0.3 e este comportamento se reflete diante de todos tipos de dados. Estes resultados, mais uma vez, induzem a pensar em ajustar os modelos fixando o parâmetro de comprimento-escala ϕ em 0.3. Porém, devido aos estudos feitos a respeito deste parâmetro no Capítulo 3, foi concluído que a modelagem diante da estimação de ϕ é mais adequada em relação aos ajustes feitos com seu valor fixado, pois determinou melhores estimativas para os demais parâmetros do modelo fatorial.

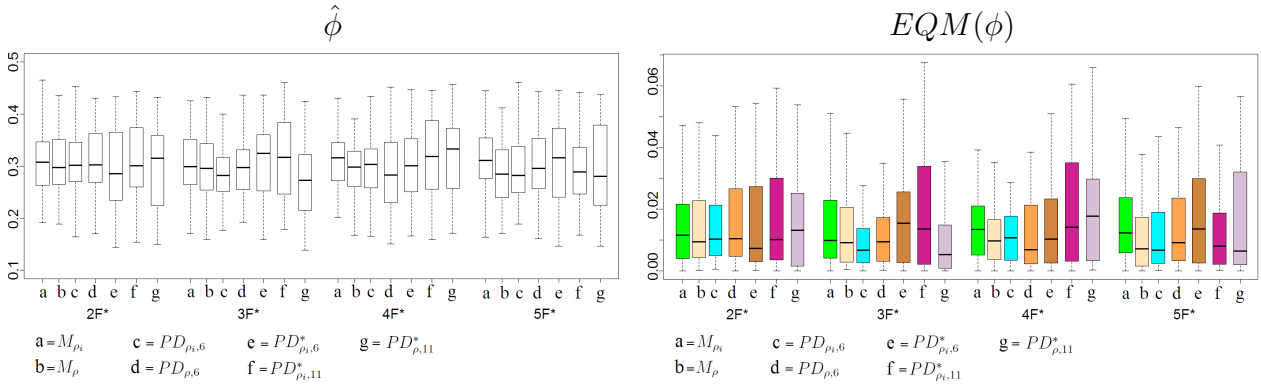

Figura 4.4: *Boxplots das médias e EQMs de ϕ dos modelos ajustados via Monte Carlo para cada tipo de conjunto de dados. O painel à esquerda exibe os boxplots com as médias a posteriori enquanto que o painel à direita apresenta os boxplots com os EQMs de ϕ. Os gráficos em verde e rosa referem-se aos EQMs para os modelos exibidos no Capítulo 3 diante das estratégias 1 e 2 para z_i, respectivamente. Os boxplots azul e laraja mostram os EQMs para os modelos dos Cenários $PD_{\rho_i,6}$ e $PD_{\rho,6}$ respectivamente. Os boxplots marron, roxo e lilás, exibem os EQMs para os modelos do Cenários $PD^*_{\rho_i,6}$, $PD^*_{\rho_i,11}$ e $PD^*_{\rho,11}$, respectivamente* .

A Figura 4.5 apresenta os *boxplots* das proporções de acerto Monte Carlo na identificação correta das interações para os modelos diante de cada cenário. As proporções foram calculadas da mesma forma como feito no Capítulo 3. Na Figura 4.5 pode-se observar proporções de acerto altas, pois a maior parte dos gráficos estão acima de 0.5 e exibem medianas em torno de 0.7. Veja que em conjuntos de dados que apresentam menos tipos de interação (grupos), pode-se verificar o bom desempenho da modelagem proposta tanto usando o PD quanto o modelo de misturas finito visto Capítulo 3. Note que o Cenário $PD_{\rho_i,6}$ mostra uma leve melhora em relação ao Cenário M_{ρ_i} diante dos tipos de dados $2F^*$, $3F^*$ e $4F^*$ exibindo *boxplots* mais altos. Veja que os Cenários $PD^*_{\rho_i,6}$, $PD^*_{\rho_i,11}$ e $PD^*_{\rho,11}$ apresentam gráficos que parecem estáveis e com medianas próximas. É importante lembrar, que o modelo nos Cenários M_{ρ_i} e M_{ρ}, têm em sua estrutura uma mistura contendo 6 componentes e o ajuste é feito em dados que apresentam poucos grupos. Pelo fato das modelagens via PD e mistura conseguirem proporcionar quantidades

de acertos altas na identificação das interações, isso indica que em dados com poucos grupos (tipos de interações), os pesos das componentes extras são nulas ou próximas a zero. Estes resultados fornecem um direcionamento de que não é preciso ajustar o modelo tendo uma noção prévia de quantas componentes existem.

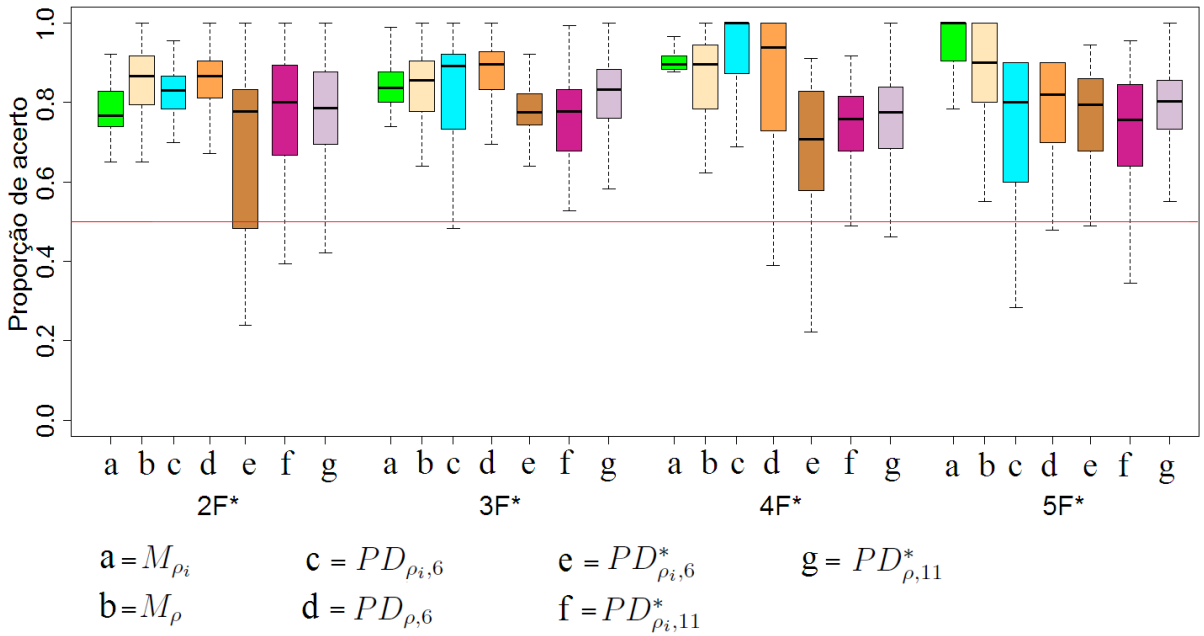

Figura 4.5: *Gráfico das proporções de acerto na identificação correta das interações obtidas via Monte Carlo em cada cenário. Os boxplots em verde e rosa mostram as proporções de acerto para os modelos exibidos no Capítulo 3 diante das estratégias 1 e 2 para z_i, respectivamente. Os boxplots azul e laraja mostram as proporções de acerto para os modelos dos Cenários $PD_{\rho_i,6}$ e $PD_{\rho,6}$ respectivamente. Os boxplots marron, roxo e lilás, exibem as proporções de acerto para os modelos dos Cenários $PD^*_{\rho_i,6}$, $PD^*_{\rho_i,11}$ e $PD^*_{\rho,11}$, respectivamente . A linha na horizontal representa o valor 0.5.*

4.4 Conclusões do Capítulo

O presente capítulo apresentou uma breve definição do processo Dirichlet e uma de suas representações, que foi explorada na abordagem proposta para os efeitos de interação do modelo fatorial. Esta ideia de usar o PD proporciona um modo alternativo

à modelagem via misturas para o agrupamento das interações, pois permite devido as propriedades do processo, criar grupos de interações estimadas. Esta nova abordagem acaba sendo mais geral e tão boa quanto a modelagem usando misturas finitas. Apesar de alguns Cenários como o $PD_{\rho_i,6}$ e $PD_{\rho,6}$ exibirem resultados não tão satisfatórios (rever Figuras 4.1 e 4.2), a modelagem via *stick-breaking* com uma quantidade maior de componentes (R=10) proporcionou melhores ajustes e estimativas para $F_{i\bullet}$ e σ^2 em comparação ao $PD_{\rho_i,6}$ e $PD_{\rho,6}$, principalmente nos dados com 5 grupos não nulos. A vantagem de utilizar o PD é a não especificação prévia do número de grupos que são formados com os efeitos de interações.

Diversos cenários foram considerados nesta etapa do trabalho. A avaliação da qualidade dos ajustes dos modelos foi feita usando razões construídas com o DIC, WAIC e LPML, assim como no Capítulo 3. Observou-se que o modelo ajustado usando o PD, em dados com poucos tipos de interação (grupos), se comporta tão bem quanto o modelo via misturas. Os EQMs dos parâmetros também apresentaram estimativas tão boas quanto a modelagem utilizando misturas. Para os efeitos de interação estimados, foram calculados as proporções em que o modelo acerta as interações reais; reveja a Figura 4.5. Notou-se que houve uma proporção de acerto alta na identificação correta das interações. Isso mostra e confirma o bom desempenho do modelo fatorial ao utilizar as abordagens propostas de agrupamento. Além do mais, este fato proporciona uma boa interpretação para dados envolvendo expressões de genes. Pois cada efeito de interação estimado nas linhas de F ($\hat{F}_{i\bullet} = \hat{F}^*_r$) está relacionado a um gene i. Neste caso, o agrupamento das interações define de certa forma um conglomerado de genes. Assim haverá um conjunto de genes afetados por um tipo de efeito de interação $\hat{F}^*_r$. Este tipo de interação representa a atuação conjunta de regiões pré-especificadas do genoma. Esta análise complementa e estende o modelo construido por Mayrink e Lucas (2013), que não abordam essa ideia de agrupamento das interações.

Portanto, a motivação de aplicações em problemas práticos mostram a impotância do tipo de modelagem que está sendo proposta aqui. No capítulo seguinte serão feitas aplicações dos modelos propostos em bases de dados reais envolvendo câncer de mama.

Capítulo 5

Aplicação a Dados Reais

Neste capítulo, será desenvolvido uma análise em 4 conjuntos de dados reais referentes ao problema de CNA em câncer de mama. Estas 4 bases estão avaliadas em Chin et al. (2006), Miller et al. (2005), Sotiriou et al. (2006) e Wang et al. (2005). A partir deste ponto do trabalho esses 4 conjuntos de dados serão referidos de forma mais simplificada como Chin, Miller, Sotiriou e Wang. O total de *microarrays* ou tamanhos amostrais de cada um dos conjuntos de dados são 118, 251, 189 e 286 amostras, respectivamente. Para melhor compreensão, considere a base de expressão de genes registrada em 118 *microarrays* relativos ao câncer de mama avaliado em Chin. Este, assim como os outros, também foi um dos conjuntos de dados analisados por Mayrink e Lucas (2013), que investigaram resultados de dois grupos de genes relacionados à partes do genoma com CNA. Nestas bases de dados são pré-especificadas 4 regiões no genoma com CNA em cromossomos diferentes. Aqui, serão analisados os pares destas regiões que irão representar os grupos G_1 e G_2. A Tabela 5.1 apresenta as localizações das 4 regiões do genoma e o número do cromossomo em cada posição. Os pares dessas posições serão avaliadas em cada um dos demais conjuntos de dados. Veja, por exemplo, que o trecho afetado pela CNA localizado na posição 35152961 do cromossomo 22, será descrito como grupo G_1. A segunda região, denotada por G_2, é localizada na posição 68771985 do cromossomo 16. Neste caso, os grupos G_1 e G_2 apresentam 50 e 42 genes, repectivamente, e correspondem ao par $(2, 4)$. A seleção desses genes é baseada em um intervalo de 2000000 para a esquerda e direita ao redor da posição localizada no genoma onde ocorre a CNA [Mayrink e Lucas (2013),

Lucas et al. (2006)].

Tabela 5.1: Regiões detectadas com CNA e número de genes antes e depois do procedimento de limpeza proposto em 2013.

			Número de genes	
Regiões	Cromossomo	Posição	Antes	Depois
1	11	117844879	38	13
2	22	35152961	50	22
3	7	101400207	45	24
4	16	68771985	42	18

* Tabela retirada de Mayrink e Lucas (2013).

Os *microarrays* selecionados para as aplicações representam 22283 genes replicados em 118, 251, 189 e 286 amostras de Chin, Miller, Sotiriou e Wang, respectivamente. 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 os tamanhos das matrizes de dados X inicialmente com 22283 linhas, selecionando os principais genes para aplicação. Neste caso, as matrizes de dados que serão exploradas após esse procedimento, terão o par das regiões (1,4) formando os grupos G_1 e G_2 com 13 e 18 genes, respectivamente. As regiões (2,4) representam os grupos G_1 e G_2 com 22 e 18 genes. Enquanto que as posições (3,4) formam G_1 e G_2 contendo 24 e 18 genes, respectivamente (reveja a Tabela 5.1). Depois deste procedimento, o grupo G_E para cada um dos pares (1,4), (2,4) e (3,4) serão compostos por 3717, 3704 e 3708 genes, respectivamente. Vale ressaltar também que estes dados foram pré-processados via RMA [Irizarry et al. (2003)] descrito brevemente no Capítulo 2. Na Figura 5.1 pode-se observar a imagem da matriz X representando a base de dados Chin, referente ao par de regiões (2,4). Para melhor visualização, a imagem exibida apresenta os valores dos níveis de expressão padronizados por linha e as colunas organizadas em relação a mediana das colunas de X. É importante lembrar que as 40 primeiras linhas de X representam o grupo $(G_1 \cup G_2)$. No painel à direita, pode-se observar uma submatriz de X contendo as expressões em G_1 e G_2. Veja no segundo painel que há um padrão de expressão da

esquerda para a direita indicando uma atuação conjunta destes genes.

