Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

# Aspectos práticos da estimação do modelo de mistura via processo de Dirichlet

Rosineide Fernando da Paz

UFSCar - São Carlos/SP
Maio/2013

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

# Aspectos práticos da estimação do modelo de mistura via processo de Dirichlet

Rosineide Fernando da Paz

Orientador: Prof. Dr. Luís A. Milan

Dissertação apresentada ao programa de pós-graduação em Estatística da Universidade Federal de São Carlos - PPG/UFSCar - como parte dos requisitos para à obtenção do título de Mestre em Estatística.

**Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar**

P348ap

Paz, Rosineide Fernando da.
Aspectos práticos da estimação do modelo de mistura via processo de Dirichlet / Rosineide Fernando da Paz. -- São Carlos : UFSCar, 2013.
69 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2013.

1. Estatística. 2. Inferência bayesiana. 3. Processos de Dirichlet. I. Título.

CDD: 519.5 (20ª)

**UNIVERSIDADE FEDERAL DE SÃO CARLOS**
**Centro de Ciências Exatas e de Tecnologia**
**Programa de Pós-Graduação em Estatística**
Via Washington Luís, Km 235 - C.P.676 - CGC 45358058/0001-40
FONE: (016) 3351-8292 – Email: ppgest@ufscar.br
13565-905 - SÃO CARLOS-SP - BRASIL

# FOLHA DE APROVAÇÃO

**Aluno(a) : Rosineide Fernando da Paz**

DISSERTAÇÃO DE MESTRADO DEFENDIDA E APROVADA EM 03/04/2013
PELA COMISSÃO JULGADORA:

Presidente _____
Prof. Dr. Luis A. Milan (DEs-UFSCar/Orientador)

1º Examinador _____
Prof. Dr. José Galvão Leite (DEs-UFSCar)

2º Examinador _____
Prof. Dr. Luís Gustavo Esteves (IME-USP)

Dedico este trabalho ao meu filho, Vinícius da Paz Monteiro, que para mim é uma fonte de força e que sofreu com a minha ausência nas horas de dedicação a este trabalho.

"O ignorante afirma, o sábio duvida, o sensato reflete."Aristóteles

# Agradecimentos

Para a elaboração deste trabalho, tive a colaboração direta ou indireta de várias pessoas, que muito contribuíram para tornar esta tarefa menos difícil e mais prazerosa. Por isso dedico aqui os meus agradecimentos.

Em primeiro lugar agradeço a Deus, que me deu a vida, discernimento e principalmente forças nas horas que estava para fraquejar. O meu reconhecimento, vai para meu marido, Amilton J. Monteiro, por ter estado sempre ao meu lado, me incentivando e me ajudando em tudo que precisei, durante todos esses anos que estamos juntos. Isso foi essencial não somente para a realização deste trabalho, mas para muitas outras conquistas que tive em minha vida. Agradeço, também, aos meus pais que me educaram e me incentivaram a lutar sempre pelos meus objetivos e a todos os membros de minha família que formam a base para minha existência.

Gostaria de destacar especial agradecimento ao meu orientador, Prof. Dr. Luís A. Milan, que teve paciência durante todo processo de desenvolvimento desta Dissertação e não deixou em nenhum momento de me incentivar a concluir este trabalho. Agradeço também a todos os docentes que de alguma forma contribuíram para o desenvolvimento deste trabalho, a todos eles o meu obrigado pelos saberes que me foram transmitidos.

Durante o desenvolvimento deste trabalho, foram-me dadas provas de amizade por colegas que muito me ajudaram e aos quais devo muito. Por isso, o meu agradecimento vai a todos os meus amigos que, de forma direta ou indireta, contribuíram para que este trabalho fosse possível, estes foram muitos e penso que todos sabem o quanto sou agradecida.

Agradeço também a todos os funcionários do Departamento de Estatística desta universidade, em especial a secretária do Programa de Pós-Graduação em Estatística, Maria Isabel de Araújo, pelos bons serviços prestados e pela dedicação a este departamento.

Meu agradecimento vai também a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo suporte financeiro concedido para a execução deste trabalho.

# Resumo

Neste trabalho, analisamos os aspectos práticos de um modelo bayesiano não paramétrico conhecido como modelo de mistura por processo de Dirichlet. Procedemos a um estudo de simulação com o objetivo de investigar a performance do modelo, no que diz respeito à classificação de dados oriundo de populações heterogêneas, em subgrupos (ou componentes). Os dados em cada componente identificado são assumidos terem uma distribuição normal, de forma que os dados de todos os componentes, juntos são assumidos serem originados de uma mistura de distribuições normais. Para verificar este desempenho, procedemos a uma análise para investigar dois aspectos. O primeiro aspecto considerado está relacionado a sensibilidade do modelo, quanto a escolha do parâmetro de locação da distribuição base adotada, normal-gama-invertida, para o processo de Dirichlet, o qual é usado como distribuição *a priori* para o modelo, como em um simples problema de Bayes. O segundo aspecto diz respeito à performance do modelo em relação ao afastamento dos parâmetros, média e variância, das distribuições dos componentes. Os resultados das simulações com estas misturas de distribuições normais, indicam sensibilidade do método para a escolha do parâmetro de locação da distribuição base normal-gama-invertida e também indicam uma boa performance, mesmo quando os componentes com distribuições normais diferem entre si apenas na variabilidade dos dados. Finalmente, aplicamos este método para três conjuntos de dados reais, sendo o último uma aplicação em dados de mistura de modelos de regressão.

**Palavras chaves:** Processo de Dirichlet; modelos com mistura de distribuições; inferência bayesiana não paramétrica; algoritmo Gibbs sampling.

# Abstract

We review the Dirichlet process mixture model and investigate its performance as a classification method. The first aspect considered is its sensibility to the choice of location parameter of the base distribution. The second aspect considers the performance of the model regarding the departure of the parameters of the component distributions. Simulation results with mixture of normal distributions indicate sensibility to location parameters choices, of the base distribution, and good performance even when components with normal distributions differ only in variances. Finally, we apply the method to three data sets.

# Sumário

# Capítulo 1

# Introdução

Em muitas aplicações estatísticas, os dados observados requerem modelos flexíveis, ou seja, modelos que possam representar distribuições que não têm forma padrão como, por exemplo, modelos com mistura de distribuições. Os modelos bayesianos não paramétricos podem ser uma boa alternativa para algumas situações deste tipo, se soubermos como aplicá-los de forma adequada. Neste trabalho, discutimos o modelo de mistura por processo de Dirichlet (DPM, do inglês *Dirichlet process mixture*) como método de classificação de observações, quanto ao componente a qual elas pertencem, em modelos com mistura de distribuições e estimação de densidades destes modelos. Este método consiste em modelar os dados observados usando uma mistura de distribuições usuais (normal, Poisson, etc.) e supõe, inicialmente, um número infinito de componentes para o modelo, ou seja, a suposição inicial é que existem infinitas distribuições compondo a mistura de distribuições.

Esta classe de modelos bayesianos não paramétricos tem base teórica que envolve um processo aleatório chamado processo de Dirichlet (DP, do inglês *Dirichlet Process*). A existência desse processo, assim como algumas propriedades utilizadas para o desenvolvimento do modelo DPM, foram desenvolvidas por Ferguson (1973).

Uma extensão do trabalho de Ferguson (1973) é vista em Antoniak (1974), em que a medida aleatória vista em Ferguson (1973) é descrita como uma mistura de distribuições. A distribuição condicional desta medida aleatória, dadas as observações, é também um processo de Dirichlet que pode ser descrito como uma mistura por processos de Dirichlet, de onde provavelmente se originou o nome *Dirichlet process mixture*. Outra referência que pode ser

consultada, neste contexto, é Ferguson (1974).

O uso do Modelo DPM tornou-se computacionalmente viável com o desenvolvimento das técnicas Markov chain Monte Carlo (MCMC), popularizadas por Gelfand & Smith (1990), que simulam valores das distribuições *a posteriori* para os parâmetros do modelo.

O primeiro algoritmo MCMC usado para ajustar o modelo DPM pode ser visto em Escobar (1994), em que é usada a representação do processo de Dirichlet em esquema de Urnas de Pólya (para mais detalhes a respeito deste modelo de urna, veja Blackwell & MacQueen (1973)) para obter um estimador do vetor de médias de uma mistura de distribuições normais, cuja informação *a priori* é um processo de Dirichlet. Por meio de um estudo de simulação, Escobar (1994) compara esse estimador aos estimadores paramétrico e não paramétrico de Bayes empírico. Uma deficiência da abordagem de Escobar (1994) é que os parâmetros do modelo são atualizados somente quando um componente da mistura é totalmente esvaziado ou quando é criado um novo componente na mistura, com o processo de agrupamento dos dados. Com isso, podem ocorrer dificuldades de convergência na distribuição *a posteriori*.

Uma extensão direta do método desenvolvido por Escobar (1994) é o trabalho de Escobar & West (1995), em que são descritos modelos para estimação de densidade usando mistura por processo de Dirichlet e faz uso de simulação para obter aproximações das distribuições *a posteriori* e preditiva.

Bush & MacEachern (1996) desenvolveram um trabalho em que é feita a atualização dos parâmetros do modelo DPM, condicionados aos componentes identificados pelo método. Esta estratégia pode ajudar a convergência do algoritmo, porém é mais interessante para o caso de modelos DPM conjugados, ou seja, quando existe conjugação entre a distribuição base adotada e a distribuição de cada observação.

O modelo DPM também pode ser usado em problemas de regressão não paramétrica. Nesta área, podemos citar o trabalho de Müller *et al.* (1996), em que são abordados problemas de suavização e ajuste de curva via inferência preditiva. Para esta abordagem, é feita uma generalização do trabalho de Escobar & West (1995) para uma classe de normais multivariadas.

Os métodos com base no algoritmo *Gibbs sampling* podem ser facilmente implementados para modelos em que a distribuição *a priori* é conjugada à verossimilhança dos dados, porém

quando se trabalha com distribuição *a priori* não conjugada, a implementação do algoritmo *Gibbs sampling* requer integração numérica, o que dificulta a aplicação do método. Escobar & West (1994) usam o método Monte Carlo para aproximar os resultados destas integrais, mas isto pode aumentar substancialmente o erro de estimação em alguns casos.

Um método que trata diretamente do caso em que se tem modelos DPM não conjugados, é tratado em MacEachern & Müller (1998), em que é apresentada uma estratégia para solucionar alguns problemas de integração, que podem surgir ao obtermos as probabilidades de transição da cadeia de Markov em um modelo não conjugado. Tal estratégia também pode ser usada para modelos DPM conjugados.

Outro trabalho importante, que também trata de modelos DPM não conjugados, é Neal (2000), onde são apresentadas duas novas classes de métodos usados para gerar valores da distribuição *a posteriori* no modelo DPM. Na primeira abordagem, o algoritmo *Metropolis-Hastings* é usado para atualizar os indicadores de componentes, estes indicadores especificam a que componente está associada cada observação. A outra abordagem é uma extensão do algoritmo *Gibbs sampling* desenvolvido por Escobar (1994). Para obtermos esta generalização, são usados parâmetros auxiliares no processo de estimação dos indicadores de componentes.

