Modelo de Mistura com Número de Componentes Desconhecido: Estimação via Método *Split-Merge*

Erlandson Ferreira Saraiva

Orientador: Prof. Dr. Luís Aparecido Milan

Tese apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Doutor em Estatística.

São Carlos

Novembro de 2009

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária/UFSCar

S243mm

Saraiva, Erlandson Ferreira.
Modelo de mistura com número de componentes desconhecido: estimação via método *split-merge* / Erlandson Ferreira Saraiva. -- São Carlos : UFSCar, 2009.
96 f.

Tese (Doutorado) -- Universidade Federal de São Carlos, 2009.

1. Estatística - análise. 2. Mistura de distribuições. 3. Inferência bayesiana. 4. MCMC. I. Título.

CDD: 519.5 (20ª)

Dedico este trabalho a meu pai Elton Oliveira Saraiva, pelo seu empenho e força de trabalho, e minha mãe Dionisia Ferreira Saraiva, pela sua ternura e mãe dedicada e compreensiva que sempre foi;

Agradeço,

aos meus pais pelo apoio e incentivo em todos os momentos, por serem pessoas honradas que com dedicação e trabalho honesto conseguiram proporcionar a mim e minha irmã Geruza Aparecida Ferreira Saraiva a oportunidade de estudar e termos sucesso em nossas carreiras acadêmicas;

à minha irmã Geruza e meu cunhado Alexandre Barbosa da Silva pelo apoio e ajuda em todos os momentos;

à minha noiva e meu grande amor Sandra Magna Lucas Gomes pelo apoio, compreensão e amor dedicado a mim nestes oito anos juntos;

ao professor Dr. Luís Aparecido Milan pela orientação, pelas idéias e principalmente pelo exemplo de dedicação e disciplina de trabalho. Agradeço ainda, as suas palavras sábias de apoio, paciência, cuidado e incentivo à pesquisa para que o trabalho que realizamos fosse concluído com exito.

ao professor Dr. José Galvão Leite pelas sugestões e esclarecimentos sobre os aspectos teóricos relacionados aos métodos estatísticos apresentados na tese;

aos professores do DEs-UFSCar, pois todos de forma direta ou indireta contribuíram para a realização desta tese;

aos funcionários do DEs-UFSCar, em especial a Maria Isabel Rinaldo Pessôa de Araujo, pelo carinho e atenção;

à amiga Juliana Cobre pelos bons momentos que passamos juntos, estudando e discutindo sobre nossas pesquisas, transformando o nosso encontro no curso de doutorado em estatística em verdadeira amizade;

aos amigos Marcelo de Paula, Adriano Kamimura Suzuki e Luís Ernesto Bueno Salasar, pelos bons momentos que passamos juntos, discutindo sobre estatística, apenas conversando sobre diversos assuntos ou jogando video game, que possibilitou o surgimento de verdadeira amizade;

à Coordenação de Aperfeiçoamente de Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

Resumo

Propomos uma abordagem bayesiana hierárquica e os algoritmos *split-merge* MCMC e *birth-split-merge* MCMC para a estimação conjunta dos parâmetros e do número de componentes de um modelo com mistura de distribuições. A proposta *split* é baseada nos dados e na distribuição *a posteriori* dos parâmetros. Nesta proposta, utilizamos probabilidades de alocação que são calculadas de acordo com os parâmetros associados a cada componente, que são gerados da distribuição *a posteriori* dado as observações previamente alocadas. As propostas *split* e *merge* são desenvolvidas para serem reversíveis e são aceitas de acordo com a probabilidade de aceitação de *Metropolis-Hastings*, para garantir a existência da distribuição estacionária. O algoritmo *birth-split-merge* apresenta as mesmas propostas *split-merge* porém este algoritmo permite que ao atualizar uma variável latente, esta seja capaz de determinar o "nascimento" (*birth*) de uma nova componente. Verificamos a performance dos algoritmos propostos utilizando dados artificiais, gerados via simulação, e dois conjuntos de dados reais. O primeiro é o bem conhecido conjunto de dados sobre a velocidade de galáxias e o segundo é um conjunto de dados de expressão gênica. A contribuição teórica presente nesta tese é o desenvolvimento de um pocesso estocástico com base nos movimentos *split-merge*, que são baseados nos dados. Ou seja, se a amostra é proveniente de uma população composta por k subpopulações, nosso método busca informações sobre as k subpopulações diretamente nos dados observados. Com isso, quando propomos o surgimento de uma nova componente esta sempre tem dados associados, i.e., determina uma partição nos dados observados, e os parâmetros são gerados da distribuição *a posteriori*, o que não ocorre nos métodos alternativos.

Abstract

We propose the split-merge MCMC and birth-split-merge MCMC algorithms to analyse mixture models with an unknown number of components. The strategy for splitting is based on data and posterior distribution. Allocation probabilities are calculated based on component parameters which are generated from the posterior distribution given the previously allocated observations. The split-merge proposals are developed to be reversible and are accepted according to Metropolis-Hastings probability. This procedure makes possible a greater change in configuration of latent variables, in a single iteration of algorithms, allow a major exploration of clusters and avoid possible local modes. As an advantage, our approach determines a quick split proposal in contrary to former split procedures which require substantial computational effort. In the birth-split-merge MCMC algorithm, the birth movement is obtained directly from the procedure to update the latent variables and occurs when an observation determine a new cluster. The performance of the method is verified using artificial data sets and two real data sets. The first real data set consist of benchmark data of velocities from distant galaxies diverging from our own while the second is *Escherichia Coli* bacterium gene expression data.

Sumário

Capítulo 1

Introdução

Modelos com mistura de distribuições são amplamente utilizados para modelar dados, em que, as observações são consideradas como sendo provenientes de uma população composta por k subpopulações, onde k pode ser conhecido ou desconhecido. Cada subpopulação é adequadamente modelada por uma densidade pertencente à uma família de distribuições paramétricas. A densidade associada a cada subpopulação é chamada de componente da mistura e é ponderada pela frequência relativa da subpopulação na população.

Além disso, modelos com mistura de distribuições fornecem uma forma conveniente de modelar dados que podem não ser adequadamente modelados por qualquer família paramétrica de distribuições padrão. Também podem ser utilizados como uma alternativa paramétrica, em relação à metodos não paramétricos, de estimação de densidades (Stephens, 2000).

Este tipo de modelagem tem uma ampla aplicação, desde classificação de clientes em estudos de *marketing* até identificação de grupos de genes e proteínas em bioinformática (Do *et al*,, 2002; Medvedovic e Sivaganesan, 2002). O objetivo desta tese é desenvolver uma abordagem bayesiana para a estimação dos parâmetros de um modelo com mistrura de distribuições em situações onde o número de componentes k é desconhecido, utilizando métodos de Monte Carlo em cadeias de Markov (MCMC).

1.1 Revisão Bibliográfica

Diebolt e Robert (1993), Diebolt e Robert (1994) and Roeder and Wasserman (1995) utilizam algoritmos baseados em métodos MCMC para estimar os parâmetros de um modelo com mistura, em situações onde o número de componentes k é conhecido.

Para situações com k desconhecido,

- Escobar e West (1995) propõem uma abordagem bayesiana semi-paramétrica utilizando o processo de Dirichlet *a priori*. Este procedimento é equivalente a considerarmos um modelo com mistura com o número de componentes $k \to \infty$.

- Richardson e Green (1997) propõem uma abordagem bayesiana utilizando o algoritmo *reversible jump* com os movimentos de "nascimento" e "morte" e *split-merge*. No movimento "nascimento" uma nova componente vazia é criada com parâmetros gerados da distribuição *a priori*. No movimento "morte" uma componente vazia é retirada do modelo. Os movimentos *split-merge* são propostos de acordo com a preservação de momentos associados à(s) componente(s) escolhida(s) para se propor um *split* ou *merge*. Para cada uma das propostas os pesos são adequadamente reescalados para somar 1. Estes movimentos não são simples de se construir para o caso multivariado.

- Stephens (2000) propõe um processo de "nascimento" e "morte" a tempo contínuo, em que, cada componente da mistura é considerada como sendo um ponto no espaço paramétrico. Para estimar k, constroi uma cadeia de Markov ergódica com apropriada distribuição estacionária. Este algoritmo requer a especificação de uma taxa de "nascimento", que é especificada de forma subjetiva. Além disso, quando ocorre o "nascimento" de uma nova componente os novos parâmetros são gerados da distribuição *a priori*. Com isso, para situações com distribuições *a priori* não informativas, a criação de uma nova componente dificilmente provoca uma partição nos dados. Embora o procedimento permita k variar, na maioria das iterações, somente uma pequena quantidade de componentes possuem dados associados.

- Jain e Neal (2004) propõem um algoritmo *split-merge* MCMC para modelos de misturas de processos de Dirichlet. Para propor o movimento *split* eles utilizam um algoritmo *Gibbs sampling* restrito. Os autores informam que a probabilidade de aceitação da proposta *split* pode ser efetada pelo número de iterações utilizadas no algoritmo *Gibbs sampling*. Este procedimento também requer um alto custo computacional. Além disso, o método integra fora os pesos e os parâmetros associados às componentes. Também não há interesse em estimar o número de componentes k, que é assumido como $k \to \infty$. Ou seja, neste método o interesse é somente identificar grupos de observações.

1.2 Propostas e Organização

Nesta tese, propomos uma nova estratégia *split-merge* para implementar uma metodologia MCMC para estimar os parâmetros e o número de componentes k, conjuntamente. A idéia presente em nossa proposta é a seguinte: se temos uma amostra proveniente de uma população composta por k subpopulações (k desconhecido), então nossa estratégia é buscar informações sobre k e os parâmetros diretamente nos dados observados. Isto é feito através do desenvolvimento de um processo estocástico com movimentos *split-merge*, que são aplicados diretamente aos dados observados. Assim, quando propomos o surgimento de uma nova componente esta sempre tem dados associados, i.e., determina uma partição nos dados, e os parâmetros são gerados da distribuição *a posteriori*. Ao contrário dos métodos alternativos (descritos acima), que propõem novas componentes e novos parâmetros sem se preocupar se esta nova componente determina uma partição nos dados observados.

Para propor o movimento *split*, utilizamos uma forma de alocação sequêncial utilizando probabilidades de alocação que são calculadas de acordo com os parâmetros das componentes, que são gerados da distribuição *a posteriori* dadas as observações previamente alocadas. Inversamente ao *split*, no movimento *merge* as observações pertencentes a duas componentes são unidas para dar origem a uma nova componente com parâmetros gerados da distribuição *a posteriori*. Estas propostas são

desenvolvidas para serem reversíveis e são aceitas de acordo com a probabilidade de aceitação de *Metropolis-Hastings* para garantir a existência da distribuição estacionária, e que esta seja a distribuição *a posteriori* conjunta dos parâmetros de interesse.

Algumas vantagens deste procedimento são:

(i) as propostas *split-merge* podem ser rápidamente propostas e testadas;

(ii) podem ser facilmente aplicadas para o caso multivariado;

(iii) quando há o surgimento de uma nova componente esta surge com base nas informações contidas nos dados observados;

(iv) evita modas locais, separando (*splitting*) ou juntando (*merging*) observações pertencentes as componentes.

Baseados nessa estratégia, descrevemos os métodos propostos através de dois algoritmos, denominados de algoritmo *split-merge* MCMC e algoritmo *birth-split-merge* MCMC. O algoritmo *birth-split-merge* MCMC apresenta as mesmas propostas *split-merge* porém este algoritmo permite que ao atualizar uma variável latente, esta seja capaz de determinar o "nascimento" (*birth*) de uma nova componente. Este algoritmo também tem o passo "morte" (*death*) que ocorre sempre que o número de observações pertencentes a uma componente "cai" para *zero*.

Verificamos a performance dos algoritmos propostos, na estimação conjunta de k e dos parâmetros das componentes, utilizando três conjuntos de dados artificiais e dois conjuntos de dados reais. Os dados artificiais são gerados via simulação. O primeiro conjunto de dados reais é o conhecido conjunto de dados sobre a velocidade de galáxias, utilizado por diversos autores, tais como, Roeder and Wasserman (1995), Escobar e West (1995), Richardson e Green (1997) e Stephens (2000). O segundo conjunto de dados reais é sobre expressão gênica, obtido do experimento realizado com a bacteria *Escherichia Coli*, descrito em Arfin *et al.* (2000).

A tese está organizada da seguinte forma. No Capítulo 2, descrevemos o modelo com mistura e a utilização das variáveis latentes. No Capítulo 3, decrevemos a abordagem bayesiana para modelos com mistura de distribuições em situações com k conhecido e desconhecido. No Capítulo 4, propomos o algoritmo *split-merge* MCMC, apresentamos os resultados obtidos na aplicação aos cinco conjuntos de dados e fazemos uma comparação com o algoritmo *reversible jump*. No Capítulo 5, propomos o algoritmo *birth-split-merge* MCMC e apresentamos os resultados obtidos na aplicação aos cinco conjuntos de dados. No Capítulo 6, fazemos uma discussão sobre as principais diferenças em relação aos métodos alternativos, descritos na revisão bibliográfica acima. No Capítulo 7 fazemos as considerações finais sobre os métodos propostos.

Capítulo 2

Modelos com Mistura de Distribuições

Modelos com mistura de distribuições são utilizados para modelar fenômenos ou experimentos cujas observações são provenientes de uma população composta por k subpopulações, onde k pode ser conhecido ou desconhecido. Nos últimos anos, este tipo de modelagem tem sido utilizada em diferentes aplicações.

Neste Capítulo, descrevemos o modelo com mistura de distribuições, sua representação através da introdução de variáveis latentes e o porque do interesse na estimação do número de componentes k.

2.1 Modelo com Mistura

Definição 1. *Qualquer combinação linear convexa*

$$\sum_{j=1}^{k} w_j f(y|\phi_j), \ \textit{com } w_j > 0 \ \textit{e} \ \sum_{j=1}^{k} w_j = 1$$

das densidades $f(y|\phi_j)$ pertencentes a uma família de distribuições indexadas pelo parâmetro ϕ_j (escalar ou vetor) é denominada de uma mistura de distribuições.

Neste texto, consideramos que $\mathbf{y} = (y_1, \ldots, y_n)$ é uma amostra aleatória proveniente de uma densidade com mistura de distribuições com k componentes,

$$f(y_i|\mathbf{w}, \boldsymbol{\phi}) = \sum_{j=1}^{k} w_j f(y_i|\phi_j), \tag{2.1}$$

onde $f(y_i|\phi_j)$ $(i = 1, \ldots, n,\ 1 \leq j \leq k$ e $k > 1)$ é a densidade de uma dada distribuição paramétrica indexada pelo parâmetro ϕ_j (escalar ou vetor), $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)$ é o vetor contendo todos os parâmetros e $\mathbf{w} = (w_1, \ldots, w_k)$, com $w_j > 0$ e $\sum_{j=1}^{k} w_j = 1$, são os pesos associados às componentes.

O modelo (2.1) assume que temos uma população heterogênea com k subpopulações. Cada subpopulação j tem tamanho proporcional a w_j, $j = 1, \ldots, k$.

A representação do modelo com mistura através de uma combinação convexa de distribuições, implica que os momentos do modelo (2.1) também são representados por uma combinação convexa dos momentos associados às densidades $f(\cdot|\phi_j)'s$,

$$E[Y^m] = \sum_{j=1}^{k} w_j E_{f(\cdot|\phi_j)}[Y^m]. \tag{2.2}$$

Este fato foi explorado por Karl Pearson (1974) para deduzir um estimador de momentos para os parâmetros de um modelo com mistura de distribuições normais com duas componentes,

$$f(y|w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = w_1 f(y|\mu_1, \sigma_1^2) + (1 - w_1) f(y|\mu_2, \sigma_2^2),$$

onde $f(y|\mu, \sigma^2)$ representa a densidade da distribuição normal com média μ e variância σ^2.

Dado $\mathbf{y} = (y_1, \ldots, y_n)$, a função de verossimilhança é dada por

$$L(\boldsymbol{\phi}, \mathbf{w}|\mathbf{y}) = \prod_{i=1}^{n} \sum_{j=1}^{k} w_j f(y_i|\phi_j). \tag{2.3}$$

Sob o enfoque bayesiano, considerando $\pi(\boldsymbol{\phi}, \mathbf{w})$ como sendo a distribuição *a priori* conjunta para $\boldsymbol{\phi}$ e $\mathbf{w}$, a distribuição *a posteriori* é dada por

$$\pi(\boldsymbol{\phi}, \mathbf{w}|\mathbf{y}) \propto \left(\prod_{i=1}^{n} \sum_{j=1}^{k} w_j f(y_i|\phi_j) \right) \pi(\boldsymbol{\phi}, \mathbf{w}). \tag{2.4}$$

Note que (2.3) e (2.4) apresentam problemas para se deduzir o estimador de máxima verossimilhança e o estimador de Bayes, pois estas envolvem a expansão da função de verossimilhança em k^n termos. Isto causa um alto custo computacional e dificilmente pode ser expressa analiticamente (Diebolt e Robert, 1994).

2.2 Variáveis latentes

Uma das primeiras ocorrências de modelagem com mistura de distribuições é encontrada no trabalho de Bertillon (1887), onde a estrutura bimodal do peso de militares recrutados na França é explicada por uma mistura de duas populações de homens, uns provenientes das planícies e outros das montanhas (Marin, Mengersen and Robert, 1992). A estrutura de mistura surge devido à perda da origem de cada observação, i.e., a origem do lugar onde cada homen foi recrutado. Assim, cada peso y_i observado é proveniente *a priori* da densidade f_1 (modela os pesos dos homens das planícies) ou f_2 (modela os pesos dos homens das montanhas) com probabilidades w_1 e $w_2 = 1 - w_1$, respectivamente.

Esta estrutura "oculta" pode ser explorada para facilitar o procedimento de estimação dos parâmetros utilizando o fato que para toda variável aleatória Y_i, proveniente de um modelo com mistura de distribuições com k componentes, é possível associar uma variável latente $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ik})$, de dimensão k, que indica a componente da qual a observação y_i é proveniente da seguinte forma,

$$Z_{ij} = \begin{cases} 1, & \text{se } y_i \text{ é proveniente da componente } j \\ 0, & \text{caso contrário,} \end{cases}$$

com $\sum_{j=1}^{k} Z_{ij} = 1$, para $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

Assim, temos que

$$y_i | z_{ij} = 1, \phi_j \sim f(y_i|\phi_j) \;\text{ e }\; Z_i \sim \mathcal{M}_k(1; w_1, \ldots, w_k) \tag{2.5}$$

onde $\mathcal{M}_k(1; w_1, \ldots, w_k)$ representa a distibuição multinomial com k modalidades e uma única observação. Dependendo do interesse, os $Z_i's$ devem ou não ser parte das quantidades de interesse.

Condicional em Z_i, $\mathbf{y} = (y_1, \ldots, y_n)$ são observações independentes provevinentes das densidades

$$P(y_i|z_{ij} = 1, \phi_j) = f(y_i|\phi_j), \tag{2.6}$$

para $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

Integrando sobre as variáveis latentes $Z_1, \ldots, Z_n$ obtemos o modelo (2.1),

$$P(y_i|\mathbf{w}, \boldsymbol{\phi}) = \sum_{j=1}^{k} P(Z_{ij} = 1)P(y_i|z_{ij} = 1, \phi_j) = \sum_{j=1}^{k} w_j f(y_i|\phi_j). \tag{2.7}$$

Com a introdução das variáveis latentes, a função de verossimilhança é dada por

$$L(\boldsymbol{\phi}, \mathbf{w}|\mathbf{y}, \mathbf{z}, k) = \prod_{j=1}^{k} \prod_{i=1}^{n} [w_j f(y_i|\phi_j)]^{z_{ij}} . \tag{2.8}$$

Assim, podemos utilizar o algoritmo EM (*Espectation-Maximization*), proposto por Dempster, Laird and Rubin (1977), como um procedimento de otimização para maximizar a função de verossimilhança em (2.8).

Na l-ésima iteração ($l = 1, \ldots, L$), cada observação "faltante" z_{ij} é substituida pelo seu valor esperado

$$E[Z_{ij}^{(l)}] = P(Z_{ij}^{(l)} = 1|\mathbf{y}, \boldsymbol{\phi}, k) = \frac{w_j^{(l)} f(y_i|\phi_j^{(l)})}{\sum_{j=1}^{k} w_j^{(l)} f(y_i|\phi_j^{(l)})}. \tag{2.9}$$

Condicional na configuração $\mathbf{z}^{(l)}$, são obtidas estimativas de máxima verossimilhança para os parâmetros $\boldsymbol{\phi}$.

Baseados no algorithmo EM, Celeux and Diebolt (1985) propõem uma versão estocástica do EM chamada de SEM, em que, ao invés de estimar os dados "faltantes" eles simulam um valor para z_{ij} da distribuição multinomial com pesos dados em (2.9).

Considerando uma distribuição *a priori* $\pi(\boldsymbol{\phi})$ para os parâmetros $\boldsymbol{\phi}$, então pela abordagem bayesiana, geramos valores para os parâmetros de sua distruibuição *a posteriori*, $\boldsymbol{\phi} \sim \pi(\boldsymbol{\phi}|\mathbf{y}, \mathbf{z})$. Este procedimerno completa o procedimento de simulação SEM. Hoje em dia este método é conhecido como *Gibbs sampling* (Diebolt e Robert, 1994).

2.3 Número de componentes

Modelos com mistura de distribuições fornecem uma conveniente e flexível família de distribuições capazes de ajustar dados que podem não ser adequadamente modelados por qualquer família paramétrica de distribuições padrão. Também podem ser utilizadas como uma alternativa paramétrica à metodos não paramétricos de estimação de densidades (Stephens, 2000).

Quando desenvolvemos uma análise utilizando um modelo com mistura, em muitas vezes, o número de componentes k pode ser desconhecido. Assim, em aplicações onde o número de componentes tem um interpretação, a inferência sobre k pode ser de interesse. Já em situações em que os modelos com mistura são utilizados puramente como uma alternativa paramétrica em relação à abordagem não-paramétrica, para estimação de densidades, um valor de k fixo pode influenciar na estimação da densidade resultante. Logo, procedimentos que permitem k variar podem ser de interesse se k tem ou não uma interpretação.

Neste caso, inferências para k podem ser interpretadas como um problema de selecionar um modelo de um conjunto de modelos competidores. Sob este enfoque, inferências para k se tornam um desafio, pois os modelos competidores são de diferentes dimensões.

Nesta tese, consideramos a abordagem bayesiana para fazermos inferências para k, pois esta fornece uma forma direta, utilizando a probabilidade *a posteriori* para k, de selecionar um dos modelos. Caso seja de interesse, também temos uma forma coerente de combinar os resultados sobre diferentes modelos, através de uma poderação de modelos. Este último caso não será discutido na tese e fica como uma proposta de trabalho futuro.

Sob o enfoque bayesiano, uma compreensiva abordagem para k-fixo, utilizando métodos MCMC é discutida por Diebolt e Robert (1994). Para k desconhecido, Richardson e Green (1997) propõem o uso do algoritmo *reversible jump* e Stephens (2000) propõe um processo de "nascimento" e "morte" a tempo contínuo. Estes dois algoritmos são descritos no Capítulo 3.

Capítulo 3

Abordagem Bayesiana para Modelos com Mistura de Distribuições

Neste Capítulo, exploramos o modelo 2.1 em situações onde o número de componentes k é conhecido e desconhecido. Descrevemos a abordagem bayesiana e métodos de Monte Carlo em cadeias de Markov (MCMC) para se fazer inferências em relação aos parâmetros de interesse.

3.1 Abordagem Bayesiana para k conhecido

Para desenvolver a abordagem bayesiana, consideramos que cada parâmetro do modelo em 2.1 tem uma interpretação em relação à modelagem com mistura de distribuições. Assim, a disponibilidade de opiniões de especialistas, em relação aos parâmetros, podem ser expressas em termos de uma distribuição *a priori* para cada parâmetro separadamente. Uma possível abordagem, é resumir a opinião dos especialistas através de distribuições *a priori* independentes. Esta não é a única abordagem possível, mas tem a vantagem de simplificar as expressões que são utilizadas no desenvolvimento dos métodos computacionais.

Assim, consideramos que a distribuição *a priori* para $\boldsymbol{\phi}$ é dada por

$$\pi(\boldsymbol{\phi}|k) = \prod_{j=1}^{k} \pi(\phi_j). \tag{3.1}$$

Para $\mathbf{w}$, consideramos a distribuição *a priori* de Dirichlet com parâmetro γ,

$$(w_1, \ldots, w_k)|\gamma, k \sim \mathcal{D}irichlet(\gamma, \ldots, \gamma). \tag{3.2}$$

Considerando que S_j é o conjunto das observações pertencentes à componente j, $S_j = \{y_i; z_{ij} = 1\}$, então de (2.8) temos que

$$L(\phi_j|S_j) = \begin{cases} \prod_{S_j} f(y_i|\phi_j), & \text{se } S_j \neq \emptyset \\ 1, & \text{se } S_j = \emptyset, \end{cases}, \tag{3.3}$$

é a função de verossimilhança para a componente j, $j = 1, \ldots, k$.

Assim, a função de verossimilhança em (2.8) pode ser reescrita sob a forma

$$L(\boldsymbol{\phi}, \mathbf{w}|\mathbf{y}, \mathbf{z}, k) = \prod_{j=1}^{k} w_j^{n_j} L(\phi_j|S_j), \tag{3.4}$$

onde n_j é o numero de observações em S_j, $j = 1, \ldots, k$.

Atualizando as distribuições *a priori* em (3.1) e (3.2) via função de verossimilhança em (3.4), a distribuição *a posteriori* para $\boldsymbol{\phi}$ é dada por

$$\pi(\boldsymbol{\phi}|\mathbf{y}, \mathbf{z}, k) = \prod_{j=1}^{k} \pi(\phi_j|S_j), \tag{3.5}$$

onde $\pi(\phi_j|S_j) \propto L(\phi_j|S_j)\pi(\phi_j)$ é a distribuição *a posteriori* para ϕ_j dado S_j, e

$$(w_1, \ldots, w_k)|\gamma, \mathbf{z}, k \sim \mathcal{D}irichlet(n_1 + \gamma, \ldots, n_k + \gamma), \tag{3.6}$$

com valor esperado

$$E[w_j|\gamma, \mathbf{c}, k] = \frac{n_j + \gamma}{n + k\gamma}, \tag{3.7}$$

$j = 1, \ldots, k$.

Note que a distribuição *a posteriori* para $\mathbf{w}$ é independente das observações $\mathbf{y}$,

$$\pi(\mathbf{w}|\mathbf{z}, \mathbf{y}) = \frac{P(\mathbf{w}, \mathbf{z}, \mathbf{y})}{P(\mathbf{z}, \mathbf{y})} = \frac{P(\mathbf{y}|\mathbf{z}, \mathbf{w})\pi(\mathbf{w}|\mathbf{z})\pi(\mathbf{z})}{P(\mathbf{y}|\mathbf{z})\pi(\mathbf{z})} = \frac{P(\mathbf{y}|\mathbf{z})\pi(\mathbf{w}|\mathbf{z})}{P(\mathbf{y}|\mathbf{z})} = \pi(\mathbf{w}|\mathbf{z}).$$

Neste ponto, destacamos dois fatos à respeito da abordagem bayesiana para modelos com mistura de distribuições. Estes fatos são importantes para o desenvolvimento dos métodos computacionais.

Fato 1. *Dada uma amostra de tamanho n, existe uma probabilidade positiva,* $(1 - w_j)^n \neq 0$*, de que a componente j não tenha observações associadas, i.e.,* $z_{ij} = 0$ *para todo* $i = 1, \ldots, n$*. Caso isto aconteça e utilizamos uma distribuição a priori imprópria para os parâmetros* ϕ_j*,* $\int \pi(\phi_j) d\phi_j = \infty$*, então após observar os dados* $\mathbf{y}$ *com a configuração* $\mathbf{z}$*, o conhecimento a posteriori sobre* ϕ_j *permanece o mesmo,* $\pi(\phi_j|\mathbf{y}, \mathbf{z}) = \pi(\phi_j)$*. Ou seja, a distribuição a posteriori também é imprópria,* $\int \pi(\phi_j|\mathbf{y}, \mathbf{z}) d\phi_j = \infty$*.*

Para evitar este problema, consideramos somente distribuições *a priori* próprias com grandes variâncias. Este é o procedimento utilizado pelo software *winBUGS* em problemas onde distribuições *a priori* impróprias não podem ser utilizadas. Diebolt e Robert (1994) discutem a possibilidade de utilizar distribuições *a priori* impróprias.