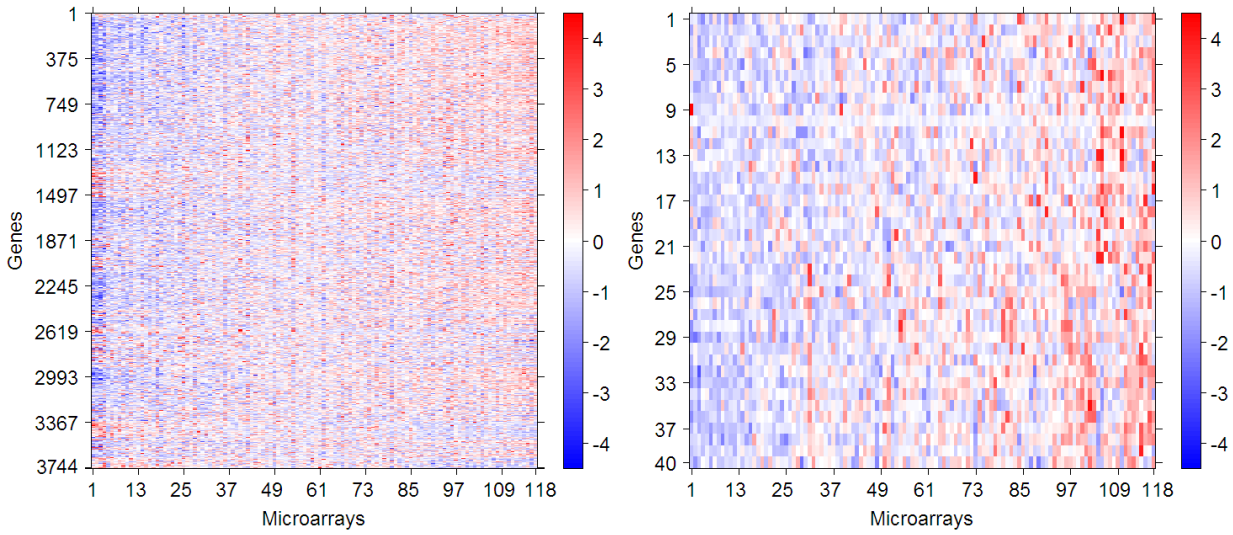

Figura 5.1: *Conjunto de dados de câncer de mama utilizados em Mayrink e Lucas (2013) e Chin et al. (2006).*

5.1 Primeira Aplicação

Nesta etapa é feita a aplicação da modelagem proposta nos Capítulos 3 e 4 para investigar um dos conjuntos de dados de câncer de mama. Inicialmente, essas aplicações serão realizadas usando a base de dados Chin com as regiões (2,4). As especificações *a priori* para α, σ^2, λ e ϕ foram as mesmas consideradas nos capítulos anteriores e encontram-se em (2.2), (2.3), (2.4) e (2.7), respectivamente. Para q_{il} e ρ_i atribuiu-se as distribuições *a priori* das Tabelas 2.2 e 3.2 que garantem a identificação do modelo fatorial e estabelecem a relação grupo-fator. Vale ressaltar que não será assumido efeito de interação para os genes em $(G_1 \cup G_2)$, pois pressupõe-se que eles serão influenciados por cada fator individualmente. Novamente, destaca-se que para o tipo de análise que está sendo desenvolvido nesta tese, os genes são agrupados segundo as interações. Por isso, inicialmente é adotado como especificação *a priori* para os $F_{i\bullet}$'s, o modelo de mistura em (3.3) considerando 51 componentes ($R = 50$). Lembrando que a primeira componente representa a interação nula. A escolha de $R = 50$ é feita devido a configuração atribuída

aos hiperparâmetros da distribuição Dirichlet, especificada para os pesos da mistura. Pois, como se quer que haja poucas interações não nulas influenciando os genes, será permitido que estes efeitos, os quais são mais relevantes, possam ser agrupados em uma quantidade razoável de grupos menor que 50.

Na Tabela 5.2 estão resumidas todas as diferentes configurações de distribuições *a priori* para os pesos ρ_i e ρ (em G_E) que serão analizadas nesta aplicaçao real. Veja que estão sendo consideradas diversas distribuições Dirichlet que se diferenciam de acordo com o primeiro hiperparâmetro que são 1, 51, 10^2, 10^3 e 10^5; ou seja, serão exploradas distribuições diferentes que colocam pesos cada vez maiores para a componente degenerada no vetor nulo; enquanto que para os demais hiperparâmetros, a partir do segundo, foi estabelecido sempre o mesmo valor que é 10^{-5} ou 10^{-10}. Esta situação faz o modelo decidir com base nos dados, quantos grupos serão formados pelos efeitos de interação mais relevante. Valores maiores do que 10^{-5} foram testados e determinaram uma quantidade de interações muito acima do que é razoável. Destaca-se também que como se trata de um conjunto de dados com uma dimensão muito grande, as distribuições Dirichlet com hiperparâmetro maiores do que 10^{-5}, a partir do segundo, seriam pouco informativas e perderiam seu efeito e importância no modelo. Neste caso, essas especificações *a priori* seriam facilmente dominadas pelos dados além de inflar o modelo com muitas interações. Nas demais colunas da Tabela 5.2 são apresentados o número de grupos e o número de interações não nulas identificadas. Veja, por exemplo, que na primeira linha foram detectados 1451 genes afetados por interações não nulas, e esses genes estão agrupados em 48 grupos diferentes. A mesma interpretação é feita para o caso ρ. Observe que ao usar a estratégia com ρ, a quantidade de grupos que são identificados geralmente é menor. Enquanto que o número de interações não nulas identificadas é maior. Uma possível explicação é que, neste caso, o parâmetro ρ é tratado como uma probabilidade global para todos os efeitos de interação. Por isso, todos os genes acabam influenciando a sua estimação. Nesta situação essas probabilidades, ρ, que não são específicas para cada gene serão únicas e acabam influenciando os genes que não seriam afetados por interação, determinando que eles terão um impacto deste tipo de efeito. Veja que a distribuição *a posteriori* condicional completa de ρ (no Apêndice B) leva em conta todos os z_i's para

sua atualização no MCMC. Ao utilizar a estratégia com ρ_i pode-se perceber que há uma influência mais local. Isso porque esta estratégia leva em conta apenas o gene i para estimar ρ_i. Neste caso a estimação de ρ_i não fica relacionada a todos os genes e isso evita o problema de se ter um gene sem interação sendo detectado com uma interação inexistente.

Tabela 5.2: Distribuições *a priori* usadas para ρ_i e ρ em G_E, número de grupos e interações não nulas identificadas.

ρ_i	Número de grupos	$F_{i\bullet}$'s não nulo
$\text{Dir}(1, 10^{-5}, \cdots, 10^{-5})$	48	1451
$\text{Dir}(51, 10^{-5}, \cdots, 10^{-5})$	46	891
$\text{Dir}(10^2, 10^{-5}, \cdots, 10^{-5})$	40	819
$\text{Dir}(10^3, 10^{-5}, \cdots, 10^{-5})$	39	607
$\text{Dir}(10^5, 10^{-5}, \cdots, 10^{-5})$	12	78
$\text{Dir}(1, 10^{-10}, \cdots, 10^{-10})$	46	1240
$\text{Dir}(51, 10^{-10}, \cdots, 10^{-10})$	44	859
$\text{Dir}(10^2, 10^{-10}, \cdots, 10^{-10})$	46	703
$\text{Dir}(10^3, 10^{-10}, \cdots, 10^{-10})$	40	432
$\text{Dir}(10^5, 10^{-10}, \cdots, 10^{-10})$	16	96
ρ	Número de grupos	$F_{i\bullet}$'s não nulo
$\text{Dir}(1, 10^{-5}, \cdots, 10^{-5})$	11	2545
$\text{Dir}(51, 10^{-5}, \cdots, 10^{-5})$	17	2247
$\text{Dir}(10^2, 10^{-5}, \cdots, 10^{-5})$	12	2462
$\text{Dir}(10^3, 10^{-5}, \cdots, 10^{-5})$	14	2456
$\text{Dir}(10^5, 10^{-5}, \cdots, 10^{-5})$	12	2286
$\text{Dir}(1, 10^{-10}, \cdots, 10^{-10})$	9	2312
$\text{Dir}(51, 10^{-10}, \cdots, 10^{-10})$	14	2486
$\text{Dir}(10^2, 10^{-10}, \cdots, 10^{-10})$	16	2547
$\text{Dir}(10^3, 10^{-10}, \cdots, 10^{-10})$	11	3066
$\text{Dir}(10^5, 10^{-10}, \cdots, 10^{-10})$	24	3223

O próximo passo é ajustar o modelo fatorial com interações, mas utilizando a abordagem via misturas por processo Dirichlet. Para esta modelagem, usou-se a construção *stick-breaking* para os pesos ρ_i e ρ truncando a representação em $R = 50$. Além disso, o parâmetro de comprimento-escala ϕ será estimado e a função de covariâncias usada no processo Gaussiano, representando a distribuição base, será a exponencial quadrática. Na Tabela 5.3 mostram-se todas as configurações do PD usado na modelagem das interações, diante do aumento do parâmetro de concentração τ variando de 0.10 à 1.00. Veja que a quantidade de grupos formados com os genes afetados por interação é geralmente menor ao utilizar a estratégia com ρ_i, assim como a quantidade de interações não nulas. Observe também que ao diminuir o valor do parâmetro de concentração τ, há uma redução na quantidade de genes afetados por interações não nulas. Além disso, há também uma redução no número de grupos formados por estes genes. Ao analizar a modelagem utilizando a estratégia com ρ, observa-se que o número de grupos formados com genes afetados por interação é maior em relação ao número de grupos identificados diante da estratégia com ρ_i. Nota-se também que ao usar um parâmetro de concentração baixo ($\tau = 0.10$ ou 0.15), não é possível identificar nenhum grupo de genes formados com efeitos de interação não nulos. Isso porque valores bastante pequenos para τ fazem com que o peso da primeira componente $\delta_{\mathbf{0}}(F_{i\bullet})$, seja muito maior do que as demais. Isso ocasiona a estimação de uma grande quantidade de efeitos de interação nulos. Neste caso, o modelo fatorial utilizado com essas configurações *a priori* para ρ e com parâmetro de concentração $\tau = 0.10$ ou 0.15, se reduz ao ajuste de um modelo fatorial sem interações, que é um caso particular da modelagem proposta aqui.

Tabela 5.3: Especificações *a priori* para ρ_i e ρ baseadas na construção via *Stick-Breaking*.

Parâmetros	Tipos de Processos	Número de grupos	$F_{i\bullet}$'s não nulo
ρ_i	$PD(1.00, N_n[\mathbf{0}, K(\lambda, \phi)])$	8	2644
	$PD(0.50, N_n[\mathbf{0}, K(\lambda, \phi)])$	9	2122
	$PD(0.20, N_n[\mathbf{0}, K(\lambda, \phi)])$	4	453
	$PD(0.15, N_n[\mathbf{0}, K(\lambda, \phi)])$	3	402
	$PD(0.10, N_n[\mathbf{0}, K(\lambda, \phi)])$	1	108
ρ	$PD(1.00, N_n[\mathbf{0}, K(\lambda, \phi)])$	18	3139
	$PD(0.50, N_n[\mathbf{0}, K(\lambda, \phi)])$	19	2574
	$PD(0.20, N_n[\mathbf{0}, K(\lambda, \phi)])$	12	2370
	$PD(0.15, N_n[\mathbf{0}, K(\lambda, \phi)])$	0	0
	$PD(0.10, N_n[\mathbf{0}, K(\lambda, \phi)])$	0	0

Na Figura 5.2, pode-se observar o gráfico feito com as médias *a posteriori* de ρ_i^* construído via $PD(0.15, N_n[\mathbf{0}, K(\lambda, \phi)])$. Veja que há um rápido decaimento dos pesos *a posteriori* para zero. Aqui o processo *stick-breaking* atribui pesos iguais a zero a partir da quinta compontente, não criando mais do que 3 grupos não nulos de interações. Note que a primeira componente é degenerada no vetor nulo e apresenta o maior peso *a posteriori* como era esperado. No painel à direita estão exibidos esses pesos das quatro primeiras componentes para melhor visualização.