Além dos trabalhos já citados anteriormente, outros trabalhos foram desenvolvidos mais recentemente, como por exemplo, Ferreira Da Silva (2007) que aplica este método em classificação de imagens de ressonância magnética do tecido do cérebro humano. Esta classificação é um requisito importante em diagnóstico, planejamento de tratamentos e em neurociência cognitiva. Outro trabalho importante é Shahbaba & Neal (2009), que apresenta um modelo de regressão não linear para classificação, usando mistura por processos de Dirichlet. Nesta abordagem, é estabelecido um relacionamento linear entre a variável resposta e a covariável dentro de cada grupo que compõe a mistura e relacionamento não linear se a mistura contém mais de uma subpopulação com diferentes coeficientes de regressão. Este modelo é aplicado em dois problemas de classificação, sendo um em detecção de classes de sequências de proteínas e o outro em detecção de doença de Parkinson.

O objetivo principal deste trabalho é analisar o modelo DPM quanto a sua precisão na obtenção de estimativas de densidades de modelos com mistura de distribuições. Para isso, analisamos o modelo DPM, explorando aspectos referentes a sua aplicação prática na

identificação de componentes em amostras oriundas de mistura de distribuições normais (entendemos por componentes de uma população, os subgrupos que compõem uma população heterogênea de onde foram extraídos os dados; neste caso, os componentes são os subgrupos que compõem a população total). Verificamos o desempenho do método no que diz respeito a capacidade de agrupamento de dados em componentes e estimação dos parâmetros do modelo de mistura de distribuições. Para avaliar este desempenho, consideramos um modelo de mistura com duas componentes normais, ou seja, um modelo composto por duas distribuições normais de parâmetros distintos. Implementamos o modelo para diversos valores de parâmetros para esta mistura de distribuições para verificarmos como o modelo se comporta a medida que afastamos as médias e as variâncias das duas componentes dessa mistura. Além disso, analisamos a sensibilidade do método quanto à escolha do parâmetro de locação da distribuição normal-gama-invertida, quando esta distribuição é usada como distribuição base do processo de Dirichlet usando como distribuição *a priori* no modelo. Com esta análise, observamos que o método é bastante sensível a escolha do parâmetro da distribuição base normal-gama-invertida e que o método tem um bom desempenho no que diz respeito a capacidade de identificação de componentes no modelo de mistura.

O trabalho está organizado da seguinte maneira. No Capítulo 2, fazemos uma revisão preliminar da teoria necessária para o desenvolvimento deste trabalho, introduzindo o conceito de conjugação, descrevendo dois métodos de simulação de valores de uma distribuição *a posteriori* e apresentando o modelo de mistura finito. No Capítulo 3, fazemos uma descrição geral do modelo DPM e abordamos aspectos relacionados as suas bases teóricas. No Capítulo 4, descrevemos o modelo DPM, considerando que os dados seguem distribuição normal e que a distribuição base é a normal-gama invertida, que pertence a família conjugada à distribuição normal com os parâmetros média ($\mu$) e variância ($\sigma^2$) desconhecidos. No Capítulo 5, fazemos uma análise do modelo DPM conjugado, visto no Capítulo 4, no que diz respeito aos aspectos práticos, tais como sensibilidade a escolha da distribuição base variando o valor do parâmetro de locação da distribuição normal-gama invertida e quanto a precisão do método quando se tem diferentes distâncias entre as médias e variâncias dos componentes da mistura. No Capítulo 6, fazemos a aplicação do modelo visto no Capítulo 4 a três conjuntos de dados reais. No Capítulo 7, fazemos uma discussão dos resultados

obtidos nos Capítulos 5 e 7.

# Capítulo 2

# Fundamentação teórica

Neste Capítulo, apresentamos alguns resultados que são utilizados ao longo deste trabalho. Iniciamos com uma introdução da metodologia bayesiana. Em seguida descrevemos algumas distribuições e finalizamos este Capítulo com a descrição de dois métodos de simulação.

## 2.1 Abordagem bayesiana

A teoria bayesiana foi introduzida por Bayes e Laplace no século XVIII. Do ponto de vista Bayesiano, diferentes graus de incerteza a respeito de quantidades de interesse (ou parâmetro $\boldsymbol{\theta}$, que pode ser um escalar ou vetor da forma $\boldsymbol{\theta}$=$(\theta_1, ..., \theta_n)$, assumindo valores em um espaço $\Theta$) podem ser representados por intermédio dos modelos probabilísticos para os parâmetros do modelo estatístico. Probabilisticamente, a informação que temos a respeito de $\boldsymbol{\theta}$ é resumida por meio de uma distribuição *a priori* ($p(\boldsymbol{\theta})$). Deste modo, podemos atualizar esta informação obtendo uma amostra de uma quantidade aleatória relacionada com este parâmetro e aplicando o teorema de Bayes. Para definirmos a regra de Bayes, considere uma amostra aleatória $(x_1, x_2, ..., x_n)$ da variável aleatória $X$, com $X_1, X_2, ..., X_n$ independentes e identicamente distribuídos conforme a distribuição de $X$. Com isso, o Teorema de Bayes conduz à relação

$$p(\boldsymbol{\theta}|x_1, x_2, ..., x_n) = \frac{\prod_{i=1}^{n} f_X(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_\Theta \prod_{i=1}^{n} f_X(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \tag{2.1}$$

Na equação (2.1), $p(\boldsymbol{\theta}|x_1, x_2, ..., x_n)$ é a distribuição *a posteriori* de $\boldsymbol{\theta}$, considerando a amostra aleatória $(x_1, x_2, ..., x_n)$ da variável aleatória (v.a.) $X$, e $f_X(.|\boldsymbol{\theta})$ é a função de densidade de probabilidade (f.d.p) da distribuição de $X$ ou função de probabilidade (f.p.) para o caso em que a v.a. $\mathbf{X}$ é discreta.

Escrevendo $\mathbf{y} = (y_1, ..., y_m) = (x_{n+1}, ..., x_{n+m})$, com $m \geq 1$ e $n \geq 1$, para denotar observações futuras da v.a. X, podemos descrever a distribuição desses valores futuros condicionados aos dados observados, $\mathbf{x} = (x_1, x_2, ..., x_n)$. Esta distribuição é denominada distribuição preditiva, que pode ser encontrada identificando a densidade conjunta,

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})},$$

em que

$$p(\mathbf{x}) = \int_\Theta \prod_{i=1}^{n} f_X(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{2.2}$$

com isso, podemos escrever

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \frac{\int_\Theta \prod_{i=1}^{n+m} f_X(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_\Theta \prod_{i=1}^{n} f_X(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \\ &= \int_\Theta \prod_{i=n+1}^{n+m} f_X(x_i|\boldsymbol{\theta}) \left( \frac{\prod_{i=1}^{n} f_X(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_\Theta \prod_{i=1}^{n} f_X(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \right) d\boldsymbol{\theta} \\ &= \int_\Theta \prod_{i=n+1}^{n+m} f_X(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \end{aligned}$$

com a expressão de $p(\boldsymbol{\theta}|\mathbf{x})$ dada em (2.1). Assim, temos que,

$$p(\mathbf{y}|\mathbf{x}) = \int_\Theta \prod_{i=1}^{m} f_X(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \tag{2.3}$$

é a densidade da distribuição preditiva de $\mathbf{y}|\mathbf{x}$.

### 2.1.1 Distribuições *a priori* conjugadas

Na análise bayesiana, dada uma função de verosimilhança $f(\mathbf{x}|\boldsymbol{\theta})$, para cada escolha de $p(\boldsymbol{\theta})$ temos integrais como em (2.2) e (2.3). A tratabilidade destas integrais está relacionada com a classe de funções a partir da qual $p(\boldsymbol{\theta})$ é escolhida.

Em termos de proporcionalidade, escrevemos a expressão vista em (2.1) da forma

$$p(\boldsymbol{\theta}|\mathbf{x}) = cf(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \propto f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{2.4}$$

com $c^{-1} = \int_\Theta \prod_{i=1}^n f_X(x_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ sendo denominada constante normalizadora. Assim, podemos escolher $p(\boldsymbol{\theta})$ de modo a ter a mesma "estrutura" de $f(\mathbf{x}|\boldsymbol{\theta})$ quando vista como uma função de $\boldsymbol{\theta}$, ou seja, de modo que a função $f(\mathbf{x}|\boldsymbol{\theta})$ seja diferente desta mesma função vista como uma função de $\boldsymbol{\theta}$ apenas por uma constante. Com isso, garantimos que tanto $p(\boldsymbol{\theta}|\mathbf{x})$ como $p(\boldsymbol{\theta})$ pertencem a mesma família de distribuições. Ou seja, se

$$p(\boldsymbol{\theta}) \in P \Rightarrow p(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \in P.$$

Então dizemos que $P$ é uma família conjugada para a família de distribuições $F = \{f(x|\theta) : \theta \in \Theta\}$.

Para mais detalhes a respeito da teoria bayesiana e conjugação, veja por exemplo, Bickel & Doksum (2001) ou Bernardo & Smith (1994).

## 2.2 Distribuições de probabilidade

Nesta Seção, vamos apresentar algumas distribuições de probabilidades usadas no nosso trabalho e que não são comuns na literatura estatística. As distribuições mais comuns, como a distribuição normal e a distribuição binomial, são assumidas serem conhecidas e são inseridas diretamente sem descrição prévia. Para uma descrição mais detalhada das distribuições de probabilidade paramétricas mencionadas neste trabalho, veja por exemplo Wilks (1962) e Bernardo & Smith (1994).

### 2.2.1 Distribuição t de Student

Uma quantidade aleatória contínua $X$ tem distribuição t de Student com parâmetros $m$, $\lambda$ e $a$, em que $m \in \mathbb{R}$, $\lambda > 0$ e $a > 0$, se sua densidade $St(x|m, \lambda, a)$ pode ser escrita

$$St(x|m, \lambda, a) = \frac{\Gamma(\frac{a+1}{2})}{\Gamma(\frac{a}{2})}\Big(\frac{\lambda}{a\pi}\Big)^{\frac{1}{2}}\Big(1 + \frac{\lambda}{a}(x-m)^2\Big)^{-\frac{a+1}{2}}, \tag{2.5}$$

com $x \in \mathbb{R}$. O parâmetro $a$ é usualmente referido como graus de liberdade desta distribuição. A distribuição t de Student tem a distribuição normal como limite quando temos $a \to \infty$. Usaremos a notação

$$X \sim St(m, \lambda, a)$$

para indicar uma v.a. com distribuição t de Student com parâmetros $m$, $\lambda$ e $a$.