Fato 2. *O modelo em 2.1 é invariante em relação à marcação j* $(j = 1, \ldots, k)$ *das componentes. Isto implica que os parâmetros das componentes são não identificáveis marginalmente, i.e., não podemos distinguir, por exemplo, a componente* 1 (ϕ_1) *da componente 2* (ϕ_2) *a partir da função de verossimilhança, devido eles serem permutáveis. Este problema é conhecido na literatura como "label switching". Como a distribuição a priori para* $\boldsymbol{\phi}$ *em (3.1) também é invariante em relação às marcações das componentes, então a distribuição a posteriori para* $\boldsymbol{\phi}$ *também é invariante. Esta simetria presente na distribuição a posteriori causa problemas para sumarizar os resultados para os parâmetros de uma componente individual, quando utilizamos métodos MCMC.*

Para garantir identificabilidade para o modelo e ser capaz de desenvolver um procedimento de simulação capaz de sumarizar os resultados para os parâmetros de cada componente, devemos considerar um único tipo de marcação. Por exemplo, podemos marcar as componentes de acordo com a ordem crescente das médias, $\mu_1 < \mu_2 < \ldots < \mu_k$, em que, a primeira componente é a que tem média μ_1, a segunda é a que tem média μ_2 e sucessivamente até a última componente que tem média μ_k. Este procedimento também pode ser feito utilizando os pesos w_j ou as variâncias σ_j^2 ou qualquer outro parâmetro de referência, $j = 1, \ldots, k$. Sob o

enfoque bayesiano esta "truncação", para garantir que o modelo seja identificável, é feita na distribuição *a priori*, i.e., a distribuição *a priori* em (3.1) agora é dada por $\pi(\boldsymbol{\phi})\mathcal{I}_{\mu_1<\mu_2<\ldots<\mu_k}$, se escolhemos marcar as componentes de acordo com as médias. Este tipo de marcação é utilizado nos algoritmos propostos nos Capítulos 4 e 5, como feito também por Richardson e Green (1997).

3.1.1 Família exponencial

Na maioria dos modelos com mistura de distribuições as densidades $f's$ pertencem à família exponencial,

$$f(y|\phi) = a(y)exp\left\{u(y)h(\phi) + b(\phi)\right\} \tag{3.8}$$

onde $a(y)$ é uma função de $\mathbb{R}$ para $\mathbb{R}^+$, $u(y)$ e $h(\phi)$ são funções de $\mathbb{R}$ e Θ para $\mathbb{R}$.

A distribuição *a priori* conjugada para ϕ tem a forma

$$\pi(\phi) \propto exp\left\{\alpha h(\phi) + \beta b(\phi)\right\} \tag{3.9}$$

onde α e β são os hiperparâmetros.

Assim, é possível associar a cada parâmetro ϕ_j de um modelo com mistura, uma distibuição *a priori* conjugada como em (3.9), com hiperparâmetros α_j e β_j, $j = 1, \ldots, k$.

De (3.5) a distribuição *a posteriori* para $\boldsymbol{\phi}$ é dada por

$$\pi(\boldsymbol{\phi}|\mathbf{y}, \mathbf{z}, k) \propto \prod_{j=1}^{k} exp\left\{\left(\alpha_j + \sum_{i=1}^{n} u(y_i)\mathcal{I}_{(z_{ij}=1)}\right)h(\phi_j) + (\beta_j + n_j)b(\phi_j)\right\}, \tag{3.10}$$

onde $n_j = \sum_{i=1}^{n} z_{ij}$.

3.1.2 Algoritmo *Gibbs sampling*

Para a estimação dos parâmetros de um modelo com mistura, em que, as distribuições *a priori* são conjugadas em relação as densidades $f's$, o algoritmo *Gibbs sampling* é o mais utilizado (Diebolt e Robert, 1993, 1994; Verdinelli and Wasserman, 1992; Chib, 1995; Escobar e West, 1995).

Este algoritmo é baseado em diversas simulações de $\mathbf{z}$, $\mathbf{w}$ e $\boldsymbol{\phi}$ condicional, um nos outros, e nas observações $\mathbf{y}$ e pode ser resumido pelos seguinte passos:

Algoritmo *Gibbs sampling*

(1) inicie fixando arbitrariamente $\mathbf{w}^{(0)}$ e $\boldsymbol{\phi}^{(0)}$;

(2) para a l-ésima iteração, $l = 1, \ldots, L$, faça:

(i) gere $\mathbf{Z}_i^{(l)} \sim \mathcal{M}_k\left(1, w_1^{,}, \ldots, w_k^{,}\right)$, onde

$$w_j^{,} = P\left(Z_{ij}^{(l)} = 1 | w_j^{(l-1)}, \phi_j^{(l-1)}, y_i\right) \propto w_j^{(l-1)} f\left(y_i | \phi_j^{(l-1)}\right),$$

para $i = 1, \ldots, n$ e $j = 1, \ldots, k$;

(ii) gere $\mathbf{w}^{(l)} \sim \mathcal{D}irichlet(n_1 + \gamma, \ldots, n_k + \gamma)$. Note que,

$$w_j^{(l)} \sim \mathcal{B}eta(n_j + \gamma, n - n_j + (k-1)\gamma);$$

(iii) gere $\phi_j^{(l)}$ da distribuição *a posteriori*

$$\pi(\phi_j^{(l)} | \mathbf{z}^{(l)}, \mathbf{y}) \propto exp\left\{\left(\alpha_j + \sum_{i=1}^{n} u(y_i)\mathcal{I}_{(z_{ij}=1)}\right) h(\phi_j) + (\beta_j + n_j) b(\phi_j)\right\},$$

para $j = 1, \ldots, k$.

Ao final das L iterações, descartamos as B primeiras iterações como um *burn in*, obtendo uma cadeia de tamanho $L_k = L - B$. Como é usual na abordagem Bayesiana, estimamos os parâmetros utilizando a média dos valores gerados

$$\widetilde{\phi}_j = \frac{1}{L_k} \sum_{l=1}^{L_k} \phi_j^{(l)} \quad \text{e} \quad \widetilde{w}_j = \frac{1}{L_k} \sum_{l=1}^{L_k} w_j^{(l)}, \tag{3.11}$$

para $j = 1, \ldots, k$.

Considerando N_{ij} como sendo o número de vezes que a observação y_i é associada a componente $j \in \{1, \ldots, k\}$ nas L_k iterações, então $P_{ij} = \frac{N_{ij}}{L_k}$ é a probabilidade *a posteriori* de y_i ser proveniente da componente j. Se $P_{ij} = \underset{1 \leq j \leq k}{max}(P_{ij})$, então consideramos que y_i é proveniente da componente j, $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

Exemplo: Considere um modelo com mistura de distribuições normais com k componentes,

$$f(y_i|\boldsymbol{\phi}, \mathbf{w}) = \sum_{j=1}^{k} w_j f(y_i|\phi_j)$$

onde $f(y_i|\phi_j)$ é a densidade da distribuição normal com média μ_j e variância σ_j^2, $\phi_j = (\mu_j, \sigma_j^2)$, $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

Uma distribuição *a priori* conjugada para os parâmetros (μ_j, σ_j^2) é dada por

$$\mu_j|\sigma_j^2, \lambda \sim \mathcal{N}\left(0, \frac{\sigma_j^2}{\lambda}\right) \quad \text{e} \quad \sigma_j^2|\tau, \nu \sim \mathcal{IG}\left(\frac{\tau}{2}, \frac{\nu}{2}\right),$$

onde λ, τ e ν são hiperparâmetros conhecidos.

As distribuições condicionais são dadas por,

$$\mu_j|\sigma_j^2, \mathbf{y}, \mathbf{c}, k \sim \mathcal{N}\left(\frac{\sum_{S_j} y_i}{n_j + \lambda}, \frac{\sigma_j^2}{n_j + \lambda}\right) \tag{3.12}$$

e

$$\sigma_j^2|\mathbf{y}, \mathbf{c}, k \sim \mathcal{IG}\left(\frac{n_j + \tau + 1}{2}, \frac{\nu + \sum_{S_j} y_i^2}{2} - \frac{\left(\sum_{S_j} y_i\right)^2}{2(n_j + \lambda)}\right), \tag{3.13}$$

para $j = 1, \ldots, k$.

Assim, $\mathbf{Z}_i^{(l)} \sim \mathcal{M}_k\left(1, w_1', \ldots, w_k'\right)$, onde

$$w_j' \propto w_j^{(l-1)} exp\left\{-\frac{1}{2\sigma_j^2}(y_i - \mu_j)^2\right\}, \tag{3.14}$$

para $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

3.2 Abordagem Bayesiana para k desconhecido

Nesta seção, consideramos o modelo 2.1 em situações onde o número de componentes k é desconhecido. Descrevemos o método *reversible jump*, proposto por Green (1995), e sua aplicação a modelos com mistura de distribuições, proposto por Richardson e Green (1997). Também descrevemos o processo de nascimento-e-morte proposto por Stephens (2000). Estes métodos são capazes de saltar entre espaços paramétricos de diferentes dimensões, correspondendo aos modelos com diferentes números de componentes na mistura.

3.2.1 Algoritmo *Reversible-jump*

Como nos algoritmos *Metropolis-Hastings*, o algoritmo *reversible jump* é baseado na proposta de um novo valor para a cadeia, que é aceito de acordo com uma probabilidade de aceitação adequada. Porém, este algoritmo permite a existência de movimentos entre modelos, que possuem espaços paramétricos de diferentes dimensões.

Por simplicidade de exposição, considere que temos os modelos $\mathfrak{M}_1, \ldots, \mathfrak{M}_m$, em que, o modelo $\mathfrak{M}_k$ $(1 \leq k \leq m)$ é indexado pelos parâmetros ϕ_k, pertencentes a um espaço paramétrico Θ_k. Considere que a distribuição *a priori* para os parâmetros ϕ_k é $\pi(\phi_k|\mathfrak{M}_k)$ e a probabilidade *a priori* para o modelo $\mathfrak{M}_k$ é $\pi(\mathfrak{M}_k)$. Suponha que $\theta_k = (\phi_k, \mathfrak{M}_k)$ é o estado atual do processo, i.e., o modelo atual é $\mathfrak{M}_k$ com parâmetros $\phi_k \in \Theta_k$.

A idéia do algoritmo *reversible jump* é propor o movimento do modelo $\mathfrak{M}_k$ para o modelo $\mathfrak{M}_{k^*}$ $(1 \leq k^* \leq m,\ k^* \neq k)$ com probabilidade $p_{k^*|k}$. Como os espaços paramétricos Θ_k ou Θ_{k^*} possuem dimensões diferentes, $dim(\Theta_k) \neq dim(\Theta_{k^*})$, é preciso completar um dos espaços, Θ_k ou Θ_{k^*}, com espaços artificiais adequados, para criar uma bijeção entre eles. Isto é feito, muitas vezes, aumentando o espaço paramétrico do modelo de menor dimensão, ou seja, se $dim(\Theta_k) < dim(\Theta_{k^*})$, geramos um vetor aleatório $\mathbf{u} \sim g(\mathbf{u})$ de dimensão $dim(\mathbf{u}) = dim(\Theta_{k^*}) - dim(\Theta_k)$ e consideramos que ϕ_{k^*} é uma transformação determinística de $(\phi_k, \mathbf{u})$, $\phi_{k^*} = T_{k\to k^*}(\phi_k, \mathbf{u})$. Green (1995) mostra que a equação de balanceamento, que garante a reversibilidade para o processo, é preservada se a probabilidade de aceitação para este movimento é dada por $\alpha[\theta_{k^*}|\theta_k] = min(1, A)$, onde

$$A = \frac{L(\phi_{k^*}|\mathbf{y}, \mathfrak{M}_{k^*})}{L(\phi_k|\mathbf{y}, \mathfrak{M}_k)} \frac{\pi(\phi_{k^*}|\mathfrak{M}_{k^*})}{\pi(\phi_k|\mathfrak{M}_k)} \frac{\pi(M_{k^*})}{\pi(M_k)} \frac{p_{k|k^*}}{p_{k^*|k}g(\mathbf{u})} \left| \frac{\partial T_{k\to k^*}(\theta_k, \mathbf{u})}{\partial(\theta_k, \mathbf{u})} \right| \tag{3.15}$$

envolve o cáculo do jacobiano da transformação $T_{k\to k^*}$, as probabilidades de movimentos $p_{k^*|k}$ e $p_{k|k^*}$ e a densidade $g(\mathbf{u})$. A probabilidade de aceitação para o movimento inverso é dada por $\alpha[\theta_k|\theta_{k^*}] = min(1, A^{-1})$.

O algoritmo *reversible jump* pode ser resumido pelos seguintes passos:

Algoritmo *reversible jump*

(1) Proponha a mudança de $(\phi_k, \mathfrak{M}_k)$ para $(\phi_{k^*}, \mathfrak{M}_{k^*})$ com probabilidade $p_{k^*|k}$;

(2) Gere $\mathbf{u} \sim g(\mathbf{u})$ com dimensão $dim(\mathbf{u}) = dim(\Theta_k^*) - dim(\Theta_k)$;

(3) Faça $\phi_{k^*} = T_{k \to k^*}(\phi_k, \mathbf{u})$;

(4) Aceite $(\phi_{k^*}, \mathfrak{M}_{k^*})$ com probabilidade $\alpha[\theta_{k^*}|\theta_k] = min(1, A)$, onde A é dado em (3.15).

Exemplo: Para ilustrar como obter a probabilidade de aceitação do algoritmo *reversible jump*, considere este simples exemplo, em que, temos apenas dois modelos $\mathfrak{M}_1$ e $\mathfrak{M}_2$. Suponha que o modelo $\mathfrak{M}_1$ é indexado pelo parâmetro $\phi \in \mathbb{R}$ e que o modelo e $\mathfrak{M}_2$ é indexado pelos parâmetros $(\phi_1, \phi_2) \in \mathbb{R}^2$.

Como temos uma diferença de 1 entre as dimensões de $\mathfrak{R}$ e $\mathfrak{R}^2$, precisamos gerar um valor $u \sim g(u)$, onde $g(u)$ é a densidade de alguma distribuição paramétrica, tal como, a distribuição uniforme no intervalo (a, b) ou a distribuição beta com parâmetros a e b entre outras. A escolha de $g(u)$ vai depender do problema que está sendo estudado e das transformações determinísticas utilizadas. Para este exemplo, consideramos as seguinte transformações,

$$\begin{aligned} (\phi_1, \phi_2) &= T_{1\to 2}(\phi, u) = (\phi - u, \phi + u) \\ (\phi, u) &= T_{2\to 1}(\phi_1, \phi_2) = \left(\frac{\phi_1 + \phi_2}{2}, \frac{\phi_1 - \phi_2}{2} \right). \end{aligned}$$

O jacobiano da transformação $T_{1\to 2}$ é dado por

$$\left| \frac{\partial T_{1\to 2}(\phi, u)}{\partial(\phi, u)} \right| = \begin{vmatrix} \frac{\partial \phi_1}{\partial \phi} & \frac{\partial \phi_1}{\partial u} \\ \frac{\partial \phi_2}{\partial \phi} & \frac{\partial \phi_1}{\partial u} \end{vmatrix} = \begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix} = 2.$$

Analogamente,

$$\left| \frac{\partial T_{2\to 1}(\phi_1, \phi_2)}{\partial(\phi_1, \phi_2)} \right| = \frac{1}{2}.$$

Fixando as probabilidades *a priori* dos modelos $\mathfrak{M}_1$ e $\mathfrak{M}_2$ em $\frac{1}{2}$ e as probabilidades de movimentos em $p_{2|1} = p_{1|2} = \frac{1}{2}$, a probabilidade de aceitação do movimento de $\mathfrak{M}_1$ para $\mathfrak{M}_2$ é dada por $min(1, A)$, onde

$$A = \frac{L(\phi_1, \phi_2|\mathbf{y}, \mathfrak{M}_2)}{L(\phi|\mathbf{y}, \mathfrak{M}_1)} \frac{\pi(\phi_1, \phi_2|\mathfrak{M}_2)}{\pi(\phi|\mathfrak{M}_1)} \frac{2}{g(u)}.$$

A probabilidade de aceitação para o movimento de $\mathfrak{M}_2$ para $\mathfrak{M}_1$ é dada por $min(1, A^{-1})$.

Definidas as probabilidades de aceitação de cada movimento, aplicamos os passos para a implementeação do algoritmo *reversible jump*, descritos acima. Ao final das L iterações descartamos um *burn in* B e calculamos, por exemplo, a probabilidade *a posteriori* para o modelo $\mathfrak{M}_1$, $P(\mathfrak{M}_1|\cdot) = \frac{N_{\mathfrak{M}_1}}{L-B}$, onde $N_{\mathfrak{M}_1}$ é o número de vezes que o modelo $\mathfrak{M}_1$ ocorre nas $L - B$ iterações.

Algoritmo *reversible jump* para modelos com mistura

Richardson e Green (1997) consideram que $\mathfrak{M}_k$ é um modelo com misturas de distribuições normais univariadas com k componentes, i.e., cada componente é modelada pela densidade da distribuição normal,

$$f(y|\phi_j) = f(y|\mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left\{-\frac{1}{2\sigma_j^2}(y-\mu_j)^2\right\}$$

e utilizam o algoritmo *reversible-jump*, RJ-MCMC, para estimar o número de componentes k.

As distribuições *a priori* utilizadas para os parâmetros μ_j e σ_j^2 são

$$\mu_j \sim \mathcal{N}(\xi, \kappa^{-1}) \;\text{ e }\; \sigma_j^{-2} \sim \mathcal{G}ama(\alpha, \beta),$$

a distribuição *a priori* para os pesos $\mathbf{w}$ é dada pelo distribuição de Dirichlet como em (3.2) com $\gamma = 1$. A distribuição *a priori* para k é dada pela distribuição uniforme $U(1, k_{max})$, onde k_{max} é um valor previamente fixado indicando o máximo valor que k pode assumir.

Restritos a uma vizinhaça de modelos $\mathfrak{M}_{k-1}$ e $\mathfrak{M}_{k+1}$, Richardson e Green (1997) utilizam o algoritmo RJ-MCMC com os movimentos de "nascimento" e "morte" e *split-merge* para estimar k.

A proposta "nascimento" consiste em adicionar uma nova componente vazia na mistura, com novos parâmetros $(w_{j^*}, \mu_{j^*}, \sigma_{j^*}^2)$ gerados das distribuições *a priori*,

$$w_{j^*} \sim \mathcal{B}eta(1, k), \quad \mu_j \sim \mathcal{N}(\xi, \kappa^{-1}), \quad \sigma_j^{-2} \sim \mathcal{G}ama(\alpha, \beta),$$

e os pesos são reescalados para somar 1 fazendo $w'_j = w_j(1 - w_{j^*})$.

A proposta "morte" consiste em remover uma componente vazia da mistura e reescalar os pesos para somar 1 fazendo $w'_j = \frac{w_j}{1-w_{j^*}}$.

As probabilidades de aceitação para as propostas de "nascimento" e "morte" são dadas por $min(1, A)$ e $min(1, A^{-1})$, respectivamente, em que

$$A = \frac{\pi(\mathfrak{M}_{k+1})}{\pi(\mathfrak{M}_k)} \frac{(k+1)w_{j^*}^{\gamma-1}(1-w_{j^*})^{n+k\gamma-k}}{\mathcal{B}(1,k)} \frac{p_{k|k+1}}{p_{k+1|k}(k_0+1)g(w_{j^*})}(1-w_{j^*})^k \quad (3.16)$$

onde $\mathcal{B}(\cdot,\cdot)$ é a função beta, k_0 é o número de componentes vazias antes do "nascimento", $g(w_{j^*})$ é a densidade da distribuição $\mathcal{B}eta(1,k)$ e $(k+1)$ surge da razão $\frac{(k+1)!}{k!}$ que é proveniente da suposição de permutabilidade para as componentes da mistura, pois como impomos uma ordenação "forçada" para os parâmetros $\phi_1, \ldots, \phi_k$, então existem $k!$ e $(k+1)!$ formas de escrever $\boldsymbol{\phi}$ e $\boldsymbol{\phi} \cup \phi_{k+1}$, respectivamente. A razão das funções de verossimilhança é 1 devido as propostas envolverem componentes vazias; a primeira divisão do lado direito de A é a razão das probabilidades *a priori* para os modelos $\mathfrak{M}_{k+1}$ e $\mathfrak{M}_k$; a segunda divisão é a razão das distribuições *a priori* de Dirichlet e a terceira divisão é a razão das probabilidades dos movimentos e o jacobiano da trasformação, em que, $k_0 + 1$ surge da probabilidade de escolher uma componente vazia para o movimento "morte".

Note que, se utilizamos distribuições *a priori* não informativas, isto pode levar a uma alta taxa de rejeição da proposta de "nascimento". Para evitar este problema Richardson e Green (1997) propõem o uso dos movimentos *split* e *merge*. No movimento *split*, uma componente é separada (*splitting*) em duas novas componentes, sob a condição que alguns momentos sejam preservados. O movimento reverso consiste em combinar (*merge*) duas componentes em uma única componente.

Seguindo a notação de Richardson e Green (1997), as propostas *split-merge* são escolhidas com probabilidades $b_k = p_{k+1|k}$ e $d_k = 1 - b_k = p_{k-1|k}$, respectivamente, dependendo de k. Na proposta *merge* duas componentes j_1 e j_2, que são adjacentes em relação aos valores de suas médias, são escolhidas aleatoriamente e é feito o *merge* destas duas compoentes em uma única componente j^*. Os parâmetros destas

nova componente são dados por

$$w_{j^*} = w_{j_1} + w_{j_2} \tag{3.17}$$

$$w_{j^*}\mu_{j^*} = w_{j_1}\mu_{j_1} + w_{j_2}\mu_{j_2} \tag{3.18}$$

$$w_{j^*}\left(\mu_{j^*}^2 + \sigma_{j^*}^2\right) = w_{j_1}\left(\mu_{j_1}^2 + \sigma_{j_1}^2\right) + w_{j_2}\left(\mu_{j_2}^2 + \sigma_{j_2}^2\right). \tag{3.19}$$

Para a proposta *split*, uma componente j^* é escolhida aleatoriamente e é feito o *split* desta componente em duas novas componentes j_1 e j_2. Como existe uma diferença de 3 entre o espaço paramétrico atual e o proposto, é preciso gerar um vetor $\mathbf{u} = (u_1, u_2, u_3)$ de dimensão 3 para propor os novos parâmetros. Richardson e Green (1997) consideram

$$u_1 \sim \mathcal{B}eta(2,2), \quad u_2 \sim \mathcal{B}eta(2,2), \quad u_3 \sim \mathcal{B}eta(1,1)$$

e as seguintes transformações determinísticas

$$w_{j_1} = w_{j^*}u_1 \quad , \quad w_{j_2} = w_{j^*}(1-u_1) \tag{3.20}$$

$$\mu_{j_1} = \mu_{j^*} - u_2\sigma_{j^*}\sqrt{\frac{w_{j_2}}{w_{j_1}}} \quad , \quad \mu_{j_2} = \mu_{j^*} + u_2\sigma_{j^*}\sqrt{\frac{w_{j_1}}{w_{j_2}}} \tag{3.21}$$

$$\sigma_{j_1}^2 = u_3(1-u_2^2)\sigma_{j^*}^2\frac{w_{j^*}}{w_{j_1}} \quad , \quad \sigma_{j_2}^2 = (1-u_3)(1-u_2^2)\sigma_{j^*}^2\frac{w_{j^*}}{w_{j_2}}. \tag{3.22}$$

que produzem os 6 novos parâmetros.

A probabilidade de aceitação da proposta *split* é $min(1, A)$, onde

$$\begin{aligned} A &= \frac{\prod_{i=1}^n [f(y_i|\phi_{j_1})]^{z_{i,j_1}} \prod_{i=1}^n [f(y_i|\phi_{j_2})]^{z_{i,j_2}}}{\prod_{i=1}^n [f(y_i|\phi_{j^*})]^{z_{i,j^*}}} \\ &\times (k+1)\frac{\pi(k+1)}{\pi(k)}\frac{w_{j_1}^{\gamma-1+n_1}w_{j_2}^{\gamma-1+n_2}}{w_{j^*\gamma-1+n_{j^*}}\mathcal{B}(\gamma,k\gamma)} \\ &\times \frac{\kappa}{2\pi}exp\left\{-\frac{1}{2}\kappa\left[(\mu_{j_1}-\xi)^2 + (\mu_{j_2}-\xi)^2 - (\mu_{j^*}-\xi)^2\right]\right\} \\ &\times \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{\sigma_{j_1}^2\sigma_{j_2}^2}{\sigma_{j^*}^2}\right)^{-\alpha-1} exp\left\{-\beta(\sigma_{j_1}^{-2} + \sigma_{j_2}^{-2} - \sigma_{j^*}^{-2})\right\} \\ &\times \frac{d_{k+1}}{b_k P_{aloc} g_{2,2}(u_1) g_{2,2}(u_2) g_{1,1}(u_3)} \\ &\times \frac{w_{j^*}|\mu_{j_1} - \mu_{j_2}|\sigma_{j_1}^2\sigma_{j_2}^2}{u_2(1-u_2^2)u_3(1-u_3)\sigma_{j^*}} \end{aligned} \tag{3.23}$$

onde a primeira linha é a razão das funções de verossimilhanças, k é o número de componentes antes do *split*, n_1 e n_2 é o número de observações alocadas na

componente j_1 e j_2, respectivamente, $\mathcal{B}(\cdot,\cdot)$ é a função beta, P_{aloc} é a probabilidade de obter a alocação j_1 e j_2 a partir de j^* e $g_{a,b}$ é a densidade da distribuição beta com parâmetros a e b. A sexta linha de (3.23) é o jacobiano da transformação (ver Apêndice A.1, pag.74). A probabilidade de aceitação para o movimento *merge* é $min(1, A^{-1})$.

O algoritmo RJ-MCMC para modelos com mistura de distribuições pode ser resumido pelos seguintes passos:

Algoritmo RJ-MCMC para modelos com mistura

(1) Inicialize todos os parâmetros;

(2) Para a l-ésima iteração, $l = 1, \ldots, L$, faça:

(i) Atualize os pesos $\mathbf{w}$;

(ii) Atualize as variáveis latentes $\mathbf{Z}$;

(iii) Escolha entre os movimento *split* ou *merge* com probabilidade b_k e d_k, respectivamente. Aceite o *split* com probabilidade $min(1, A)$ e o *merge* com probabilidade $min(1, A^{-1})$, onde A é dado em (3.23);

(iv) Escolha entre os movimento "nascimento" ou "morte" com probabilidade $p_{k+1|k}$ e $p_{k-1|k}$, respectivamente. Aceite o "nascimento" com probabilidade $min(1, A)$ e o movimento "morte" com probabilidade $min(1, A^{-1})$, onde A é dado em (3.16);

Ao final das L iterações, consideramos um *burn in* B e calculamos o número de vezes $N_{k=j}$ que $k = j$ nas $L - B$ iterações. Assim, $P_{k=j} = \frac{N_{k=j}}{L-B}$ é a probabilidade *a posteriori* para $k = j$ e $\widetilde{k} = \underset{1 \leq j \leq k_{max}}{\arg\max}(P_{k=j})$ é a estimativa para o número de componentes, $j = 1, \ldots, k_{max}$.

3.2.2 Processo de nascimento-e-morte

Stephens (2000) considera cada componente da mistura como sendo um ponto no espaço paramétrico e adapta a teoria de simulação de um processo pontual para

construir uma cadeia de Markov, tal que, a distribuição *a posteriori* dos parâmetros seja a distribuição estacionária.

Este processo é baseado na construção de um processo de "nascimento" e "morte" a tempo contínuo (Preston, 1976; Ripley, 1977; Geyer e Möller, 1994; Grenander e Miller, 1994), em que novas componentes "nascem" a uma taxa $\beta(\boldsymbol{\theta})$ previamente fixada, onde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) = ((w_1, \phi_1), \ldots, (w_k, \phi_k)) = (\mathbf{w}, \boldsymbol{\phi})$ é o estado atual do processo (modelo com mistura com k componentes), e "morrem" a uma taxa $\delta(\boldsymbol{\theta})$ que é determinada segundo a equação de balanceamento para garantir que o processo tenha a distribuição estacionária.

Stephens (2000) utiliza os movimentos "nascimento" e "morte" somente para modificar o número de componentes no modelo, e estes movimentos não consideram os dados completos, i.e., não consideram as marcações das componentes, $\mathbf{z}$. A log-verossimilhança utilizada é dada por

$$log\left\{L(\boldsymbol{\phi}, \mathbf{w}, k|\mathbf{y})\right\} = \sum_{i=1}^{n} log\left\{\sum_{j=1}^{k} w_j f(y_i|\phi_j)\right\}.$$

Dado que estamos no estado $\boldsymbol{\theta}$, o tempo que o processo permanece neste estado tem distribuição exponencial com média $\frac{1}{\beta(\boldsymbol{\theta})+\delta(\boldsymbol{\theta})}$. Ao final deste tempo o processo salta para um novo estado, de acordo com o tipo de evento, "nascimento" ou "morte", que ocorrem segundo as probabilidades,

$$P(nascimento) = \frac{\beta(\boldsymbol{\theta})}{\beta(\boldsymbol{\theta}) + \delta(\boldsymbol{\theta})} \quad \text{e} \quad P(morte) = \frac{\delta(\boldsymbol{\theta})}{\beta(\boldsymbol{\theta}) + \delta(\boldsymbol{\theta})}. \tag{3.24}$$

Se no tempo t ocorre um "nascimento", o novo peso w_{k+1} e os novos parâmetros ϕ_{k+1} são amostrados da distribuição conjunta $h(w_{k+1}, \phi_{k+1})$, e os pesos são reescalados para somar 1. Este novo estado é denotado por $\boldsymbol{\theta} \cup (w_{k+1}, \phi_{k+1})$.