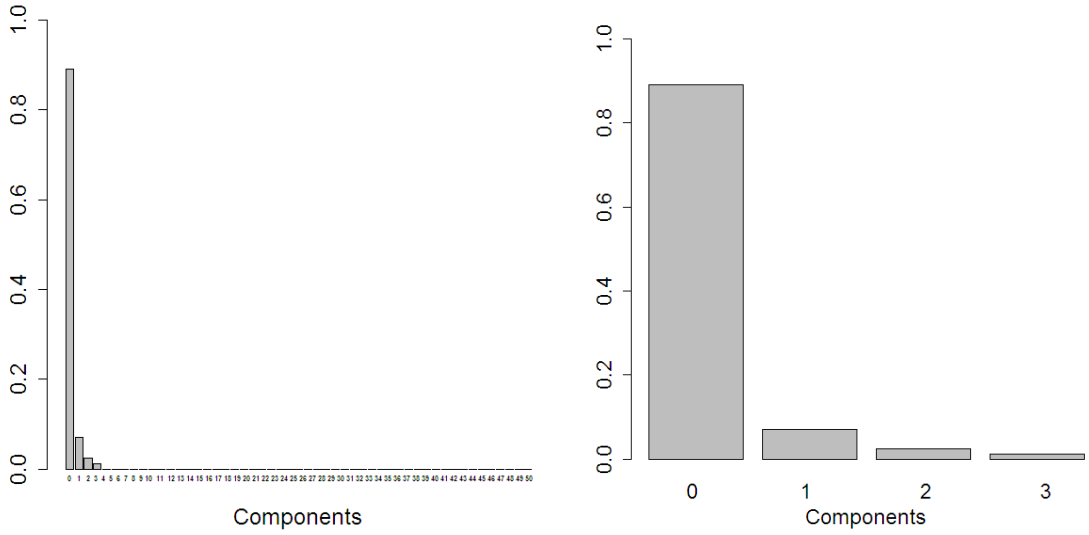

Figura 5.2: *Gráfico da média a posteriori dos pesos de cada componente criada pelo processo stick-breaking considerando* $PD\left(0.15, N_n\left[\mathbf{0}, K(\lambda, \phi)\right]\right)$.

A Figura 5.3 exibe gráficos de superfícies representando os três tipos de efeitos de interação identificados para os 402 genes. O modelo considerado aqui é com ρ_i e construído via $PD\left(0.15, N_n\left[\mathbf{0}, K(\lambda, \phi)\right]\right)$. Veja que estas representações das superfícies, na Figura 5.3, justificam a nomenclatura de processo Gaussiano que está sendo usada nas expressões (2.5), (3.3) e (4.5). Observe que o formato das superfícies são diferentes indicando efeitos distintos de interações para os diferentes grupos de genes. Essas superficies são construídas usando a média *a posteriori* da interação F_r^*. É importante ressaltar que é possível também construír intervalos de credibilidade para essas superfícies, indicando que podem haver variações neste intervalo. Os painéis da direita na Figura 5.3 representam os gráficos de imagem e contorno dessas superfícies. Veja que no Painel (a) as regiões em amarelo indicam os maiores valores de interação e encontram-se dispersos na imagem. Diferentemente do que acontece com os efeitos de interação exibidos nos Painéis (b) e (c), que parecem mostrar algum tipo de padrão similar aos dois. Por exemplo, no Painel (c) os valores do efeito de interação são maiores para os pares (λ_{1j},λ_{2j}) em que λ_{1j} é alto e λ_{2j} é baixo. Observe que nas superfícies os eixos na vertical não estão na mesma as escalas, porque os efeitos de interação são diferentes e aqui não está sendo feita uma análise comparativa entre interação real e estimada.

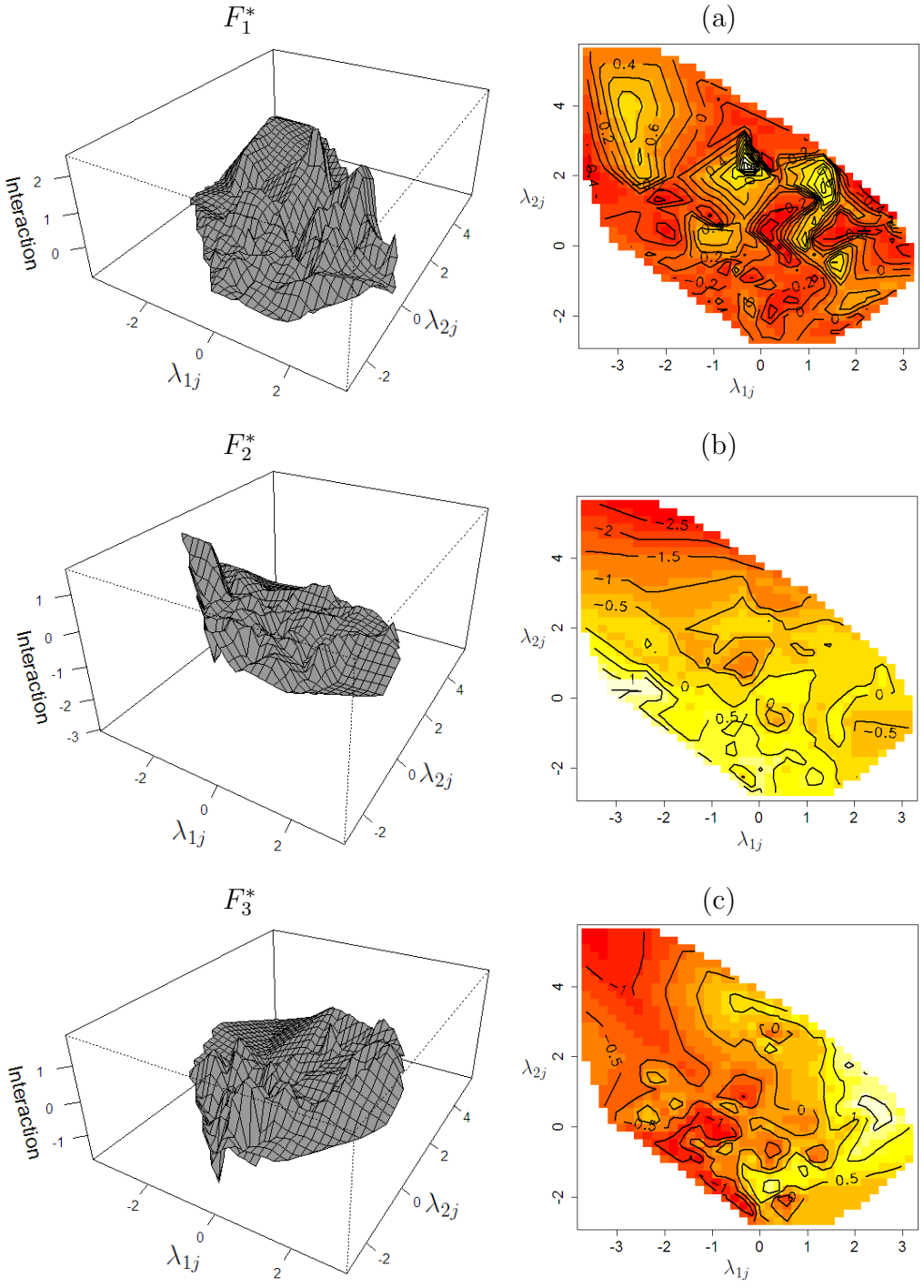

Figura 5.3: *Gráficos de superfície e contorno dos três tipos de interação identificadas e estabelecidas para os 402 genes agrupados utilizando a modelagem via* $PD\left(0.15, N_n\left[\mathbf{0}, K(\lambda, \phi)\right]\right)$.

A Figura 5.4 apresenta a matriz $\hat{F}$ das interações. O Painel (a) mostra a matriz $\hat{F}$ completa com todos os efeitos de interação identificados para os genes. Veja que no topo desta matriz há somente interações nulas. Isso serve para relembrar ao leitor da suposição inicial estabelecida neste trabalho, em que a interação dos fatores não afeta os genes dos grupos G_1 e G_2 localizados no topo desta matriz. Veja que a maioria das linhas está em branco indicando que há uma maior quantidade de efeitos de interação nulos.

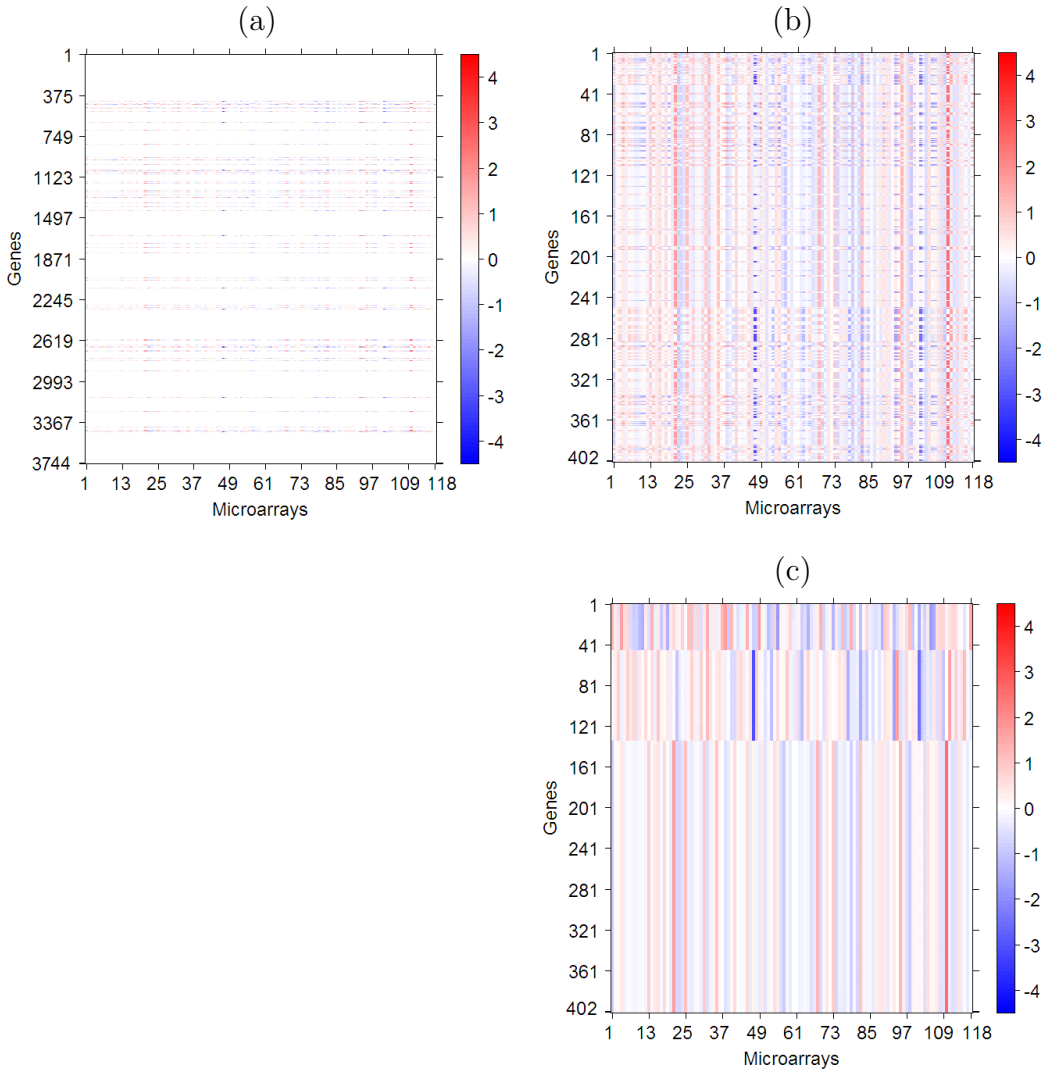

Figura 5.4: *Gráficos de imagem da matriz $\hat{F}$. O painel (a) exibe a matriz completa com todos os efeitos de interação estimados. O painel (b) mostra somente os casos onde foram detectados efeitos de interação não nulos. O painel (c) apresenta essas interações não nulas ordenadas por linha em relação à mediana.*

Ainda na Figura 5.4 é exibido no Painel (b) apenas as 402 interações não nulas sem qualquer tipo de organização das linhas. O Painel (c) mostra as mesmas 402 interações porém elas foram organizadas em termos da mediana para permitir uma melhor visualização dos efeitos de interação por grupos. Note que, após essa organização mencionada, fica nitidamente visível os tamanhos dos 3 grupos de genes formados por esses 402 efeitos de intaração.

Na Figura 5.5 pode-se observar o gráfico com as médias *a posteriri* das cargas referentes aos grupos G_1 e G_2 juntamente com seus intervalos HPD. Veja que a maioria das cargas α_{i1} em G_1 são negativas, enquanto que a maioria das cargas α_{i2} em G_2 são positivas. Este é um resultado esperado, pois mostra a direção e o efeito de cada fator associado aos genes desses grupos. É importante destacar que se houvesse uma variação intensa entre os sinais das cargas dentro de G_1 e G_2, isso indicaria um problema na associação dos fatores com estes grupos, sinais contrários indicam comportamentos diferentes do fator dentro do grupo. Esses resultados exibidos para as cargas são semelhantes aos encontrados por Mayrink e Lucas (2013) quando utilizaram o modelo com todas as interações distintas.