### 2.2.2 Distribuição multinomial

Em um experimento aleatório, cujos resultados possíveis são os eventos $A_1, ..., A_k$, considere que cada evento $A_j$ ocorra com probabilidade $\theta_j$, $j = 1, ..., k$, com $\sum_{j=1}^{k} \theta_j = 1$. Em $n$ repetições independentes do experimento seja $\mathbf{X} = (X_1, ..., X_k)$ definido por

$$X_j = \text{número de ocorrências de } A_j \text{ nas } n\text{repetições}, j = 1, 2, ..., k.$$

Prova-se que a f.p. conjunta $Mu_k\big((x_1, ..., x_k)|n, \boldsymbol{\theta}\big)$ de $\mathbf{X}$ é

$$Mu_k\big((x_1, ..., x_k)|n, \boldsymbol{\theta}\big) = \binom{n}{x_1, x_2, ..., x_k} \theta_1^{x_1} \theta_2^{x_2} ... \theta_k^{x_k} \tag{2.6}$$

se $0 \leq x_j \leq n$ com $\sum_{j=1}^{k} x_j = n$, $n \in \mathbb{N}$, e zero, caso contrário. Esta função de probabilidade caracteriza a distribuição multinomial de $X$, com parâmetros $n$ e $\boldsymbol{\theta}$. Usaremos a notação

$$X \sim Mu_k(n, \boldsymbol{\theta})$$

para indicar uma v.a. com distribuição multinomial com parâmetros $n$ e $\boldsymbol{\theta}$.

#### Exemplo de variável aleatória com distribuição multinomial

Uma urna tem 3 bolas pretas, 2 verdes e 5 brancas. Retiramos 6 bolas com reposição. Qual a probabilidade de sair 2 bolas pretas, 1 bola verde e 3 brancas?

Vamos definir a v.a. $\mathbf{X} = (X_1, X_2, X_3)$, como

$$\begin{aligned} X_1 &= \text{número de ocorrências de bolas pretas} \\ X_2 &= \text{número de ocorrências de bolas verdes} \\ X_3 &= \text{número de ocorrências de bolas brancas} \end{aligned}$$

As probabilidades de ocorrência de bola preta, verde e branca, são dadas, respectivamente, por

$$\begin{aligned} P(\text{bola preta}) &= \frac{3}{10}, \\ P(\text{bola verde}) &= \frac{2}{10} = \frac{1}{5}, \\ P(\text{bola branca}) &= \frac{5}{10} = \frac{1}{2}. \end{aligned}$$

Então,

$$Mu_3((X_1 = 2, X_2 = 1, X_3 = 3)|6, (3/10, 1/5, 1/2)) = \frac{6!}{2!1!3!}(3/10)^2(1/5)^1(1/2)^3 = 0,045.$$

Se $k = 2$ obtemos a distribuição binomial, ou seja, a distribuição multinomial é uma generalização da distribuição binomial.

### 2.2.3 Distribuição beta

Uma v.a. contínua $X$ tem distribuição beta com parâmetros $\nu_1$ e $\nu_2$ ($\nu_1 > 0, \nu_2 > 0$) se sua f.d.p. $Be(x|\nu_1, \nu_2)$, pode ser escrita como,

$$Be(x|\nu_1, \nu_2) = \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} x^{\nu_1 - 1}(1 - x)^{\nu_2 - 1} \tag{2.7}$$

se $0 < x < 1$, e zero, caso contrário.

Prova-se que

$$\frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} = \int_0^1 x^{\nu_1 - 1}(1 - x)^{\nu_2 - 1} dx \tag{2.8}$$

e que a esperança e a variância da v.a. X com distribuição beta com parâmetros $\nu_1$ e $\nu_2$ são dadas, respectivamente, por

$$E[X] = \frac{\nu_1}{\nu_1 + \nu_2} \quad \text{e} \quad Var[X] = \frac{\nu_1 \nu_2}{(\nu_1 + \nu_2)^2(\nu_1 + \nu_2 + 1)}.$$

A integral definida vista no lado direito da equação (2.8) é chamada de função beta, denotada por $B(\nu_1, \nu_2)$. Se em (2.7) $\nu_1 = \nu_2 = 1$, temos a densidade da distribuição uniforme no $[0, 1]$.

A Figura 2.1 mostra os gráficos das densidades da distribuição beta para diferentes valores de parâmetros, onde podemos observar o gráfico da densidade da distribuição uniforme no $[0, 1]$ para o caso de $\nu_1 = \nu_2 = 1$.

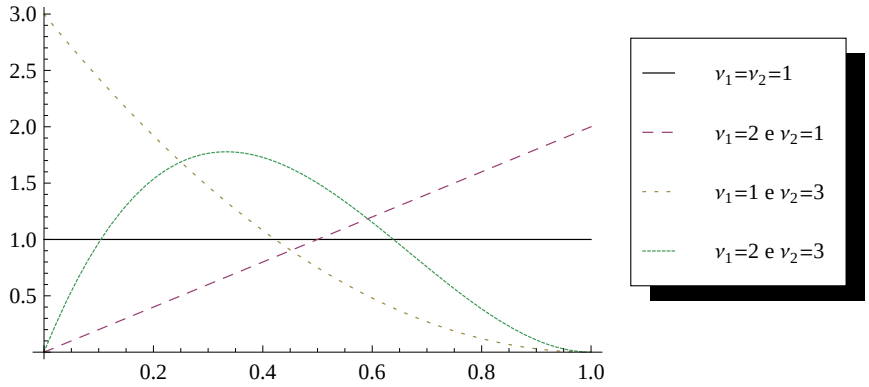


Figura 2.1: Gráficos das densidades da distribuição beta com vetores de parâmetros $(1,1)$, $(2,1)$, $(1,3)$ e $(2,3)$.

### 2.2.4 Distribuição de Dirichlet

O caso $k$ variado para o modelo Beta, cuja f.d.p. é dada em (2.7), é a distribuição de Dirichlet com vetor de parâmetros $\nu = (\nu_1, ..., \nu_{k+1})$, $(\nu_j > 0, j = 1, ..., k+1)$, cuja f.d.p. $Di_k(x_1, ..., x_k|\nu)$, é dada por

$$Di_k(x_1, ..., x_k|\nu) = \frac{\Gamma(\nu_1 + ... + \nu_{k+1})}{\Gamma(\nu_1)...\Gamma(\nu_{k+1})} x_1^{\nu_1 - 1} ... x_k^{\nu_k - 1} (1 - \sum_{j=1}^{k} x_j)^{\nu_{k+1} - 1}, \tag{2.9}$$

que está definida em qualquer ponto do simplex $S_k = \{(x_1, ..., x_k) \in \mathbb{R}^k : 0 \leq x_j \leq 1, j = 1, ..., k, \sum_{j=1}^{k} x_j \leq 1\}$ e assume zero caso contrário. Se $k = 1$ temos $Di_k(x_1|\nu_1, \nu_2)$ que é a densidade da distribuição beta dada em (2.7), com isso (2.9) se reduz a f.d.p. da distribuição beta $Be(x|\nu_1, \nu_2)$.

Prova-se que a média e a variância da $j$-ésima componente do vetor da v.a. com distribuição de Dirichlet são dadas respectivamente por

$$E[X_j] = \frac{\nu_j}{\sum_{j=1}^{k+1} \nu_j} \quad \text{e} \quad Var[X_j] = \frac{E[X_j](1 - E[X_j])}{1 + \sum_{j=1}^{k+1} \nu_j}.$$

Para indicar uma v.a. de dimensão $k$ seguindo uma distribuição de Dirichlet com vetor de parâmetros $\nu$, escrevemos brevemente,

$$\mathbf{X} \sim Di_k(\nu).$$

As Figuras 2.2, 2.3 e 2.4 mostram os gráficos das densidades da distribuição de Dirichlet bidimensional, com vetores de parâmetros $(1, 1, 1)$, $(1, 2, 2)$ e $(3, 2, 2)$.

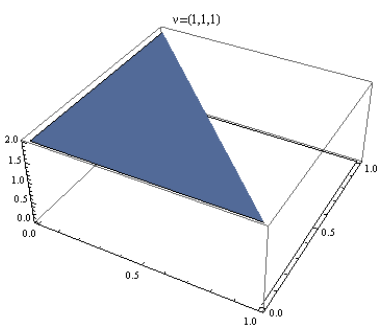


Figura 2.2: Gráfico da densidade da distribuição de Dirichlet com vetor de parâmetros $(1, 1, 1)$.

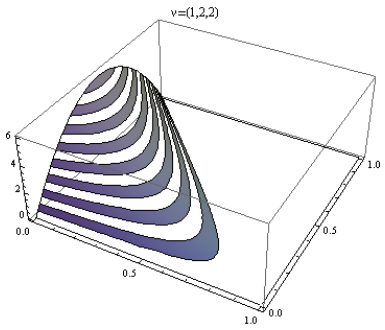


Figura 2.3: Gráfico da densidade da distribuição de Dirichlet com vetor de parâmetros $(1, 2, 2)$.

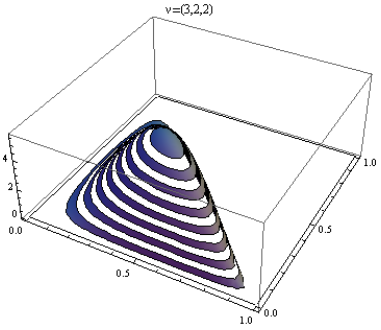


Figura 2.4: Gráfico da densidade da distribuição de Dirichlet com vetor de parâmetros $(3, 2, 2)$.

### 2.2.5 Modelo com mistura de distribuições

Um vetor ou uma variável aleatória $Y$, assumindo valores no espaço $\mathcal{Y}$, tem sua distribuição dada por uma mistura de distribuições se sua f.d.p. pode ser representada (ou f.p. se $\mathcal{Y}$ é um espaço discreto) por

$$f(y|\boldsymbol{\omega}, k) = \sum_{j=1}^{k} \omega_j f_j(y), \tag{2.10}$$

com $y \in \mathcal{Y}$, $k \geq 1$, $\omega_j > 0$ com $\omega_1 + ... + \omega_k = 1$ para $j = 1, ..., k$ e cada $f_j(.) > 0$, para $j = 1, ..., k$, com $\int_{\mathcal{Y}} f_j(y)dy = 1$.

As funções $f_j(.)$ para $j = 1, ..., k$ são chamadas de densidades componentes da mistura e $\omega_j$ para $j = 1, ..., k$ são chamados de proporções ou pesos da mistura.

Nesta dissertação, vamos considerar que $f_j(.)$ para $j = 1, ..., k$ são densidades que pertencem a uma mesma família de distribuições, mas com valores de parâmetros, $\theta_1, ..., \theta_k$, diferentes. Assim podemos escrever a equação dada em (2.10) da forma

$$f(y|\boldsymbol{\theta}, \boldsymbol{\omega}, k) = \sum_{j=1}^{k} \omega_j f(y|\theta_j), \tag{2.11}$$

com $\theta_j$ denotando o parâmetro da $j$-ésima densidade, $f_j(.|\theta_j)$. Aqui, $\theta_j$ pode ser um escalar ou vetor e o vetor $\boldsymbol{\theta}$ representa a coleção de parâmetros $\theta_j$, $j = 1, ..., k$. Se $k = \infty$, dizemos que a mistura de distribuições é infinita, caso contrário temos uma mistura de distribuições finita.