Inversamente, se $\boldsymbol{\theta}$ representa uma configuração com $k+1$ componentes, $\boldsymbol{\theta} \cup (w_{k+1}, \phi_{k+1})$, uma componente j "morre" a uma taxa,

$$\delta_j(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta}|\mathbf{y})}{L(\boldsymbol{\theta} \cup (w_j, \phi_j)|\mathbf{y})} \frac{\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta} \cup (w_j, \phi_j))} \frac{\beta(\boldsymbol{\theta}) h(w_j, \phi_j)}{(k+1)(1-w_j)^{k+1}}. \tag{3.25}$$

que é obtida diretamente da condição de que a equação de balanceamento seja satisfeita, i.e.,

$$L(\boldsymbol{\theta}|\mathbf{y})k!\pi(\boldsymbol{\theta})\beta(\boldsymbol{\theta})h(w_j, \phi_j) = L(\boldsymbol{\theta} \cup (w_j, \phi_j)|\mathbf{y})(k+1)!\pi(\boldsymbol{\theta} \cup (w_j, \phi_j))\delta_j(\boldsymbol{\theta})(1-w_j)^{k+1}$$

onde $k!$ e $(k+1)!$ sugem da suposição de permutabilidade para as componentes da mistura e $(1-w_j)^{k+1}$ é o jacobiano da renormalização dos pesos.

Note que, neste processo o movimento proposto é sempre aceito; a probabilidade de aceitação, usuais em métodos MCMC, é substituída por diferentes tempos que o processo permanece em um determinado estado; e que uma componente j pouco provável deve "morrer" rapidamente.

Para simplificar a implementação do método, Stephens (2000) propõe gerar w_{k+1} e ϕ_{k+1} das distribuições *a priori*, $w_{k+1} \sim \mathcal{B}eta(1,k)$ e $\phi_{k+1} \sim \pi(\phi_{k+1})$. Assim, a taxa de morte em (3.25) é dada por

$$\delta_j(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta}|\mathbf{y})}{L(\boldsymbol{\theta} \cup (w_j, \phi_j)|\mathbf{y})} \frac{\beta(\boldsymbol{\theta})}{(k+1)(1-w_j)^{k+1}}. \tag{3.26}$$

A taxa total de morte é

$$\delta(\boldsymbol{\theta}) = \sum_{j=1}^{k} \delta_j(\boldsymbol{\theta}). \tag{3.27}$$

O algoritmo para implementar o processo de "nascimento" e "morte" pode ser resumido pelos seguintes passos:

(1) Inicie com a configuração $\boldsymbol{\theta} = ((w_1, \phi_1), \ldots, (w_k, \phi_k))$;

(2) Fixe a taxa de nascimento $\beta(\boldsymbol{\theta})$;

(3) Para a l-ésima iteração, $l = 1, \ldots, L$, faça:

   (i) Calcule a taxa de morte $\delta_j(\boldsymbol{\theta})$ para cada componente j, dada em (3.26);

   (ii) Calcule a taxa total de morte $\delta(\boldsymbol{\theta})$ dada em (3.27);

   (iii) Gere o tempo para o próximo salto da distribuição exponencial com média $\frac{1}{\beta(\boldsymbol{\theta})+\delta(\boldsymbol{\theta})}$;

   (iv) Escolha o tipo de salto de acordo com as probabilidade em (3.24);

Stephens (2000) estima o número de componentes k utilizando a probabilidade $P_{k=j} = \frac{N_{k=j}}{L-B}$, onde $N_{k=j}$ é o número de vezes que $k = j$ nas $L-B$ iterações, em que, B é o *burn in*.

Capítulo 4

Algoritmo *Split-Merge* MCMC

Neste capítulo, propomos um novo algoritmo, denominado por algoritmo *split-merge* MCMC, SM-MCMC, para análise de modelos com mistura de distribuições onde o número de componentes k é desconhecido.

Ao contrário do algoritmo RJ-MCMC e do processo de nascimento-e-morte, que propõem novas componentes a partir da proposta de novos parâmetros, nossa estrátegia é propor uma partição nos dados (*split*) e condicional nessa partição propor os parâmetros para as novas componentes. Esta proposta de partição nos dados é feita de forma sequencial, em que, uma observação é alocada para uma de duas novas componentes de acordo com probabilidades de alocação que são calculadas de acordo com os parâmetros gerados da distribuição *a posteriori* dadas as observações previamente alocadas. Para garantir a reversibilidade para o processo, temos o movimento inverso, em que, as observações de duas componentes são agrupadas (*merge*) para determinar uma nova componente. Estas propostas determinam um aumento de uma unidade no número de componentes pela proposta *split* e a diminuição de uma unidade pela proposta *merge* e são aceitas conforme a probabilidade de aceitação *Metropolis-Hastings*, para garantir a existência da distribuição estacionária.

Para simplificar a notação utilizada nas expressões abaixo, considere um conjunto de variáveis indicadoras $\mathbf{c} = \{c_1, \ldots, c_n\}$, tais que, $c_i = j$ se $z_{ij} = 1$, i.e., $c_i = j$ indica que a observação y_i pertence à componente j, $y_i \in S_j$, $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

Note que, dados $\mathbf{w}$ e k a variável latente $\mathbf{Z}_i$ segue a distribuição mutinomial e

uma individual variável indicadora c_i segue uma distribuição discreta com $P(c_i = j|\mathbf{w}, k) = w_j$, para $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

A distribuição conjunta de $\mathbf{c} = \{c_1, \ldots, c_n\}$ dados $\mathbf{w}$ e k é

$$\pi(\mathbf{c}|\mathbf{w}, k) = \prod_{j=1}^{k} w_j^{n_j}, \tag{4.1}$$

onde n_j é o número de $c_i's = j$.

A distribuição conjunta de todas as variáveis é dada por

$$P(\mathbf{y}, \boldsymbol{\phi}, \mathbf{c}, \mathbf{w}, k) = L(\boldsymbol{\phi}, \mathbf{c}, \mathbf{w}, k|\mathbf{y})\pi(\boldsymbol{\phi}|\mathbf{c}, \mathbf{w}, k)\pi(\mathbf{c}|\mathbf{w}, k)\pi(\mathbf{w}|k)\pi(k) \tag{4.2}$$

onde, $L(\cdot|\mathbf{y})$ é a função de verossimilhança dada em (3.4) e a distribuição *a priori* para $\boldsymbol{\phi}$ e $\mathbf{w}$, $\pi(\boldsymbol{\phi}|\mathbf{c}, \mathbf{w}, k)$ e $\pi(\mathbf{w}|k)$, são dadas em (3.1) e (3.2), respectivamente.

A distribuição *a priori* para k é dada pela distribuição uniforme discreta $U(1, k_{max})$, onde k_{max} é um valor previamente fixado que indica o máximo valor que k pode assumir.

Como nossa estratégia é baseada nas propostas de separar (*split*) ou agrupar (*merge*) observações, para em seguida propor os novos parâmetros para as novas componentes, consideramos a probabilidade *a priori* de $\mathbf{c}$ dado apenas k, pois pelas propostas temos uma modificação na configuração $\mathbf{c}$ e uma mudança de k para $k + 1$ ou $k - 1$ dependendo da proposta. Esta probabilidade é obtida utilizando a conjugação da distribuição multinomial com a distribuição de Dirichlet, integrando fora os pesos $\mathbf{w}$, i.e.,

$$\begin{aligned}
\pi(\mathbf{c}|k) &= \int \pi(\mathbf{c}|\mathbf{w}, k)\pi(\mathbf{w}|k)d\mathbf{w} \qquad (4.3)\\
&= \int \left[\prod_{j=1}^{k} w_j^{n_j}\right] \left[\frac{\Gamma(k\gamma)}{[\Gamma(\gamma)]^k} \prod_{j=1}^{k} w_j^{\gamma-1}\right] d\mathbf{w}\\
&= \frac{\Gamma(k\gamma)}{[\Gamma(\gamma)]^k} \int \prod_{j=1}^{k} w_j^{n_j+\gamma-1} d\mathbf{w}\\
&= \frac{\Gamma(k\gamma)}{[\Gamma(\gamma)]^k \Gamma(n + k\gamma)} \prod_{j=1}^{k} \Gamma(n_j + \gamma).
\end{aligned}$$

Assim, temos o seguinte modelo Bayesiano hierárquico conjunto,

$$\begin{aligned} \mathbf{y}|\boldsymbol{\phi}, \mathbf{c}, k &\sim \prod_{j=1}^{k}\prod_{i=1}^{n} f(y_i|\phi_j)^{\mathcal{I}_{\{c_i=j\}}} \\ \boldsymbol{\phi}|k &\sim \prod_{j=1}^{k} \pi(\phi_j) \\ \mathbf{c}|k &\sim \pi(\mathbf{c}|k) \\ k &\sim \pi(k), \end{aligned} \tag{4.4}$$

onde $\mathcal{I}_{\{c_i=j\}} = 1$ se $c_i = j$ e $\mathcal{I}_{\{c_i=j\}} = 0$ caso contrário.

A função de verossimilhança é

$$L(\boldsymbol{\phi}, \mathbf{c}, k|\mathbf{y}) = \prod_{j=1}^{k} L(\phi_j|S_j), \tag{4.5}$$

onde $L(\phi_j|S_j)$ é dada em (3.3).

Atualizando a distribuição *a priori* conjunta para $(\boldsymbol{\phi}, \mathbf{c}, k)$ via função de verossimilhança (4.5), a distribuição *a posteriori* é dada por

$$\pi(\boldsymbol{\phi}, \mathbf{c}, k|\mathbf{y}) \propto L(\boldsymbol{\phi}, \mathbf{c}, k|\mathbf{y})\pi(\boldsymbol{\phi}|k)\pi(\mathbf{c}|k)\pi(k). \tag{4.6}$$

4.1 Algoritmo *Split-Merge* MCMC

Para estimar todos os parâmetros do modelo, propomos o algoritmo SM-MCMC. Os parâmetros $\boldsymbol{\phi}$ e $\mathbf{w}$ são atualizados via algoritmo *Gibbs sampling*, gerando valores de suas distribuições *a posteriori*, enquanto $(\mathbf{c}, k)$ são gerados via algoritmo *Metropolis-Hastings*. Como é usual em *Metropolis-Hastings*, dado o estado atual $(\boldsymbol{\phi}, \mathbf{c}, k)$, propomos um movimento para $(\boldsymbol{\phi}^*, \mathbf{c}^*, k^*)$, que é aceito com probabilidade $\alpha[(\boldsymbol{\phi}^*, \mathbf{c}^*, k^*)|(\boldsymbol{\phi}, \mathbf{c}, k)] = min(1, A)$,

$$A = \frac{L(\boldsymbol{\phi}^*, \mathbf{c}^*, k^*|\mathbf{y})}{L(\boldsymbol{\phi}, \mathbf{c}, k|\mathbf{y})} \frac{\pi(\boldsymbol{\phi}^*|k^*)}{\pi(\boldsymbol{\phi}|k)} \frac{\pi(\mathbf{c}^*|k^*)}{\pi(\mathbf{c}|k)} \frac{\pi(k^*)}{\pi(k)} \frac{q[(\boldsymbol{\phi}, \mathbf{c}, k)|(\boldsymbol{\phi},^* \mathbf{c}^*, k^*)]}{q[(\boldsymbol{\phi}^*, \mathbf{c}^*, k^*)|(\boldsymbol{\phi}, \mathbf{c}, k)]} \tag{4.7}$$

onde, $q(\cdot|\cdot)$ é a proposta de transição, discutida abaixo.

Utilizamos dois tipos de movimentos diretamente na configuração das variáveis latentes $\mathbf{c}$, denominados por *split* (sp) e *merge* (mg). Sejam $P_{sp|k}$ e $P_{mg|k}$ as probabilidades de propor um *split* e um *merge*, respectivamente, com $P_{sp|k} + P_{mg|k} = 1$.

Como estas probabilidades são condicionais em k, fixamos

$$P_{sp|k} = \begin{cases} 1, \text{ se } k = 1 \\ 0.5, \text{ se } 2 \leq k \leq k_{max} - 1 \\ 0, \text{ se } k = k_{max} \end{cases},$$

e

$$P_{mg|k} = \begin{cases} 0, \text{ se } k = 1 \\ 0.5, \text{ se } 2 \leq k \leq k_{max} - 1 \\ 1, \text{ se } k = k_{max} \end{cases},$$

pois se $k = 1$, então podemos propor somente o movimento *split*, $P_{sp|k} = 1$; se $k = k_{max}$, então podemos propor somente o movimento *merge*, $P_{mg|k} = 1$. Para $2 \leq k \leq (k_m - 1)$, consideramos que as propostas *split* e *merge* são igualmente prováveis, $P_{sp|k} = P_{mg|k} = \frac{1}{2}$.

Com o objetivo de desenvolver um procedimento para atualizar as variáveis latentes e propor o movimento *split*, consideramos a probabilidade condicional *a priori* para uma única variável latente c_i dadas todas as outras, $\mathbf{c}_{-i} = (c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n)$. Utilizando a integral Dirichlet como em (4.3), para todas as variáveis latentes, com exceção de c_i, a probabilidade condicional *a priori* para $c_i|\mathbf{c}_{-i}$ é dada por

$$P(c_i = j|\mathbf{c}_{-i}, k) = \frac{n_{j,-i} + \gamma}{k\gamma + n - 1}, \tag{4.8}$$

onde $n_{j,-i}$ é o número de observações pertencentes a S_j, exceto a observação y_i, $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

4.1.1 Proposta *Split*

Seja $\boldsymbol{\theta} = (\boldsymbol{\phi}, \mathbf{c}, k)$ o estado atual. Dada a escolha de um *split*, considere $k_2 \leq k$ como sendo o número de componentes com $n_j \geq 2$, $j = 1, \ldots, k$. Assim, selecionamos uma componente j, com $n_j \geq 2$, com probabilidade $P_{j|k_2} = \frac{1}{k_2}$ e propomos o *split* das observações $y_i \in S_j$ em duas novas componentes j_1 e j_2, onde o conjunto das observações pertencentes a j_1 e j_2 são denotadas por S_{j_1} e S_{j_2}, respectivamente. Esta proposta é implementada pelos seguintes passos:

(i) Inicie selecionando duas observações $y_{v'}, y_{v''} \in S_j$;

(ii) Faça $S_{j_1} = \{y_{v'}\}$, $S_{j_2} = \{y_{v''}\}$ e $n_{j_1} = n_{j_2} = 1$;

(iii) Considere $S_j^* = \{S_j\} \setminus \{S_{j_1}\} \cup \{S_{j_2}\}$;

(iv) Gere $\phi_{j_1}^*$ e $\phi_{j_2}^*$ das distribuições *a posteriori*

$$\pi(\phi_{j_1}|S_{j_1}) \propto L(\phi_{j_1}|S_{j_1})\pi(\phi_{j_1}) \quad \text{e} \quad \pi(\phi_{j_2}|S_{j_2}) \propto L(\phi_{j_1}|S_{j_1})\pi(\phi_{j_1});$$

(v) Selecione uma observação $y_v \in S_j^*$. Dados $\phi_{j_1}^*$ e $\phi_{j_2}^*$, aloque y_v na componente j_1 com probabilidade (ver Apêndice B.1, pag.76)

$$P_{j_1}(y_v) = \frac{(n_{j_1} + \gamma) f(y_v|\phi_{j_1}^*)}{(n_{j_1} + \gamma) f(y_v|\phi_{j_1}^*) + (n_{j_2} + \gamma) f(y_v|\phi_{j_2}^*)}; \tag{4.9}$$

(vi) Gere uma variável indicadora $\mathcal{I}_{sp} \sim Bernoulli(P_{j_1}(y_v))$. Se $\mathcal{I}_{sp} = 1$, então $y_v \in S_{j_1}$. Faça $S_{j_1} = \{S_{j_1}\} \cup \{y_v\}$ e $n_{j_1} = n_{j_1} + 1$. Caso contrário, faça $S_{j_2} = \{S_{j_2}\} \cup \{y_v\}$ e $n_{j_2} = n_{j_2} + 1$;

(vii) Repita os passos (iii) a (vi) até $S_j^* = \emptyset$.

A probabilidade da configuração S_{j_1} e S_{j_2} é

$$P_{aloc} = \prod_{y_v \in S_{j_1}} P_{j_1}(y_v) \prod_{y_v \in S_{j_2}} (1 - P_{j_1}(y_v)),$$

onde $P_{j_1}(y_{v'}) = 1$ e $P_{j_1}(y_{v''}) = 0$.

Condicional em S_{j_1} e S_{j_2} gere candidatos ϕ_{j_1} e ϕ_{j_2} das distribuições *a posteriori* $\pi(\phi_{j_1}|S_{j_1})$ e $\pi(\phi_{j_2}|S_{j_2})$, respectivamente.

Assim, temos uma nova configuração $\mathbf{c}^{sp}$, um novo conjunto de parâmetros $\boldsymbol{\phi}^{sp}$ e o número de componentes k aumenta em uma unidade. Esta proposta de transição é denotada por $\boldsymbol{\theta}^{sp}|\boldsymbol{\theta}$, onde $\boldsymbol{\theta}^{sp} = (\boldsymbol{\phi}^{sp}, \mathbf{c}^{sp}, k+1)$, e sua probabilidade é dada por

$$q[\boldsymbol{\theta}^{sp}|\boldsymbol{\theta}] = P_{sp|k} P_{j|k_2} P_{aloc} \pi(\phi_{j_1}|S_{j_1}) \pi(\phi_{j_2}|S_{j_2}). \tag{4.10}$$

4.1.2 Proposta *Merge*

Dada a escolha de um *merge*, selecionamos duas componentes j_1 e j_2. Aqui, estabelecemos o critério que um *merge* é feito sempre para duas componentes que são

adjacentes em relação aos valores de sua médias (ver Apêndice B.2, pag.76). Desta forma, temos o movimento reverso para a proposta *split* e uma convergência mais rápida para a distribuição de equilíbrio.

A probabilidade de selecionar as componentes j_1 e j_2 para um *merge* é

$$P_{j_1,j_2|k} = P_{j_1|k}P_{j_2|j_1,k} + P_{j_2|k}P_{j_1|j_2,k} = \begin{cases} 1, \text{ if } k = 2 \\ \frac{3}{2k}, \text{ if } k > 2, \ j_1, j_2 \in \{1, k\} \\ \frac{1}{k}, \text{ if } k > 2, \ j_1, j_2 \notin \{1, k\} \end{cases}, \tag{4.11}$$

onde $P_{b_1|k}$ é a probabilidade de escolher a componente b_1 e $P_{b_2|b_1,k}$ é a probabilidade condicional de escolher a componente b_2 dado que a componente b_1 foi previamente selecionada.

Faça o *Merge* de j_1 e j_2, determinando a componente j com $S_j = \{S_{j_1}\} \cup \{S_{j_2}\}$. Gere um candidato ϕ_j de sua distribuição *a posteriori* $\pi(\phi_j|S_j)$.

A proposta *merge* determina uma nova configuração $\mathbf{c}^{mg}$, um novo conjunto de parâmetros $\boldsymbol{\phi}^{mg}$ e o número de componentes k diminui uma unidade. Esta proposta de transição é denotada por $\boldsymbol{\theta}^{mg}|\boldsymbol{\theta}$, onde $\boldsymbol{\theta}^{mg} = (\boldsymbol{\phi}^{mg}, \mathbf{c}^{mg}, k-1)$, e sua probabilidade é dada por

$$q[\boldsymbol{\theta}^{mg}|\boldsymbol{\theta}] = P_{mg|k}P_{j_1,j_2|k}\pi(\phi_j|S_j). \tag{4.12}$$

Note que, dado o estado atual $\boldsymbol{\theta}$, a probabilidade de propor um *split* da componente j nas componentes j_1 e j_2, $q[\boldsymbol{\theta}^{sp}|\boldsymbol{\theta}]$, é equivalente a estar no estado com j_1 e j_2 juntas como uma única componente j, $\boldsymbol{\theta}^{mg}$, e propor a volta para o estado atual $\boldsymbol{\theta}$, isto é,

$$q[\boldsymbol{\theta}^{sp}|\boldsymbol{\theta}] = q[\boldsymbol{\theta}|\boldsymbol{\theta}^{mg}] = P_{sp|k-1}P_{j|k_2}P_{aloc}\pi(\phi_{j_1}|S_{j_1})\pi(\phi_{j_2}|S_{j_2}). \tag{4.13}$$

Como existe somente uma forma de juntar duas componentes em uma única componente, precisamos calcular a correspondente probabilidade, P_{aloc}, de gerar o estado original, com as duas componentes separadas. Isto é feito como na proposta *split*, descrita acima, porém agora utilizando os parâmetros ϕ_{j_1} e ϕ_{j_2}, que são conhecidos, para calcular a probabilidade $P_{j_1}(\cdot)$ dada em (4.9).

Analogamente a (4.13),

$$q[\boldsymbol{\theta}^{mg}|\boldsymbol{\theta}] = q[\boldsymbol{\theta}|\boldsymbol{\theta}^{sp}] = P_{mg|k+1}P_{j_1,j_2}\pi(\phi_j|S_j). \tag{4.14}$$

4.1.3 Probabilidades de transição

Definidas as propostas *split* e *merge*, calculamos a probabilidade de aceitação A, dada em (4.7), de cada proposta.

Para isto, considere a razão das densidades *a posteriori*

$$P^r = \frac{\pi(\phi_j|S_j)}{\pi(\phi_{j_1}|S_{j_1})\pi(\phi_{j_2}|S_{j_2})} = \frac{L(\phi_j|S_j)\pi(\phi_j)}{L(\phi_{j_1}|S_{j_1})\pi(\phi_{j_1})L(\phi_{j_2}|S_{j_2})\pi(\phi_{j_2})}\frac{I_{j_1}I_{j_2}}{I_j},$$

onde $I_d = \int L(\phi_d|S_d)\pi(\phi_d)d\phi_d$ para $d \in \{j, j_1, j_2\}$.

Assim, a probabilidade de transição para a proposta *split* é dada por

$$\frac{q[\boldsymbol{\theta}|\boldsymbol{\theta}^{sp}]}{q[\boldsymbol{\theta}^{sp}|\boldsymbol{\theta}]} = \frac{P_{mg|k+1}}{P_{sp|k}}\frac{P_{j_1,j_2|k+1}}{P_{j|k_2}}\frac{P^r}{P_{aloc}} = Q^{sp}P^r \tag{4.15}$$

onde,

$$Q^{sp} = \begin{cases} \frac{1}{2P_{aloc}}, \text{ se } k = 1 \\ \left(\frac{1}{2}\right)^{1-\mathcal{I}_{\{k=k_{max}-1\}}} \frac{3k_2}{k+1}\frac{1}{P_{aloc}}, \text{ se } 2 \leq k \leq k_{max}-1,\ j_1, j_2 \in \{1, k\} \\ 2^{\mathcal{I}_{\{k=k_{max}-1\}}} \frac{k_2}{k+1}\frac{1}{P_{aloc}}, \text{ se } 2 \leq k \leq k_{max}-1,\ j_1, j_2 \notin \{1, k\} \end{cases}, \tag{4.16}$$

onde $\mathcal{I}_{\{k=k_{max}-1\}} = 1$ se $k = k_{max} - 1$ e $\mathcal{I}_{\{k=k_{max}-1\}} = 0$ caso contrário.

A probabilidade de transição para a proposta *merge* é dada por

$$\frac{q[\boldsymbol{\theta}|\boldsymbol{\theta}^{mg}]}{q[\boldsymbol{\theta}^{mg}|\boldsymbol{\theta}]} = \frac{P_{sp|k-1}}{P_{mg|k}}\frac{P_{j|k_2}}{P_{j_1,j_2|k}}\frac{P_{aloc}}{P^r} = Q^{mg}\frac{1}{P^r} \tag{4.17}$$

onde,

$$Q^{mg} = \quad = \begin{cases} 2P_{aloc}, \text{ se } k = 2 \\ (2)^{1-\mathcal{I}_{\{k=k_{max}\}}} \frac{k}{3k_2}P_{aloc}, \text{ se } 3 \leq k \leq k_{max}, j_1, j_2 \in \{1, k\} \\ \left(\frac{1}{2}\right)^{\mathcal{I}_{\{k=k_{max}\}}} \frac{k}{k_2}P_{aloc}, \text{ se } 3 \leq k \leq k_{max},\ j_1, j_2 \notin \{1, k\} \end{cases}, \tag{4.18}$$

onde $\mathcal{I}_{\{k=k_{max}\}} = 1$ se $k = k_{max}$ e $\mathcal{I}_{\{k=k_{max}\}} = 0$ caso contrário.

Considerando a proposta *split*, a razão das funções de verossimilhança é dada por

$$\frac{L(\boldsymbol{\theta}^{sp}|\mathbf{y})}{L(\boldsymbol{\theta}|\mathbf{y})} = \frac{L(\phi_{j_1}|S_{j_1})L(\phi_{j_2}|S_{j_2})}{L(\phi_j|S_j)}, \tag{4.19}$$

e a razão das distribuições *a priori*, com $\gamma = 1$, é dada por

$$\frac{\pi(\boldsymbol{\theta}^{sp})}{\pi(\boldsymbol{\theta})} = \frac{\pi(\phi_{j_1})\pi(\phi_{j_2})}{\pi(\phi_j)}\frac{\Gamma(n_{j_1}+1)\Gamma(n_{j_2}+1)}{\Gamma(n_j+1)}\frac{k(k+1)}{n+k}\frac{\pi(k+1)}{\pi(k)}, \tag{4.20}$$

onde $(k+1)$ suge da suposição de permutabilidade para as componentes da mistura.

A probabilidade de aceitação para a proposta *split* é $\alpha[\boldsymbol{\theta}^{sp}|\boldsymbol{\theta}] = min(1, A^{sp})$, dada pelo produto de (4.19), (4.20) e (4.15), onde

$$A^{sp} = \frac{I_{j_1} I_{j_2}}{I_j} \frac{\Gamma(n_{j_1}+1)\Gamma(n_{j_2}+1)}{\Gamma(n_j+1)} \frac{k(k+1)}{n+k} Q^{sp}. \tag{4.21}$$

A probabilidade de aceitação para a proposta *merge* é $\alpha[\boldsymbol{\theta}^{mg}|\boldsymbol{\theta}] = min(1, A^{mg})$ onde

$$A^{mg} = \frac{I_j}{I_{j_1} I_{j_2}} \frac{\Gamma(n_j+1)}{\Gamma(n_{j_1}+1)\Gamma(n_{j_2}+1)} \frac{n+k-1}{k(k-1)} Q^{mg}. \tag{4.22}$$

4.1.4 Atualização das variáveis latentes

Dada a aceitação ou não de uma proposta *split* ou *merge*, atualizamos as variáveis latentes $\mathbf{c}$ condicional ao atual valor de $\boldsymbol{\phi}$.

Dados y_i, ϕ_j e $\mathbf{c}_{-i}$ a probabilidade condicional *a posteriori* de $c_i = j$ é dada por

$$\begin{aligned} p(c_i = j|y_i, c_{-i}, \phi_j, k) &\propto p(c_i = j|\mathbf{c}_{-i}, k) p(y_i|c_i = j, \mathbf{c}_{-i}, \phi_j) \\ &= b \frac{n_{j,-i}+1}{k\gamma + n - 1} f(y_i|\phi_j), \end{aligned} \tag{4.23}$$

onde $p(c_i = j|\mathbf{c}_{-i}, k)$ é dada em (4.8) e $b = \left[\sum_{j=1}^{k} \frac{n_{j,-i}+1}{k\gamma+n-1} f(y_i|\phi_j)\right]^{-1}$ é a constante normalizadora.

4.1.5 Comentários

Para garantir identificabilidade em nossa proposta, consideramos uma marcação das componentes de acordo as médias μ_j de cada componente, que são consideradas em ordem crescente de valor, $\mu_1 < \mu_2 < \ldots < \mu_k$. Assim, a primeira componente é a que tem média μ_1, a segunda é a que tem média μ_2 e sucessivamente até a última componente que tem média μ_k (Richardson e Green, 1997).

Condicional neste tipo de marcação das componentes, temos que ao propor novos parâmetros ϕ_{j_1}, ϕ_{j_1} na proposta *split*, devemos checar se a condição de adjacência é satisfeita, i.e., se $\mu_{j_1-1} < \mu_{j_1} < \mu_{j_2} < \mu_{j_2+1}$. Caso não seja, devemos rejeitar a proposta pois os o movimentos *split-merge* podem não ser reversíveis.