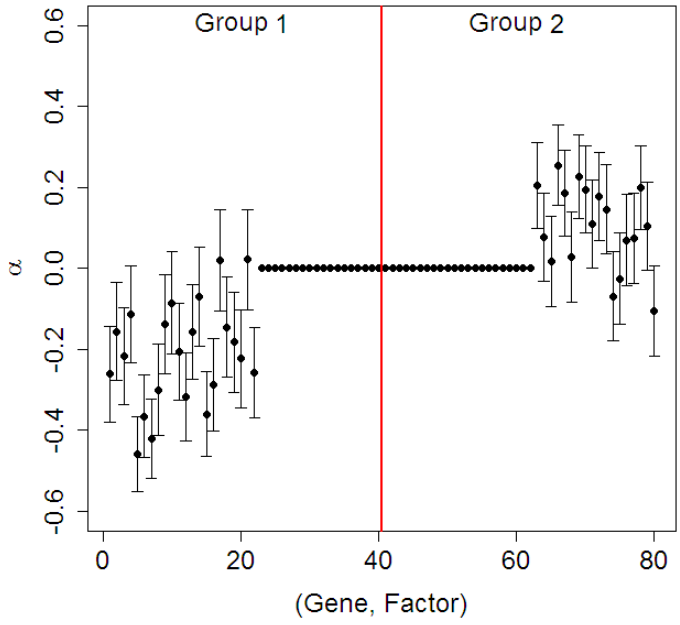

Figura 5.5: *Gráficos com as médias a posteriori (círculo) e intervalos HPD de 95% de credibilidade sendo representado pelo segmento de reta na vertical.*

Considerando as avaliações feitas anteriormente e as diferentes versões do modelo usando as estatégias com ρ_i e ρ, sugere-se ao usuário, diante destas modelagens propostas, que utilizem a abordagem estabelecida mediante a estratégia com ρ_i. A justificativa para isso é que ela se mostrou mais maleável do que a estratégia com ρ, que pareceu estar inflando com mais interações o modelo fatorial. Uma das suposições estabelecidas na aplicação real é de que haja poucas interações associadas aos genes e isto está de acordo com a estratégia ao utilizar ρ_i. Além disso, vale ressaltar que um modelo com muitos efeitos de interação requer a estimação de quantidades maiores de parâmetros, tornando-se um modelo não parcimonioso. A modelagem usando ρ fornece mais interações do que se é esperado, além da necessidade de mais interpretações e explicações para os componentes do modelo. Outro fato impotante é que a estimação de ρ está relacionado a todo os genes e este acaba atribuindo efeitos de interação onde realmente não existem. Logo, o ajuste feito usando a abordagem com ρ_i estabelece uma análise mais parcimoniosa, destacando para o usuário as interações mais importantes e relevantes.

5.2 Segunda Aplicação

Nesta seção será apresentada uma análise envolvendo as quatro bases de dados mencionadas no início deste capítulo. Esta etapa consiste em mostrar alguns resultados referentes à três versões dos modelos estudados na Seção 5.1, que utilizam misturas finita e o PD para $F_{i\bullet}$ considerando ρ_i. O algoritmo MCMC apresentou 4000 iterações e os tempos computacionais (em horas), referentes ao banco de dados Wang (maior base) com as regiões (1,4), foram 18,79h para a modelagem via mistura com $\rho_i \sim \text{Dir}(10^3, 10^{-10}, \cdots, 10^{-10})$, e 17,09h para a modelagem via PD com $\tau = 0.10$. As aplicações foram feitas na maquina com processador Intel Core $i7$ com 16GB de mémoria RAN e sistema operacional Ubunto 16 A execução do código é feita com exclusividade, sem nenhum outro programa executado em paralelo.

Assim como abordado por Mayrink e Lucas (2013), serão identificados o número de genes comuns que pertencem a interseção do grupo G_E dos diferentes conjuntos de dados. Espera-se que haja uma alta interseção entre os genes destes conjuntos de dados, pois

está sendo analizado o mesmo tipo de câncer (mama).

A Tabela 5.4 apresenta uma comparação dois a dois dos dados reais mostrando o número de genes afetados por interação para os quatro casos. Ela está divida em quatro partes. As três primeiras correspondem a cada um dos pares formados com duas regiões. Por exemplo, a diagonal principal (localizada na primeira parte no topo da Tabela 5.4) refere-se ao número de genes afetados por interação em cada um dos conjuntos de dados analizados individualmente. Fora desta diagonal principal estão as quantidades de genes afetados por interação encontrados na interseção dois a dois destes dados, isto é, ao avaliar as regiões (1,4) na base de dados Chin foram identificados os mesmos 100 genes afetados por interação encontrados na base de dados Miller. A última parte da Tabela 5.4 exibe o número de grupos formados por interações não nulas encontradas considerando cada região. Neste caso, considere novamente a base de dados Chin e as regiões (1,4). Nela foram identificadas 590 efeitos de interação não nulos agrupados em 44 grupos distintos. Vale destacar que Mayrink e Lucas (2013) conseguiram identificar poucos genes na interseção dois a dois destes quatro conjuntos de dados, variando de 1 à 9, 7 à 14 e 2 à 11 genes detectados com os pares de regiões (1,4), (2,4) e (3,4), respectivamente. Vale ressaltar que o modelo ajustado pelos autores para identificar estes genes encontrados na interseção dos dados, era o modelo fatorial considerando a interação multiplicativa dos fatores com superfície em formato de sela. Diferente do que foi feito pelos autores o modelo especificado nesta seção estabelece vários formatos de superficies de efeitos de interação. Os resultados que estão propostos aqui identificam quantidades maiores de genes afetados por interação e que pertencem a interseção dois a dois de G_E nestas quatro bases de dados. Aqui este número de genes encontrados na interseção, usando o modelo de misturas com $\rho_i \sim \text{Dir}(10^3, 10^{-10}, \cdots, 10^{-10})$, varia de 12 à 119, 31 à 118 e 23 à 84 genes identificados usando os pares de regiões (1,4), (2,4) e (3,4), respectivamente.

Tabela 5.4: Comparação dois a dois dos conjuntos de dados reais. As três primeiras partes na tabela exibem o número genes afetados por interação e identificados na interseção de G_E para as bases de dados em cada par de regiões. A modelagem é feita via misturas finitas utilizando $\rho_i \sim \text{Dir}(10^3, 10^{-10}, \cdots, 10^{-10})$. Na base da tabela apresentam-se o número de grupos não nulos para cada par de regiões.

Regiões (1,4)	Chin	Miller	Sotiriou	Wang
Chin	590	100	119	59
Miller	100	252	83	18
Sotiriou	119	83	314	12
Wang	59	18	12	227
Regiões (2,4)	Chin	Miller	Sotiriou	Wang
Chin	432	118	76	41
Miller	118	383	100	31
Sotiriou	76	100	313	32
Wang	41	31	32	202
Regiões (3,4)	Chin	Miller	Sotiriou	Wang
Chin	498	84	80	42
Miller	84	311	69	23
Sotiriou	80	69	286	29
Wang	42	23	29	193

Dados	Número de grupos não nulos e regiões		
	(1,4)	(2,4)	(3,4)
Chin	44	40	43
Miller	38	45	44
Sotiriou	35	40	39
Wang	36	37	35

A Tabela 5.5 exibe o número de genes afetados por interações identificadas em G_E considerando as interseções três a três dos conjuntos de dados. Essa análise é feita

para cada par de regiões e o modelo ajustado aqui é o de misturas finita com $\rho_i \sim \text{Dir}(10^3, 10^{-10}, \cdots, 10^{-10})$. A última linha na Tabela 5.5 mostra o número de genes comumente afetados por interações identificadas nas quatro bases de dados. Por exemplo, ao avaliar as quatro bases com as regiões (1,4), foi detectado 1 gene afetado por interação, sendo ele comum aos quatro conjuntos de dados. Além deste gene, também foi identificado um outro apresentando 1 efeito de interação comum às quatro bases de dados, isso ocorreu para as regiões (3,4).

Tabela 5.5: Interseção três a três dos conjuntos de dados. Quantidades de genes simultaneamente afetados por interação em três bases de dados. A linha final faz a comparação dos quatro conjuntos de dados. A modelagem é feita via misturas finita utilizando $\rho_i \sim \text{Dir}(10^3, 10^{-10}, \cdots, 10^{-10})$.

Dados	Regiões e número de interações não nulas		
	(1,4)	(2,4)	(3,4)
Chin, Miller, Sotiriou	34	25	23
Chin, Miller, Wang	6	12	5
Miller, Sotiriou, Wang	3	7	6
Chin, Sotiriou, Wang	5	1	4
Chin, Miller, Sotiriou, Wang	1	4	1

Na Figura 5.6 pode-se observar o efeito de interação que afeta este gene comum nas quatro bases de dados de câncer de mama para as regiões (1,4). Veja que as superfícies representando a interação são semelhantes e parecem apresentar algum tipo de padrão exibindo uma certa inclinação. Os painéis (a), (b) (c) e (d) mostram os gráficos de imagem e contorno dessas superfícies identificadas em cada base de dados. Observe que as regiões em amarelo, representando valores grandes da interação, encontram-se nas extremidades acima e à direita na imagem. Note que este efeito de interação é maior quando os pares $(\lambda_{1j}, \lambda_{2j})$ apresentam valores grandes e é menor ao observar os valores dos pares $(\lambda_{1j}, \lambda_{2j})$ baixos.

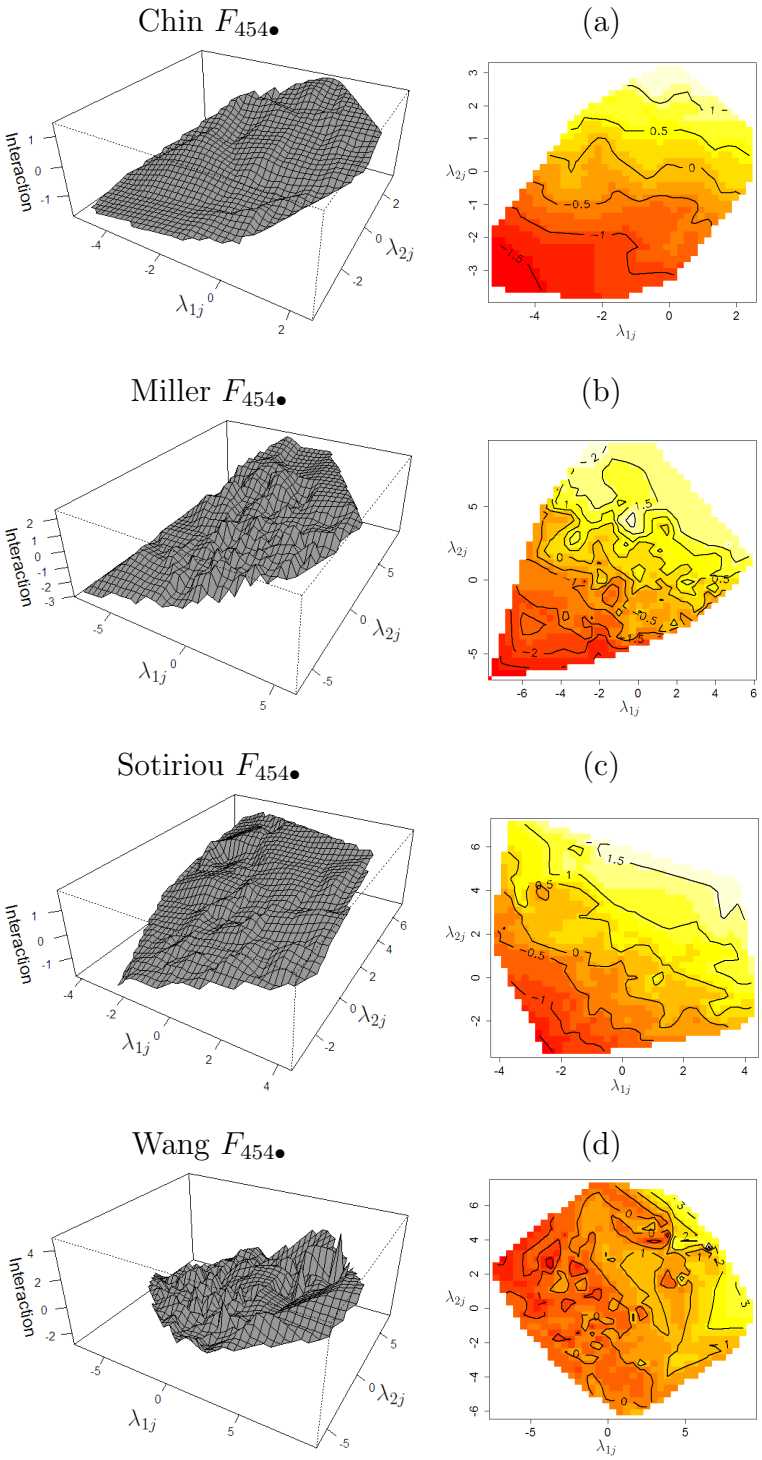

Figura 5.6: *Gráficos de superfície e contorno do efeito de interação detectado para o mesmo gene nos quatro conjuntos de dados. A modelagem é feita com* $\rho_i \sim Dir(10^3, 10^{-10}, \cdots, 10^{-10})$.

A Figura 5.7 exibe os gráficos com as médias *a posteriori* e intervalos HPD das cargas relacionadas aos grupos G_1 e G_2 para os quatro conjuntos de dados considerando o par de regiões (1,4). Note que a maioria das cargas α_{il}, relacionadas ao fator l, apresentam estimativas com o mesmo sinal. Este fato é observado em quase todos os dados e para a maioria das cargas relacionadas aquele grupo. Isso mostra a direção e influência de cada fator relacionado a estes grupos de genes afetados pela CNA.