Para indicar uma v.a. $X$ seguindo uma mistura de $k$ distribuições, em que cada uma é representada pela densidade $f_j(x|\theta)$ para $j = 1, ..., k$, escrevemos, brevemente,

$$\mathbf{X} \sim \sum_{j=1}^{k} \omega_j f_j(x|\theta).$$

A estrutura da mistura eventualmente surge devido à perda da origem de cada observação em relação ao componente (ou subgrupo da população) ao qual ela pertence. Assim, observações desta população pertencem a uma amostra que pode ser também heterogênea, assim podemos dizer que a amostra pode ser composta por subgrupos aos quais vamos nos referir como componentes da amostra. Se estamos interessados em classificar os dados segundo esses componentes, para cada observação $y_i$ de uma amostra $(y_1, y_2, ..., y_n)$ de $Y$, podemos associar uma variável $Z_i = (Z_{i1}, Z_{i2}, ..., Z_{ik})$, que indica o $j$-ésimo componente ao qual pertence esta observação. Neste contexto, vamos dizer que cada densidade $(f_j(.|\theta_j))$ em (2.11), é denominada $j$-ésima densidade componente da mistura de distribuições, além de ser a densidade que representa a distribuição do $j$-ésimo componente (ou subgrupo) da população. Assim, a variável indicadora, $Z_{ij}$, é escrita como

$$Z_{ij} = \begin{cases} 1, & \text{se, } \quad y_i \quad \text{pertence ao componente } j \\ 0, & \text{se,} \qquad\qquad\quad \text{caso contrário} \end{cases}$$

de modo que, $\sum_{j=1}^{k} Z_{ij} = 1$. Com isso, $\omega_{ij} = p(Z = z_j)$ é a probabilidade da observação $y_i$ ser distribuída de acordo com a distribuição representada pela função de densidade $f(.|\theta_j)$ e dizemos que $\omega_{ij}$ é o "peso de associação" da observação $y_i$ à $j$-ésima densidade componente da mistura de distribuições. Usando a regra de Bayes, vamos computar o peso de associação

de cada observação $y_i$ a uma densidade componente $j$, dado os parâmetros $\boldsymbol{\theta}$, como

$$\omega_{ij} = p(Z_{ij} = 1|y_i, \boldsymbol{\theta}) = \frac{\omega_j f(y_i|\theta_j)}{\sum_{m=1}^{k} \omega_m f(y_i|\theta_m)} \tag{2.12}$$

com $1 \leq j \leq k$ e $1 \leq i \leq n$. Assim, o vetor $Z_i = (Z_{i1}, Z_{i2}, ..., Z_{ik})$, tem distribuição multinomial de parâmetros 1 e $(\varphi_{i1}, \varphi_{i2}, ..., \varphi_{ik})$, com $\varphi_{ij} = \frac{\omega_j f(y_i|\theta_j)}{\sum_{j=1}^{k} \omega_j f(y_i|\theta_j)}$. A densidade condicional de $Y_i$ dado que a variável aleatória $Z_{ij} = 1$ é dada por

$$f_{Y_i}(y_i|z_{ij} = 1) = f(y_i|\theta_j).$$

Dada uma amostra aleatória $\mathbf{y} = (y_1, y_2, ..., y_n)$ de $Y$ cuja distribuição é representada pela f.d.p. dada em (2.11), a função de verossimilhança dos dados é dada por

$$L(\boldsymbol{\theta}, \boldsymbol{\omega}|\mathbf{y}) = \prod_{i=1}^{n} \left( \sum_{j=1}^{k} \omega_j f(y_i|\theta_j) \right). \tag{2.13}$$

Com a inclusão da variável indicadora $Z_{ij}$ a função de verossimilhança fica simplificada,

$$L(\boldsymbol{\theta}, \boldsymbol{\omega}|\mathbf{y}, z) = \prod_{i=1}^{n} \prod_{j=1}^{k} \Big[ \omega_j f(y_i|\theta_j) \Big]^{z_{ij}}.$$

Vamos denotar um conjunto de $n$ vetores indicadores $Z_i = (Z_{i1}, Z_{i2}, ..., Z_{ik})$, $i = 1, 2, ...n$, por $Z_n$. Deste modo a especificação da distribuição *a priori* pode ser escrita como

$$p(\boldsymbol{\theta}, \boldsymbol{\omega}, Z_n, k) = p(\boldsymbol{\theta}|\boldsymbol{\omega}, z_n, k) p(\boldsymbol{\omega}|z_n, k) p(z_n|k) p(k),$$

que combinada com a verossimilhança dos dados fornece a densidade que representa a distribuição *a posteriori*

$$p(\boldsymbol{\theta}, \boldsymbol{\omega}, z_n, k|\mathbf{y}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \Big[ \omega_j f(y_i|\theta_j) \Big]^{z_{ij}} p(\boldsymbol{\theta}, \boldsymbol{\omega}, z_n, k).$$

Uma abordagem bayesiana não paramétrica para estimação de densidade de modelos com mistura de distribuições é o tema principal desse trabalho e será tratado nos Capítulos seguintes.

Outra forma de classificarmos as observações é por meio da inclusão de identificadores de componentes, $z_i$'s ao modelo, definidos, tais que

$$z_i = j \text{ se } y_i \text{ pertence ao componente } j \ , \text{ para } j = 1, ..., k \text{ e } i = 1, ..., n. \tag{2.14}$$

Neste trabalho, esta é a notação utilizada para identificar componentes (ou subgrupos) na amostra. Estes componentes serão uma estimativa dos componentes da população de onde foram extraídos estes dados.

Para uma análise mais detalhada a respeito de modelos com mistura de distribuições finita e também aplicações, veja, por exemplo, Titterington *et al.* (1985).

## 2.3 *Gibbs sampling*

O método *Gibbs sampling* pertence a classe de métodos denominada Markov Chain Monte Carlo (MCMC), que é sustentada pelas propriedades da cadeia de Markov (CM). Este método explora as distribuições condicionais completas *a posteriori* por meio de algoritmo iterativo que foi proposto inicialmente por Geman & Geman (1984) para o problema de reconstrução de imagens. Após a publicação do trabalho de Gelfand & Smith (1990), esta técnica tornou-se mais conhecida e utilizada em problemas práticos da inferência estatística. Um exemplo é Sheng *et al.* (2005) em que esta técnica é empregada no campo da bioinformática.

Na obtenção de densidades marginais, muitas vezes o cálculo de integrais pode não ser trivial ou até mesmo analiticamente impossível, mas se as distribuições condicionais completas forem conhecidas, empregamos o método *Gibbs sampling* para obtermos uma aproximação da densidade conjunta, evitando cálculos muito complicados em obtenção de densidades marginais.

Se o objetivo é obter amostras de um conjunto de variáveis aleatórias $(\theta_1, \theta_2, ..., \theta_p)$, cujas distribuições condicionais completas $p(\theta_r|\theta_j; j \neq r)$ (para $r = 1, ...p$) sejam conhecidas, especificando valores iniciais $(\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_p^{(0)})$ para o contador de iterações da CM e $t = 0$, o algoritmo *Gibbs sampling* pode ser representado pelo seguinte esquema,

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

A medida que t cresce ($t \to \infty$) a distribuição de $(\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_p^{(t)})$ converge para $p(\theta_1, \theta_2, ..., \theta_p)$. Mais detalhes a respeito do método *Gibbs sampling*, veja, por exemplo, Geman & Geman (1984).

## 2.4 Algoritmo *Metropolis-Hastings*

O algoritmo de *Metropolis-Hastings* (MH) também pertence a classe de métodos MCMC e permite simular valores $\theta$ de uma função de densidade de probabilidade $p(\boldsymbol{\theta})$. A idéia geral do algoritmo MH consiste em gerar valores $\theta^*$ a partir de uma distribuição proposta $q(.|.)$ e estes valores são aceitos com uma dada probabilidade. A cada tempo $t > 0$ um novo candidato é gerado. O valor gerado da distribuição proposta será aceito e adicionado à sequência $\theta^{(t+1)} = \theta^*$, com probabilidade dada por,

$$\alpha(\theta, \theta^*) = \min\{1, \frac{p(\theta^*)q(\theta|\theta^*)}{p(\theta)q(\theta^*|\theta)}\},$$

se, $p(\theta)q(\theta^*|\theta) > 0$ e $\alpha(\theta, \theta^*) = 1$ se, caso contrário.

O agoritmo MH pode ser descrito como segue,

1. Inicializar o contador de iterações $t = 0$ e atribuir um valor inicial arbitrário a $\theta^{(0)}$.

2. Gerar um novo $\theta^*$ da distribuição $q(.|\theta)$.

3. Calcular a probabilidade de aceitação $\alpha(\theta, \theta^*)$.

4. Gerar um valor $u$ de uma distribuição Uniforme$(0, 1)$.

5. Se $u < \alpha$ então o novo valor é aceito e faz-se $\theta^{(t+1)} = \theta^*$, caso contrário faz-se $\theta^{(t+1)} = \theta^t$.

6. Incrementar $t$, $t = t + 1$.

7. Voltar ao passo 2 e continuar as iterações até atingir um tamanho L suficientemente grande para considerar convergência.

Para o processo inferencial, desconsiderar os M primeiros valores gerados e obter uma amostra de tamanho L-M.

Em aplicações bayesianas, a distribuição de interesse é a distribuição *a posteriori*.

O algoritmo *Gibbs sampling*, visto anteriormente, pode ser entendido como um caso particular do algoritmo *Metropolis-Hastings*, em que a probabilidade de aceitação é sempre 1. Uma diferença dentre estes dois algoritmos é de que no algoritmo *Gibbs sampling* cada passo é univariado, ou seja, é gerada uma variável de cada vez para compor o vetor que será incluído na sequência com probabilidade 1. Com o algoritmo *Metropolis-Hastings* podemos gerar um conjunto de v.a. que poderá ser aceito ou não com uma dada probabilidade, ou seja, o movimento do algoritmo é multivariado. Uma desvantagem do algoritmo *Gibbs sampling* é de que as v.a. geradas são correlacionadas, este problema pode ser solucionado adicionando "saltos" ao algoritmo, descartando alguns vetores ou valores gerados a cada número determinado de gerações.

Para maiores detalhes sobre este algoritmo veja, por exemplo, Chib & Greenberg (1995).

# Capítulo 3

# Metodologia Bayesiana não Paramétrica

Modelos bayesianos não paramétricos são modelos de probabilidade em espaços funcionais. Este tipo de modelo é usado para evitar dependência excessiva a pressupostos feitos a respeito dos parâmetros, para tornar modelos paramétricos mais robustos e para definir análise de sensibilidade e diagnóstico de modelos paramétricos, incorporando-os em um abrangente e mais flexível modelo não paramétrico. Uma referência que fornece mais informações a respeito deste tema, é Quintana & Müller (2004).

Neste Capítulo, apresentamos um modelo bayesiano não paramétrico conhecido como modelo de mistura por processo de Dirichlet, cuja estratégia computacional utilizada no processo de inferência é construída com base no esquema de urna de Pólya de Blackwell & MacQueen (1973).

## 3.1 Processo de Dirichlet

Ferguson (1973) define um processo aleatório chamado processo de Dirichlet (DP, do inglês *Dirichlet Process*) e mostra propriedades importantes deste processo. Neste trabalho, apresentamos uma definição simplificada do processo de Dirichlet de Ferguson (1973) e também algumas propriedades deste processo que são necessárias para o desenvolvimento deste trabalho. Para mais detalhes sobre o assunto tratado neste Capítulo, podemos consultar, por exemplo, Ferguson (1974), Ghosh & Ramamoorthi (2003) e Teh (2003).