A mesma verificação é feita para a proposta *merge*, i.e., se escolhemos as componentes j_1 e j_2 para um merge, então a média μ_j da nova componente $S_j = \{S_{j_1}\} \cup \{S_{j_2}\}$ deve satisfazer a condição $\mu_{j_1-1} < \mu_j < \mu_{j_2+1}$.

Com isso, temos que o processo estocástico definido pelos movimentos *split* e *merge* é reversível por construção, $A^{mg} = \frac{1}{A^{sp}}$. Logo, a equação de balanceamento $L(\theta|\mathbf{y})\pi(\boldsymbol{\theta})q[\boldsymbol{\theta}^*|\boldsymbol{\theta}] = L(\boldsymbol{\theta}^*|\mathbf{y})\pi(\boldsymbol{\theta}^*)q[\boldsymbol{\theta}|\boldsymbol{\theta}^*]$, necessária para garantir a existência da distribuição invariante (Preston, 1976; Ripley, 1976; Geyer e Møller, 1994; Tierney, 1994), que é proporcional a $L(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\theta})$, é satisfeita. Dada uma vizinhança ao redor do estado atual, existe uma probabilidade positiva de que a cadeia se mova para esta vizinhança, isto é, a cadeia é aperiódica. Como a cadeia pode se mover para outro valor de k em cada iteração e cada observação tem probabilidade positiva de ser alocada em uma das componentes, a cadeia também é irredutível.

4.1.6 Algoritmo:

Definidas as probabilidades de aceitação para as propostas *split* e *merge*, expressamos o método proposto como um algoritmo.

Algoritmo SM-MCMC:

(1) Inicialize todos os parâmetros;

(2) Para a l-ésima iteração, $l = 1, \ldots, L$, Faça:

- **(i)** Escolha entre *split* ou *merge* com probabilidades $P_{sp|k}$ e $P_{mg|k}$, respectivamente;
- **(ii)** Aceite a proposta com probabilidade $\alpha[\boldsymbol{\theta}^*|\boldsymbol{\theta}]$, onde o sinal $*$ é sp ou mg;
- **(iii)** Atualize $\mathbf{c}$ utilizando as probabilidade condicionais *a posteriori* em (4.23);
- **(iii)** Atualize $\boldsymbol{\phi}$ e $\mathbf{w}$
  - **(a)** Gere $\boldsymbol{\phi}$ da distribuição *a posteriori* em (3.5);
  - **(b)** Considere w_j como em (3.7);

Para estimar k, consideramos um *burn in* B das L primeiras iterações e calculamos o número de vezes $N_{k=j}$ que $k = j$ nas $L - B$ iterações. Assim, $P_{k=j} = \frac{N_{k=j}}{L-B}$ é a probabilidade *a posteriori* para $k = j$ e $\widetilde{k} = \underset{1 \leq j \leq k_{max}}{\arg\max}(P_{k=j})$ é a estimativa para o número de componentes, $j = 1, \ldots, k_{max}$.

Dado $\widetilde{k}$, seja $L_0 \leq L$ o número de iterações em que o número de componentes é $k = \widetilde{k}$. Para estimar os parâmetros ϕ_j e w_j, consideramos um *burn in* B_0 das L_0 iterações, obtendo uma cadeia de tamanho $L_{\widetilde{k}} = L_0 - B_0$. Como é usual na abordagem Bayesiana, utilizamos a média dos valores gerados como estimativa, i.e.,

$$\widetilde{\phi}_j|\widetilde{k} = \frac{1}{L_{\widetilde{k}}} \sum_{l=1}^{L_{\widetilde{k}}} \phi_j^{(l)} \quad \text{e} \quad \widetilde{w}_j|\widetilde{k} = \frac{1}{L_{\widetilde{k}}} \sum_{l=1}^{L_{\widetilde{k}}} w_j^{(l)}. \tag{4.24}$$

A probabilidade *a posteriori* de cada observação ser proveniente da componente j é $P_{ij} = \frac{N_{ij}}{L_{\widetilde{k}}}$, onde N_{ij} é o número de vezes que a observação y_i é associada à componente $j \in \{1, \ldots, \widetilde{k}\}$ nas $L_{\widetilde{k}}$ iterações. Se $P_{ij} = \underset{1 \leq j \leq \widetilde{k}}{max}(P_{ij})$, então consideramos que y_i é proveniente da componente j, $i = 1, \ldots, n$ e $j = 1, \ldots, \widetilde{k}$.

4.2 Análise de dados

Nesta seção, aplicamos o método proposto, SM-MCMC, a cinco conjuntos de dados. Os três primeiros são dados artificiais, obtidos via simulação. O quarto é o bem conhecido conjunto de dados sobre a velocidade de galáxias e o quinto é um conjunto de dados sobre expressão gênica, utilizado por Arfin *et al.* (2000).

Como em Richardson e Green (1997) e Stephens (2000), modelamos os conjuntos de dados utilizando uma mistura de distribuições normais, i.e.,

$$f(y_i|\boldsymbol{\phi}, \mathbf{w}) = \sum_{j=1}^{k} w_j f(y_i|\phi_j)$$

onde $f(y_i|\phi_j)$ é a densidade da distribuição normal com média μ_j e variância σ_j^2, $\phi_j = (\mu_j, \sigma_j^2)$, $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

Seguindo a mesma linha de Casella, Robert and Wells (2000), utilizamos distribuições *a priori* conjugadas para os parâmetros μ_j e σ_j^2,

$$\mu_j|\sigma_j^2, \lambda \sim \mathcal{N}\left(0, \frac{\sigma_j^2}{\lambda}\right) \quad \text{e} \quad \sigma_j^2|\tau, \nu \sim \mathcal{IG}\left(\frac{\tau}{2}, \frac{\nu}{2}\right),$$

onde λ, τ e ν são hiperparâmetros conhecidos.

As distribuições condicionais são dadas por,

$$\mu_j|\sigma_j^2, \mathbf{y}, \mathbf{c}, k \sim \mathcal{N}\left(\frac{\sum_{S_j} y_i}{n_j+\lambda}, \frac{\sigma_j^2}{n_j+\lambda}\right) \tag{4.25}$$

e

$$\sigma_j^2|\mathbf{y}, \mathbf{c}, k \sim \mathcal{IG}\left(\frac{n_j+\tau+1}{2}, \frac{\nu+\sum_{S_j} y_i^2}{2} - \frac{\left(\sum_{S_j} y_i\right)^2}{2(n_j+\lambda)}\right), \tag{4.26}$$

para $j = 1, \ldots, k$.

A constante normalizadora $I_d = \int L(\phi_d|S_d)\pi(\phi_d)d\phi_d$, $d \in \{j, j_1, j_2\}$, é dada por

$$I_d = \left[\frac{1}{\nu\pi}\right]^{\frac{n_d}{2}} \left[\frac{\lambda}{n_d+\lambda}\right]^{\frac{1}{2}} \frac{\Gamma(\frac{n_d+\tau}{2})}{\Gamma(\frac{\tau}{2})} \left[1 + \frac{\sum_{S_d} y_i^2}{\nu} - \frac{(\sum_{S_d} y_i)^2}{\nu(n_d+\lambda)}\right]^{-\frac{n_d+\tau}{2}}. \tag{4.27}$$

Os hiperparâmetros utilizados foram fixados de forma que as distribuições *a priori* sejam não informativas. Fixamos $\lambda = 0.1$, $\tau = 0.1$, $\nu = 0.1$, $\gamma = 1$ and $k_{max} = 10$. Para verificar a obtenção de convergência, geramos duas cadeias e utilizamos o diagnóstico de Gelman e Rubin (Gelman e Rubin, 1992).

Aplicamos o algoritmo SM-MCMC com $L = 100.000$ e um *burn in* $B = 10.000$. Este número de iterações foi considerado suficiente pois, para os 5 conjunto de dados utilizados, o diagnóstico de Gelman-Rubin para todos os parâmetros é ≤ 1.01. Iniciamos o algoritmo com uma componente, i.e., $c_1 = \ldots = c_n = 1$ e parâmetros $\mu_1 = \overline{y}$ e $\sigma_1^2 = s^2$, onde $\overline{y}$ e s^2 são a média e a variância das observações $\mathbf{y}$.

4.2.1 Dados artificiais 1

Para geração do primeiro conjunto de dados artificiais, consideramos um modelo com mistura de distribuições normais com $k = 2$ componentes,

$$f(y_i|\boldsymbol{\phi}, \mathbf{w}) = w_1 f(y_i|\phi_1) + (1-w_1) f(y_i|\phi_2),$$

onde $\phi_1 = (\mu_1, \sigma_1^2)$ e $\phi_2 = (\mu_2, \sigma_2^2)$.

Para geração dos dados, fixamos $n = 100$, $w_1 = 0.5$, $w_2 = 0.5$, $\mu_1 = 0$, $\mu_2 = 4$ e $\sigma_1^2 = \sigma_2^2 = 1$.

O procedimento de simulação é dado pelos seguintes passos:

1. Gere n observações da seguinte forma: para $i = 1, \ldots, n$, gere $U_i \sim \mathcal{U}(0,1)$; se $u_i \leq w_1$, gere $Y_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$; se $u_i > w_1$, gere $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$;

2. Para registrar de qual componente cada observação é gerada, considere o vetor $G = (G_1, \ldots, G_n)$, tal que, $G_i = 1$ se $u_i \leq w_1$ e $G_i = 2$ se $u_i > w_1$;

3. Aplique o algoritmo SM-MCMC;

Os resultados obtidos estão apresentados nas Tabelas 4.1 e 4.2. A Tabela 4.1 mostra a probabilidade *a posteriori* para o número de componentes k. O máximo *a posteriori* é obtido em $\widetilde{k} = 2$, que tem probabilidade $P(k = 2|\cdot) = 0.9874$.

A Tabela 4.2 mostra as estimativas para os parâmetros (média dos valores gerados) e os intervalos de credibilidade empíricos (quantis 0.025 e 0.975 dos valores gerados). Como esperado, os verdadeiros valores dos parâmetros pertencem aos intervalos de 95% de credibilidade. O Apêndice C.1 (pag.78) mostra os gráficos ergódicos dos valores gerados para os parâmetros, indicando a convergência.

Tabela 4.1: Probabilidade *a posteriori* para k.

k	1	2	3	≥ 4
$P(k\|\cdot)$	0.0001	0.9874	0.0124	0.0001

Tabela 4.2: Estimativas para os parâmetros.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. G. R.
μ_1	-0.0416	(-0.3312, 0.2757)	1
μ_2	3.9814	(3.6790, 4.2557)	1.01
σ_1	0.9187	(0.5603, 1.4986)	1
σ_2	0.7976	(0.4801, 1.3206)	1.01
w_1	0.5083	(0.4706, 0.5392)	1.01
w_2	0.4917	(0.4608, 0.5294)	1.01

4.2.2 Dados artificiais 2

Para geração do segundo conjunto de dados artificiais, consideramos um modelo com mistura de distribuições normais com $k = 3$ componentes,

$$f(y_i|\boldsymbol{\phi}, \mathbf{w}) = w_1 f(y_i|\phi_1) + w_2 f(y_i|\phi_2) + w_3 f(y_i|\phi_3),$$

onde $\phi_1 = (\mu_1, \sigma_1^2)$, $\phi_2 = (\mu_2, \sigma_2^2)$ e $\phi_3 = (\mu_3, \sigma_3^2)$.

Fixamos $n = 100$, $w_1 = 0.3$, $w_2 = 0.4$, $w_3 = 0.3$, $\mu_1 = -4$, $\mu_2 = 0$, $\mu_3 = 4$ e $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$.

O procedimento de simulação é dado pelos seguintes passos:

1. Gere n observações da seguinte forma: para $i = 1, \ldots, n$, gere $U_i \sim \mathcal{U}(0, 1)$; se $u_i \leq w_1$, gere $Y_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$; se $w_1 < u_i \leq w_1 + w_2$, gere $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$; se $u_i > w_1 + w_2$, gere $Y_i \sim \mathcal{N}(\mu_3, \sigma_3^2)$;

2. Para registrar de qual componente cada observação é gerada, considere o vetor $G = (G_1, \ldots, G_n)$, tal que, $G_i = 1$ se $u_i \leq w_1$, $G_i = 2$ se $w_1 < u_i \leq w_1 + w_2$ e $G_i = 3$ if $u_i > w_1 + w_2$;

3. Aplique o algoritmo SM-MCMC;

Os resultados estão apresentados nas Tabelas 4.3 e 4.4. O maximo *a posteriori* para k é obtido em $\widetilde{k} = 3$, com $P(k = 3|\cdot) = 0.9154$. Os verdadeiros valores dos parâmetros pertencem aos intervalos de 95% de credibilidade. No Apêndice C.2 (pag.79) temos os gráficos ergódicos dos valores gerados para os parâmetros, indicando a convergência.

Tabela 4.3: Probabilidade *a posteriori* para k.

k	1	2	3	≥ 4
$P(k\|\cdot)$	0.0202	0.0407	0.9154	0.0238

Tabela 4.4: Estimativas para os parâmetros.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. G. R.
μ_1	-4.1317	(-4.6601, -3.3386)	1
μ_2	0.0513	(-0.3134, 0.4156)	1
μ_3	3.9702	(3.6391, 4.2682)	1.01
σ_1	1.1055	(0.7485, 1.7849)	1
σ_2	0.9786	(0.6901, 1.4305)	1.01
σ_3	0.7574	(0.5569, 1.0550)	1
w_1	0.2620	(0.2233, 0.3301)	1
w_2	0.4385	(0.2718, 0.5049)	1
w_3	0.2995	(0.2718, 0.3204)	1.01

4.2.3 Dados artificiais 3

Para geração do terceiro conjunto de dados artificiais, consideramos um modelo com mistura de distribuições normais com $k = 5$ componentes,

$$f(y_i|\boldsymbol{\phi}, \mathbf{w}) = \sum_{j=1}^{k=5} w_j f(y_i|\phi_j),$$

onde $\phi_j = (\mu_j, \sigma_j^2)$, $j = 1, \ldots, 5$.

Para geração do dados fixamos $n = 100$, $\mu_1 = -15$, $\mu_2 = -6$, $\mu_3 = 0$, $\mu_4 = 5$, $\mu_5 = 12$, $\sigma_j^2 = 1$ and $w_j = 0.2$, for $j = 1, \ldots, 5$.

Os resultados estão apresentados nas Tabelas 4.5 e 4.6. Na Tabela 4.5 estão as probabilidades *a posteriori* para k. O máximo *a posteriori* é obtido em $\widetilde{k} = 5$, que têm probabilidade $P(k = 5|\cdot) = 0.9219$. A Tabela 4.6 mostra as estimativas para os parâmetros e seus respectivos intervalos de 95% de credibilidade. Como esperado, os verdadeiros valores dos parâmetros pertencem aos intervalos de credibilidade. O Apêndice C.3 (pag.80) mostra os gráficos ergódicos dos valores gerados para os parâmetros (μ_j, σ_j), $j = \{1, 3, 5\}$, também indicando a convergência. Para (μ_j, σ_j), $j = \{2, 4\}$ os graficos ergódicos são similares.

Tabela 4.5: Probabilidade *a posteriori* para k.

k	1	2	3	4	5	≥ 6
$P(k\|\cdot)$	0.0001	0.0002	0.0030	0.0403	0.9219	0.0345

Tabela 4.6: Estimativas para os parâmetros.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. G. R.
μ_1	-15.0841	(-15.7387, -14.4273)	1
μ_2	-5.7788	(-6.2316, -5.3222)	1
μ_3	-0.0019	(-0.4249, 0.4372)	1
μ_4	4.5590	(3.0465, 5.3495)	1.01
μ_5	11.9655	(11.5643, 12.3691)	1
σ_1	1.5410	(1.1595, 2.0870)	1.01
σ_2	0.9256	(0.6673, 1.3285)	1.01
σ_3	0.9565	(0.6650, 1.3607)	1.01
σ_4	1.1250	(0.6198, 2.3764)	1
σ_5	1.0300	(0.7916, 1.3632)	1.01
w_1	0.2191	(0.1910, 0.2190)	1
w_2	0.1714	(0.1714, 0.2117)	1
w_3	0.2420	(0.1905, 0.2571)	1.01
w_4	0.1106	(0.0952, 0.2119)	1
w_5	0.2571	(0.1954, 0.2571)	1

4.2.4 Dados de velocidades de galáxias

Aplicamos o método proposto ao conhecido conjunto de dados sobre a velocidade de galáxias, previamente analisados por Roeder and Wasserman (1995), Escobar e West (1995), Richardson e Green (1997) e Stephens (2000). Seguindo a mesma linha destes autores, consideramos que a velocidade das galáxias são realizações de variáveis aleatórias distribuídas segundo um modelo com mistura de distribuições

normais com k (desconhecido) componentes.

Estimativas para o número de componentes k para este conjunto de dados vão desde $k = 3$ em Roeder and Wasserman (1995) e Stephens (2000) à $k = 5$ ou $k = 6$ em Richardson e Green (1997) e $k = 7$ em Escobar e West (1995).

Os resultados obtidos com a aplicação do algoritmo SM-MCMC estão apresentados nas Tabelas 4.7 e 4.8 e Figura 4.1. As probabilidades *a posteriori* para o número de componentes k são mostradas na Tabela 4.7, com um máximo em $\widetilde{k} = 3$, $P(k = 3|\cdot) = 0.9842$. A Tabela 4.8 mostra a estimativa e o intervalo de 95% de credibilidade para cada um dos parâmetros. O Apêndice C.4 (pag.81) mostra o gráfico ergódico dos valores gerados para os parâmetros. A Figura 4.1 mostra o histograma do conjunto de dados e a densidade estimada pelo método proposto.

Tabela 4.7: Probabilidade *a posteriori* para k.

k	1	2	3	≥ 4
$P(k\|\cdot)$	0	0.0053	0.9842	0.0105

Tabela 4.8: Estimativas para os parâmetros.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. G. R.
μ_1	9.6950	(9.2336, 10.1573)	1
μ_2	21.3297	(20.7646, 21.8735)	1.01
μ_3	31.0525	(23.2596, 35.3912)	1
σ_1	0.5896	(0.3597, 1.0176)	1
σ_2	2.1732	(1.7234, 2.6034)	1
σ_3	3.2168	(1.1664, 7.6744)	1.01
w_1	0.0939	(0.0941, 0.1419)	1
w_2	0.8330	(0.0471, 0.8588)	1
w_3	0.0731	(0.0471, 0.2235)	1.01

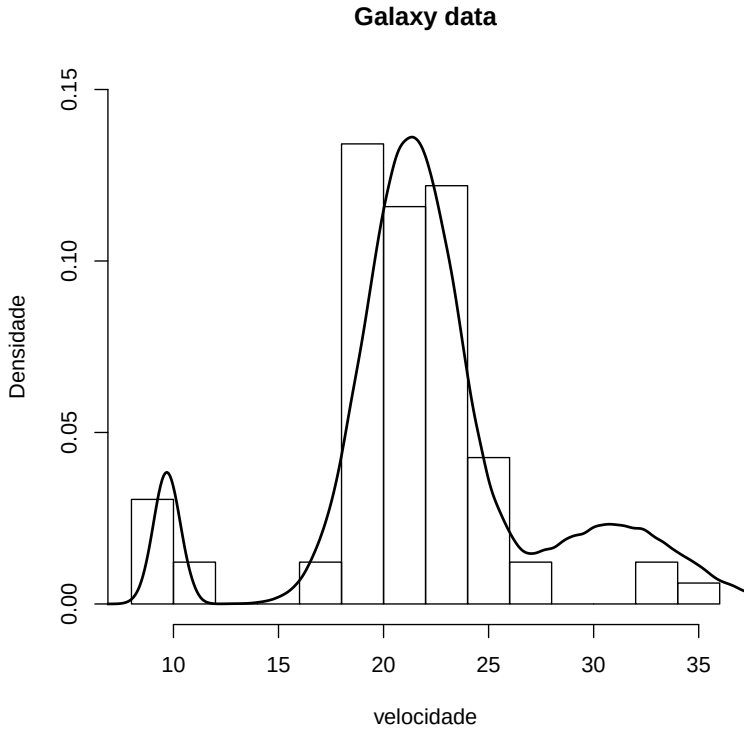

Figura 4.1: Histograma dos dados sobre a velocidade de galáxias e densidade estimada pelo algoritmo SM-MCMC.

4.2.5 Dados de expressão gênica

Segundo Baldi e Long (2001), dados de expressão gênica podem ser analisados sob pelo menos 3 níveis de crescente dificuldade. No primeiro nível de análise, os genes são analisados separadamente, um gene por vez, onde o objetivo é identificar se uma diferença de níveis de expressão observados, entre uma condição de tratamento em relação a um controle, é significante ou não. No segundo nível de análise o objetivo é identificar grupos de genes com níveis de expressão gênica similares. No terceiro nível de análise o objetivo é identificar as relações existentes entre genes e proteínas, que são responsáveis pelas caracteristicas observaveis (fenótipos).

Nesta seção, consideramos o segundo nível de análise de dados de expressão gênica com o objetivo de identificar grupos de genes com níveis de expressão gênica similares, utilizando o algoritmo SM-MCMC. Para isto, assumimos que os níveis de expressão gênica observados são realizações de variáveis aleatórias distribuídas segundo um modelo com mistura de distribuições normais com k componentes, em que, k é desconhecido.

Os dados utilizados, descritos em Arfin *et al.* (2000), são observações obtidas do experimento realizado com a bactéria *Escherichia Coli*. Este conjunto de dados é composto por $n = 434$ genes, em que, para cada gene i, temos 5 medidas de níveis de expressão em controle (c), denotado $\mathbf{y}_{i_c} = \{y_{1i_c}, ..., y_{5i_c}\}$, e 5 medidas de níveis de expressão em tratamento (t), denotado por $\mathbf{y}_{i_t} = \{y_{1i_t}, ..., y_{5i_t}\}$.

Seja $y_i = \overline{y}_{i_t} - \overline{y}_{i_c}$ o efeito de tratamento observado para o gene i e $\mathbf{y} = \{y_1, \ldots, y_n\}$ o conjunto de todos os efeitos de tratamento observados. Assumimos que genes com efeito de tratamento similares, isto é, efeitos de tratamento gerados segundo uma mesma mesma distribuição normal, determinam um grupo de genes. Esta distribuição normal é diferente da distribuição normal associada a um outro grupo.

Considerando k (desconhecido e finito) como sendo o número de grupos e $\mathbf{w} = (w_1, \ldots, w_k)$, tal que , w_j é a probabilidade *a priori* do gene g pertencer ao grupo j, temos o seguinte modelo com mistura de distribuições normais

$$f(y_i|\mu_j, \sigma_j^2, w_j, k) = \sum_{j=1}^{k} w_j f(y_i|\mu_j, \sigma_j^2), \tag{4.28}$$

onde $f(y_i|\mu_j, \sigma_j^2)$ é a densidade da distribuição normal com parâmetros (μ_j, σ_j^2) e w_j é o peso associado a componente j, $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

As distribuições *a priori* para os parâmetros e os hiperparâmetros utilizados são como os descritos no inicio da seção.

Os resultado obtidos estão apresentados nas Tabelas 4.9 e 4.10 e na Figura 4.2. Na Tabela 4.9 estão as probabilidades *a posteriori* para o número de componentes k, com um máximo em $\widetilde{k} = 3$, $P(k = 3|\cdot) = 0.9111$. A Tabela 4.10 mostra as estimativas para os parâmetros e os intervalos de credibilidade. O Apêndice C.5

(pag.82) mostra o gráfico ergódico dos valores gerados para os parâmetros, indicando a convergência.

A Figura 4.2 mostra a média de controle *versus* a média de tratamento dos 434 genes e os 3 grupos identificados. Triangulos $\triangledown$ representam o grupo G_1, $\bullet$ representam o grupo G_2 e os triangulos $\triangle$ representam o grupo G_3. O grupo G_1 é composto por 61 genes, o grupo G_2 por 359 genes e o grupo G_3 por 14 genes.

Observando a Figura 4.2 podemos interpretar os resultados da seguinte forma: (i) genes em G_3 apresentam níveis de expressão de tratamento maiores do que os de controle; (ii) genes em G_1 apresentam níveis de expressão de tratamento menores do que os de controle e (iii) genes em G_2 apresentam diferenças não significantes, $0 \in IC_{\mu_2}$.

Tabela 4.9: Probabilidade *a posteriori* k.

k	1	2	3	≥ 4
$P(k\|\cdot)$	0.0001	0.0723	0.9111	0.0165

Tabela 4.10: Estimativas para os parâmetros.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. G. R.
μ_1	-0.2065	(-0.4787, -0.0285)	1
μ_2	0.0372	(-0.0082, 0.0674)	1
μ_3	0.6670	(0.1604, 1.3501)	1
σ_1	0.2519	(0.1447, 0.3465)	1.01
σ_2	0.1516	(0.1156, 0.1817)	1.01
σ_3	0.6199	(0.3922, 0.9466)	1
w_1	0.2720	(0.0709, 0.5400)	1
w_2	0.6772	(0.0160, 0.8810)	1.01
w_3	0.0507	(0.0160, 0.1030)	1

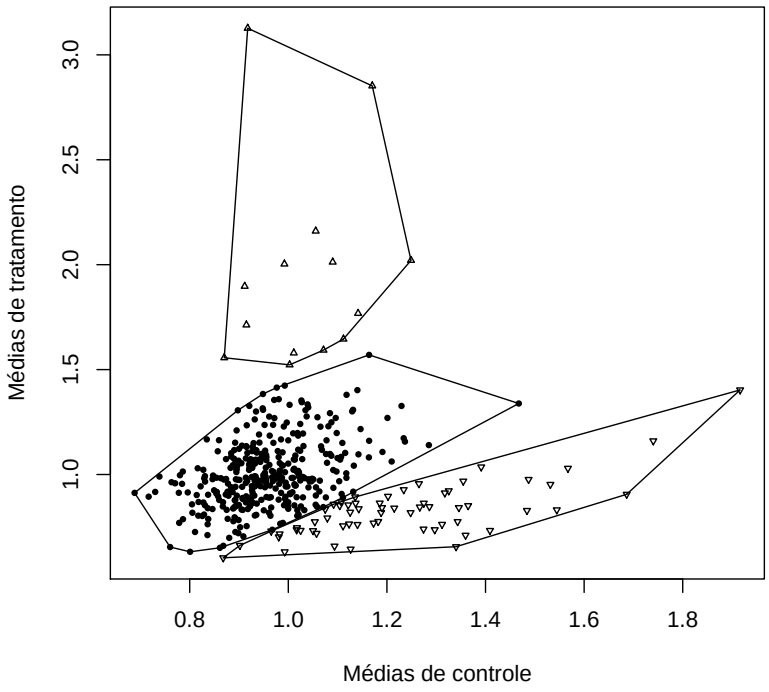

Figura 4.2: Médias de controle *versus* médias de Tratamento e grupos identificados. $\triangledown$ representam o grupo G_1, $\bullet$ o grupo G_2 e $\triangle$ o grupo G_3, SM-MCMC.

4.3 Comparação com o algoritmo *reversible-jump*

Nesta seção, aplicamos o algoritmo RJ-MCMC aos cinco conjunto de dados utilizados na Seção 4.2. Utilizamos a mesma modelagem com mistura de distribuições normais e as mesmas distribuições *a priori*. Os hiperparâmetros são os mesmos, $\lambda = 0.1$, $\tau = 0.1$, $\nu = 0.1$ e $\gamma = 1$. O número de iterações, L, o *burn in*, B, e o número de cadeias foram fixados como na aplicação do SM-MCMC.

Comparamos somente com o RJ-MCMC pois este método segue a mesma linha de nosso trabalho, i.e., considera a marcação das componentes (dados completos), ao contrário do método de Stephens (2000) que não utiliza a marcação das compoentes.

4.3.1 Dados artificiais 1

Os resultados obtidos para os dados artificiais 1 estão apresentados nas Tabelas 4.11 e 4.12. Na Tabela 4.11 estão as probabilidades *a posteriori* para k. O máximo *a posteriori* é obtido em $\widetilde{k} = 2$, que tem, probabilidade $P(k = 2|\cdot) = 0.5770$. A Tabela 4.12 mostra as estimativas para os parâmetros, os intervalos de credibilidade empíricos e os valores do diagnóstico de Gelman e Rubin. Os verdadeiros valores dos parâmetros pertencem aos intervalos de 95% de credibilidade.