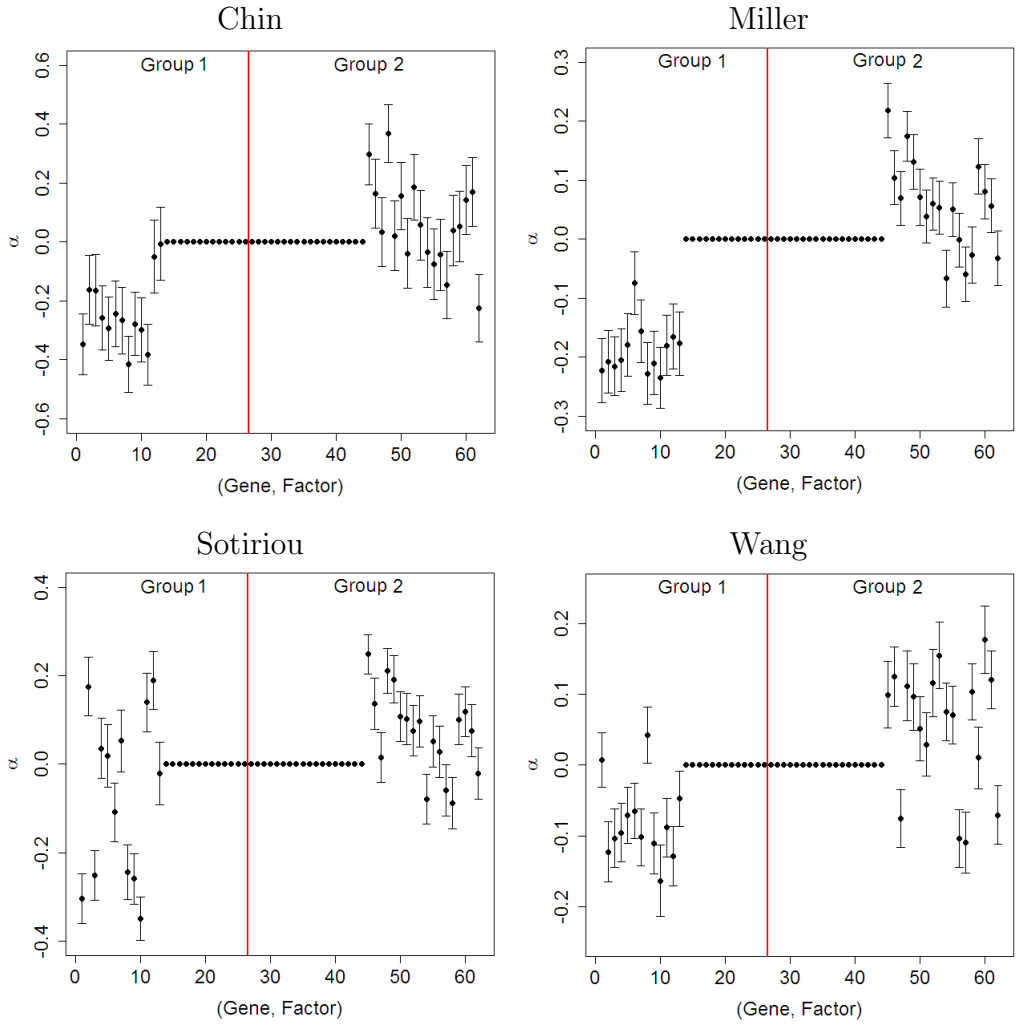

Figura 5.7: *Gráficos com as médias a posteriori (círculo) e intervalos HPD de 95% de credibilidade sendo representado pelo segmento de reta na vertical. Resultados relacionados ao par das regiões (1,4). A modelagem é feita com* $\rho_i \sim Dir(10^3, 10^{-10}, \cdots, 10^{-10})$.

Análises envolvendo o modelo com ρ_i construído via $PD(0.15, N_n[\mathbf{0}, K(\lambda, \phi)])$ também foram feitas. Os resultados obtidos ao considerar este tipo de modelagem são semelhantes ao do modelo com misturas finita. A Tabela 5.6 mostra o número de genes afetados por interações e que estão na interseção dois a dois das quatro bases de dados. A Tabela 5.6 também está divida em 4 partes, assim como a Tabela 5.4. Na diagonal principal das três primeiras partes estão o número de genes afetados por interação em cada conjunto de dados. Fora desta diagonal está o número de genes afetados por interação que pertencem a interseção entre os grupos G_E dos dados analizados dois a dois. Na última parte da Tabela 5.6 estão o número de grupos não nulos formados com o agrupamento das interações diante de cada par de regiões. Veja que o número de grupos não nulos formados com as interações são bem menores do que os apresentados na Tabela 5.4. Note que o número de genes aferados por interação é maior nos dados Miller, Sotiriou e Wang ao ajustar o modelo considerando $PD(0.15, N_n[\mathbf{0}, K(\lambda, \phi)])$, entretanto, com a base de dados Chin, a modelagem mostrou uma quantidade de genes afetados por interação sendo menor em relação ao modelo de misturas finita com $\rho_i \sim \text{Dir}(10^3, 10^{-10}, \cdots, 10^{-10})$. Além disso, na Tabela 5.6, ao analisar o número de genes encontrados na interseção dois a dois destes dados, encontrou-se uma quantidade maior destes genes em relação aos que foram apresentados na Tabela 5.4.

Tabela 5.6: Comparação dois a dois dos conjuntos de dados reais. As três primeiras partes na tabela exibem o número genes afetados por interação e identificados na interseção de G_E para as bases de dados em cada par de regiões. A modelagem é feita usando o $PD(0.15, N_n[\mathbf{0}, K(\lambda, \phi)])$ com ρ_i contruído via *stick-breaking*. Na base da tabela apresentam-se o número de grupos não nulos para cada par de regiões.

Regiões (1,4)	Chin	Miller	Sotiriou	Wang
Chin	327	280	244	267
Miller	280	1120	583	561
Sotiriou	244	583	1275	501
Wang	267	561	501	1041
Regiões (2,4)	Chin	Miller	Sotiriou	Wang
Chin	402	250	275	329
Miller	250	714	340	489
Sotiriou	275	340	1194	706
Wang	329	489	706	1909
Regiões (3,4)	Chin	Miller	Sotiriou	Wang
Chin	287	161	238	198
Miller	161	654	343	372
Sotiriou	238	343	1606	562
Wang	198	372	562	1183

Dados	Número de grupos não nulos e regiões		
	(1,4)	(2,4)	(3,4)
Chin	3	3	4
Miller	6	4	6
Sotiriou	6	6	5
Wang	8	6	6

Na Tabela 5.7 são apresentados o número de genes afetados por interação e detectados na interseção três a três dos quatro conjuntos de dados considerando cada par de regiões.

Observe que a modelagem feita com $PD(0.15, N_n[\mathbf{0}, K(\lambda, \phi)])$ identificou quantidades maiores de genes afetados em relação ao número encontrado via modelo de misturas finito. Veja que foram identificados no mínimo 140 genes afetados por interação comuns para três de quantro bases de dados referente as regiões (3,4). Note também que ao avaliar a interseção dos quatro dados foram identificados para os pares (1,4), (2,4) e (3,4), as quantidades de 215, 199 e 131 genes afetados por interação, respectivamente. Estas quantidades são bem maiores em relação as que foram encontradas e exibidas na Tabelas 5.5.

Tabela 5.7: Número genes afetados por interação e identificados na interseção três a três dos conjuntos de dados. A linha final apresenta o número de genes detectados com interação comum nas quatro bases de dados. A modelagem é feita usando o $PD(0.15, N_n[\mathbf{0}, K(\lambda, \phi)])$ com ρ_i contruído via *stick-breaking*.

Dados	Regiões e número de interações não nulas		
	(1,4)	(2,4)	(3,4)
Chin, Miller, Sotiriou	227	206	151
Chin, Miller, Wang	254	237	140
Miller, Sotiriou, Wang	382	289	253
Chin, Sotiriou, Wang	221	233	170
Chin, Miller, Sotiriou, Wang	215	199	131

A Tabela 5.8 exibe os resultados encontrados ao ajustar o modelo com PD considerando agora o parâmetro de concentração $\tau = 0.10$. Veja que ao reduzir o valor deste parâmetro, o número de genes afetados por interação diminui em relação a modelagem com $\tau = 0.15$, assim como a quantidade de grupos formados com as interações não nulas. Isso é um resultado esperado devido a metodologia apresentada e discutida no final da Seção 4.2 do Capítulo 4. Observe também que o número de genes encontrados na interseção dois a dois dos dados Miller, Sotiriou e Wang, ainda é maior quando comparado ao número de genes detectados ao usar o modelo de misturas finito. Entretanto, o número de grupos formados com os efeitos de interação não nulos é menor. Compare os

resultados das Tabelas 5.4 e 5.8.

Tabela 5.8: Comparação dois a dois dos conjuntos de dados reais. As três primeiras partes na tabela exibem o número genes afetados por interação e identificados na interseção de G_E para as bases de dados em cada par de regiões. A modelagem é feita usando o $PD(0.10, N_n[\mathbf{0}, K(\lambda, \phi)])$ com ρ_i contruído via *stick-breaking*. Na base da tabela apresentam-se o número de grupos não nulos para cada par de regiões.

Regiões (1,4)	Chin	Miller	Sotiriou	Wang
Chin	140	105	76	109
Miller	105	541	165	228
Sotiriou	76	165	786	267
Wang	109	228	267	962
Regiões (2,4)	Chin	Miller	Sotiriou	Wang
Chin	108	99	87	99
Miller	99	621	270	333
Sotiriou	87	270	754	276
Wang	99	333	276	889
Regiões (3,4)	Chin	Miller	Sotiriou	Wang
Chin	49	31	25	43
Miller	31	498	259	291
Sotiriou	25	259	826	268
Wang	43	291	268	805

Dados	Número de grupos não nulos e regiões		
	(1,4)	(2,4)	(3,4)
Chin	3	1	1
Miller	4	4	3
Sotiriou	3	3	4
Wang	5	4	4

Na Tabela 5.9 pode-se observar o número de genes encontrados na interseção três a três dos dados. Veja que o número de genes afetados por interação comum à três das quatro bases de dados é menor em relação aos mostrados na Tabela 5.7. Note que ao considerar o par de regiões (3,4), 21 genes afetados por interação foram encontrados na interseção dos dados Chin, Miller e Sotiriou. Além disso, ao avaliar as bases Miller, Sotiriou e Wang foram identificados 191 genes afetados por interação e comuns a esses três dados. Ao analisar a interseção dos quatro conjuntos de dados foram detectados 47, 82 e 21 genes afetados por interação para as regiões (1,4), (2,4) e (3,4), respectivamente. Essas quantidades são ainda maiores do que as encontradas quando utilizado o modelo de misturas finito (veja a Tabela 5.5) e são menores em relação à modelagem usando PD com parâmetro $\tau = 0.15$ exibidas na Tabela 5.7.

Tabela 5.9: Número genes afetados por interação e identificados na interseção três a três dos conjuntos de dados. A linha final apresenta o número de genes detectados com interação comum nas quatro bases de dados. A modelagem é feita usando o $PD(0.10, N_n[\mathbf{0}, K(\lambda, \phi)])$ com ρ_i contruído via *stick-breaking*.

Dados	Regiões e número de interações não nulas		
	(1,4)	(2,4)	(3,4)
Chin, Miller, Sotiriou	56	83	21
Chin, Miller, Wang	81	96	31
Miller, Sotiriou, Wang	114	197	191
Chin, Sotiriou, Wang	62	85	24
Chin, Miller, Sotiriou, Wang	47	82	21

A Tabela 5.10 exibe algumas estimativas *a posteriori* para o parâmetro de comprimento escala ϕ da função de covariâncias Gaussiana. Os resultados apresentados são referentes ao ajuste dos modelos para as regiões (2,4). Os resultados com as estimativas de ϕ para os outros pares estão no Apêndice D. Observa-se que algumas medianas são maiores que a média sugerindo uma distribuição com certa assimetria. Veja que ao ajustar o modelo de misturas na base de dados Miller, obteve-se uma estimativa média

de 0.25 enquanto que no modelo via PD com $\tau = 0.15$ e $\tau = 0.10$, este valor fica entre de 0.19 e 0.38, respectivamente. Vale lembrar que o espaço paramétrico de ϕ varia entre 0.1 à 0.5. Note também que, para as bases Chin e Sotiriou, a modelagem via misturas apresenta estimativas para a mediana semelhantes aos resultados encontrados com a modelagem via $PD(0.15, N_n[\mathbf{0}, K(\lambda, \phi)])$.

Tabela 5.10: Estimativas *a posteriori* para o parâmetro de comprimento-escala ϕ da função de covariâncies Gaussiana. Resultados referentes ao ajuste do modelo de mistura finito e com PD para o par de regiões (2,4).

Especificações de ρ_i	Dados	Estimativas *a posteriori*			
		Mediana	Média	Desvio Padrão	Intervalo HPD
$\text{Dir}(10^3, 10^{-10}, \cdots, 10^{-10})$	Chin	0.3430	0.3257	0.1033	[0.1739 ; 0.4949]
	Miller	0.2198	0.2531	0.0974	[0.1339 ; 0.4682]
	Sotiriou	0.3676	0.3728	0.0613	[0.2653 ; 0.4854]
	Wang	0.3667	0.3589	0.0784	[0.2230 ; 0.4906]
$PD(0.15, N_n[\mathbf{0}, K(\lambda, \phi)])$	Chin	0.3467	0.3506	0.0788	[0.2284 ; 0.5000]
	Miller	0.1913	0.1925	0.0487	[0.1048 ; 0.2743]
	Sotiriou	0.3694	0.3242	0.1312	[0.1123 ; 0.4956]
	Wang	0.3255	0.3295	0.0441	[0.2525 ; 0.4183]
$PD(0.10, N_n[\mathbf{0}, K(\lambda, \phi)])$	Chin	0.2669	0.2928	0.1171	[0.1293 ; 0.4830]
	Miller	0.3832	0.3813	0.0594	[0.2665 ; 0.4829]
	Sotiriou	0.1554	0.1911	0.0854	[0.1001 ; 0.3569]
	Wang	0.3321	0.3083	0.0850	[0.1576 ; 0.4661]

Devido aos estudos realizados e exibidos no Capítulo 3 para o parâmetro ϕ, chegou-se conclusão de que é preferível ajustar o modelo fatorial fazendo a sua estimação. Pois foi mostrado, por meio de simulações, uma melhora nas estimativas dos demais parâmetros determinando uma modelagem mais adequada aos dados. Concluiu-se que fazer a estimação de ϕ é mais vantajoso, porque permite que usuário não corra o risco de fixar o parâmetro em um valor errado, ocasionando um ajuste ruim do modelo fatorial com interações.