O processo de Dirichlet é um caso particular de um processo estocástico que pode ser in-

terpretado como uma distribuição de probabilidade. Este processo é utilizado em um modelo bayesiano não paramétrico conhecido como modelo de mistura por processo de Dirichlet.

Nesta abordagem bayesiana não paramétrica, utilizamos o processo de Dirichlet como distribuição *a priori*, especificando uma distribuição *a priori* sobre o espaço de medidas de probabilidade, definidas em um espaço mensurável $(\Theta, \mathbb{A})$, sendo $\Theta$ um conjunto não enumerável, podendo ser, por exemplo, o conjunto dos números reais, e $\mathbb{A}$ é uma $\sigma$-álgebra sobre o conjunto $\Theta$.

Para definir o processo de Dirichlet, consideremos $A = \{A_1, A_2, ..., A_k\}$, $k = 1, 2, ...$, uma partição mensurável qualquer de um espaço contínuo $\Theta$. Com isso, um processo estocástico $G$, indexado pelos elementos da partição particular $A$, é um processo de Dirichlet sobre $(\Theta, \mathbb{A})$ com parâmetro $G_0$, se para qualquer partição de $\Theta$ o vetor aleatório $(G(A_1), ..., G(A_k))$, tem distribuição de Dirichlet com vetor de parâmetros $(G_0(A_1), ..., G_0(A_k))$. Os $A_j$'s podem ser, por exemplo, intervalos disjuntos dos números reais $(\mathbb{R})$. Neste caso, a função de distribuição $G$ associa, a cada intervalo do conjunto $\mathbb{R}$ um valor no intervalo $[0, 1]$, de modo que a soma desses valores para todo conjunto $\mathbb{R}$ somam 1. O parâmetro $G_0$ é uma distribuição definida no mesmo espaço da $G$ a qual vamos nos referir como distribuição base do processo de Dirichlet.

Temos que a função de distribuição $G$ é uma variável aleatória, pois ela associa cada elemento de uma partição aleatória do espaço $\Theta$, a um valor no intervalo $[0, 1]$. Assim, a variável aleatória $G$ assume valores no conjunto de todas as possíveis medidas de probabilidade, que são definidas no espaço $(\Theta, \mathbb{A})$, como pode ser visto em Antoniak (1974).

Para darmos uma definição mais restrita do processo de Dirichlet, vamos incluir um novo parâmetro $\alpha > 0$ ao processo, de forma a podermos escrever que, se

$$\big(G(A_1), ..., G(A_k)\big) \sim Di_{k-1}\big(\alpha G_0(A_1), \alpha G_0(A_2), ..., \alpha G_0(A_k)\big) \tag{3.1}$$

então a função de distribuição $G$ segue um processo de Dirichlet com parâmetros $\alpha$ e $G_0$. Vamos escrever brevemente

$$G \sim DP(\alpha, G_0)$$

para indicar que a função de distribuição $G$ é definida por um processo de Dirichlet com

parâmetros $\alpha$ e $G_0$, com $G_0$ sendo uma função de distribuição definida no mesmo espaço da $G$.

Além de provar a existência do processo de Dirichlet, Ferguson (1973) também mostra que as propriedades válidas para a distribuição de Dirichlet, também valem para o processo de Dirichlet. Usando este fato, vamos discutir agora, algumas propriedades do DP que são essenciais para o desenvolvimento deste trabalho.

Para verificarmos as propriedades 3.1, 3.2, 3.3 e 3.4, vamos considerar uma função de distribuição $G$ definida no espaço contínuo $\Theta$, seguindo um processo de Dirichlet, e uma partição mensurável qualquer $B = \{B_1, ..., B_k\}$ $(k = 1, 2, ...)$ deste espaço.

**Propriedade 3.1.** *A esperança de uma distribuição $G$ dada por um processo de Dirichlet, é uma distribuição de probabilidade.*

Considerando os elementos da partição particular $B = \{B_1, B_2, ..., B_k\}$, temos que a esperança da $j$-ésima componente do vetor $(G(B_1), G(B_2), ..., G(B_k))$ com distribuição de Dirichlet é dada por

$$\begin{aligned} E[G(B_j)] &= \frac{\alpha G_0(B_j)}{\alpha G_0(\Theta)} \\ &= \frac{G_0(B_j)}{G_0(\Theta)} \\ &= G_0(B_j) \end{aligned}$$

para $j = 1, 2, ..., k$ $(k = 1, 2, ...)$. Deste modo, para todo $A \in \mathbb{A}$ que satisfaz a condição (3.1), $E[G(A_j)] = G_0(A_j)$, $A_j \in A$ para $j = 1, 2, ..., k$. Então dizemos que a distribuição $G_0$ é a esperança do processo de Dirichlet. Deste modo, escrevemos

$$E[G] = G_0.$$

**Propriedade 3.2.** *Se $G \sim DP(\alpha, G_0)$, temos que o parâmetro $\alpha$ é dito ser um parâmetro de concentração deste processo, pois controla a dispersão das distribuições obtidas por meio do DP em torno da distribuição base $G_0$.*

Usando as propriedades da distribuição de Dirichlet, temos que,

$$\begin{aligned} Var[G(B_j)] &= \frac{E[G(B_j)]\Big(1 - E[G(B_j)]\Big)}{(\alpha+1)} \\ &= \frac{G_0(B_j)\Big(1 - G_0(B_j)\Big)}{(\alpha+1)} \end{aligned}$$

para $j = 1, 2, ..., k$ $(k = 1, 2, ...)$.

O parâmetro $\alpha$ pode ser visto como uma variação inversa, ou seja, quanto maior seu valor, menor será a variabilidade em torno de $G_0(B_j)$. Assim,

$$\text{se } \alpha \to \infty \Rightarrow G(B_j) \to G_0(B_j).$$

Ou seja, a medida que $\alpha$ cresce existe uma convergência pontual em distribuição de $G(B_j)$ para $G_0(B_j)$, $B_j \in B$ para $j = 1, 2, ..., k$, para todo $B \in \mathbb{A}$ satisfazendo a condição (3.1). Por outro lado, quanto menor $\alpha$, maior será a variabilidade em torno de $G_0(B_j)$, ou seja, o processo ira se concentrar menos em torno de $G_0(B_j)$ para $j = 1, 2, ..., k$ $(k = 1, 2, ...)$. Vale ressaltar que este fato não implica que, $G \to G_0$ quando $\alpha \to \infty$, mas apenas que $G(B_j) \to G_0(B_j)$ quando $\alpha \to \infty$, para $j = 1, 2, ..., k$ $(k = 1, 2, ...)$..

**Propriedade 3.3.** *Se a função de distribução $G$ segue um processo de Dirichlet, com vetor de parâmetros $(\alpha, G_0)$ e $\{\theta_1, ..., \theta_n\}$ é um conjunto formado por realizações de uma variável aleatória que segue distribuição $G$, temos que $G$ dado este conjunto de observações segue ainda um processo de Dirichlet com parâmetros $G^*$ e $\alpha^*$, em que*

$$G^*(B_j) = \frac{\alpha G_0(B_j)}{(\alpha+n)} + \frac{n_j}{(\alpha+n)} \text{ e } \alpha^* = \alpha + n. \tag{3.2}$$

*Para todo $B_j$ que compõem o vetor $B$ com $B = (B_1, B_2, ..., B_k) \in \mathbb{A}$ satisfazendo a condição* (3.1).

Como a função de distribuição $G$ está definida em $\Theta$, qualquer realização de uma variável aleatória com distribuição $G$ pertence a $\Theta$, ou seja, $\theta_i \in \Theta$ para $i = 1, ..., n$. Vamos considerar ainda, a distribuição *a priori* $G \sim DP(\alpha, G_0)$. Com isso, tomemos $B_1, B_2, ..., B_k \in \mathbb{A}$, $k \geq 1$, satisfazendo a condição (3.1). Ou seja,

$$\Big(G(B_1), ..., G(B_k)\Big) \sim Di_{k-1}\Big(\alpha G_0(B_1), ..., \alpha G_0(B_k)\Big),$$

com $B_1, B_2, ..., B_k$ formando uma partição de $\Theta$. Agora seja $n_j$, $j = 1, ..., k$, o número de vezes que ocorre $\theta_i \in B_j$, $i = 1, ..., n$, temos que

$$\sum_{j=1}^{k} n_j = n \quad e \quad \sum_{j=1}^{k} G(B_j) = 1,$$

com $n_1, ..., n_k$ independentes. Deste modo,

$$\Big(n_1, ..., n_k\Big) \sim Mu_k\Big(n, G(B_1), ..., G(B_k)\Big).$$

Então, para $i = 1, ..., n$, temos

$$p(\theta_i \in B_j | G) = \prod_{j=1}^{k} G(B_j)^{\delta_{\theta_i}(B_j)}$$

em que,

$$\delta_{\theta_i}(B_j) = \begin{cases} 1 & \text{se} \quad \theta_i \in B_j \\ 0 & \text{se} \quad \theta_i \notin B_j. \end{cases}$$

Além disso,

$$\begin{aligned} p\Big(\theta_1, ..., \theta_n | G(B_1), ..., G(B_k)\Big) &= \prod_{i=1}^{n} p\Big(\theta_i | G(B_1), ..., G(B_k)\Big) \\ &= \prod_{j=1}^{k} \prod_{i=1}^{n} G(B_j)^{\delta_{\theta_i}(B_j)} \\ &= \prod_{j=1}^{k} G(B_j)^{\sum_{i=1}^{n} \delta_{\theta_i}(B_j)} \end{aligned}$$

em que, $\theta_i \in B_j$ para algum $j = 1, ..., k$.