Tabela 4.11: Probabilidade *a posteriori* para k.

k	1	2	3	4	5	≥ 6
$P(k\|\cdot)$	0.0576	0.5770	0.2627	0.0774	0.0197	0.0057

Tabela 4.12: Estimativas para os parâmetros.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. G. R.
μ_1	-0.1640	(-0.9831, 1.3009)	1.19
μ_2	3.3265	(1.8040, 4.2341)	1.02
σ_1	0.8227	(0.2031, 1.9722)	1.17
σ_2	1.4042	(0.7032, 2.4288)	1.02
w_1	0.3728	(0.0196, 0.5980)	1.01
w_2	0.6272	(0.4020, 0.9804)	1.01

Comparado ao algoritmo SM-MCMC o algoritmo RJ-MCMC também tem como máximo *a posteriori* o modelo com $k = 2$. Porém, a probabilidade *a posteriori* para $k = 2$ no algoritmo RJ-MCMC é menor do que a probabilidade *a posteriori* obtida com o algoritmo SM-MCMC (0.5770 contra 0.9874). Ou seja, o algoritmo SM-MCMC identificou com maior probabilidade *a posteriori* o modelo utilizado para a geração dos dados. Os intervalos de 95% credibilidade para os parâmetros,

obtidos com o SM-MCMC, possuem uma amplitude menor do que os obtidos com o RJ-MCMC. As estimativas para os parâmetros são mais próximas dos verdadeiros valores utilizados para geração dos dados em relação as estimativas obtidas com o algoritmo RJ-MCMC.

Ao contrário do algoritmo SM-MCMC, o diagnóstico de Gelman-Rubin para os parâmetros μ_1 e σ_1 indicam a não obtenção de convergência (ver Tabela 4.12). O Apêndice D.1 (pag.83) mostra o gráfico ergódico para os parâmetros. Em relação a classificação das observações entre as componentes identificadas os métodos apresentam resultados similares, como pode ser observado na Figura 4.3 que mostra os grupos gerados e os grupos idenficados pelos algoritmos SM-MCMC e RJ-MCMC.

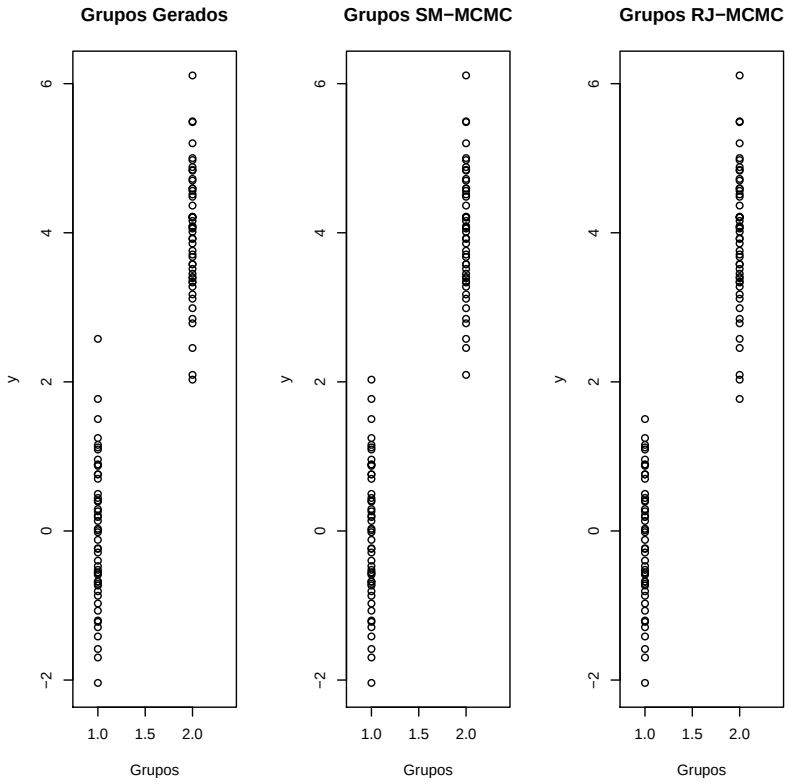

Figura 4.3: Grupos gerados e grupos identificados pelos algoritmos SM-MCMC e RJ-MCMC - Dados artificiais 1.

4.3.2 Dados artificiais 2

Os resultados obtidos para o segundo conjundo de dados artificiais, estão apresentados nas Tabelas 4.13 e 4.14 que mostram, respectivamente, a probabilidade *a posteriori* para k e as estimativas para os parâmetros.

Ao contrário do algoritmo SM-MCMC que identifica o modelo com $k = 3$ com maior probabilidade *a posteriori* ($P(k = 3|\cdot) = 0.9154$), o algoritmo RJ-MCMC identifica o modelo com $k = 2$ com maior probabilidade *a posteriori* ($P(k = 2|\cdot) = 0.4287$). Isto ocorre devido ao algoritmo RJ-MCMC não conseguir separar as observações geradas das componentes 2 e 3 em duas componentes, como pode ser observado na Figura 4.5 que mostra os grupos gerados e os grupos idenficados pelos algoritmos SM-MCMC e RJ-MCMC. Isto é, o algoritmo RJ-MCMC com as funções de transição em (3.20), (3.21) e (3.22) não consegue propor um movimento *split* com parâmetros de forma adequada para que haja a separação das observações geradas das componentes 2 e 3 em duas componentes. Isto poderia ser evitado se utilizássemos uma função de transição que fosse capaz de propor parâmetros que levassem a uma separação das observações geradas das componentes 2 e 3. Logo, a escolha da função de transição é um importante aspecto para que o algoritmo RJ-MCMC tenha um bom desempenho em modelos com mistura.

O diagnóstico de Gelman e Rubin para σ_1, σ_2, w_1 e w_2 são ≤ 1.05 indicando a convergência, porém, para μ_1, μ_2 o diagnóstico é 1.12 e 1.10, respectivamente. O Apêndice D.2 (pag.84) mostra o gráfico ergódico dos valores gerados para os parâmetros.

Comparando com os resultados obtidos com o algoritmo SM-MCMC proposto, temos um melhor desempenho do SM-MCMC, pois este identifica o verdadeiro modelo com maior probabilidade *a posteriori* (Tabela 4.3), as estimativas para os parâmetros (Tabela 4.4) e a classificação das observações em relação às componentes (Figura 4.5) são satisfatórias. Ou seja, nossa estratégia de buscar informações sobre as k componentes e seus parâmetros diretamente nos dados observados é mais adequada do que propor o surgimento de novas componentes e novos parâmetros, através de uma função de transição, sem se preocupar se esta nova componente

determina uma partição nos dados observados.

Tabela 4.13: Probabilidade *a posteriori* para k.

k	1	2	3	4	5	≥ 6
$P(k\|\cdot)$	0.0552	0.4287	0.2778	0.1437	0.0606	0.0340

Tabela 4.14: Estimativas para os parâmetros.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. G. R.
μ_1	-1.5012	(-4.8640, 0.4122)	1.12
μ_2	0.7799	(-0.2208, 4.0244)	1.10
σ_1	1.5078	(0.1672, 3.6548)	1.04
σ_2	2.1648	(0.1858, 3.7081)	1.01
w_1	0.3781	(0.0196, 0.9706)	1.01
w_2	0.6219	(0.0294, 0.9804)	1.01

4.3.3 Dados artificiais 3

As Tabelas 4.15 e 4.16 apresentam os resultados para o terceiro conjunto de dados artificiais. O máximo *a posteriori* é obtido em $\widetilde{k} = 2$, $P(k = 2|\cdot) = 0.6274$. Ou seja, para este exemplo o algoritmo RJ-MCMC também não apresenta um bom desempenho em identificar as componentes da mistura. O RJ-MCMC não consegue identificar nenhuma componente em relação aos dados gerados, ou seja, o algoritmo não propõe novas componentes com parâmetros adequados para que haja uma satisfatória partição nos dados gerados.

Novamente, o algoritmo SM-MCMC proposto apresenta melhor desempenho, pois identifica o verdadeiro modelo com maior probabilidade *a posteriori* (Tabela 4.5), as estimativas para os parâmetros (Tabela 4.6) e a classificação das observações

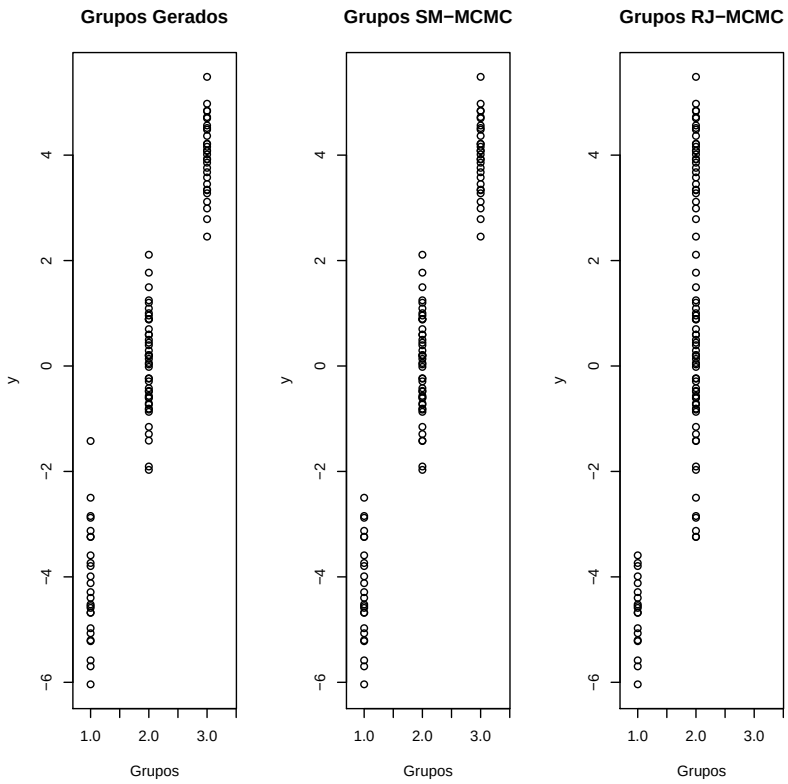

Figura 4.4: Grupos gerados e grupos identificados pelos algoritmos SM-MCMC e RJ-MCMC - Dados artificiais 2.

em relação às componentes (Figura 4.5) são satisfatórias. Aumentando o número de componentes da mistura, de $k = 2$ para $k = 3$ e para $k = 5$, pior é o desempenho do algoritmo RJ-MCMC. Já o algoritmo SM-MCMC identifica os modelos utilizados para gerar os dados com maior probabilidade *a posteriori*.

Tabela 4.15: Probabilidade *a posteriori* para k.

k	1	2	3	4	5	≥ 6
$P(k\|\cdot)$	0.0072	0.6274	0.1785	0.0796	0.0453	0.0620

Tabela 4.16: Estimativas para os parâmetros.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. G. R.
μ_1	-1.2913	(-5.7930, 0.3204)	1.16
μ_2	0.3159	(-0.4920, 2.1995)	1.13
σ_1	7.9905	(0.2374, 11.8833)	1.10
σ_2	2.7926	(0.1633, 11.3527)	1.08
w_1	0.7091	(0.0588, 0.9510)	1.05
w_2	0.2908	(0.0490, 0.9412)	1.06

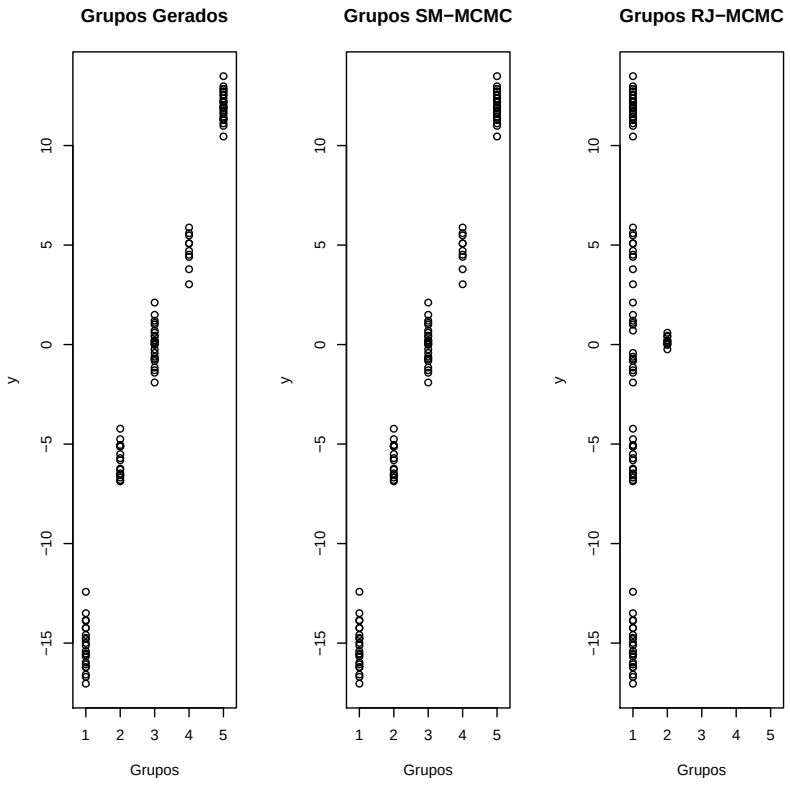

Figura 4.5: Grupos gerados e grupos identificados pelos algoritmos SM-MCMC e RJ-MCMC - Dados artificiais 3.

4.3.4 Dados de velocidades de galáxias

Para os dados de velocidade de galáxias, as probabilidades *a posteriori* para k são mostradas na Tabela 4.17. O máximo *a posteriori* é obtido em $k = 2$ com probabili-

dade *a posteriori* $P(k = 2|\cdot) = 0.5555$. A Tabela 4.18 mostra as estimativas para os parâmetros. A Figura 5.1 mostra o histograma do conjunto de dados e a densidade estimada pelo RJ-MCMC. O Apêndice D.4 (pag.86) mostra o gráfico ergódico dos valores gerados para os parâmetros.

Comparando as Figuras 4.2 e 4.6, densidades estimadas pelo SM-MCMC e RJ-MCMC, respectivamente, podemos notar um melhor desempenho do algoritmo SM-MCMC, pois a densidade estimada ajusta melhor os dados observados. Também, no SM-MCMC temos convergêcia, segundo o diagnóstico de Gelman e Rubin, para todos os parâmetros, o que não ocorre no RJ-MCMC, como pode ser observado pelo diagnóstico de Gelman e Rubin na Tabela 4.18.

Tabela 4.17: Probabilidade *a posteriori* para k.

k	1	2	3	4	5	≥ 6
$P(k\|\cdot)$	0.0480	0.5555	0.3212	0.0647	0.0093	0.0014

Tabela 4.18: Estimativas para os parâmetros.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. G. R.
μ_1	14.8141	(-3.2257, 21.4359)	1.25
μ_2	21.8448	(20.3359, 23.0560)	1.14
σ_1	5.1486	(0.2555, 11.1933)	1.06
σ_2	3.1589	(1.5256, 9.2737)	1.02
w_1	0.2584	(0.0119, 0.7857)	1.03
w_2	0.8267	(0.2024, 0.9881)	1.30

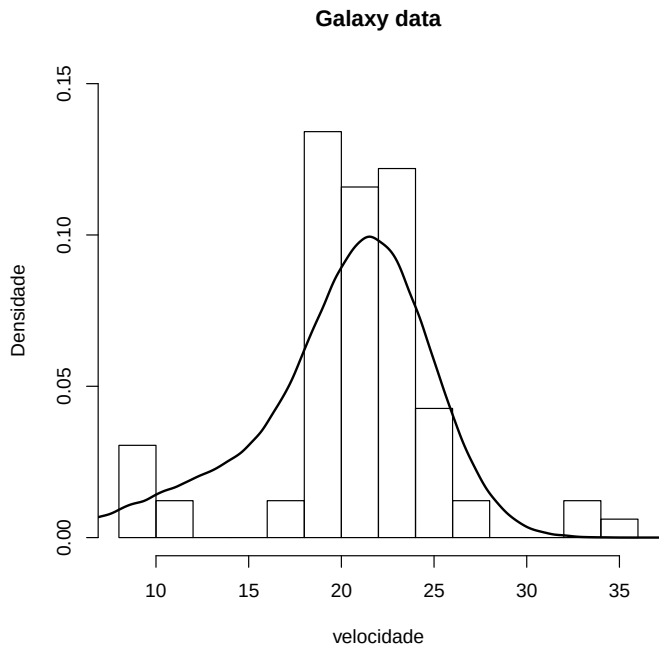

Figura 4.6: Histograma dos dados sobre a velocidade de galáxias e densidade estimada pelo algoritmo RJ-MCMC.

4.3.5 Dados de expressão gênica

As Tabelas 4.19 e 4.20 e a Figura 4.7 mostram os resultados obtidos com a aplicação do algoritmo RJ-MCMC aos dados de expressão gênica. Na Tabela 4.19 estão as probabilidades *a posteriori* para o número de componentes k, com um máximo em $\widetilde{k} = 3$, $P(k = 3|\cdot) = 0.7389$. A Tabela 4.20 mostra as estimativas para os parâmetros e os intervalos de credibilidade para os parâmetros das três componentes identificadas. Ao contrário do SM-MCMC, onde o diagnóstico de Gelman e Rubin é ≤ 1.01 para todos os parâmetros, o diagnóstico de Gelman e Rubin para os parâmetros μ_j e σ_j no RJ-MCMC são > 1.05. O Apêndice D.5 (pag.87) mostra o gráfico ergódico dos valores gerados para os parâmetros.

A Figura 4.7 mostra a média de controle *versus* a média de tratamento dos 434 genes e os 3 grupos identificados. Os triângulos $\triangledown$ representam o grupo G_1, $\bullet$ representam o grupo G_2 e os triângulos $\triangle$ representam o grupo G_3. O grupo G_1 é composto por 50 genes, o grupo G_2 por 370 genes e o grupo G_3 por 14 genes.

Em relação ao algoritmo SM-MCMC, a probabilidade *a posteriori* obtida para $k = 3$ no algoritmo RJ-MCMC é menor do que a obtida com o algoritmo SM-MCMC (0.7389 contra 0.9111). As estimativas para os parâmetros são similares. Logo, para este conjunto de dados temos que os algoritmos SM-MCMC e RJ-MCMC apresentam resultados similares.

O grupo G_1 no SM-MCMC apresenta 11 genes a mais do que o grupo G_1 no RJ-MCMC. Estes genes são os que apresentam efeito de tratamento y na fronteira dos grupos G_1 e G_2 e estão em destaque, "pontos" maiores em negrito, na Figura 4.7.

Tabela 4.19: Probabilidade *a posteriori* k.

k	1	2	3	4	5	≥ 6
$P(k\|\cdot)$	0.0013	0.1598	0.7389	0.0905	0.0090	0.0006

Tabela 4.20: Estimativas para os parâmetros.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. G. R.
μ_1	-0.2744	(-0.5409, -0.0199)	1.12
μ_2	0.0384	(-0.0028, 0.0700)	1.06
μ_3	0.6475	(0.1249, 1.3016)	1.06
σ_1	0.2358	(0.1325, 0.3528)	1.07
σ_2	0.1654	(0.1200, 0.2357)	1.06
σ_3	0.6183	(0.3722, 0.9367)	1.25
w_1	0.2201	(0.0320, 0.5927)	1.02
w_2	0.7229	(0.0160, 0.9291)	1.02
w_3	0.0571	(0.0160, 0.1053)	1.01

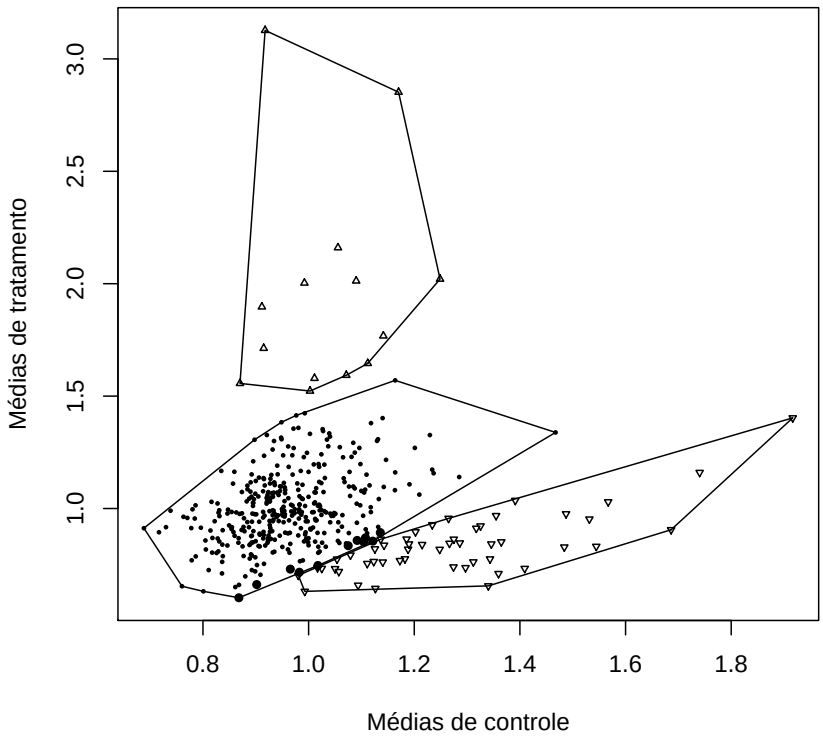

Figura 4.7: Médias de controle *versus* médias de Tratamento e grupos identificados. $\triangledown$ representam o grupo G_1, $\bullet$ o grupo G_2 e $\triangle$ o grupo G_3, RJ-MCMC.

Capítulo 5

Algoritmo *Birth-Split-Merge* MCMC

Neste capítulo, propomos o algoritmo *birth-split-merge*, BSM-MCMC, em que, além dos movimentos *split-merge* descritos no capítulo anterior, consideramos que ao atualizar uma variável indicadora c_i esta é capaz de determinar o "nascimento" (*birth*) de uma nova componente.

Para isto, considere a abordagem Bayesiana descrita no Capítulo 3, com distribuição *a priori* para $\boldsymbol{\phi}$ dada como em (3.1, pag.12) e para $\mathbf{w}$ a distribuição *a priori* de Dirichlet, porém agora ao invez de parâmetros γ, considere parâmetros $\frac{\gamma}{k}$,

$$(w_1, \ldots, w_k)|\gamma, k \sim Dirichlet\left(\frac{\gamma}{k}, \ldots, \frac{\gamma}{k}\right). \tag{5.1}$$

Assim, de (4.3, pag.26), a probabilidade *a priori* conjunta para $\mathbf{c}$ dado k, é dada por

$$\pi(\mathbf{c}|k) \;=\; \frac{\Gamma(\gamma)}{\left[\Gamma\left(\frac{\gamma}{k}\right)\right]^k} \frac{\prod_{j=1}^{k} \Gamma\left(n_j + \frac{\gamma}{k}\right)}{\Gamma(n+\gamma)}. \tag{5.2}$$

De (4.8, pag.28), a probabilidade condicional *a priori* para c_i dado $\mathbf{c}_{-i}$ é

$$P(c_i = j|\mathbf{c}_{-i}, k) = \frac{n_{j,-i} + \frac{\gamma}{k}}{\gamma + n - 1}, \tag{5.3}$$

onde $n_{j,-i}$ é o número de observações pertencentes a S_j, exceto a observação y_i, $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

5.1 Limite $k \to \infty$

Para obter o movimento *birth* de forma direta, exploramos o limite $k \to \infty$ e deduzimos as probabilidades condicionais *a posteriori* para as variáveis latentes $\mathbf{c} = (c_1, \ldots, c_n)$. Com $k \to \infty$ temos um modelo equivalente ao modelo de misturas de processos de Dirichlet (Ferguson, 1973; Antoniak , 1974; Escobar e West, 1995; Neal, 1998).

Note que para $k \to \infty$, dada a amostra $\mathbf{y} = (y_1, \ldots, y_n)$ e uma configuração $\mathbf{c} = (c_1 \ldots, c_n)$ para as variáveis indicadoras, temos uma quantidade k^+ de componentes com observações associadas, $n_j > 0$, e uma infinidade de componentes vazias, $n_j = 0$. Assim, consideramos que a distribuição *a posteriori* para $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{k^+})$, dada em (3.5, pag.12), é obtida substituindo k pelo número de componentes k^+ que tem observações associadas.

Considerando (5.3) com $k \to \infty$, a probabilidade condicional *a priori* de $c_i = j$, para uma componente j com $n_{j,-i} > 0$, é dada por

$$P(c_i = j|\mathbf{c}_{-i}) = \frac{n_{j,-i}}{\gamma + n - 1} \tag{5.4}$$

e a probabilidade condicional *a priori* de $c_i \neq c_{i'},\ \forall\ i \neq i', i, i' = 1 \ldots, n$, é dada por

$$P(c_i \neq c_{i'},\ \forall\ i \neq i'|\mathbf{c}_{-i}) = \frac{\gamma}{\gamma + n - 1}. \tag{5.5}$$

Note que (5.4) é a probabilidade condicional *a priori* da observação y_i pertencer à componente j, que é composta por pelo menos uma outra observação, $n_{j,-i} > 0$, e (5.5) é a probabilidade condicional *a priori* da observação y_i determinar o "nascimento" (*birth*) de uma nova componente. Estas probabilidades são proporcionais ao número de observações pertencentes à componente j (exceto a observação y_i) e ao hiperparâmetro α, respectivamente.

Para atualizar as probabilidades *a priori* em (5.4) e (5.5) via função de verossimilhança, consideramos a função de verossimilhança para uma única observação y_i quando $c_i = j$, para algum j com $n_{j,-i} > 0$, e quando $c_i \neq c_{i'}, \forall i \neq i',\ i, i' = 1, \ldots, n$. Se $c_i = j$, para algum j com $n_{j,-i} > 0$, então as observações y_i pertencentes à componente j, $y_i \in S_j$, são modeladas pela densidade $f(y_i|\phi_j)$. Caso contrário, se

$c_i \neq c_{i'}, \forall i \neq i'$, consideramos que a função de verossimilhança para as componentes não representadas (componentes com $n_{j,-i} = 0$), é dada pela função de verossimilhança marginal,

$$q_0(y_i) = \int f(y_i|\phi_j)\pi(\phi_j)d\phi_j, \tag{5.6}$$

para algum j com $n_{j,-i} = 0$, $i = 1, \ldots, n$ e $j = 1, \ldots, k \to \infty$.

Assim, as probabilidades condicionais *a posteriori*, são dadas por

$$P(c_i = j|\mathbf{c}_{-i}, \phi_j, y_i) = b\frac{n_{j,-i}}{\gamma + n - 1}f(y_i|\phi_j) \tag{5.7}$$

e

$$P(c_i \neq c_{i'},\ \forall\ i \neq i'|\mathbf{c}_{-i}, y_i) = b\frac{\gamma}{\gamma + n - 1}q_0(y_i). \tag{5.8}$$

onde

$$b = \left(\gamma q_0(y_i) + \sum_{j;n_{j,-i}>0} n_{j,-i}f(y_i|\phi_j)\right)^{-1}$$

é a constante normalizadora, $i, i' = 1, \ldots, n$ e $i \neq i'$.

Note que (5.7) é a probabilidade condicional *a posteriori* da observação y_i pertencer à componente j, com $n_{j,-i} > 0$, e (5.8) é a probabilidade condicional *a posteriori* da observação y_i determinar o "nascimento" de uma nova componente. Neste último caso, uma densidade $f(\cdot|\phi_j)$, com ϕ_j gerado da distribuição *a posteriori* $\pi(\phi_j|y_i)$, é associada a esta nova componente. Esta última probabilidade define o movimento *birth* do algoritmo proposto neste capítulo.

Obtida uma configuração para as variáveis latentes $\mathbf{c}$, via probabilidades (5.7) e (5.8), temos uma partição dos dados $\mathbf{y}$ em k^+ componentes não vazias. A função de verossimilhança é definida pelas componentes que tem observações associadas,

$$L(\boldsymbol{\phi}, \mathbf{c}|\mathbf{y}, k^+) = \prod_{S_j \neq \emptyset} L(\phi_j|S_j), \tag{5.9}$$

onde $L(\phi_j|S_j) = \prod_{S_j} f(y_i|S_j)$ é a função de verossimilhança para a componente j, com $n_j > 0$.

Atualizando a distribuição *a priori* conjunta para $(\boldsymbol{\theta}, \mathbf{c})$, agora com $k \to \infty$, via função de verossimilhnaça (5.9), a distribuição *a posteriori* é dada por

$$\pi(\boldsymbol{\phi}, \mathbf{c}|\mathbf{y}, k^+) \quad \propto \quad L(\boldsymbol{\phi}, \mathbf{c}|\mathbf{y}, k^+)\pi(\boldsymbol{\phi})\pi(\mathbf{c}). \tag{5.10}$$

5.2 Algoritmo *birth-split-merge* MCMC

Utilizando as probabilidades em (5.7) e (5.8) MacEachern (1994), MacEachern, Clyde e Liu (1999), Neal (1998) e Medvedovic e Sivaganesan (2002) propõem o uso do algoritmo *Gibbs sampling* para atualizar as variáveis indicadoras. Porém, alguns autores, tais como Celeux, Hurn and Robert (2000) e Jain e Neal (2004), informam que este algoritmo pode ser ineficiente em situações onde, por exemplo, existam duas componentes com médias próximas. Pois o algoritmo pode considerar estas duas componentes como uma única componente, com parâmetros que são uma média entre os parâmetros das duas componentes. Isto ocorre, devido às probabilidades (5.7) e (5.8) atualizarem somente uma variável indicadora por vez.