5.3 Conclusões do Capítulo

Este capítulo apresentou uma aplicação das duas modelagens propostas em 4 conjuntos de dados referentes ao câncer de mama. Para cada um dos dados analizados foram pré-especificados quatro posições no genoma com CNA. As bases de dados possuem 22283 genes estruturados e replicados em *microarrays*. Os tamanhos de amostrais (número de *microarrays*) variam de 118 à 286. Para reduzir o custo computacional foi realizado um procedimento de seleção desses genes descrito na Seção E do material suplementar de Mayrink e Lucas (2013). Aqui foram selecionados 3744, 3748 e 3750 genes para análises dos pares formados com as regiões (1,4), (2,4) e (3,4), respectivamente.

Na primeira aplicação foram feitas avaliações das modelagens apenas no conjunto de dados Chin com os pares (2,4). Duas estratégias foram usadas para os pesos do modelo de mistura finito e da mistura via PD, atribuídos para as interações. Após várias análises e verificações de como o modelo fatorial se comporta diante de diferentes especificações *a priori*, foi sugerido ajustar o modelo considerando a estratégia com ρ_i, pois esta abordagem mostra uma influência mais local, levando em conta apenas o gene i para sua estimação, enquanto que a estratégia de modelagem com ρ, a qual está relacionada a todos os genes, acaba atribuindo efeitos de interação onde realmente não exite. Esta probabilidade global ρ determina um modelo não parcimonioso com respeito a quantidade de interações. Isto é, estima-se mais parâmentros de interação demandando mais interpretações e explicações. Desta forma, a modelagem com ρ_i é preferível, pois estabelece uma análise parcimoniosa, neste contexto, exibindo os efeitos de interação mais importantes e relevantes.

No caso de escolher entre um dos modelos propostos, o usuário pode decidir entre duas opções: A primeira, se quiser poucas interações em uma quantidade grande de grupos, neste caso é preferível usar a modelagem via misturas finita. A segunda, seria optar por mais interações em poucas quantidades de grupos, neste caso, é sugerido usar a modelagem via PD. Essa escolha é feita em conjunto com um especialista da área de genética que estuda estes prolemas de CNA.

Os resultados após a primeira aplicação foram satisfatórios e esseciais para uma se-

gunda análise feita aqui, a qual envolveu o agrupamento das interações e a identificação dos genes afetados por elas, que são comuns aos quatro conjuntos de dados. O ajuste do modelo fatorial via misturas finita, diante de cada par de regiões, forneceu quantidades de grupos maiores em relação à modelagem via PD. Entretanto, o modelo com misturas finita conseguiu identificar poucos genes afetados por interação e que são comuns aos quatro conjuntos de dados, diferentemente da modelagem usada com PD que identificou quantidades grandes destes genes comuns às quatro bases de dados. É natural que, quando há um número maior de genes afetados por interação, haja também uma interseção não vazia entre os grupos G_E das quatro bases de dados, já que está sendo avaliado o mesmo tipo de câncer. Veja que cada tipo de modelo identificou diferentes quantidades de genes afetados por interação em diferentes conjuntos de dados. Desta maneira pode ser equivocado escolher um dos modelos baseado apenas na quantidade de genes encontrados na interseção dos grupos G_E. Pois o que está sendo verificado aqui é se há uma coerência neste tipo de modelagem. Se fosse encontrada uma interseção vazia (não contendo genes comuns) entre os dados, então haveria um problema de um modelo não plausível. Portanto, a identificação dos mesmos genes nas interseções dos dados para o mesmo tipo de câncer tem bastante importância, pois estes resultados reforçam a ideia de que a modelagem proposta é coerente para o estudo de interações.

Capítulo 6

Conclusões

Inicialmente, neste trabalho, foi exibido um breve resumo sobre análise de expressão de genes e o problema de CNA, envolvendo trechos do genoma com grupos distintos de genes. Foi apresentado o modelo fatorial com interações proposto em 2013. A partir de simulações foi verificado como o modelo se comporta diante da estimação do parâmetro de comprimento escala ϕ, presente na função de covariâncias definida no processo Gaussiano usado para modelar as interações. Diversos cenários foram feitos e analisados considerando situações onde o valor de ϕ é fixado ou estimado. Na modelagem apresentada por Mayrink e Lucas (2013) este parâmetro estava fixo em vários valores para uma análise de sensibilidade. Por isso, uma das contribuições desta tese foi fazer sua estimação. Os resultados mostraram que há uma vantagem na modelagem com a estimação de ϕ, pois proporcionou melhores estimativas para a maioria dos demais parâmetros (α, σ^2 e F), resultando em um ajuste melhor em relação a modelagem quando ϕ é fixado no valor errado. Isso mostrou que a estimação de ϕ tem grande relevância, pois em uma situação prática, em que seu valor é desconhecido, ajustar o modelo fatorial com interações considerando sua estimação pode ser uma boa opção.

O segundo estudo, desenvolvido no Capítulo 3, apresentou uma proposta de modelagem para as interações que estende o modelo de Mayrink e Lucas (2013), em que os autores consideram todas as interações não nulas sendo iguais. Na verdade duas situações extremas são abordadas no artigo referência de 2013: uma estabelece que todas as interações não nulas são iguais e a outra considera que todas são diferentes. Estas

abordagens acabam sendo uma limitação para o modelo de 2013, pois se restringem a dois casos extremos. Isso faz pensar em desenvolver um modelo que considera uma situação intermediária com a finalidade de formar grupos com as interações estimadas, em que os efeitos de interação dentro de um grupo são iguais mas diferentes entre grupos. Esta ideia proposta e desenvolvida aqui é apresentada com duas abordagens para estabelecer o agrupamento das interações. A primeira atribuiu-se às interações um modelo de misturas finito considerando várias componentes com pontos de massa. Este tipo de aboragem é uma extensão do modelo proposto por Mayrink e Lucas (2013), que consideram apenas duas componentes degeneradas, uma no vetor nulo e outra em um efeito de interação estimado e compartilhado por diversos genes. Na modelagem proposta neste trabalho, vários efeitos de interação não nulos são estimados e alocados para uma certa quantidade de grupos. Neste caso, as interações não nulas são estimadas por meio de um processo Gaussiano usando a função de covariâncias exponencial quadrática.

Muitos cenários foram simulados via Monte Carlo e avaliados afim de estudar este tipo de agrupamento proposto para as interações. Além do mais, casos em que o valor de ϕ é fixado ou estimado também foram considerados nas análises de sensibilidade. A avaliação da qualidade das estimativas dos parâmetros foi feita por meio do EQM e a qualidade dos ajustes a partir de razões construídas com os critérios DIC, WAIC e LPML. Os resultados inidicaram que a modelagem proposta com a estimação de ϕ apresenta uma vantagem em relação à modelagem considerando o valor de ϕ fixado. Essas análises revelaram o bom desempenho do modelo fatorial com agrupamento das interações, indicando que em situações onde não se conhece o valor real de ϕ, ajustar o modelo diante de sua estimação é a melhor escolha. Além disso, foi avaliado no estudo Monte Carlo a proporção de acerto que o modelo tem em identificar corretamente cada tipo de interação real. Simulações em que a matriz de dados foi gerada contendo poucos tipos de interações (grupos) foram úteis para mostrar a capacidade que o modelo ajustado, com muitos grupos ou componentes, tem em se adequar aos dados. A modelagem realizada apresentou também uma alta proporção de acerto ao identificar corretamente as interações. Estes resultados estabeleceram que não é preciso ter uma noção de quantos tipos de interação (grupos) existem, basta ajustar um modelo com uma quantidades de grupos grande. Por

conta desta questão foi desenvolvido no Capítulo 4 uma segunda abordagem propondo o agrupamento dos efeitos de interação a partir do processo Dirichlet. Esta nova abordagem representa outra contribuição deste trabalho, pois essa metodologia não foi abordada pelos autores em 2013. Neste caso, foi utilizada a propriedade de agrupamento do PD para as interações por meio de modelos de misturas.

Uma das vantagens do PD é a determinação automática do número de grupos, isto é, não há necessidade de uma especificação prévia do número de componentes de um modelo de misturas, tornando esta ideia mais atrativa e flexivél para a modelagem de agrupamento das interações. Nesta tese, a representação de Sethuraman (1994) do PD foi utilizada para modelagem dos efeitos de interação. O modelo de misturas via PD é visto como um modo alternativo à modelagem via misturas finita. No estudo simulado com vários cenários, exibido no Capítulo 4, é feito uma comparação das duas abordagens de agrupamentos. Observou-se que a abordagem usando o PD se ajustou tão bem quanto à modelagem via misturas para o agrupamento das interações em alguns cenários. No Capítulo 5, estas abordagens foram utilizadas em conjuntos de dados reais. Em uma primeira aplicação foi mostrado um estudo da base de dados Chin, em que foram testados várias configurações da distribuição *a priori* para os pesos ρ_i e ρ. A abordagem usando misturas finitas estabeleceu uma quantidade maior de grupos formados com as interações em relação à modelagem via PD. Além disso, também foram testados diferentes valores do parâmetro de concentração τ, com as duas estratégias ρ_i e ρ, mostrando a importância destes parâmetros na identificação (estimação) do número de interações e grupos. Os resultados dos modelos ajustados com ρ_i mostraram ser mais satisfatórios para as duas abordagens propostas, pois selecionam melhor e destacam as interações mais importantes.

Na segunda aplicação, foram identificados os genes afetados por interação que são comuns nos quatro conjuntos de dados avaliados. Interseções dois à dois e três à três dos grupos G_E também foram analizadas. Como era de se esperar alguns genes foram identificados nestas interseções obtidas com o ajuste dos modelos de agrupamento. Isso é uma situação natural já que está sendo avaliado o mesmo tipo de câncer (mama). A modelagem via misturas apresentou poucos genes comuns aos quatro conjuntos de dados, entretanto o número de grupos formados em cada caso foi bem maior do que o resultado

apresentado pela modelagem via PD. A modelagem via PD indicou quantidades maiores de genes comuns aos quatro dados, mas fornecendo um número menor de grupos. Em relação a escolha de um dos dois modelos propostos nesta tese, cabe ao usuário decidir, se for preferível poucas quantidades de interações e muitos grupos, então é sugerido utilizar o modelo com agrupamento das interações via mistura finita. No caso do usuário optar por mais interações e poucos grupos, então a modelagem via PD é preferível. É importante tomar essa decisão junto com um especialista da área de genética, para verificar qual modelagem seria mais plausível de se utilizar nestes estudos envolvendo prolemas de CNA.

Estas aplicações exibidas aqui, em problemas práticos, mostram a importância do tipo de modelagem que foi desenvolvida nesta tese. A motivação sugiu das complexas relações de interações entre conjuntos de genes (GRN). Portanto, os modelos propostos permite explicar partes destas associações estabelecidas pelos agrupamentos das interações, que representam a atuação conjunta de diversos genes, afetados pela CNA, influenciando grupos em diferentes partes do genoma.

6.1 Trabalhos futuros

Como trabalhos futuros relacionados a esta Tese pretende-se desenvolver os seguintes pontos:

- Revisitar o estudo sobre o parâmetro de comprimento-escala investigando resultados sob uma outra parametrização: $K(\lambda, \phi) = \upsilon^2 \exp\left\{-\frac{\phi^*}{2}\left\|\lambda_{\bullet j_1} - \lambda_{\bullet j_2}\right\|^2\right\}$, em que $\phi^* = \frac{1}{\phi^2}$. Veja que ϕ^* tem outra escala relativo a $\phi \in (0.1, 0.5)$. Neste caso, poderia ser observado se haveria alguma melhora dos resultados ou se o algoritmo Metropolis-Hastings teria um comportamento mais fácil de ser adaptado para boas qualidades de convergência. Além disso, poderia ser proposto outros tipos de distribuições geradoras de candidatos, pois com a nova parametrização, ϕ^* estaria em outra escala.

- Fazer o ajuste dos modelos propostos utilizando outras funções de covariâncias, tais

como a exponencial potência ou da classe Matérn. Estas funções apresentam um parâmetro extra κ que também é responsável pela suavização da superficie gerada pelo processo Gaussiano. Note que, estas funções englobam a função Gaussiana como um caso particular quando o parâmetro de suavização extra é $\kappa = 2$ na função exponencial potência ou $\kappa \to \infty$ na função de covariâncias Matérn.

- Desenvolver um pacote no `R` para o ajuste da modelagem de agrupamentos porposta nesta tese. Isso contribui para aumentar o número de leitores e de citações para deste trabalho.

- Fazer a investigação e aplicação da modelagem proposta em outras áreas do conhecimento além da genética.

Apêndice A: Verossimilhança e condicionais completas do modelo estimando ϕ no Capítulo 2.