A distribuição condicional de $\Big(G(B_1), ..., G(B_k)\Big)$ dado $(\theta_1, ..., \theta_n)$ é facilmente encontrada devido a conjugação entre a distribuição de Dirichlet e a distribuição multinomial, ou seja

$$\begin{aligned} p\Big(G(B_1), ..., G(B_k) | \theta_1, ..., \theta_n\Big) &\propto p\Big(\theta_1, ..., \theta_n | G(B_1), ..., G(B_k)\Big) p\Big(G(B_1), ..., G(B_k)\Big) \\ &\propto \prod_{j=1}^{k} G(B_j)^{\sum_{i=1}^{n} \delta_{B_j}(\theta_i)} \prod_{j=1}^{k} G(B_j)^{\alpha G_0(B_j) - 1} \\ &\propto \prod_{j=1}^{k} G(B_j)^{\alpha G_0(B_j) + \sum_{i=1}^{n} \delta_{\theta_i}(B_j) - 1}. \end{aligned}$$

Então,

$$\Big(G(B_1),...,G(B_k)\Big)|\theta_1,...,\theta_n \sim Di_{k-1}\Big(\alpha G_0(B_1)+\sum_{i=1}^{n}\delta_{B_1}(\theta_i),...,\alpha G_0(B_k)+\sum_{i=1}^{n}\delta_{B_k}(\theta_i)\Big). \tag{3.3}$$

Se $n_j = \sum_{i=1}^{n}\delta_{\theta_i}(B_j)$ é o número de vezes que ocorre $\theta_i$ em $B_j$, a expressão (3.3) pode ser reescrita como

$$\Big(G(B_1),...,G(B_k)\Big)|\theta_1,...,\theta_n \sim Di_{k-1}\Big(\alpha G_0(B_1)+n_1,...,\alpha G_0(B_k)+n_k\Big). \tag{3.4}$$

Vamos considerar uma distribuição $G^*$ e um escalar $\alpha^*$, tomados de tal forma que $\alpha^* G^*(B_j) = \alpha G_0(B_j) + n_j$. Além disso, tomemos

$$\begin{aligned} G^*(B_j) &= E[G(B_j)|\theta_1,...,\theta_n] \\ &= \tfrac{\alpha G_0(B_j)+n_j}{(\alpha G_0(\Theta)+n)} \\ &= \tfrac{\alpha G_0(B_j)+n_j}{(\alpha+n)} \\ &= \tfrac{\alpha G_0(B_j)}{(\alpha+n)} + \tfrac{n_j}{(\alpha+n)},\ \text{para } j=1,...,k \end{aligned} \tag{3.5}$$

e

$$\alpha^* = \alpha + n.$$

Com isso, a Equação (3.4), pode ser reecrita da forma

$$\Big(G(B_1),...,G(B_k)\Big)|\theta_1,...,\theta_n \sim Di_{k-1}\Big(\alpha^* G^*(B_1),...,\alpha G^*(B_k)\Big). \tag{3.6}$$

Uma vez que a expressão (3.6) vale para qualquer partição finita e mensurável de $\Theta$, a distribuição de $G$ dado um conjunto de observações $\{\theta_1,...,\theta_n\} \subseteq \Theta$, é também um processo de Dirichlet com parâmetros $G^*$ e $\alpha^*$. Ou seja,

$$G|\theta_1,...,\theta_n \sim DP\big(G^*,\alpha^*\big). \tag{3.7}$$

Como pode ser observado na expressão (3.5), a distribuição $G^*(B_j)$ é uma média ponderada entre $G_0(B_j)$ e a distribuição empírica $n_j = \frac{\sum_{i=1}^{n}\delta_{\theta_i}(B_j)}{n}$. A distribuição $G_0$ tem seu peso controlado pelo valor de $\alpha$ e a distribuição empírica tem seu peso controlado pelo número de observações $n$. Assim, $\alpha$ pode ser interpretado como uma medida de intensidade associada à distribuição base $G_0$.

**Propriedade 3.4.** *O parâmetro $G^*$ da distribuição de $G$ dado um conjunto de realizações de uma variável aleatória que segue distribuição $G$, vista na Propriedade* (3.5)*, também é a distribuição de $\theta_{n+1}$ dado este conjunto de observações.*

Considerando ainda que,

$$\begin{aligned} G &\sim DP(\alpha, G_0) \\ \theta_i|G &\sim G \end{aligned}$$

para $i = 1, ..., n$. Então, a probabilidade de $\theta_{n+1} \in B_j$, $j = 1, ..., k$, condicionada aos valores $\theta_1, ..., \theta_n$ e $G$, é dada por,

$$\begin{aligned} p(\theta_{n+1} \in B_j|\theta_1, ..., \theta_n) &= \int p(\theta_{n+1} \in B_j|G) dP_{_G}(G|\theta_1, ..., \theta_n) \\ &= \int G(B_j) dP_{_G}(G|\theta_1, ..., \theta_n) \\ &= E[G(B_j)|\theta_1, ..., \theta_n] \\ &= \frac{\alpha G_0(B_j)}{\alpha + n} + \frac{\sum_{i=1}^{n} \delta_{\theta_i}(B_j)}{\alpha + n}, \end{aligned}$$

que é a distribuição $G^*$ vista na Equação (3.5). Com isso,

$$\theta_{n+1}|\theta_1, ..., \theta_n \sim \frac{\alpha G_0}{\alpha + n} + \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}. \tag{3.8}$$

em que $\delta_\theta$ é uma distribuição de massa no ponto $\theta$.

Se $\alpha \to 0$, a distribuição de $\theta_{n+1}$ dado $\theta_1, ..., \theta_n$ é dada apenas pela distribuição empírica, por outro lado, a medida que o número de observações aumenta, $n \gg \alpha$, esta distribuição é dominada pela distribuição empírica que, por sua vez, é uma boa aproximação para a distribuição $G^*$. Isto fornece uma propriedade de consistência para o DP (Teh, 2003).

Blackwell (1973) e Antoniak (1974) mostram que valores gerados de um processo de Dirichlet são discretos, isto é, existe probabilidade não nula de que sejam gerados valores pontuais ao contrário de modelos contínuos em que a medida de probabilidade de um ponto é sempre zero. Portanto, dada uma amostra $\theta_1, ..., \theta_n$ extraída de uma distribuição $G(.)$, existe probabilidade positiva de existirem valores coincidentes, isto pode ser observado por meio da equação (3.8), em que um novo valor de $\theta$ é gerado de uma distribuição base

$G_0$ com probabilidade proporcional a $\alpha/(\alpha+n)$ ou assume um dos valores anteriores com probabilidade igual a $1/(\alpha+n)$.

O processo de Dirichlet tem várias representações; dentre elas, pode ser citada o processo de restaurante Chinês, (veja por exemplo, Blei *et al.* (2010)), a construção stick-breaking (veja por exemplo, Ishwaran & James (2001)) e o esquema de Urna de Pólya de Blackwell & MacQueen (1973). O processo de Dirichlet pode também ser obtido generalizando uma de suas representações. Neste trabalho, a representação considerada é o esquema de Urna de Pólya, proposto por Blackwell & MacQueen (1973).

## 3.2 O esquema de Urna de Pólya

O esquema de urna de Pólya é um procedimento em que, a cada passo uma bola é extraída de uma urna e em seguida é devolvida com outra da mesma cor. Se considerarmos uma urna com $\alpha$ bolas, sendo que $\alpha m_j$ ($\alpha m_j$ um número inteiro positivo) são da cor $j$, $j = 1, ..., k$, e $p(X_i = j)$ a probabilidade de se obter uma bola de cor $j$ no passo $i$, $i = 1, 2, ...$, então,

$$\begin{aligned} p(X_1 = j) &= \frac{\alpha m_j}{\sum_{i=1}^{n} \alpha m_i} = \frac{\alpha m_j}{\alpha} = m_j \\ p(X_2 = j|x_1) &= \frac{\alpha m_j + \delta_{x_1}}{\alpha+1} \\ p(X_3 = j|x_1, x_2) &= \frac{\alpha m_j + \delta_{x_1} + \delta_{x_2}}{\alpha+1} \\ p(X_{N+1} = j|x_{1:N}) &= \frac{\alpha m_j + \sum_{i=1}^{N} \delta_{x_i}}{\alpha+N}, \end{aligned} \tag{3.9}$$

em que $\delta_{x_i}$ é um ponto de massa dando probabilidade 1 para $x_i$ se $x_i = j$.

Podemos observar por meio das expressões (3.9) e (3.8) a equivalência do DP com o esquema de Urna de Pólya. Um método alternativo para a obtenção do DP é considerar o limite das cores no esquema de Urna de Pólya a um espaço contínuo. Este procedimento pode ser visto em Blackwell & MacQueen (1973) e forma a base dos algoritmos para realização de inferência sobre o DP, que são tratados neste trabalho.

## 3.3 Modelos de mistura por Processos de Dirichlet

Um modelo de mistura por processos de Dirichlet (ou modelos DPM, do inglês *Dirichlet Process Mixture Model*) é um modelo bayesiano não paramétrico em que se admite

um processo de Dirichlet, em vez de uma distribuição paramétrica *a priori*. Escobar (1994) introduziu métodos MCMC em versões simplificadas para modelos de mistura por DP, estendendo depois para estimação de densidade de mistura de distribuições normais uni-variadas em Escobar & West (1995), em que a distribuição base é de uma família conjugada para verossimilhança dos dados.

Para descrever o modelo DPM, vamos considerar inicialmente um vetor de variáveis aleatórias permutáveis $\mathbf{Y} = (Y_1, ..., Y_n)$ e $\mathbf{y} = (y_1, ..., y_n)$ um vetor de observações de $\mathbf{Y}$, de modo que

$$Y_i \sim \sum_{j=1}^{k} w_j F(y_i|\phi_j), \tag{3.10}$$

com $i = 1, ..., n$, em que $F(.|\phi)$ é uma densidade ou distribuição de probabilidade de uma família de distribuições amostrais e $\phi$ assume valores em um conjunto $\{\phi_1, ..., \phi_n\}$, podendo ser um escalar ou um vetor. Neste caso, temos que a variável aleatória $Y$ segue um modelo de mistura finito, como visto no Capítulo 2.2.5, em que $w_j$, $j = 1, ..., k$, $k \leq n$, é a probabilidade da $i$-ésima observação ter distribuição cuja densidade é a $j$-ésima densidade componente desta mistura distribuições. Neste caso, podemos assumir uma distribuição *a priori* Dirichlet para $w_j$, $j = 1, ..., k$, e supor os parâmetros $\phi_i$'s independentes com distribuição *a priori* $G_0$. Incluindo os identificadores de componentes, $z_i$'s, tais que

$$z_i = j \text{ se } y_i \text{ pertence ao componente } j \text{ , para } j = 1, ..., k \text{ e } i = 1, ..., n, \tag{3.11}$$

podemos escrever o modelo visto em (3.10) de forma hierárquica, de modo que,

$$\begin{aligned} Y_i|z_i, \phi_i &\sim F(\phi_{z_i}) \\ z_i|w_i, ..., w_k &\sim Mu_k(w_i, ..., w_k) \\ w_i, ..., w_k &\sim Di_k(\alpha/k, ..., \alpha/k) \\ \phi_i &\overset{\text{iid}}{\sim} G_0(\phi), \end{aligned} \tag{3.12}$$

Neste modelo, a notação $Y_i|z_i, \phi_i \sim F(\phi_i)$ indica que $Y_i$ dado $z_i$ e $\phi_i$ segue uma distribuição paramétrica da forma $F(.)$, esta distribuição pode ser, por exemplo, a distribuição normal. Integrando sobre a distribuição *a priori* Dirichlet (veja por exemplo Neal, 2000), eliminamos

a proporção da mistura, $w_i$, e obtemos a distribuição condicional para $z_i$, ou seja,

$$p(z_i = z|z_1, ..., z_{i-1}) = \frac{n_{iz} + \alpha/k}{i - 1 + \alpha}, \tag{3.13}$$

em que $n_{iz}$ indica o número de elementos atribuídos previamente ao componente $z$. Como podemos observar pela equação (3.13), a medida que $n_{iz}$ aumenta a probabilidade de um novo elemento ser atribuído ao componente $z$ também aumenta. Tomando o limite quando $k$ vai para o infinito ($k \to \infty$) em (3.13) obtemos

$$\begin{aligned} p(z_i = z|z_1, ..., z_{i-1}) &\to \tfrac{n_{iz}}{i-1+\alpha} \\ p(z_i \neq z_j \text{ para todo } j < i|z_1, ..., z_{i-1}) &\to \tfrac{\alpha}{i-1+\alpha}. \end{aligned} \tag{3.14}$$

Fazendo $\theta_i = \phi_{z_i}$, temos que a probabilidade condicional para $\theta_i$ é similar ao modelo visto em (3.8), ou seja,

$$\theta_i|\theta_1, ..., \theta_{i-1} \sim \frac{\alpha G_0}{\alpha + i - 1} + \frac{\sum_{j=1}^{i-1} \delta_{\theta_i}}{\alpha + i - 1}. \tag{3.15}$$

em que $\delta_\theta$ é uma distribuição de massa no ponto $\theta$.