Para evitar este problema e permitir uma maior mudança na configuração das variáveis indicadoras em uma única iteração do algoritmo, propomos o uso dos movimentos *split-merge* descritos no capítulo anterior.

5.2.1 Propostas *split-merge*

Como para $k \to \infty$, o número de componentes não pode ser representada explicitamente, então somente para desenvolver o procedimento de simulação e descrever o algoritmo proposto, fixamos um valor máximo k_{max} para k (Neal, 1998). Como, dada uma amostra $\mathbf{y} = (y_1, \ldots, y_n)$ de tamanho n, o número máximo de componentes não vazias é $k^+ = n$, fixamos $k_{max} = n$. Assim, para cada iteração l, $l = 1 \ldots, L$ do algoritmo descrito abaixo, a variável latente $c_i = j$ para $j \in \{1, \ldots, k_{max}\}$ com $n_j > 0$ ou $c_g = k_{max} + 1$ se a observação y_i determina o nascimento de uma nova componente.

As propostas *split-merge* são as descritas no Capítulo 4, porém, agora o estado atual é $\boldsymbol{\theta} = (\boldsymbol{\phi}, \mathbf{c})$ com k^+ componentes não vazias. A probabilidade de escolher entre o movimento *split* ou *merge* é

$$P_{sp|k^+} = \begin{cases} 1, \text{ se } k^+ = 1 \\ 0.5, \text{ se } 2 \leq k^+ \leq k_{max} - 1 \\ 0, \text{ se } k^+ = k_{max} \end{cases},$$

e

$$P_{mg|k^+} = \begin{cases} 0, \text{ se } k^+ = 1 \\ 0.5, \text{ se } 2 \leq k^+ \leq k_{max} - 1 \\ 1, \text{ se } k^+ = k_{max} \end{cases},$$

respectivamente.

A proposta *split* determina a configuração $\boldsymbol{\theta}^{sp} = (\boldsymbol{\phi}^{sp}, \mathbf{c}^{sp})$, com $k^{sp^+} = k^+ + 1$ componentes não vazias. Os passos para propor o movimento *split* são como os descrito no Capítulo 4 (pag.29), porém agora com $k \to \infty$, a probabilidade de alocação é dada por

$$P_{j_1}(y_v) = \frac{n_{j_1} f(y_v|\phi^*_{j_1})}{n_{j_1} f(y_v|\phi^*_{j_1}) + n_{j_2} f(y_v|\phi^*_{j_2})}. \tag{5.11}$$

A proposta *merge* determina a configuração $\boldsymbol{\theta}^{mg} = (\boldsymbol{\phi}^{mg}, \mathbf{c}^{mg})$, com $k^{mg^+} = k^+ - 1$ componentes não vazias. A probabilidade de selecionar as componentes j_1 e j_2 para um *merge* é dada em (4.11, pag.30).

Estas propostas são aceitas de acordo com a probabilidade de aceitação *Metropolis-Hastings* $\alpha[\boldsymbol{\theta}^*|\boldsymbol{\theta}] = min(1, A)$, onde

$$A = \frac{L(\boldsymbol{\theta}^*|\mathbf{y}, k^*)}{L(\boldsymbol{\theta}|\mathbf{y}, k^+)} \frac{\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta})} \frac{\pi(\mathbf{c}^*)}{\pi(\mathbf{c})} \frac{q[\boldsymbol{\theta}|\boldsymbol{\theta}^*]}{q[\boldsymbol{\theta}^*|\boldsymbol{\theta}]} \tag{5.12}$$

em que, $q[\cdot|\cdot]$ são as propostas de transição *split-merge*, dadas em (4.10 e 4.12, pag.29 e 30, respectivamente) substituindo k por k^+ e o sinal $*$ é sp^+ ou mg^+.

Considerando a propostas *split*, a razão das funções de verossimilhanças é dada por

$$\frac{L(\boldsymbol{\theta}^{sp}|\mathbf{y}, \widetilde{k}^{sp^+})}{L(\boldsymbol{\theta}|\mathbf{y}, k^{sp^+})} = \frac{L(\phi_{j_1}|S_{j_1}) L(\phi_{j_2}|S_{j_2})}{L(\phi_j|S_j)}. \tag{5.13}$$

e a razão das distribuições *a priori*, com $k \to \infty$, é dada por

$$\frac{\pi(\boldsymbol{\theta}^{sp})}{\pi(\boldsymbol{\theta})} = \frac{\pi(\phi_{j_1})\pi(\phi_{j_2})}{\pi(\phi_j)} \frac{\Gamma(n_{j_1})\Gamma(n_{j_2})}{\Gamma(n_j)}. \tag{5.14}$$

A probabilidade de aceitação para a proposta *split* é $\alpha[\boldsymbol{\theta}^{sp^+}|\boldsymbol{\theta}] = min(1, A^{sp})$,

$$A^{sp} = \frac{I_{j_1} I_{j_2}}{I_j} \frac{\Gamma(n_{j_1})\Gamma(n_{j_2})}{\Gamma(n_j)} Q^{sp^+}, \tag{5.15}$$

onde $I_d = \int L(\phi_d|S_d)\pi(\phi_d)d\phi_d$, $d \in \{j, j_1, j_2\}$, e Q^{sp^+} é dada em (4.16, pag.31), substituindo k por k^+.

A probabilidade de aceitação para a proposta *merge* é $\alpha[\boldsymbol{\theta}^{mg^+}|\boldsymbol{\theta}] = min(1, A^{mg})$ onde

$$A^{mg} = \frac{I_j}{I_{j_1} I_{j_2}} \frac{\Gamma(n_j)}{\Gamma(n_{j_1})\Gamma(n_{j_2})} Q^{mg^+}, \tag{5.16}$$

onde Q^{mg^+} é dada em (4.18, pag.31), substituindo k por k^+.

5.2.2 Comentários

Para garantir identificabilidade para o modelo, utilizamos a marcação das componentes não vazias de acordo com a condição de adjacência $\mu_1 < \mu_2 < \ldots < \mu_{k^+}$. Logo, esta deve ser verificada para as propostas *split-merge*.

Para garantir a adjacência ao propor um movimento *birth*, fazemos a seguinte restrição: se na $(l-1)$-ésima iteração do algoritmo $c_i^{(l-1)} = j$ e na l-ésima iteração $c_i^{(l)} = k_{max} + 1$ (proposta de nascimento de uma nova componente), verifique se o valor gerado para a média $\mu_{k_{max}+1}$ dessa nova componente satifaz a condição $\mu_{k_{max}+1} > \mu_j$, para todo $j \in \{1, \ldots, k^+\}$. Se satisfaz, aceite a proposta *birth* e faça a marcação $c_i^{(l)} = k^+ + 1$ e $\mu_{k^+ + 1} = \mu_{k_{max}+1}$. Caso não satisfaça, mantenha $c_i^{(l)} = c_i^{(l-1)}$.

Sempre que o número de observações pertencentes a uma componente cai para "zero", esta componente é retirada do estado atual. Denominamos este movimento por *death*, devido este ser o movimento inverso a um movimento *birth*.

5.2.3 Algoritmo

Definidas as probabilidades dos movimentos *birth-split-merge* e as probabilidades de aceitação para os movimentos *split* e *merge*, expressamos o método proposto como um algoritmo.

Algoritmo BSM-MCMC:

(1) Inicialize todos os parâmetros;

(2) Para l-ésima iteração, $l = 1, \ldots, L$, faça:

(i) Escolha entre *split* ou *merge* com probabilidades $P_{sp|k^+}$ e $P_{mg|k^+}$, respectivamente;

(ii) Aceite a proposta com probabilidade $\alpha[(\boldsymbol{\theta}^*, \mathbf{c}^*)|(\boldsymbol{\theta}, \mathbf{c})]$, onde o sinal $*$ é entre sp^+ ou mg^+;

(iii) Atualize $\mathbf{c}$ utilizando as probabilidades condicionais *a posteriori* em (5.7) e (5.8):

(a) se $c_i^{(l)} = j$, para algum $j \in \{1, \ldots, k_{max}\}$ com $n_{j,-i} > 0$, então a observação y_i pertence a componente j. Faça $n_j = n_{j,-i} + 1$;

(b) se $c_i^{(l)} = k_{max}+1$, gere $\phi_{k_{max}+1}$ da distribuição *a posteriori* $\pi(\phi_{k_{max}+1}|y_i)$ e verifique se a condição $\mu_{k_{max}+1} > \mu_j$, para todo $j \in \{1, \ldots, k^+\}$, é satisfeita. Se satisfaz, faça $k^{+(l)} = k^{+(l-1)} + 1$, $c_i^{(l)} = k^{+(l)}$, $n_{k^{+(l)}} = 1$ e $\mu_{k^{+(l)}} = \mu_{k_{max}+1}$. Se não satisfaz, mantenha $c_i^{(l)} = c_i^{(l-1)}$;

(iv) Atualize os parâmetros $\boldsymbol{\phi}$ e $\mathbf{w}$:

(a) gere um valor para ϕ_j de sua distribuição *a posteriori* $\pi(\phi_j|S_j)$;

(b) considere w_j como sendo o valor esperado $E[w_j|\mathbf{c}] = \frac{n_j}{n}$,

para $j = 1, \ldots, k^+$;

Ao final das L iterações, consideramos *burn in* B e calculamos o número de vezes $N_{k^+=j}$ que $k^+ = j$ nas $L - B$ iterações, $j = 1, \ldots, k_{max}$. Definindo $P_{k^+=j} = \frac{N_{k^+=j}}{L-B}$ como sendo a probabilidade *a posteriori* para $k^+ = j$, $j = 1, \ldots, k_{max}$, então $\widetilde{k} = \underset{1 \le j \le k_{max}}{\arg\max}(P_{k^+=j})$ é a estimativa para o número de componentes.

Dado $\widetilde{k}$, a estimativa para os parâmetros é dada por

$$\widetilde{\phi}_j|\widetilde{k} = \frac{1}{L_{\widetilde{k}}} \sum_{l=1}^{L_{\widetilde{k}}} \phi_j^{(l)} \text{ and } \widetilde{w}_j|\widetilde{k} = \frac{1}{L_{\widetilde{k}}} \sum_{l=1}^{L_{\widetilde{k}}} w_j^{(l)}, \tag{5.17}$$

onde $L_{\widetilde{k}}$ é o número de iterações, tal que, $k^+ = \widetilde{k}$.

A probabilidade *a posteriori* de cada observação ser proveniente da componente j é $P_{ij} = \frac{N_{ij}}{L_{\widetilde{k}}}$, onde N_{ij} é o número de vezes que a observação y_i é associada à componente $j \in \{1, \ldots, \widetilde{k}\}$ nas $L_{\widetilde{k}}$ iterações. Se $P_{ij} = \underset{1 \le j \le \widetilde{k}}{max}(P_{ij})$, então consideramos que y_i é proveniente da componente j, $i = 1, \ldots, n$ e $j = 1, \ldots, \widetilde{k}$.

5.3 Análise de dados

Nesta seção aplicamos o algoritmo BSM-MCMC aos cinco conjuntos de dados utilizados no Capítulo 4.

Utilizamos a mesma modelagem com mistura de distribuições normais e as mesmas distribuições *a priori*. Os hiperparâmetros são os mesmos, $\lambda = 0.1$, $\tau = 0.1$, $\nu = 0.1$ e $\gamma = 1$. Para verificar a obtenção de convergência, geramos duas cadeias e utilizamos o diagnóstico de Gelman e Rubin (Gelman e Rubin, 1992).

Aplicamos o algoritmo BSM-MCMC com $L = 100.000$ e um *burn in* $B = 10.000$. Este número de iterações foi considerado suficiente pois para os 5 conjuntos de dados utilizados o diagnóstico de Gelman-Rubin para todos os parâmetros são $\leq$ 1.01, indicando a obtenção de convergência. Iniciamos o algoritmo com uma única componente, como feito no algoritmo SM-MCMC e RJ-MCMC.

5.3.1 Dados artificiais 1

Os resultados obtidos para o primeiro conjunto de dados simulados estão apresentados nas Tabelas 5.1 e 5.2.

Na Tabela 5.1 estão as probabilidades *a posteriori* para k e na Tabela 5.2 estão as estimativas para os parâmetros e os intervalos empíricos de 95% de credibilidade.

Os resultados obtidos, pelos dois algoritmos propostos, SM-MCMC e BSM-MCMC, para a probabilidade *a posteriori* do número de componentes k e as estimativas para os parâmetros são similares, como pode ser obsevado nas Tabelas 5.1, 5.2 e 4.1 e 4.2. O Apêndice E.1 (pag.88) mostra o gráfico ergódico dos valores gerados para os parâmetros, indicando a convergência. Como no algoritmo SM-MCMC, a probabilidade *a posteriori* para k no algoritmo BSM-MCMC é maior do que a obtida com o algoritmo RJ-MCMC. As estimativas dos parâmetros são mais próximas dos verdadeiros valores utilizados para a geração dos dados e os intervalos de credibilidade possuem uma amplitude menor.

Tabela 5.1: Probabilidade *a posteriori* para k.

k	1	2	3	≥ 4
$P(k\|\cdot)$	0.0004	0.9494	0.0485	0.0017

Tabela 5.2: Estimativas para os parâmetros.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. G. R.
μ_1	-0.0385	(-0.3327, 0.2864)	1.01
μ_2	3.9824	(3.6746, 4.2593)	1.01
σ_1	0.9269	(0.5596, 1.5375)	1.01
σ_2	0.7985	(0.4776, 1.3380)	1.01
w_1	0.5086	(0.4706, 0.5490)	1.01
w_2	0.4912	(0.4510, 0.5294)	1.01

5.3.2 Dados artificiais 2

Para o segundo conjundo de dados artificiais, os resultados obtidos estão apresentados nas Tabelas 5.3 e 5.4 que mostram, respectivamente, a probabilidade *a posteriori* para k e as estimativas para os parâmetros.

Os dois algoritmos propostos identificam o modelo com $k = 3$ com maior probabilidade *a posteriori* (Tabelas 4.3 e Tabela 5.3), ao contrário do algoritmo RJ-MCMC que encontra o modelo com $k = 2$ com maior probabilidade *a posteriori* (Tabela 4.13). Ou seja, o algoritmo BSM-MCMC também apresenta um melhor desempenho em relação ao algoritmo RJ-MCMCM neste conjunto de dados. As estimativas e os intervalos de 95% de credibilidade obtidos são similares aos do algoritmo SM-MCMC. O Apêndice E.2 (pag.89) mostra o gráfico ergódico dos valores gerados para os parâmetros, indicando a convergência.

Tabela 5.3: Probabilidade *a posteriori* para k.

k	1	2	3	≥ 4
$P(k\|\cdot)$	0.0488	0.0914	0.8183	0.0415

Tabela 5.4: Estimativas para os parâmetros.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. G. R.
μ_1	-4.1361	(-4.6822, -3.3476)	1.01
μ_2	0.0556	(-0.3267, 0.4380)	1
μ_3	3.9695	(3.6300, 4.2783)	1
σ_1	1.1000	(0.7404, 1.7789)	1.01
σ_2	0.9945	(0.6881, 1.5149)	1
σ_3	0.7588	(0.5533, 1.0628)	1.01
w_1	0.2614	(0.2136, 0.3301)	1
w_2	0.4403	(0.2621, 0.5146)	1
w_3	0.2982	(0.2621, 0.3301)	1

5.3.3 Dados artificiais 3

Os resultados para o terceiro conjunto de dados artificiais estão apresentados nas Tabelas 5.5 e 5.6 que mostram, respectivamente, a probabilidade *a posteriori* para k e as estimativas para os parâmetros. O Apêndice E.3 (pag.90) mostra o gráfico ergódico dos valores gerados para os parâmetros, indicando a covergência.

Como no algoritmo SM-MCMC o máximo *a posteriori* é obtido em $k = 5$ que tem probabilidade $P(k = 5|\cdot) = 0.8773$. Ao contrário do algoritmo RJ-MCMC que identifica o modelo com $k = 2$ com maior probabilidade *a posteriori*. As estimativas e os intervalos de 95% de credibilidade obtidos são similares aos do algoritmo SM-MCMC. Ou seja, os algoritmos propostos apresentam resultados similares.

Tabela 5.5: Probabilidade *a posteriori* para k.

k	1	2	3	4	5	≥ 6
$P(k\|\cdot)$	0.0001	0.0019	0.0126	0.0895	0.8773	0.0186

Tabela 5.6: Estimativas para os parâmetros.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. G. R.
μ_1	-15.0832	(-15.7389, -14.4271)	1
μ_2	-5.7799	(-6.2319, -5.3238)	1
μ_3	0.0011	(-0.4279, 0.4484)	1
μ_4	4.5571	(2.8331, 5.3637)	1.01
μ_5	11.9651	(11.5645, 12.3685)	1.01
σ_1	1.5400	(1.1599, 2.0901)	1
σ_2	0.9250	(0.6657, 1.3310)	1.01
σ_3	0.9608	(0.6570, 1.3765)	1
σ_4	1.1306	(0.6156, 2.5053)	1
σ_5	1.0325	(0.7916, 1.3660)	1
w_1	0.2190	(0.2190, 0.2190)	1.01
w_2	0.1713	(0.1810, 0.1714)	1.01
w_3	0.2417	(0.1810, 0.2571)	1
w_4	0.1109	(0.0952, 0.1714)	1.01
w_5	0.2569	(0.2571, 0.2571)	1.01

5.3.4 Dados de velocidades de galáxias

Para os dados de velocidade de galáxias, as probabilidades *a posteriori* para k são mostradas na Tabela 5.7. Como no algoritmo SM-MCMC o máximo *a posteriori* é obtido em $k = 3$ também com alta probabilidade *a posteriori*, $P(k = 3|\cdot) = 0.9798$.

A Tabela 5.8 mostra as estimativas para os parâmetros. Os valores obtidos são similares aos do algoritmo SM-MCMC (ver Tabela 4.8). O Apêndice E.4 (pag.91) mostra o gráfico ergódico dos valores gerados para os parâmetros, indicando a convergência. A Figura 5.1 mostra o histograma do conjunto de dados e a densidade estimada pelo BSM-MCMC.

Tabela 5.7: Probabilidade *a posteriori* para k.

k	1	2	3	≥ 4
$P(k\|\cdot)$	0.0001	0.0013	0.9798	0.0189

Tabela 5.8: Estimativas para os parâmetros.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. G. R.
μ_1	9.6948	(9.2315, 10.1578)	1
μ_2	21.3629	(20.8266, 21.8918)	1
μ_3	31.9979	(25.0482, 35.6158)	1
σ_1	0.5884	(0.3601, 1.0110)	1.01
σ_2	2.1958	(1.8147, 2.6092)	1
σ_3	2.8201	(1.1318, 7.0733)	1.01
w_1	0.0940	(0.0939, 0.0942)	1.01
w_2	0.8479	(0.0471, 0.8588)	1
w_3	0.0579	(0.0471, 0.1294)	1.01

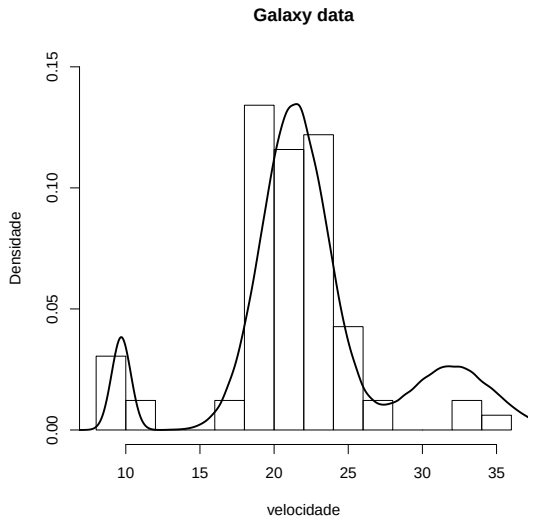

Figura 5.1: Histograma dos dados de velocidades de galáxias e densidade estimada pelo BSM-MCMC.

5.3.5 Dados de expressão gênica

As Tabelas 5.9 e 5.10 e a Figura 5.2 mostram os resultados obtidos com a aplicação do algoritmo BSM-MCMC aos dados de expressão gênica, obtidos do experimento realizado com a bactéria *Escherichia Coli*. Na Tabela 5.7 estão as probabilidades *a posteriori* para o número de componentes k, com um máximo em $\widetilde{k} = 3$, $P(k = 3|\cdot) = 0.6946$. Em relação ao algoritmo SM-MCMC, a probabilidade *a posteriori* obtida para $k = 3$ é menor (0.6946 contra 0.9111). Em relação ao algoritmo RJ-MCMC, a probabilidade *a posteriori* para $k = 3$ também é menor (0.6946 contra 0.7389).

A Tabela 5.8 mostra as estimativas para os parâmetros e os intervalos de credibilidade para os parâmetros das três componentes identificadas. O Apêndice E.5 (pag.92) mostra o gráfico ergódico dos valores gerados para os parâmetros, indicando a covergência.

A Figura 5.2 mostra a média de controle *versus* a média de tratamento dos 434 genes e os 3 grupos identificados. Triangulos $\triangledown$ representam o grupo G_1, $\bullet$ representam o grupo G_2 e os triangulos $\triangle$ representam o grupo G_3. O grupo G_1 é composto por 54 genes, o grupo G_2 por 366 genes e o grupo G_3 por 14 genes.

Em relação ao algoritmo SM-MCMC, o grupos G_1 no BSM-MCMC apresenta 7 genes a menos. Estes 7 genes estão na fronteira dos grupos identificados. A média de controle *versus* a média tratamento destes 7 genes estão em destaque, "pontos" maiores em negrito na Figura 5.2. Em relação ao RJ-MCMC o grupo G_1 apresenta 4 genes a mais.

Tabela 5.9: Probabilidade *a posteriori* para k.

k	1	2	3	≥ 4
$P(k\|\cdot)$	0	0.0172	0.6946	0.2882

Tabela 5.10: Estimativas para os parâmetros.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. G. R.
μ_1	-0.2258	(-0.5124, -0.0146)	1
μ_2	0.0367	(0.0049, 0.0669)	1.01
μ_3	0.7920	(0.2137, 1.7212)	1
σ_1	0.2515	(0.1362, 0.3778)	1
σ_2	0.1571	(0.1162, 0.1913)	1.01
σ_3	0.6184	(0.3729, 0.9934)	1
w_1	0.2582	(0.0595, 0.5652)	1.01
w_2	0.7001	(0.0092, 0.8993)	1.01
w_3	0.0415	(0.0092, 0.0887)	1

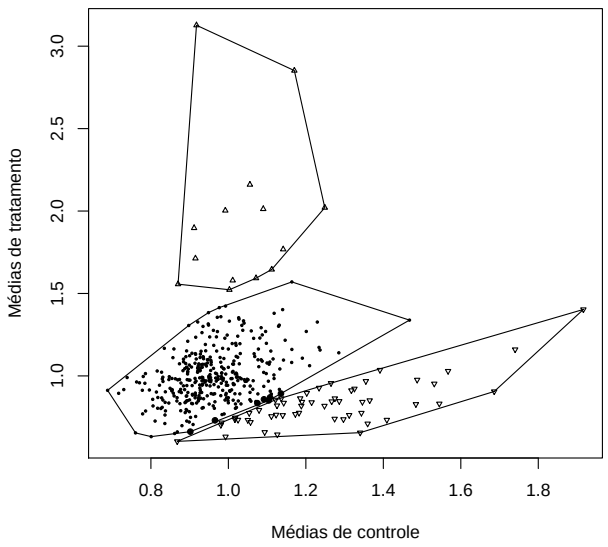

Figura 5.2: Médias de controle *versus* médias de Tratamento e grupos identificados. $\triangledown$ representam o grupo G_1, $\bullet$ o grupo G_2 e $\triangle$ o grupo G_3.

Capítulo 6

Discussão

Nesta seção, fazemos uma discussão sobre algumas das vantagens dos algoritmos propostos, SM-MCMC e BSM-MCMC, em relação aos procedimentos propostos por Richardson e Green (1997), Stephens (2000) e Jain e Neal (2004).

Richardson e Green (1997) propõem o algoritmo RJ-MCMC para modelos com mistura de distribuições normais univariadas, em que, a construção da proposta *split* satisfaz algumas condições em relação aos momentos da componente escolhida para se propor este movimento. Porém, a construção de propostas *split* e o cáculo da probabilidade de aceitação não é simples para o caso multivariado. Como visto nos exemplos com dados artificiais, a escolha da funções de transição é um importante ponto para que o RJ-MCMC tenha um bom desempenho. Caso a função de transição não seja adequada o RJ-MCMC pode não conseguir separar as observações em diferentes componentes, como observado nos dados artificiais 2 ($k = 3$) e 3 ($k = 5$). Ao contrário do algoritmo RJ-MCMC, nossa proposta *split* pode ser mais facilmente implementada, tanto para o caso univariado como para o caso multivariado. As propostas *split-merge* são baseadas nos dados observados e na distribuição *a posteriori* dos parâmetros. Como visto nos exemplos com dados artificiais $1, 2$ e 3 este proce- dimento é mais eficiente na estimação conjunta de k e dos parâmetros associados às componentes. Uma outra vantagem de nosso método é que removemos a necessidade de calcular o jacobiano da transformação, tornando a probabilidade de aceitação mais simples de se calcular e simplificando a implementação computa-

cional.

Stephens (2000) considera os parâmetros do modelo com mistura como sendo um processo pontual marcado, em que, cada ponto representa uma componente da mistura e propõe um processo de "nascimento" e "morte" para estimação de k. Este procedimento permite k variar pelo "nascimento" de uma nova componente e "morte" de uma componente existente. Um "nascimento" ocorre a uma taxa constante, que é fixada de forma subjetiva a partir de conhecimentos prévios, e "morrem" a uma taxa que é calculada de acordo com a equação de balanceamento, necessária para garantir que o processo tenha a distribuição estacionária. Em relação aos métodos MCMC usuais, neste método a probabilidade de aceitação é substituída por diferentes tempos que o processo permanece em um estado. Este tempo segue uma distribuição exponencial com parâmetro que é definido pela taxa de "nascimento" e "morte". Ao final do tempo, um "nascimento" ou "morte" necessáriamente ocorre, independentemente, se o modelo atual tem ou não alta probabilidade. Dada a escolha de um "nascimento", novos parâmetros são gerados da distribuição de "nascimento", que não depende dos dados. Este procedimento determina uma nova componente no modelo, porém se os parâmetros gerados não são adequados nenhuma observação será associada a esta nova componente. Logo, a escolha da distribuição da qual os novos parâmetros serão gerados é um importante passo para que este procedimento tenha uma boa performance. Como Stephens sugere gerar os novos parâmetros da distribuição *a priori* esta boa performance pode ficar comprometida se esta distribuição for não informativa. Se o movimento "morte" é feito em uma componente com observações associadas, então estas observações são alocadas entre as outras componentes. Em nossa proposta, quando ocorre a criação de novas componentes, estas possuem observações associados e os novos parâmetros são gerados da distribuição *a posteriori*, o que torna a proposta mais interessante do ponto de vista da dinâmica do processo. Além disso, as propostas *split-merge* não dependem de taxas fixadas a partir de conhecimentos prévios. A utilização de probabilidades de aceitação, que dependem somente das observações pertencentes à(s) componente(s) escolhidas para um *split* ou *merge*, permite que o modelo permaneça em um estado

que tem alta probabilidade *a posteriori*.

O algoritmo *split-merge*, proposto por Jain e Neal (2004) para modelos de misturas de processos de Dirichlet tem interesse somente em identificar os grupos de observações. Eles integram fora os pesos e os parâmetros das componentes da mistura. A estratégia *split* é baseada, primeiramente, em uma separação aleatória das observações pertencentes a uma componente e em seguida eles propõem o uso de um algoritmo *Gibbs sampling* restrito para melhorar este *split*. O número de iterações utilizadas no algoritmo *Gibbs sampling* deve ser previamente fixado e segundo os autores este número de iterações pode afetar a probabilidade de aceitação da proposta. Esta estratégia para propor o *split* requer um alto custo computacional. Ao contrário do algoritmo de Jain e Neal (2004), nossas propostas consideram o interesse em k, nos parâmetros das componentes e a presença destas quantidades nas propostas *split-merge*. Além disso, nossa estratégia para propor um *split* é mais rapidamente implementada e testada, possibilitando uma maior eficiência computacional.