É assumido que as observações X_{ij} são condicionalmente independentes dado os parâmetros. 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: $(X_{\bullet j} \mid \alpha, \lambda, F, \sigma^2) \sim N_m(\alpha\lambda_{\bullet j}+F_{\bullet j}, D)$, 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: $(X_{i\bullet}^\top \mid \alpha, \lambda, F, \sigma^2) \sim N_n(\lambda^\top \alpha_{i\bullet}^\top + F_{i\bullet}^\top, \sigma^2 I_n)$ e

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

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.

A partir da regra de Bayes são obtidos os seguintes resultados:

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

$$B = \tfrac{1}{2}\left[X_{i\bullet}X_{i\bullet}^\top - 2\alpha_{i\bullet}\lambda(X_{i\bullet}^\top - F_{i\bullet}^\top) - 2F_{i\bullet}X_{i\bullet}^\top + F_{i\bullet}F_{i\bullet}^\top + \alpha_{i\bullet}\lambda\lambda^\top\alpha_{i\bullet}^\top\right] + b.$$

- 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}, calcula-se 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 $F_{i\bullet}^\top$ será $\delta_{\mathbf{0}}(F_{i\bullet})$.

- Se $z_i = 1$, a condicional completa de $F_{i\bullet}^\top$ será a $N_n(M_{F_{i\bullet}}, V_{F_{i\bullet}})$ com $V_{F_{i\bullet}} = \left[\dfrac{1}{\sigma_i^2} I_n + K^{-1}(\lambda, \phi)\right]^{-1}$ e $M_{F_{i\bullet}} = V_{F_{i\bullet}} \left[\dfrac{1}{\sigma_i^2}\left(X_{i\bullet}^\top - \lambda^\top \alpha_{i\bullet}^\top\right)\right]$.

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

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

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

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

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

sendo $V_\lambda = \left[\alpha^\top D^{-1}\alpha + I_L\right]^{-1}$ e $M_\lambda = V_\lambda\left[\alpha^\top D^{-1}(X_{\bullet j} - F_{\bullet j})\right]$. Este núcleo não permite reconhecer uma distribuição de probabilidade, então será necessário um método para amostragem indireta desta condicional completa. Considera-se aqui o algoritmo Metropolis-Hastings com passeio aleatório para gerar candidatos.

- Para ϕ tem-se a condicional completa:

$$\begin{aligned} p(\phi \mid \alpha, \lambda, F, \sigma_i^2, z, X) &\propto p(X \mid \alpha, \lambda, F, \sigma^2)\ p(F \mid \lambda, z)\ p(\phi) \\ &\propto \left\{\prod_{i=1}^m \left[N_n(F_{i\bullet}^\top \mid M_{F_{i\bullet}^\top}(\phi), V_{F_{i\bullet}^\top}(\phi))\right]^{z_i}\right\} p(\phi), \end{aligned}$$

sendo $p(\phi)$ a densidade da distribuição $U(0.1, 0.5)$. Novamente para gerar desta distribuição condicional completa utiliza-se o algoritmo Metropolis-Hastings com passeio aleatório.

Apêndice B: Distribuições condicionais completas das modelagens de agrupamento.

As distribuições *a posteriori* condicionais completa de α e σ^2, usadas neste estudo, são as mesmas vistas no Apêndice A.

- $(F_r^* \mid \alpha, \lambda, \sigma^2, F_{-\{i\bullet\}}, z, X) \sim N_n(M_{F_r^*}, V_{F_r^*})$ com $V_{F_r^*} = \left[\left(\sum_{i=1}^{m} \frac{z_{ir}}{\sigma_i^2}\right) I_n + K^{-1}(\lambda, \phi)\right]^{-1}$ e $M_{F_r^*} = V_{F_r^*}\left[\sum_{i=1}^{m} \frac{z_{ir}}{\sigma_i^2}\left(X_{i\bullet}^\top - \lambda^\top \alpha_{i\bullet}^\top\right)\right]$.

  Para avaliar a significância de $F_{i\bullet}^\top = F_r^*$, considere os casos:

- Se $z_{i0} = 1$ então $F_{i\bullet} = \mathbf{0}$.

- Se $z_{ir} = 1$ então $F_{i\bullet}^\top = F_r^*$, com $r = 1, 2, \cdots, R$.

- Atualize o vetor z_i por meio de $(z_i \mid \rho_i^*) \sim \text{Mult}(1, \rho_i^*)$

- Usando a primeira estratégia de especificação *a priori* para o vetor z_i: Atualiza-se o vetor ρ_i com: $(\rho_i \mid z_i) \sim \text{Dir}(z_{i0} + \nu_0, z_{i1} + \nu_1, \cdots, z_{iR} + \nu_R)$. Atualize os pesos *a posteriori* $\rho_i^* = (\rho_{i0}^*, \rho_{i1}^*, \cdots, \rho_{iR}^*)$ do modelo de mistura via:

  $\rho_{ir}^* = p(z_{ir} = 1 \mid \alpha, \lambda, F_{-\{i\bullet\}}, \sigma^2, \rho_{ir}, X) = \frac{\rho_{ir} \exp\{Q_r\}}{\sum_{s=0}^{R} \rho_{is} \exp\{Q_s\}}$, com

  $Q_r = -\frac{1}{2\sigma_i^2}\left[F_r^* F_r^{*\top} - 2F_r^*\left(X_{i\bullet}^\top - \lambda^\top \alpha_{i\bullet}^\top\right)\right]$, $r = 0, 1, \cdots, R$. Considere $\exp\{Q_0\} = 1$.

- Usando a segunda estratégia de especificação *a priori* para o vetor z_i: Atualiza-se o vetor ρ com: $(\rho \mid z) \sim \text{Dir}(m_0 + \nu_0, m_1 + \nu_1, \cdots, m_R + \nu_R)$, sendo $m_r = \sum_{i=1}^{m} z_{ir}$

com $r = 0, 1, \cdots, R$. Atualize os pesos *a posteriori* do modelo de mistura via:

$\rho^*_{ir} = p(z_{ir} = 1 \mid \alpha, \lambda, F_{-\{i\bullet\}}, \sigma^2, \rho_r, X) = \dfrac{\rho_r \exp\{Q_r\}}{\sum_{s=0}^{R} \rho_s \exp\{Q_s\}}$, com

$Q_r = -\dfrac{1}{2\sigma_i^2}\left[F_r^* F_r^{*\top} - 2F_r^*\left(X_{i\bullet}^\top - \lambda^\top \alpha_{i\bullet}^\top\right)\right]$, $r = 0, 1, \cdots, R$. Considere $\exp\{Q_0\} = 1$.

- Para a construção dos pesos via processo *stick-breaking* pode-se usar a primeira estratégia de especificação *a priori*:

$$\begin{aligned} p(V_{ir} \mid F, z, \rho, \lambda, \phi) &\propto p(F, z \mid F^*, \rho, \lambda, \phi)\ p(V_{ir}) \\ &\propto \left\{\prod_{i=1}^{m}\prod_{r=0}^{R}\left[\rho_{ir}\delta_{F_r^*}(F_{i\bullet})\right]^{z_{ir}}\right\} p(V_{ir}) \\ &\propto \left\{\prod_{r=0}^{R}\left[\rho_{ir}\delta_{F_r^*}(F_{i\bullet})\right]^{z_{ir}}\right\} p(V_{ir}), \end{aligned}$$

após alguns cálculos chega-se em:

$$(V_{ir} \mid F, z, \rho, \lambda, \phi) \sim \text{Beta}(z_{ir} + 1, \sum_{s=r+1}^{R} z_{is} + \tau).$$

Usando a segunda estratégia de especificação *a priori* tem-se:

$$(V_r \mid F, z, \rho, \lambda, \phi) \sim \text{Beta}(m_r + 1, \sum_{s=r+1}^{R} m_s + \tau).$$

- Para ϕ tem-se a condicional completa:

$$\begin{aligned} p(\phi \mid \alpha, \lambda, F, \sigma_i^2, z, X) &\propto p(X \mid \alpha, \lambda, F, \sigma^2)\ p(F \mid \lambda, z)\ p(\phi) \\ &\propto \left\{\prod_{r=0}^{R}\prod_{i=1}^{m}\left[N_n(F_{i\bullet}^\top \mid M_{F_{i\bullet}^\top}(\phi), V_{F_{i\bullet}^\top}(\phi))\right]^{z_{ir}}\right\} p(\phi), \end{aligned}$$

com $p(\phi)$ sendo a densidade da distribuição $U(0.1, 0.5)$. Novamente utiliza-se o algoritmo Metropolis-Hastings com passeio aleatório para gerar desta distribuição condicional completa.

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

$$\begin{aligned} p(\lambda_{\bullet j} \mid \alpha, \lambda_{-\{\bullet j\}}, F, \sigma_i^2, X) &\propto p(X \mid \alpha, \lambda, F, \sigma^2)\ p(F_1^*, F_2^*, \cdots, F_R^* \mid \lambda, z_i)\ p(\lambda_{\bullet j}) \\ &\propto N_L(\lambda_{\bullet j} \mid M_\lambda, V_\lambda)\, |K(\lambda)|^{-\sum_{r=1}^{R}\frac{z_{ir}}{2}} \\ &\times \exp\left\{-\frac{1}{2}\sum_{r=1}^{R} z_{ir} F_r^* K^{-1}(\lambda, \phi) F_r^{*\top}\right\}, \end{aligned}$$

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

Apêndice C: Critérios de comparação.

O *Deviance Information Criterion* (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 é calculada 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:

$$D_{\hat{\theta}}(x) = D(x, \hat{\theta}(x)), \tag{C.1}$$

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:

$$\hat{D}_{avg}(x) = \frac{1}{S} \sum_{s=1}^{S} D(x, \theta^{(s)}). \tag{C.2}$$

Para Gelman et al. (2003) a média em (C.2) é um melhor resumo do erro do modelo que a discrepância da estimativa pontual em (C.1). 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 é obtido por:

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

com $\hat{D}_{avg}$ e $D_{\hat{\theta}}$ definidos em (C.2) e (C.1), respectivamente. Valores baixos do DIC indicam melhor ajuste. Para mais detalhes veja Gelman et al. (2003).

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

$$\widehat{\text{lpd}} = \log\left[\frac{1}{S}\sum_{s=1}^{S} p(x_i \mid \theta^{(s)})\right], \tag{C.3}$$

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

$$\hat{p}_{WAIC} = \sum_{i=1}^{m} V_{s=1}^{S}\left(\log p(x_i \mid \theta^{(\mathbf{s})})\right), \tag{C.4}$$

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 (C.3) e (C.4) pode-se calcular o WAIC (valores altos indicam melhor ajuste do modelo) como segue:

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

Outra abordagem para avaliação e seleção de modelos é usar a distribuição preditiva para obter a medida chamada *Conditional Predictive Ordinate* (CPO). Os CPO's são densidades de validação cruzadas que sugerem quais valores de observaçõs x_i são prováveis quando o modelo é ajustado para todas as observações, exceto a i-ésima. O CPO fornece uma medida para cada observação e quando somados os logaritmos de cada uma, isso proporciona a medida *Log Pseudo Marginal Likelihood* (LPML). Pode ser calculado um CPO para cada x_i usando apenas um MCMC ao explorar as amostras da distribuição *a posteriori* e a verossimilhança. O estimador Monte Carlo para calcular o CPO é:

$$\widehat{CPO}_i = \left[\frac{1}{S}\sum_{s=1}^{S}\frac{1}{p(x_i \mid \theta^{(s)})}\right]^{-1}, \tag{C.5}$$

sendo S o número de iterações do algoritmo MCMC utilizado para calcular a média harmónica em (C.5). Valores altos de $\widehat{CPO}_i$ indicam melhores ajustes. Finalmente, o

LPML pode ser calculado da seguinte maneira:

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

Apêndice D: Tabelas e gráficos extras das aplicações.

Tabela D.1: Estimativas *a posteriori* para o parâmetro de comprimento-escala ϕ da função de covariâncies Gaussiana. Resultados referentes ao ajuste do modelo de mistura finito e com PD para o par de regiões (1,4).

Especificações de ρ_i	Dados	Estimativas *a posteriori*			
		Mediana	Média	Desvio Padrão	Intervalo HPD
$\text{Dir}(10^3, 10^{-10}, \cdots, 10^{-10})$	Chin	0.4478	0.4448	0.0325	[0.3819 ; 0.4997]
	Miller	0.2842	0.2896	0.1199	[0.1246 ; 0.4923]
	Sotiriou	0.2392	0.2456	0.0816	[0.1089 ; 0.3917]
	Wang	0.2498	0.2412	0.0962	[0.1004 ; 0.3789]
$PD(0.15, N_n[\mathbf{0}, K(\lambda, \phi)])$	Chin	0.1709	0.1770	0.0490	[0.1006 ; 0.2804]
	Miller	0.2242	0.2211	0.0687	[0.1001 ; 0.3377]
	Sotiriou	0.2259	0.2275	0.0730	[0.1106 ; 0.3830]
	Wang	0.2132	0.2116	0.0502	[0.1250 ; 0.3033]
$PD(0.10, N_n[\mathbf{0}, K(\lambda, \phi)])$	Chin	0.3688	0.3574	0.0770	[0.2254 ; 0.4870]
	Miller	0.3362	0.3232	0.0842	[0.1325 ; 0.4587]
	Sotiriou	0.4708	0.4657	0.0242	[0.4218 ; 0.4998]
	Wang	0.3200	0.3321	0.0709	[0.2230 ; 0.4648]

Tabela D.2: Estimativas *a posteriori* para o parâmetro de comprimento-escala ϕ da função de covariâncies Gaussiana. Resultados referentes ao ajuste do modelo de mistura finito e com PD para o par de regiões (3,4).