Por outro lado se considerarmos, para $\mathbf{Y}$, o modelo

$$\begin{aligned} Y_i|\theta_i &\sim F(\theta_i) \\ \theta_i|G &\overset{\text{iid}}{\sim} G(\theta) \\ G|\alpha, G_0 &\sim DP(\alpha G_0), \end{aligned} \tag{3.16}$$

temos que cada $Y_i$ tem a distribuição que é representada aqui por $F(\theta_i)$, os parâmetros $\theta_i$'s tem distribuição dada por $G$ e $G$ segue um processo de Dirichlet com distribuição base $G_0$ e parâmetro de concentração $\alpha$. Integrando este modelo em $G$, podemos obter uma representação para a distribuição condicional completa de $\theta_i$,

$$\theta_i|\theta_1, ..., \theta_{i-1} \sim \frac{\alpha}{i - 1 + \alpha} G_0 + \frac{1}{i - 1 + \alpha} \sum_{j=1}^{i-1} \delta_{\theta_j}, \tag{3.17}$$

Como resultado temos que tomando o limite $k \to \infty$ no modelo visto em (3.12) obtemos o modelo visto em (3.16) devido a correspondência entre a probabilidade condicional para $\theta_i$ na equação (3.17) e as sugeridas pela equação (3.14).

Dado que os $\theta_i$'s são permutáveis, Blackwell & MacQueen (1973) mostra que uma distribuição *a priori* para $\theta_i$ dado $\theta_1, ..., \theta_n$ exceto o $i-$ésimo ($\theta_{-i} = (\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_n)$) é

dada por

$$\theta_i|\theta_{-i} \sim \frac{\alpha}{\alpha+n-1}G_0 + \frac{1}{\alpha+n-1}\sum_{\substack{j=1\\j\neq i}}^{n}\delta_{\theta_j}, \tag{3.18}$$

que é baseada na relação entre o processo de Dirichlet e a generalização do esquema de urna de Pólya. Assim,

$$p(\theta_i|\theta_{-i}) = \frac{\alpha}{\alpha+n-1}G_0 + \frac{1}{\alpha+n-1}\sum_{\substack{j=1\\j\neq i}}^{n}\delta_{\theta_j} \tag{3.19}$$

é a densidade da distribuição condicional completa para o vetor de parâmetros $\boldsymbol{\theta}= (\theta_1, \theta_2, ...\theta_n)$.

Técnicas MCMC são usadas para obter estimativa da densidade da distribuição *a posteriori*, se a formulação do esquema de Urna de Polya visto em (3.9) é usada. Para obter a densidade que representa as distribuições condicionais completas *a posteriori*, combinamos a densidade da distribuição *a priori* dada em (3.19) com a verossimilhança $F(y_i|\theta_i)$, dada em (3.16), de cada observação. Ou seja,

$$p(\theta_i|\theta_{-i}, y_i) \propto F(y_i|\theta_i)p(\theta_i|\theta_{-i}). \tag{3.20}$$

Como mostra Escobar (1994), podemos escrever

$$p(\theta_i|\theta_{-i}) \propto \frac{\alpha}{\alpha+n-1}F(y_i|\theta_i)G_0(\theta_i) + \frac{1}{\alpha+n-1}\sum_{\substack{j=1\\j\neq i}}^{n}F(y_i|\theta_j)\delta_{\theta_j}, \tag{3.21}$$

A expressão $F(y_i|\theta_i)G_0(\theta)$, vista no primeiro termo da equação dada em (3.21), é escrita na forma da distribuição *a posteriori* para os parâmetros, ou seja

$$p(\theta_i|y_i) = q_0^{-1}F(y_i|\theta_i)G_0(\theta_i),$$

em que $q_0$ é a distribuição marginal dos dados dada por,

$$q_o = \int_{\theta_i} F(y_i|\theta)G_0(\theta_i)d\theta_i. \tag{3.22}$$

Multiplicando e dividindo a equação dada em (3.21) por $q_0$ dada em (3.22) e normalizando por meio de uma constante $b$, temos

$$p(\theta_i|\theta_{-i}, y_i) = b^{-1}\left(\alpha q_0 p(\theta_i|y_i) + \sum_{\substack{j=1\\j\neq i}}^{n} q_{i,j}\delta_{\theta_j}\right). \tag{3.23}$$

Em (3.23), $q_{i,j} = F(y_i|\theta_j)$ (com $F(y_i|\theta_j)$ sendo a verossimilhança dos dados avaliada no ponto amostral $y_i$ para $j = 1, 2, ..., n$ e $j \neq i$) e $b$ é dada por,

$$b = \alpha \int_{\theta_i} q_0 p(\theta_i|y_i) d\theta_i + \sum_{\substack{j=1 \\ j \neq i}}^{n} q_{i,j} \delta_{\theta_j}.$$

Se a distribuição base $G_0$ for uma distribuição *a priori* conjugada para a verossimilhança dos dados, dizemos que o modelo é conjugado, neste caso um algoritmo *Gibbs-Sampling* é usado para gerar valores da distribuição *a posteriori* dada em (3.23). Pois, neste caso, $q_0$ é resultado de uma integral que é analiticamente tratável e de fácil resolução. Para mais detalhes a respeito de modelos de mistura por processo de Dirichlet, veja, por exemplo, Neal (2000).

### 3.3.1 Modelos DPM conjugados

Uma importante característica na estrutura do modelo de mistura por processos de Dirichlet diz respeito ao fato de $G(.)$ ser uma distribuição discreta sobre o pressuposto do processo de Dirichlet. Então, dado qualquer amostra de tamanho $n$ de valores gerados de $G(.)$, o próximo valor poderá ser idêntico a outro já existente na amostra. Então, com probabilidade positiva a amostra de $n$ elementos poderá ser reduzida para $k < n$ valores distintos de $\theta$. Além disso, a medida que os $\theta_i$'s são atualizados, implicitamente também se atualiza as suas configurações $z_i$'s. Bush & MacEachern (1996) propõem um algoritmo Gibbs Sampling que utiliza estes indicadores $z_i$'s. Como já visto na Seção anterior, a variável $z_i$ indica que classe latente está associada com a observação $y_i$.

Vamos denotar por $\boldsymbol{\theta}^* = (\theta_1^*, ..., \theta_k^*)$ o conjunto dos distintos $\theta_i$'s no vetor de parâmetros $\boldsymbol{\theta}$. Seja ainda $\mathbf{z} = (z_1, ..., z_n)$ indicadores de componentes definidos tais que,

$$z_i = j \text{ se } \theta_i = \theta_j^* \text{ para } j = 1, ..., k \text{ e } i = 1, ..., n. \tag{3.24}$$

Com isto uma componente (ou grupo) é formado pelo conjunto das observações $y_i$'s ou correspondentes $\theta_i$'s com configuração de indicadores $z_i$'s idênticos, denota-se por $n_j$ o tamanho da j-ésima componente, ou seja,

$$n_j = \#\{i : z_i = j\}.$$

Assim, $n_j$ é a quantidade de vezes que $z_i = j$ para $j = 1, ..., k$. As componentes da mistura são formadas naturalmente pela estrutura do modelo e são enumeradas a medida que elas surgem, ou seja, à primeira componente é atribuído o número 1, à segunda componente é atribuído o número 2, e assim por diante.

Com isso, a expressão (3.23) pode ser escrita na forma,

$$(\theta_i|\boldsymbol{\theta}^{-i}, \mathbf{z}^{-\mathbf{i}}, y) = \begin{cases} \theta_1^* & \text{com probabilidade proporcional a } n_1^- q_{i,1} \\ \theta_2^* & \text{com probabilidade proporcional a } n_2^- q_{i,2} \\ \vdots & \\ \theta_k^* & \text{com probabilidade proporcional a } n_k^- q_{i,k} \\ \sim P(.) & \text{com probabilidade proporcional a } \alpha q_0, \end{cases} \tag{3.25}$$

em que $n_j^-$ é o número de elementos no componente $j$ sem a $i$-ésima observação, $\mathbf{x}^{-i}$ indica que do vetor original $\mathbf{x}$ foi excluído o i-ésimo elemento, $k$ é o número de componentes formados na amostra e $P(.)$ representa a distribuição *a posteriori* para os parâmetros $\theta_i$'s.

**Algoritmo 1**

1. Escolher valores iniciais $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ com variáveis indicadoras $\mathbf{z} = (z_1, ..., z_n)$.

2. Gerar novas amostras da seguinte maneira

   - Para $r = 1, .., R$ faça

     – Para $i = 1, ..., n$ faça

     Gerar $\theta_i$ de $\theta_i|\boldsymbol{\theta}^{-i}, \mathbf{z}^{-\mathbf{i}}, y$ como descrito na equação (3.25), atualizar $z_i$ com base no novo valor de $\theta_i$.

     – Fim

     $k^r$ = número de distintos valores no vetor $\theta^r$

     $\theta^*$ = indica o conjunto de valores distintos de $\theta_i^r$'s

     Para $j = 1, ..., k$, gerar um novo valor de $\theta_j^*|$ todos $y_i$, atualizar $\boldsymbol{\theta}$ com base no vetor atualizado $\boldsymbol{\theta}^*$ e $\mathbf{z}$.

   - Fim

3. $k^* = R^{-1}\sum_1^R k^r$

Para facilitar notação, vamos referir a estimativa de $k$ ($k^*$) simplesmente por $k$ estimado.

### 3.3.2 Modelos DPM não conjugados

Se a distribuição base do processo $G_0$ e a verossimilhança não formam um par conjugado, a integral $q_0$ em (3.22) muitas vezes é analiticamente intratável o que torna o algoritmo descrito acima inviável. Neste caso algum método numérico deve ser usado para aproximar o valor da integral em $q_0$. MacEachern & Müller (1998) propõem uma aproximação para $\int_\theta f(y_i|\theta)G_0(\theta)d(\theta)$ usando o método de aproximação Monte Carlo para integração, mas em vez de considerar um conjunto de $m$ gerações de $G_o$, eles consideram apenas uma geração para fazer a aproximação desta integral. No algoritmo, eles substituem $q_0$ por $f(y_i|\theta')$, na qual $\theta'$ é uma geração aleatória da distribuição base $G_0$ e $f(y_i|\theta')$ é a verossimilhança dos dados. Com esta substituição o problema de calcular a integral em $q_0$ é eliminado.