Em relação aos algoritmos propostos, SM-MCMC e BSM-MCMC, os resultados obtidos para os conjuntos de dados utilizados são similares. A diferença é no tempo computacional dos métodos, pois no algoritmo BSM-MCMC temos que $k_{max} = n$. Se fixarmos $k_{max} = n$ no SM-MCMC os tempos de simulação são equivalentes entre os dois algoritmos. As probabilidades *a posteriori* para k, nos cinco conjuntos de dados utilizados, são menores no BSM-MCMC em relaçao ao SM-MCMC. Porém, no algoritmo BSM-MCMC, o modelo com mistura infinita, $k \to \infty$, pode ter algumas vantagens em relação à abordagem com mistura finita: (i) em muitas aplicações pode ser mais apropriado não limitar o número de componentes, pois não conhecemos o valor de k; (ii) o número de componentes pode ser estimado, como para o caso com um valor limite k_{max} para k e (iii) uma nova observação pode ser alocada em uma das componentes existentes ou determinar uma nova componente, o que não ocorre no SM-MCMC.

Capítulo 7

Considerações finais e propostas futuras

Nesta tese, propomos os algoritmos *split-merge* MCMC e *birth-split-merge* MCMC para modelos com mistura de distribuições com número de componentes desconhecido. Modelamos conjuntamente os parâmetros, o número de componentes e as variáveis latentes. Inferências sobre as quantidades de interesse são feitas com base no desenvolvimento de um processo estocástico com movimentos *split-merge*. Estes movimentos são aplicados diretamente aos dados observados, são desenvolvidos para serem reversíveis e são aceitos de acordo com a probabilidade de aceitação de *Metropolis-Hastings*.

A estratégia *split-merge* evita modas locais separando ou juntado observações pertencentes às componentes. Com isso, temos uma maior exploração dos grupos de observações. No movimento *split*, observações são alocadas para uma de duas novas componentes, baseados em probabilidades que são calculadas de acordo com os parâmetros gerados da distribuição *a posteriori* dadas as observações previamente alocadas. Este procedimento torna os algoritmos mais eficientes computacionalmente, em relação aos métodos alternativos, pois o movimento pode ser rapidamente proposto e testado.

Verificamos a performance dos algoritmos propostos utilizando três conjuntos de dados artificiais, gerados de um modelo com mistura de distribuições normais

com $k = 2$, $k = 3$ e $k = 5$ componentes. Os resultados obtidos mostram um boa performance dos algoritmos na estimação de k e dos parâmetros das componentes.

Na aplicação aos dados de expressão gênica, três grupos com diferentes comportamentos, em relação aos níveis de expressão observados, são identificados pelos métodos. Isto pode ajudar biólogos e geneticistas a estudarem possíveis relações existentes entre os genes pertencentes a um mesmo grupo.

Comparamos os resultados com o algoritmo RJ-MCMC e os métodos propostos apresentam um melhor desempenho. Obtivemos convergência, segundo o critério de Gelman e Rubin, para todos os parâmetros. Comparamos somente com o algoritmo RJ-MCMC pois este segue a linha de nosso trabalho, utilizando a marcação das componentes através das variáveis latentes nas propostas *split-merge*.

Para finalizar, destacamos que a contribuição teórica presente nesta tese é o desenvolvimento de um processo estocástico com base nos movimento *split-merge*, que são baseados nos dados. Utilizando estes movimentos, nosso método busca informações sobre k e os parâmetros $\boldsymbol{\phi}$ diretamente nos dados observados $\mathbf{y}$. Com isso, quando propomos o surgimento de uma nova componente esta sempre tem dados associados, i.e., temos uma partição nos dados, e os parâmetros são gerados da distribuição *a posteriori*. Ao contrário dos métodos *reversible jump* e processo de nascimento-e-morte, que propõem novas componentes com novos parâmetros sem se preocuparem se estas novas componentes determinam uma partição nos dados.

Como proposta de pesquisa futura, destacamos o interesse em situações com distribuições *a priori* não conjugadas e formas alternativas para evitar o problema de não identificabilidade presente na definição de um modelo com mistura na forma de uma combinação convexa de distribuições.

Apêndices

Apêndice A: Jacobiano - RJ-MCMC

Apêndice A.1: Jacobiano - proposta *split*

$$|Jac^{sp}| = \begin{vmatrix} u_1 & w_{j^*} & 0 & 0 & 0 & 0 \\ 1-u_1 & -w_{j^*} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -\sigma_{j^*}\sqrt{\frac{w_{j_2}}{w_{j_1}}} & \frac{-u_2\sqrt{\frac{w_{j_2}}{w_{j_1}}}}{2\sigma_{j^*}} & 0 \\ 0 & 0 & 1 & \sigma_{j^*}\sqrt{\frac{w_{j_1}}{w_{j_2}}} & \frac{u_2\sqrt{\frac{w_{j_1}}{w_{j_2}}}}{2\sigma_{j^*}} & 0 \\ 0 & \frac{u_3(-1+u_2^2)\sigma_{j^*}^2}{u_1^2} & 0 & \frac{-2u_3u_2\sigma_{j^*}^2}{u_1} & \frac{-u_3(-1+u_2^2)}{u_1} & -\frac{(-1+u_2^2)\sigma_{j^*}^2}{u_1} \\ 0 & \frac{(-1+u_3)(-1+u_2^2)\sigma_{j^*}^2}{(-1+u_1)^2} & 0 & \frac{-2(-1+u_3)u_2\sigma_{j^*}^2}{-1+u_1} & \frac{-(-1+u_3)(-1+u_2^2)}{-1+u_1} & -\frac{(-1+u_2^2)\sigma_{j^*}^2}{-1+u_1} \end{vmatrix}$$

$$|Jac^{sp}| = \frac{w_{j^*}\left(\sigma_{j^*}^2\right)^2(-1+u_2^2)\left[\sqrt{\frac{w_{j_1}}{w_{j_2}}}+\sqrt{\frac{w_{j_2}}{w_{j_1}}}\right]}{u_1(-1+u_1)}$$

De (3.21), temos que

$$\sqrt{\frac{w_{j_1}}{w_{j_2}}} = \frac{\mu_{j_2}-\mu_{j^*}}{u_2\sigma_{j^*}} \quad \text{e} \quad \sqrt{\frac{w_{j_1}}{w_{j_2}}} = -\frac{\mu_{j_1}-\mu_{j^*}}{u_2\sigma_{j^*}}.$$

Logo,

$$\sqrt{\frac{w_{j_1}}{w_{j_2}}}+\sqrt{\frac{w_{j_1}}{w_{j_2}}} = \frac{|\mu_{j_1}-\mu_{j_2}|}{u_2\sigma_{j^*}}$$

Assim, substituindo em $|Jac^{sp}|$, temos

$$|Jac^{sp}| = \frac{w_{j^*}\sigma_{j^*}^2(1-u_2^2)|\mu_{j_1}-\mu_{j_2}|}{u_1(1-u_1)u_2}.$$

De (3.22), temos que

$$(1-u_2^2) = \frac{w_{j_1}\sigma_{j_1}}{u_3\sigma_{j^*}^2 w_{j^*}}.$$

Novamente, substituindo em $|Jac^{sp}|$ temos

$$|Jac^{sp}| = \frac{w_{j^*}\sigma_{j^*}^2 w_{j_1}\sigma_{j_1}^2|\mu_{j_1}-\mu_{j_2}|}{u_1(1-u_1)u_3\sigma_{j^*}u_2 w_{j^*}} = \frac{w_{j_1}\sigma_{j_1}^2|\mu_{j_1}-\mu_{j_2}|}{u_1(1-u_1)u_2}.$$

De (3.20)

$$u_1 = \frac{w_{j_1}}{w_{j^*}} \quad \text{e} \quad 1-u_1 = \frac{w_{j_2}}{w_{j^*}}.$$

Logo,

$$|Jac^{sp}| = \frac{w_{j^*}^2 w_{j_1}\sigma_{j_1}^2|\mu_{j_1}-\mu_{j_2}|}{w_{j_1}w_{j_2}u_2u_3} = \frac{w_{j^*}^2\sigma_{j_1}^2|\mu_{j_1}-\mu_{j_2}|}{w_{j_2}u_2u_3}.$$

De (3.22)

$$w_{j_2} = \frac{(1-u_3)(1-u_2^2)\sigma_{j^*}^2 w_{j^*}}{\sigma_{j_2}^2}.$$

Portanto,

$$|Jac^{sp}| = \frac{w_{j^*}^2\sigma_{j_1}^2\sigma_{j_2}^2|\mu_{j_1}-\mu_{j_2}|}{(1-u_3)(1-u_2^2)\sigma_{j^*}^2 w_{j^*}u_2u_3} = \frac{w_{j^*}|\mu_{j_1}-\mu_{j_2}|\sigma_{j_1}^2\sigma_{j_2}^2}{u_2(1-u_2^2)u_3(1-u_3)\sigma_{j^*}^2}.$$

Apêndice B: Algoritmo SM-MCMC

Apêndice B.1: Probabilidade de alocação

Restritos ao conjunto S_{j_1}, temos que, analogamente à (4.8), a probabilidade *a priori* de uma observação y_v pertencer a este conjunto é

$$p(c_v|\mathbf{c}_{j_1}) \propto n_{j_1} + \gamma,$$

onde $\mathbf{c}_{j_1}$ é o conjunto de variáveis indicadoras, tal que, $c_i = j_1$.

Dados $\phi^*_{j_1}$, $\phi^*_{j_2}$, $\mathbf{c}_{j_1}$, $\mathbf{c}_{j_2}$ e a observação y_v, a probabilidade condicional *a posteriori* de $c_v = j_1$ é dada por

$$\begin{aligned} P_{j_1}(y_v) &= P(c_v = j_1|\mathbf{c}_{j_1}, \mathbf{c}_{j_2}, \phi^*_{j_1}, \phi^*_{j_2}, y_v) \\ &= \frac{P(c_v = j_1|\mathbf{c}_{j_1})P(y_v|\phi^*_{j_1})}{P(c_v = j_1|\mathbf{c}_{j_1})P(y_v|\phi^*_{j_1}) + P(c_v = j_2|\mathbf{c}_{j_2})P(y_v|\phi^*_{j_2})} \\ &= \frac{(n_{j_1} + \gamma)f(y_v|\phi^*_{j_1})}{(n_{j_1} + \gamma)f(y_v|\phi^*_{j_1}) + (n_{j_2} + \gamma)f(y_v|\phi^*_{j_2})} \end{aligned}$$

Apêndice B.2: Escolha das componentes j_1 e j_2 para o *merge*

Considere que μ_{j_1} e μ_{j_2} são as médias das componentes j_1 and j_2. As componentes j_1 e j_2 são adjacentes em relação aos valores de suas médias se $\mu_{j_1} < \mu_{j_2}$ com nenhum outro $\mu_j \in [\mu_{j_1}, \mu_{j_2}]$ ou $\mu_{j_2} < \mu_{j_1}$ com nenhum outro $\mu_j \in [\mu_{j_2}, \mu_{j_1}]$.

Assim, temos que:

(i) Se $k = 2$, $P_{j_1,j_2|k} = 1$, pois só temos duas componentes;

(ii) Se $k > 2$, escolhemos $j_1 = 1$ com probabilidade $\frac{1}{k}$. Dado que $j_1 = 1$, então $j_2 = 2$ com probabilidade 1. Assim, a configuração $\{j_1 = 1, j_2 = 2\}$ tem probabilidade $\frac{1}{k}$. Porém, esta configuração é equivalente à $\{j_1 = 2, j_2 = 1\}$. Logo, a probabilidade de escolher $j_1 = 2$ é $\frac{1}{2}$. Dado que $j_1 = 2$, $j_2 = 1$ com probabilidade $\frac{1}{2}$, pois assumimos que j_2 poderia assumir os valores $j_1 - 1 = 1$ ou $j_1 + 1 = 3$ com mesma probabilidade. Assim, $P_{j_1,j_2|k} = P_{1,2|k} + P_{2,1|k} = \frac{3}{2k}$. Note que, esta mesma probabilidade é obtida para a configuração $\{j_1 = k, j_2 = k-1\} = \{j_1 = k-1, j_2 = k\}$. Logo, $P_{j_1,j_2|k} = \frac{3}{2k}$ se $k > 2$ e $j_1, j_2 \in \{1, k\}$;

(iii) Se $j_1, j_2 \notin \{1, k\}$, $P_{j_1,j_2|k} = \frac{1}{2k} + \frac{1}{2k} = \frac{1}{k}$.

Assim, probabilidade de escolher j_1 e j_2 para um *merge* é dada por

$$P_{j_1,j_2|k} = P_{j_1|k}P_{j_2|j_1,k} + P_{j_2|k}P_{j_1|j_2,k} = \begin{cases} 1, \text{ if } k = 2 \\ \frac{3}{2k}, \text{ if } k > 2, \ j_1, j_2 \in \{1, k\} \\ \frac{1}{k}, \text{ if } k > 2, \ j_1, j_2 \notin \{1, k\} \end{cases},$$

Para o caso multivariado, escolhemos uma coordenada aleatoriamente. Condicional na coordenada escolhida, calculamos $P_{j_1,j_2|k}$ como descrito acima.

Apêndice C: Gráficos ergódicos - Análise de dados - SM-MCMC

Apêndide C.1: Gráficos ergódicos - dados artificiais 1

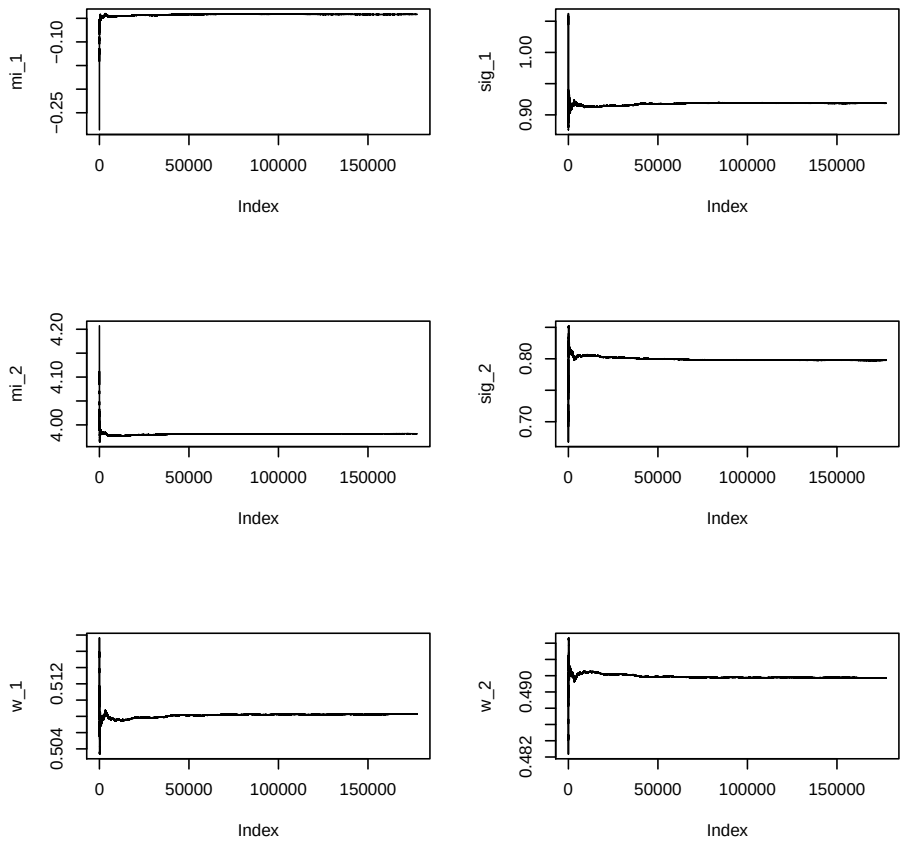

Figura 7.1: Gráfico ergódico para a média dos valores gerados para μ_j, σ_j e w_j, $j = 1, 2$.

Apêndide C.2: Gráficos ergódicos - dados artificiais 2

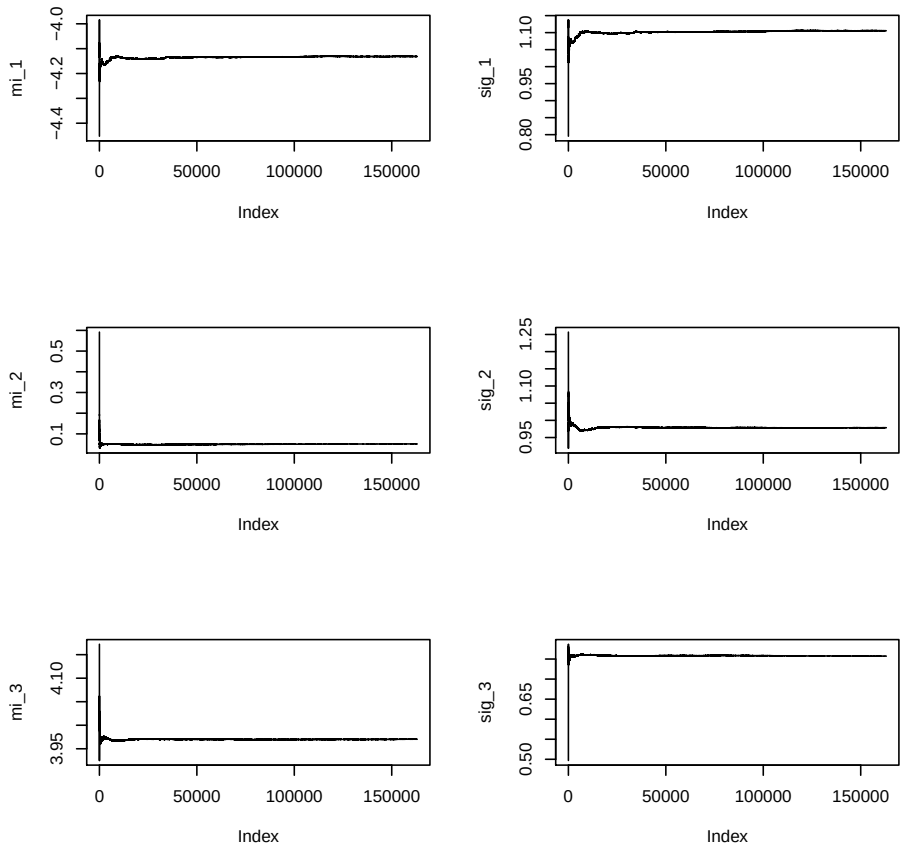

Figura 7.2: Gráfico ergódico para a média dos valores gerados para μ_j e σ_j, $j = 1, 2, 3$.

Apêndide C.3: Gráficos ergódicos - dados artificiais 3

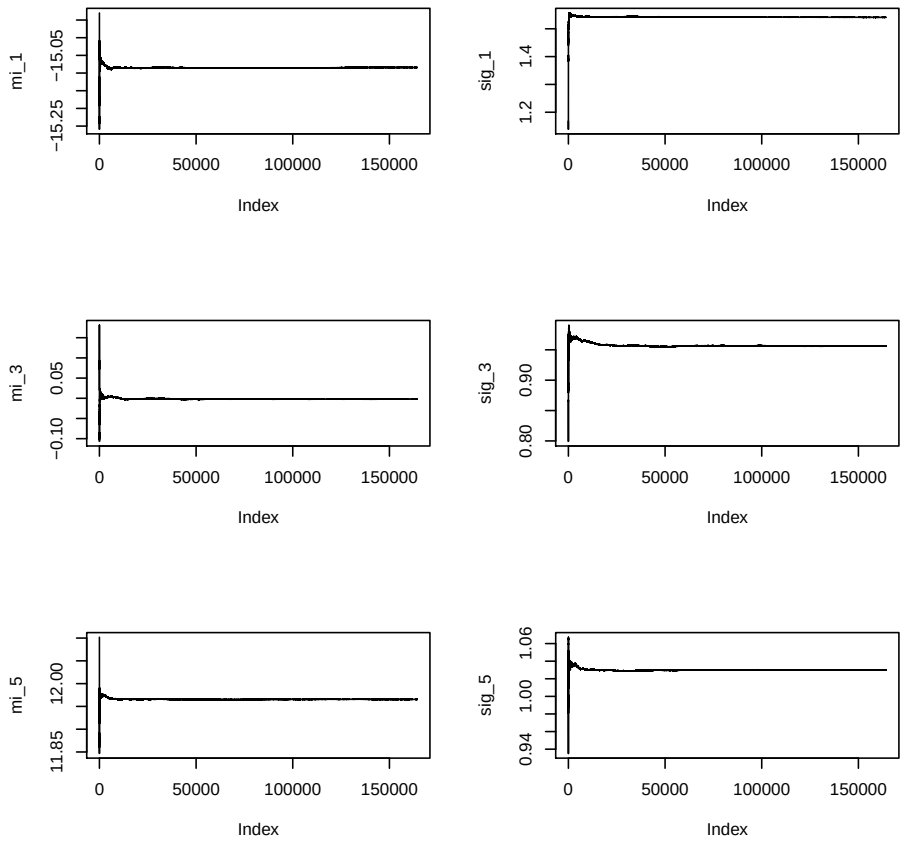

Figura 7.3: Gráfico ergódico para a média dos valores gerados para μ_j e σ_j, $j \in \{1, 3, 5\}$.

Apêndide C.4: Gráficos ergódicos - dados de galáxias

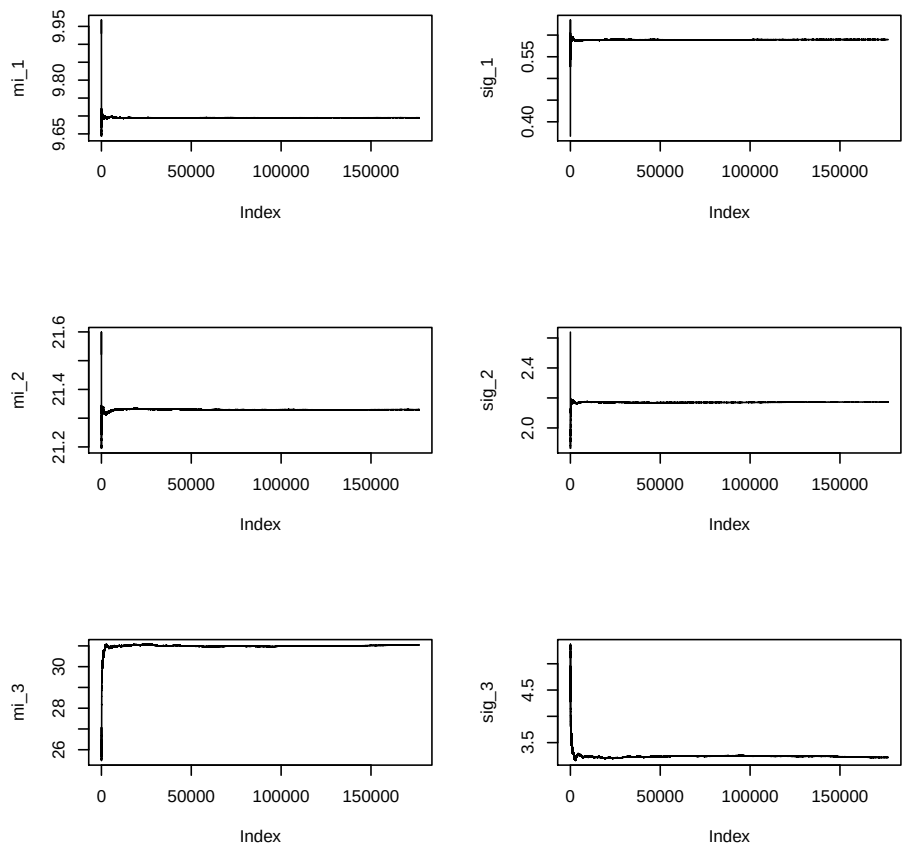

Figura 7.4: Gráfico ergódico para a média dos valores gerados para μ_j e σ_j, $j = 1, 2, 3$.

Apêndide C.5: Gráficos ergódicos - dados de expressão gênica

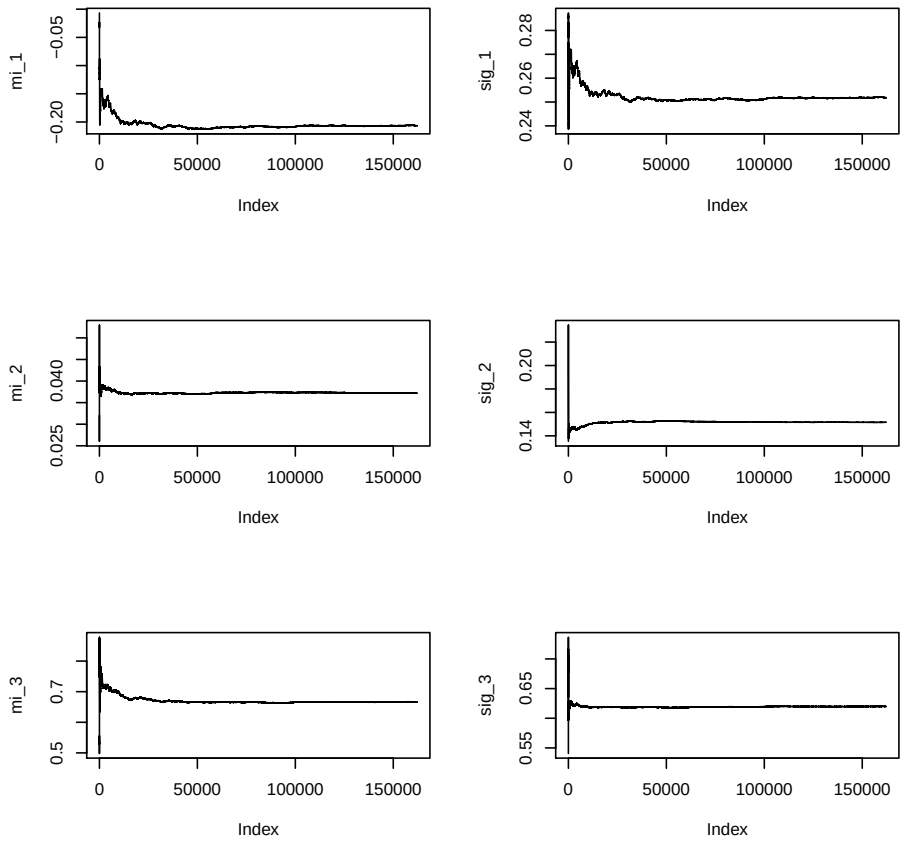

Figura 7.5: Gráfico ergódico para a média dos valores gerados para μ_j e σ_j, $j = 1, 2, 3$.

Apêndice D: Gráficos ergódicos - Análise de dados - RJ-MCMC

Apêndide D.1: Gráficos ergódicos - dados artificiais 1

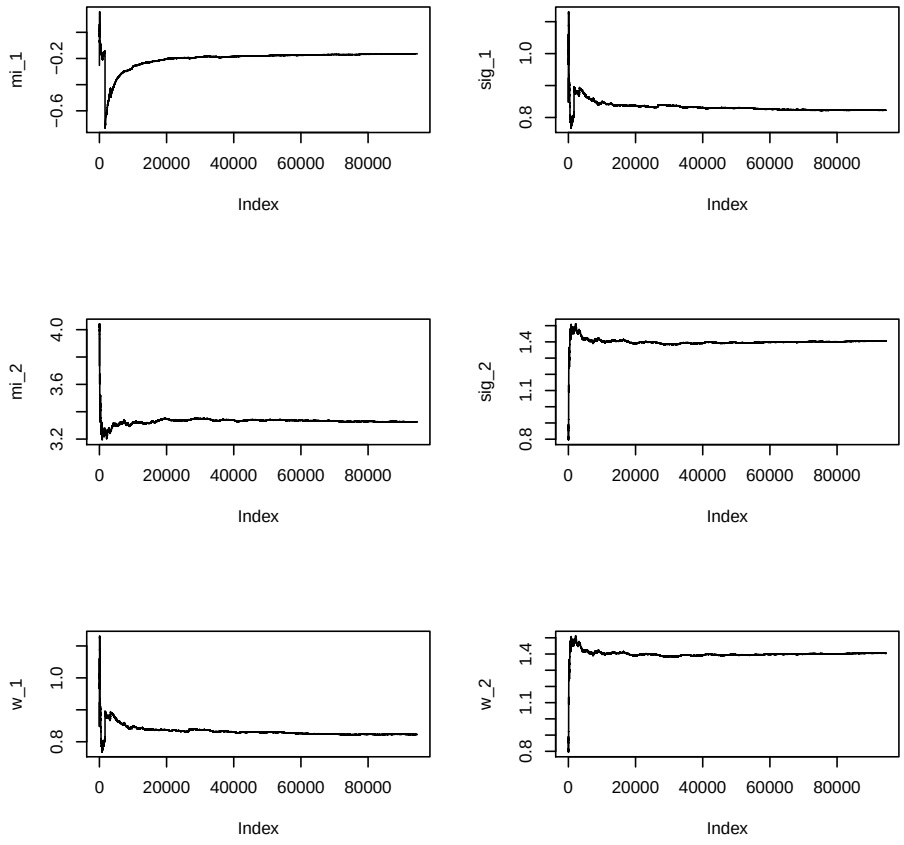

Figura 7.6: Gráfico ergódico para a média dos valores gerados para μ_j, σ_j e w_j, $j = 1, 2$.