Especificações de ρ_i	Dados	Estimativas *a posteriori*			
		Mediana	Média	Desvio Padrão	Intervalo HPD
$\text{Dir}(10^3, 10^{-10}, \cdots, 10^{-10})$	Chin	0.2101	0.2102	0.0555	[0.1224 ; 0.3265]
	Miller	0.3249	0.3486	0.0923	[0.2028 ; 0.4940]
	Sotiriou	0.2311	0.2221	0.0530	[0.1089 ; 0.3011]
	Wang	0.4237	0.4143	0.0616	[0.3121 ; 0.4998]
$PD(0.15, N_n[\mathbf{0}, K(\lambda, \phi)])$	Chin	0.2743	0.2741	0.0666	[0.1499 ; 0.3827]
	Miller	0.4479	0.4402	0.0435	[0.3618 ; 0.4999]
	Sotiriou	0.1648	0.1902	0.0812	[0.1001 ; 0.3415]
	Wang	0.4245	0.4128	0.0531	[0.2901 ; 0.4947]
$PD(0.10, N_n[\mathbf{0}, K(\lambda, \phi)])$	Chin	0.1657	0.1692	0.0395	[0.1001 ; 0.2412]
	Miller	0.2142	0.2119	0.0497	[0.1227 ; 0.3034]
	Sotiriou	0.2252	0.2302	0.0890	[0.1013 ; 0.3856]
	Wang	0.3608	0.3616	0.0649	[0.1810 ; 0.4463]

A Figura D.1 pode-se observar o efeito de interação que afeta o gene comum nas quatro bases de dados de câncer de mama para as regiões (3,4).

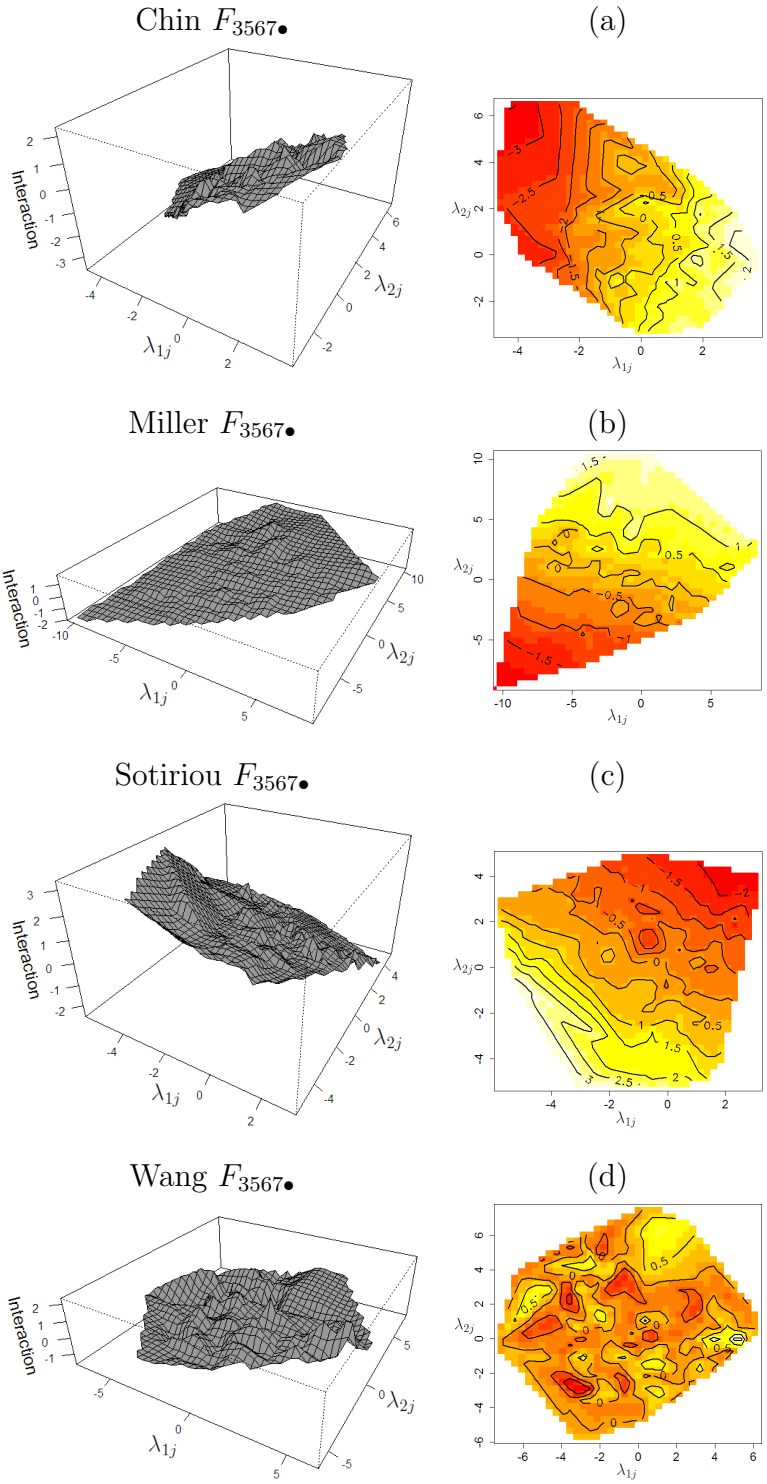

Figura D.1: *Gráficos de superfície e contorno do efeito de interação detectado para o mesmo gene nos quatro conjuntos de dados. A modelagem é feita com* $\rho_i \sim Dir(10^3, 10^{-10}, \cdots, 10^{-10})$.

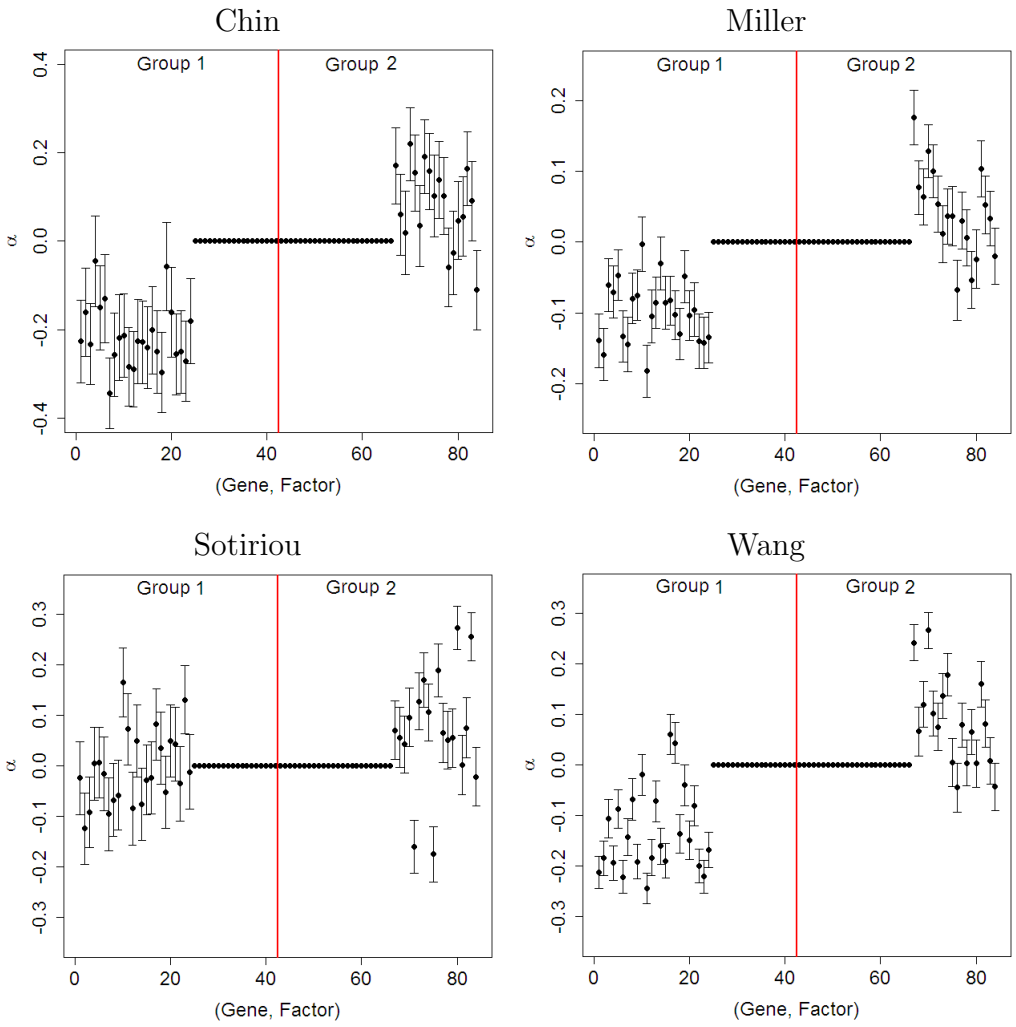

Figura D.2: *Gráficos com as médias a posteriori (círculo) e intervalos HPD de 95% de credibilidade sendo representado pelo segmento de reta na vertical. Resultados relacionados ao par das regiões (3,4). A modelagem é feita com* $\rho_i \sim Dir(10^3, 10^{-10}, \cdots, 10^{-10})$.

Referências Bibliográficas

Affymetrix (2001), *Statistical algorithms reference guide*, Affymetrix Technical Report, http://www.affymetrix.com/estore/.

Amorim, E. C. (2016), "Influência de funções de covariâncias sobre o modelo fatorial latente esparso com interações," Master's thesis, Departamento de Estatística, Universidade Federal de Minas Gerais.

Antoniak, C. E. (1974), "Mixture of Dirichlet processes with applications to Bayesian nonparametric problems," *The Annals of Statistics*, 2, 1152–1174.

Blackwell, D. e MacQueen, J. B. (1973), "Ferguson distributions via Pólya urn schemes," *The Annals of Statistics*, 1, 353–355.

Blei, D. M., Griffiths, T. L., e Jordan, M. I. (2010), "The nested chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," *Journal of the ACM*, 57, 1–30.

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.

Celeux, G., Hurn, M., e Robert, C. P. (1999), "Computational and inferential difficulties with mixture posterior distriburions," *Journal of the American Statistical Association*, 95, 957–970.

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.

Core Team (2019), *R: A Language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Eddelbuettel, D. (2013), *Seamless R and C++ interations with Rcpp*, vol. 64, Springer, New York.

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), "RcppArmadillo: Accelerating R with high-performance C++ linear algebra," *Computational Statistics and Data Analysis*, 71, 1054–1063.

Emmert-Streib, F., Dehmer, M., e Haibe-Kains, B. (2014), "Gene regulatory networks and their applications: understanding biological and medical problem in terms of networks," *Frontiers in Cell and Developmental Biology*.

Escobar, M. D. e West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, 1, 209–230.

Ferguson, T. S. (1974), "Prior distributions on spaces of probability measures," *The Annals of Statistics*, 2, 615–629.

Fruhwirth-Schnatter, S. (2001), "Markov Chain Monte Carlo estimation of classical and dynamic switching and mixture models," *Journal of the American Statistical Association*, 96, 194–209.

Fruhwirth-Schnatter, S. e Lopes, H. (2009), "Parsimonious Bayesian factor analysis when the number of factors is unknown," Tech. rep.

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.

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, third edn.

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.

Ghosh, J. e Ramamoorthi, R. (2003), *Bayesian nonparametrics*, Springer, Springer Series in Statistics, 1 edn.

Gonçalves, F. B., Gamerman, D., e Soares, T. M. (2013), "Simultaneous multifactor DIF analysis and detection in Item Response Theory," *Computational Statistics & Data Analysis*, 59, 144–160.

Green, P. J. (1995), "Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

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.

Jasra, A., Holmes, C. C., e Stephens, A. (2005), "Markov Chain Monte Carlo methods and the label switchig problem in Bayesian mixture modeling," *Statistical Science*, 20, 50–67.

Jeliazkov, I. e Yang, X.-S. (2014), *Bayesian inference in the social sciences*, John Wiley & Sons.

Lopes, H. F. e West, M. (2004), "Bayesian model assessment in factor analysis," *Statistica Sinica*, 14, 41–67.

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.

McEachern, S. N. e Muller, P. (1998), "Estimating mixture of Dirichlet process models," *Journal of Computational and Graphical Statistics*, 7, 223–238.

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. (2000), "Markov Chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

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.

Rasmussen, C. E. e Williams, C. K. I. (2005), *Gaussian processes for machine learning (Adaptive computation and machine learning)*, The MIT Press.

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

Sethuraman, J. (1994), "A constructive definition of the Dirichlet process prior," *Statistica Sinica*, 2, 639–650.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Kains, B. H., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Vijver, M. J. V. D., Bergh, J., Piccart, M., e Delorenzi, M. (2006), "Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis," *Journal of the National Cancer Institute*, 98, 262–272.

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, Serie B*, 64, 583–639.

Stein, M. L. (1999), *Interpolation of spatial data*, Springer Series in Statistics, Springer-Verlag, New York, Some theory for Kriging.

Stephens, M. (2000), "Dealing with label switching in mixture models," *Journal of the Royal Statistical Society, Series B*, 62, 795–809.

Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwert, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Gelder, M. E. M. V., Jatkoe, T., Berns, E. M. J. J., Atkins, D., e Foekens, J. A. (2005), "Gene expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet*, 365, 671–679.

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.