Apêndide D.2: Gráficos ergódicos - dados artificiais 2

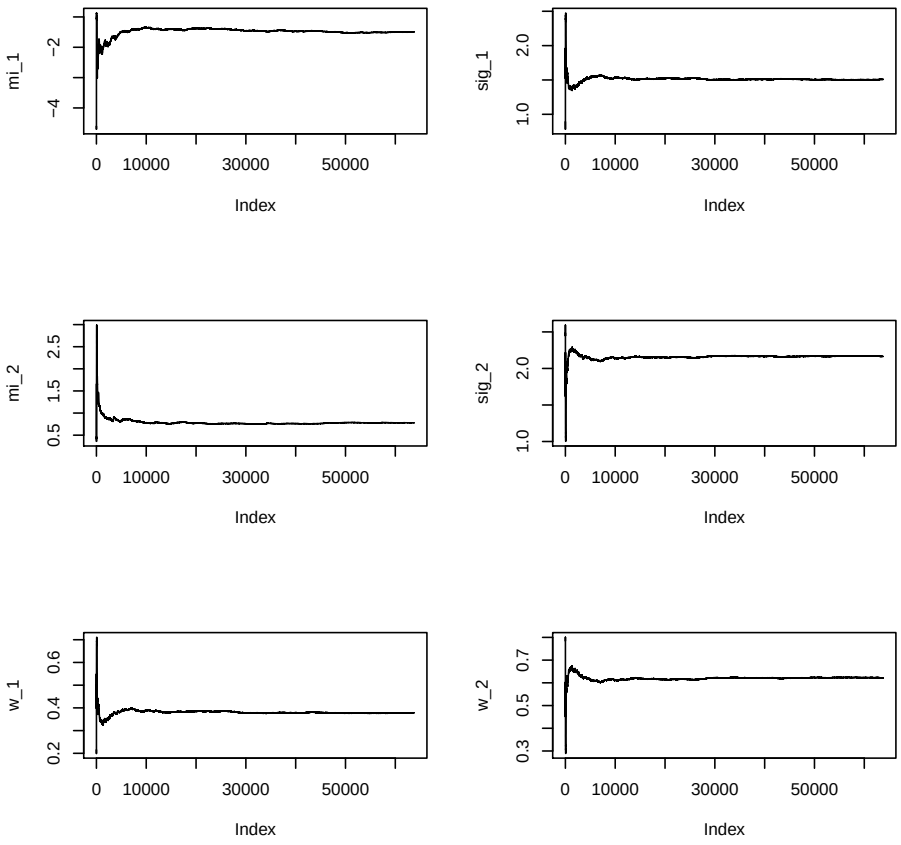

Figura 7.7: Gráfico ergódico para a média dos valores gerados para μ_j e σ_j, $j = 1, 2, 3$.

Apêndide D.3: Gráficos ergódicos - dados artificiais 3

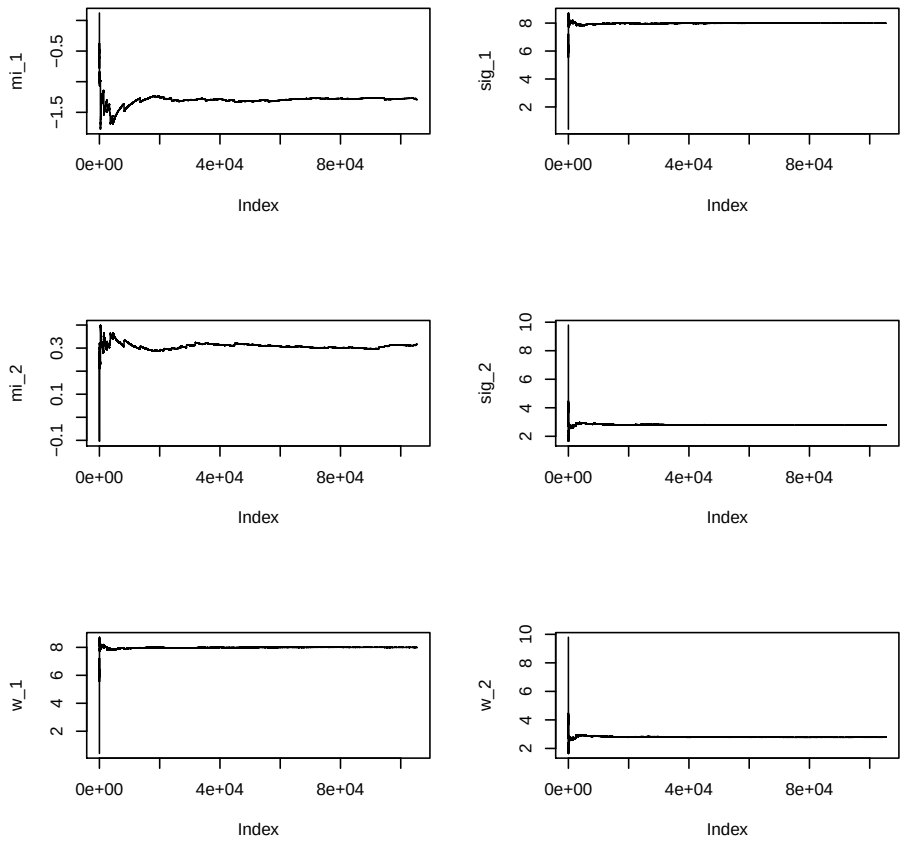

Figura 7.8: Gráfico ergódico para a média dos valores gerados para μ_j e σ_j, $j \in \{1, 3, 5\}$.

Apêndide D.4: Gráficos ergódicos - dados de galáxias

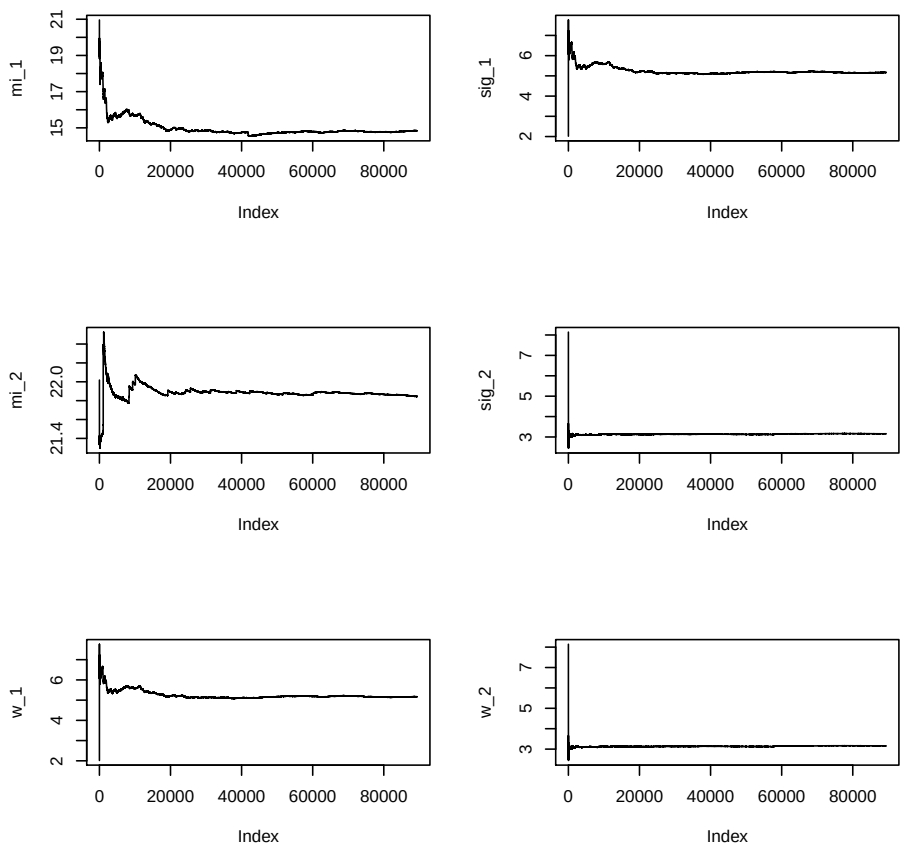

Figura 7.9: Gráfico ergódico para a média dos valores gerados para μ_j e σ_j, $j = 1, 2, 3$.

Apêndide D.5: Gráficos ergódicos - dados de expressão gênica

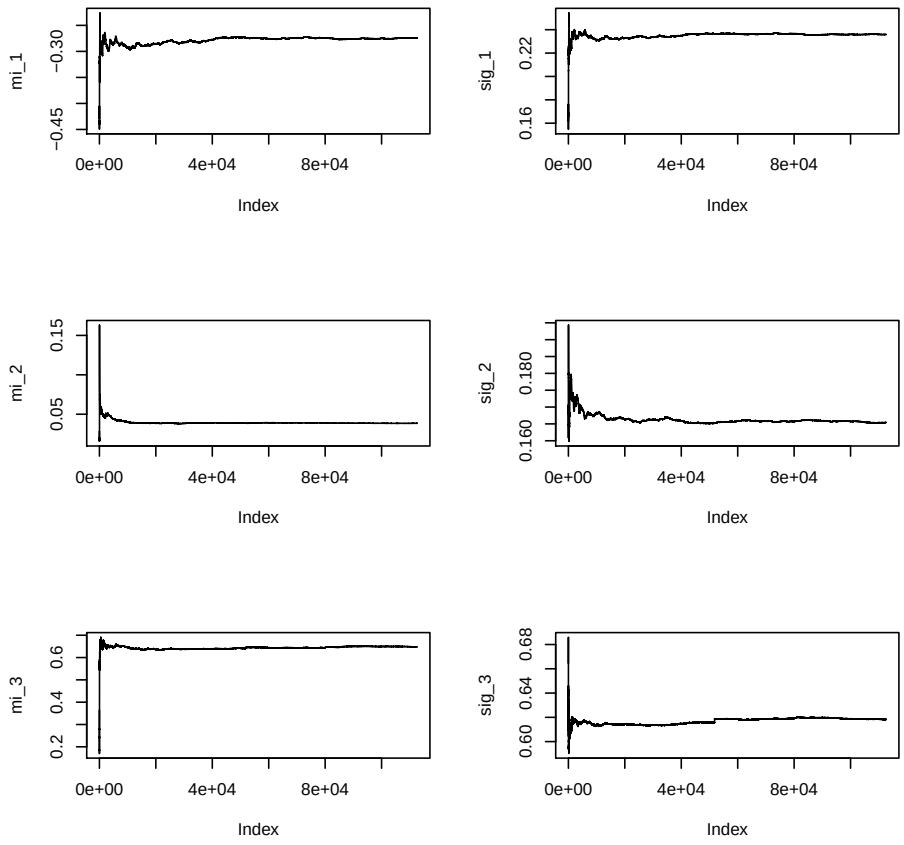

Figura 7.10: Gráfico ergódico para a média dos valores gerados para μ_j e σ_j, $j = 1, 2, 3$.

Apêndice E: Gráficos ergódicos - Análise de dados - BSM-MCMC

Apêndide E.1: Gráficos ergódicos - dados artificiais 1

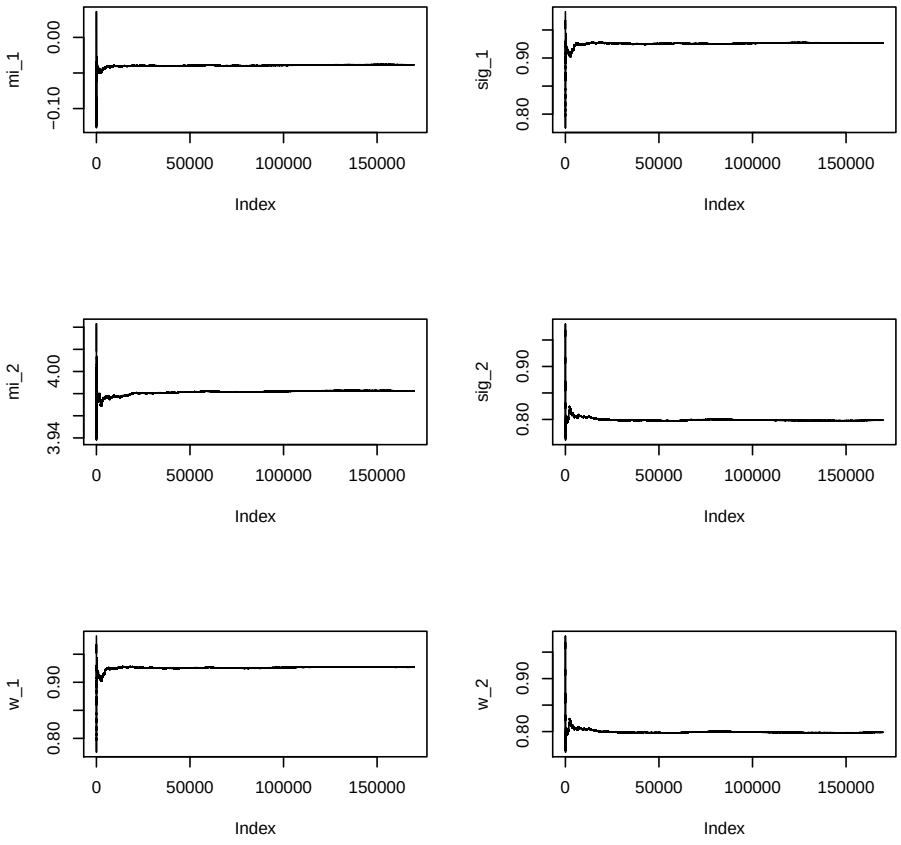

Figura 7.11: Gráfico ergódico para a média dos valores gerados para μ_j, σ_j e w_j, $j = 1, 2$.

Apêndide E.2: Gráficos ergódicos - dados artificiais 2

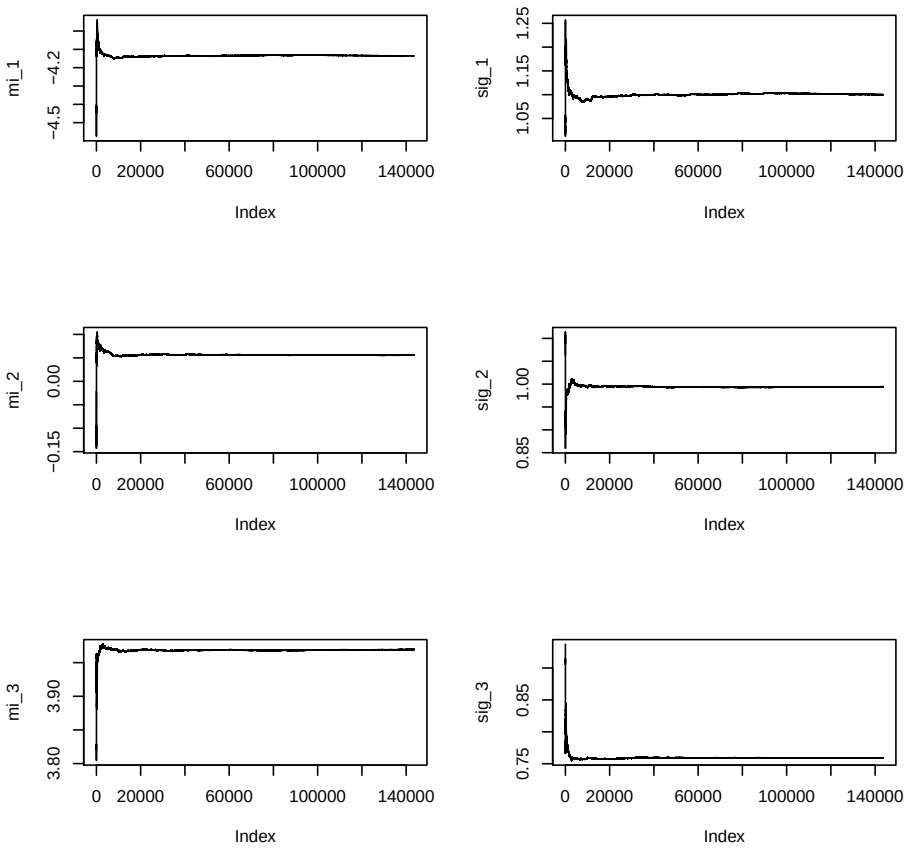

Figura 7.12: Gráfico ergódico para a média dos valores gerados para μ_j e σ_j, $j = 1, 2, 3$.

Apêndide E.3: Gráficos ergódicos - dados artificiais 3

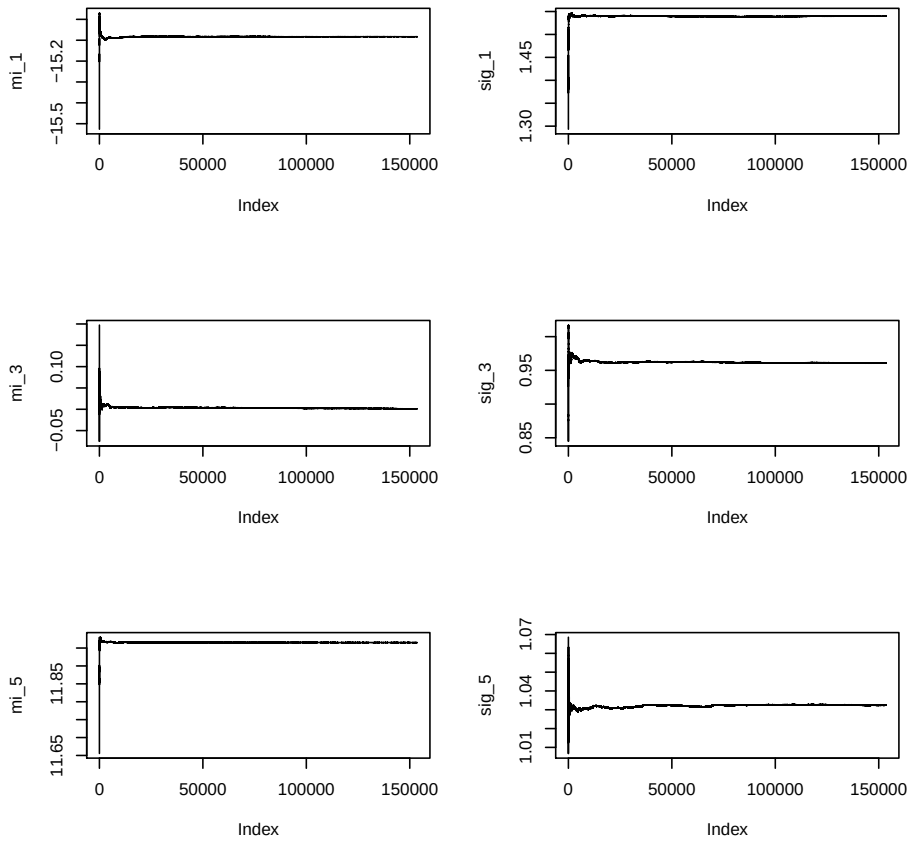

Figura 7.13: Gráfico ergódico para a média dos valores gerados para μ_j e σ_j, $j \in \{1, 3, 5\}$.

Apêndide E.4: Gráficos ergódicos - dados de galáxias

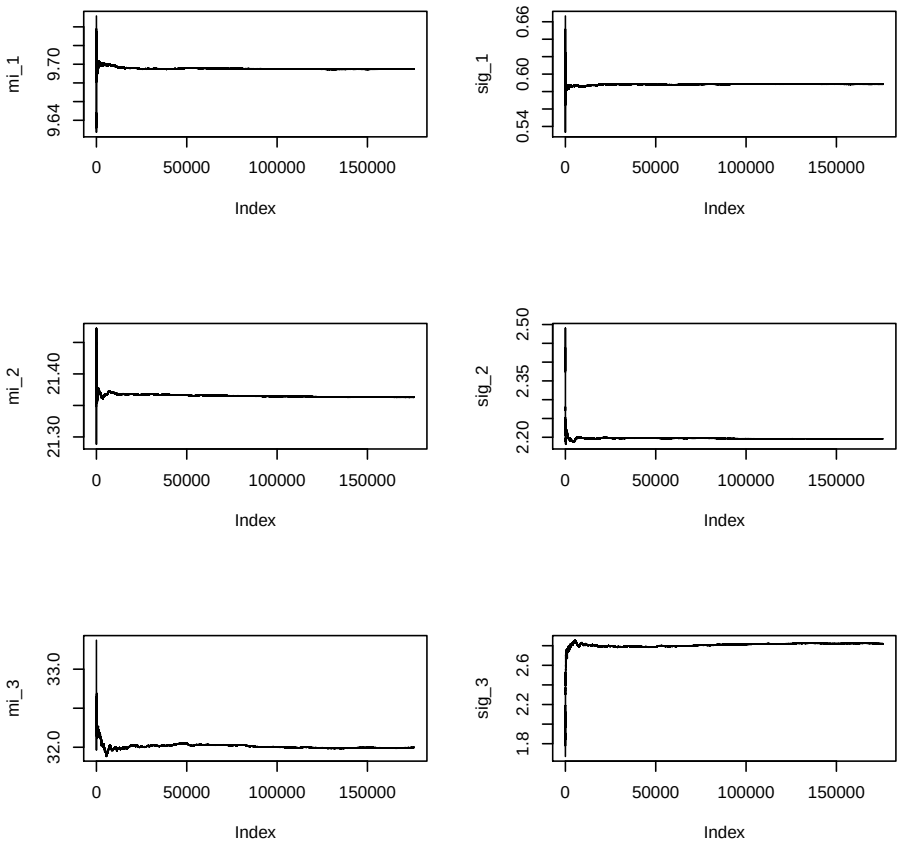

Figura 7.14: Gráfico ergódico para a média dos valores gerados para μ_j e σ_j, $j = 1, 2, 3$.

Apêndide E.5: Gráficos ergódicos - dados de expressão gênica

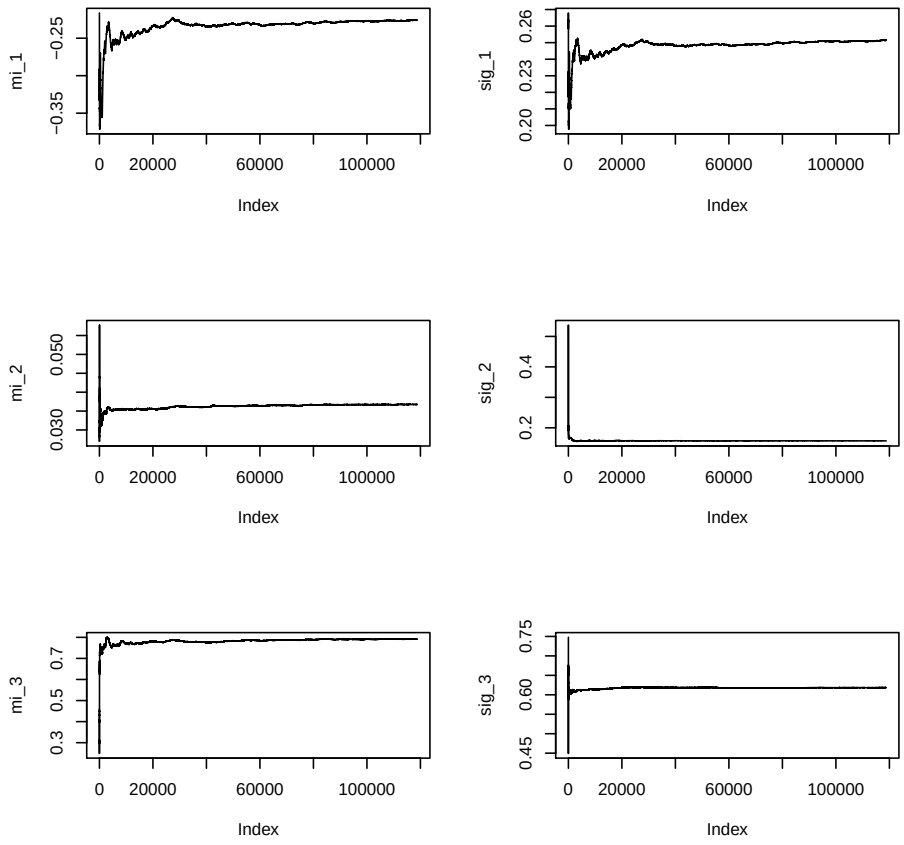

Figura 7.15: Gráfico ergódico para a média dos valores gerados para μ_j e σ_j, $j = 1, 2, 3$.

Referências Bibliográficas

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

Arfin, S. M., Long, A. D., Ito, E. T., Tolleri, L., Riehle, M. M., Paegle, E. S. and Hatfield, G. W. (2000). Global gene expression profiling in *Escherichia Coli* K12. *The Journal of Biological Chemistry*, **275**, 29672-29684.

Baldi, P. and Long, D. A. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509-519.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distribution via Pólya urn schemes. *The Annals of Statistics*, **1**, 353-355.

Casella, G., Robert, C., and Wells, M. (2000). Mixture models, latent variables and partitioned importance sampling. *Technical Report 2000-03, CREST, INSEE, Paris.*

Celeux, G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics*, **2**, 73-82.

Celeux, G. and Hurn, M. and Robert, C. P. (2000). Computational and inferential diffculties with mixture posterior distributions. *Journal of the American Statistical Association*, **95**, 957-970.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Jounal of the American Statistical Association*, **90**, 1313-1321.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, B, **39**, 1-38.

Diebolt, J. and Robert, C. (1993). Discussion of "Bayesian computations via the Gibbs sampler". *Jounal of the Royal Statistical Society*, B, **55**, 71-72.

Diebolt, J. and Robert, C. (1994). Estimation of finite mixture distribution through Bayesian sampling. *Journal of the Royal Statistical society*, B, **56**, 163-175.

Do, K.A; Müller, P. Tang, F.(2002). A Bayesian Mixture for Differential Gene Expression. http://odin.mdacc.tmc.edu/ pm/pap/DMT02.pdf.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577-511.

Ferguson, S. T. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, **2**, 209-230.

Fox, R. J. and Dimmic, M. W. (2006). A two-sample Bayesian t-test for microarray data. *BMC Bioinformatics*, **7**, 118:126.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequence. *Statistical science*, **7**, 457-472.

Geyer, C. J. and Møller, J. (1994). Simulation procedure and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, **21**, 359-373.

Gopalan, R.; Berry, D. A. (1998). Bayesian Multiple Comparisons Using Dirichlet Process priors. *Journal of the American Statistical Association*, **93**, 1130-1139.

Green, P. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, **82**, 711-732.

Jain, S. and Neal, R.M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, **13**, 158-182.

Lavine, M. and West, M. (1992). A Bayesian method for classification and discrimination. *Canadian Journal of Statistics*, **20**, 451-461.

Marin, J. M., Mengersen, K. and Robert, C. P. (1992). Bayesian Modelling and Inference on Mixtures of Distributions.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian Infinite Mixture Model Based clustering of Gene Expression Profiles. *Bioinformatics*, **18**, 1194-1206.

McLachlan, G. J., and Basford, K. E. (1988). Mixture Models: inference and applications to clustering, Marcel Dekker: New York.

Neal, R. M. (1998). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Technical Report 4915. Department of Statistics,university of Toronto. http://cs.toronto.edu/ redford/mixmc.abstract.html.

Preston, C. J. (1976). Spatial birth-and-death processes. *Bulletim of the Institute of International Statistics*, **46**, 371-391.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixture with an unknown number of components. *Journal of the Royal Statistical Society*, B, **59**, 731-792,

Ripley, B. D. (1976). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society*, B, **39**, 172-212.

Robert, C. (1996). Mixture of distributions: Inference and estimation, chapter 24 (pp. 441-464) of *practical Markov Chain Monte Carlo*, W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds. Chapman and Hall, London.

Roeder, K. and Wasserman, L. (1995). Practical Bayesian density estimation using mixture of normals. *Technical report 633, Department of Statistics, Carnegie Mellon University*, 187-220.

Stephens, M. (2000). Bayesian Analysis of mixture models with an unknown number of components-an alternative to reversible jump method. *The Annals of Statistics*, **28**, 40-74.

Tierney, L. (1995). Markov chains for exploring posterior distributions, *The Annals of Statistics*, **22**, 1701-1762.

Tintterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). Statistical Analysis of Finite Mixture Distribution. Wiley, Chichester.

Verdinelli, I. and Wasserman, L. (1992). Bayesian analysis of outliers problems using the Gibbs sampler. *Statistical Computation*, **1**, 105-117.

West, M., Müller, P., and Escobar, M. D. (1994). Hierarchical Priors and Mixture Models, With Application in Regression and Density Estimation in *Aspects of Uncertainty: A Tribute to D. V. Lindley*, eds, A. F. M. Smith and P. Freeman, New York: Wiley, pp. 363-386.

Zhaohui, S. Q. (2006) Clustering microarray gene expression data using weighted chinese restaurant process. *Bioinformatics*, **22**, 1988-1997.