Vamos considerar inicialmente um conjunto de parâmetros $\{\theta_1^*, ..., \theta_k^*, \theta_{k+1}^*, ..., \theta_n^*\}$, com a mesma configuração de indicadores $z_i$ vista em (3.24), consideremos que $\boldsymbol{\theta}_c^* = \{\theta_1^*, ..., \theta_k^*\}$ é utilizado para identificar componentes não vazios, ou seja, os valores em $\boldsymbol{\theta}_c^*$ estão associadas a pelo menos uma observação cada, formando ao todo $k$ componentes. Enquanto $\boldsymbol{\theta}_v^* = \{\theta_{k+1}^*, ..., \theta_n^*\}$ são valores com potencial para identificar algum componente, mas ainda não foram utilizados, estes valores não estão associados às observações e são gerados da distribuição base $G_0$, assim para $j = 1, ..., k$ temos $n_j > 0$ e para $j = k+1, ..., n$ temos $n_j = 0$.

**Algoritmo 2**

1. Para $i = 1, ..., n$.

   - Se $n_{z_i} > 1$ ($n_{z_i}$ é o número de elementos em $\mathbf{z}$ que são iguais a $z_i$), então atualiza-se $z_i$ de modo que

$$\begin{aligned} p(z_i = j|z^{i-}, \theta^*) &\propto n_j^- f(y_i|\theta_j^*), \text{para } j = 1, ..., k^- \\ p(z_i = k^- + 1|z^{i-}, \theta^*) &\propto \tfrac{\alpha}{k^-+1} f(y_i|\theta_{k^-+1}^*), \end{aligned} \tag{3.26}$$

em que $k^-$ é o número de distintos $z_j$ para $j \neq i$ com $z_j \in \{1, ..., k^-\}$, neste caso $k^- = k$.

- Se $n_{z_i} = 1$, com probabilidade $\frac{k-1}{k}$ o valor de $z_i$ não se altera e com probabilidade $\frac{1}{k}$ temos $z_i = k$, neste caso $z_i$ é atualizado usando a Expressão (3.26).

  Este procedimento evita que ocorra lacunas com a remoção de $z_i$ e por este motivo este algoritmo é conhecido por modelo *no gaps*.

2. Para $j = 1, ..., k$ atualiza-se $\boldsymbol{\theta}^*|z, y$, ou seja, dentro de cada grupo j existente, atualiza-se $\theta_j^*$ usando a distribuição *a posteriori* segundo o teorema de Bayes. E para $j = k+1, ..., n$ $\theta_j^*$ é atualizado gerando-se valores da distribuição base $G_0$.

O algoritmo 2, como foi descrito neste trabalho, foi proposto por MacEachern & Müller (1998).

# Capítulo 4

# Modelo DPM Conjugado Normal-Normal (DPMN-N)

Neste Capítulo, discutimos a estimação da densidade de modelos de mistura, considerando o método bayesiano não paramétrico visto no Capítulo 2, para o caso em que cada observação é assumida ser originada de uma distribuição normal e é escolhida uma distribuição base $G_0$ conjugada a verossimilhança dos dados. Por motivo de simplificação, vamos nos referir ao modelo DPM, nestas condições, como modelo DPMN-N que significa modelo DPM conjugado normal-normal. A distribuição base adotada para o modelo é a distribuição normal-gama-invertida.

## 4.1 Modelo

O modelo de mistura por processo de Dirichlet é discutido em Escobar & West (1995), em que é assumida distribuição normal para os dados e o processo de Dirichlet é usado para descrever a incerteza acerca da distribuição dos parâmetros. O modelo DPM é usado considerando para o modelo hierárquico visto em (3.16), uma distribuição normal, para a $i$-ésima observação, com vetor de parâmetros $\theta_i = (\mu_i, \sigma_i^2)$ composto pelo parâmetro de locação $\mu_i$ e parâmetro de escala $\sigma_i^2$, para $i = 1, ..., n$. Deste modo o modelo hierárquico em (3.16) pode ser escrito como

$$\begin{aligned} Y_i|\mu_i,\sigma_i^2 &\sim N(\mu_i,\sigma_i^2) \\ (\mu_i,\sigma_i^2)|G &\sim G \\ G|\alpha,G_0 &\sim DP(\alpha,G_0), \end{aligned} \tag{4.1}$$

em que $N(\mu_i,\sigma_i^2)$ representa a distribuição normal com vetor de parâmetros $(\mu_i,\sigma_i^2)$. Como já foi visto visto no Capítulo 2, as distribuições condicionais, neste caso, são obtidas por meio da expressão

$$p(\theta_i|\theta_{-i}) = \frac{\alpha}{\alpha+n-1}G_0 + \frac{1}{\alpha+n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}\delta_{\theta_j} \tag{4.2}$$

descrita, nesta dissertação, em (3.19), que combinada com a verossimilhança dos dados fornece a distribuição condicional *a posteriori* para os parâmetros do modelo, como descrito em (3.23).

Agora, devemos computar $q_0$ e $f(\mu_i,\sigma_i^2|y_i)$, para completar os termos da equação (3.23). Pelo que foi visto no Capítulo 2, é necessário especificar uma distribuição *a priori* para $\mu_i$ e $\sigma_i^2$, que é a distribuição base $G_0$, vamos utilizar uma distribuição *a priori* da família conjugada para o modelo normal, que é a normal-gama-invertida.

Escrevendo $\phi_i = \sigma_i^{-2}$, a densidade conjunta de $(\mu_i,\phi_i)$ pode então ser escrita como $p(\mu_i,\phi_i) = p(\mu_i|\phi_i)p(\phi_i)$. Deste modo escrevemos

$$G_0 \rightarrow \begin{cases} \mu_i|\phi_i \sim N\Big(m,(\eta\phi_i)^{-1}\Big) \\ \phi_i \sim Ga\Big(\frac{a}{2},\frac{ab}{2}\Big), \end{cases} \tag{4.3}$$

com $m\in\mathbb{R}$, $\eta$, $a$, $b$ números reais positivos e $Ga(\mu_i,\sigma_i^2)$ representando a distribuição gama. A distribuição conjunta para $(\mu_i,\sigma^2=\phi_i^{-1})$ é chamada de distribuição normal-gama-invertida, $Ig(.)$, com parâmetros $(m,\eta,a,b)$.

Do mesmo modo, escrevemos a densidade da distribuição *a posteriori* conjunta para $(\mu_i,\phi_i)$ como, $f(\mu_i,\phi_i|y_i) = f(\mu_i|\phi_i,y_i)f(\phi_i|y_i)$.

Considerando $\phi_i$ fixo, o núcleo da função de densidade de probabilidade para $\mu_i|\phi_i,y_i$, é obtida combinando a verossimilhança dos dados com a distribuição *a priori* normal vista na primeira parte da expressão (4.3). Assim,

$$f(\mu_i|\phi_i, y_i) \propto \exp\left\{-\frac{1}{2}\phi_i(\eta+1)\left(\mu_i - \frac{\eta m + y_i}{\eta+1}\right)^2\right\}.$$

Este núcleo é da f.d.p. da distribuição normal com parâmetros $\left(\frac{\eta m+y_i}{1+\eta}, \frac{1}{\phi_i(1+\eta)}\right)$. Então,

$$\mu_i|\phi_i, y_i \sim N\left(\frac{\eta m + y_i}{\eta+1}, \frac{1}{\phi_i(\eta+1)}\right).$$

A densidade da distribuição de $\phi_i|y_i$, é obtida por integração como,

$$\begin{aligned}
f(\phi|y_i) &\propto \int f(y_i|\mu_i,\phi_i)G_0(\mu_i,\phi_i)d\mu_i \\
&\propto \int \phi_i^{\frac{1}{2}} \exp\left\{-\tfrac{\phi_i}{2}(y_i-\mu_i)^2\right\}\phi_i^{\frac{1}{2}} \exp\left\{-\tfrac{\phi_i\eta}{2}(\mu_i-m)^2\right\}\phi_i^{\frac{a}{2}-1}\exp\left\{-\tfrac{ab\phi_i}{2}\right\}d\mu_i \\
&\propto \phi_i^{\frac{a+1}{2}-1}\exp\left\{-\tfrac{ab\phi_i}{2}\right\}\int \phi_i^{\frac{1}{2}}\exp\left\{-\tfrac{\phi_i}{2}\left(y_i^2 - 2y_i\mu_i + \mu_i^2 + \mu_i^2\eta - 2\mu_i m\eta + m^2\eta\right)\right\}d\mu_i \\
&\propto \phi_i^{\frac{a+1}{2}-1}\exp\left\{-\tfrac{\phi_i}{2}\left[ab + y_i^2 + m^2\eta\right]\right\}\int \phi_i^{\frac{1}{2}}\exp\left\{-\tfrac{\phi_i}{2}\left[\mu_i^2(\eta+1) - 2\mu_i(y_i+m\eta)\right]\right\}d\mu_i \\
&\propto \phi_i^{\frac{a+1}{2}-1}\exp\left\{-\tfrac{\phi_i}{2}\left[ab + y_i^2 + m^2\eta - \tfrac{(y_i+m\eta)^2}{(\eta+1)}\right]\right\} \\
&\propto \phi_i^{\frac{a+1}{2}-1}\exp\left\{-\tfrac{\phi_i}{2}\left[ab + \tfrac{\eta(y_i-m)^2}{(\eta+1)}\right]\right\}.
\end{aligned}$$

O resultado obtido é o núcleo da função de densidade de probabilidade da distribuição gama, $Ga(.)$, com parâmetros $\left(\frac{1}{2}(a+1), \frac{1}{2}\left(ab + \frac{\eta(y_i-m)^2}{(\eta+1)}\right)\right)$. Ou seja,

$$\phi_i|y_i \sim Ga\left(\frac{1}{2}(a+1), \frac{1}{2}\left(ab + \frac{\eta(y_i-m)^2}{(\eta+1)}\right)\right).$$

Com isso, concluímos que a distribuição conjunta *a posteriori* para $(\mu_i, \sigma^2 = \phi_i^{-1})$ é também a distribuição normal-gama-invertida. Para mais informações a respeito de conjugação na família de distribuições normais, veja por exemplo, Bernardo & Smith (1994).

Para completar os termos da equação (3.23), falta encontrar a representação analítica

para $q_0$. Neste caso, basta resolver analiticamente a integral em (3.22). Ou seja,

$$\begin{aligned}
q_o &= \int\int f(y_i|\mu_i,\phi_i)G_0(\mu_i,\phi_i)d\mu_i d\phi_i \\
&= \int \frac{(\frac{ab}{2})^{\frac{a}{2}}\phi_i^{\frac{a}{2}-1}}{\Gamma(\frac{a}{2})}\exp\left\{\frac{-\phi_i ab}{2}\right\}\int\frac{\phi_i^{\frac{1}{2}}}{\sqrt{2\pi}}\exp\left\{\frac{-\phi_i}{2}(y_i-\mu_i)^2\right\}\frac{(\eta\phi_i)^{\frac{1}{2}}}{\sqrt{2\pi}}\exp\left\{\frac{-\phi_i\eta}{2}(\mu_i-m)^2\right\}d\mu_i d\phi_i \\
&= \int\frac{\eta^{\frac{1}{2}}(\frac{ab}{2})^{\frac{a}{2}}\phi_i^{\frac{a}{2}}}{\Gamma(\frac{a}{2})2\pi}\exp\left\{\frac{-\phi_i}{2}(ab+y_i^2+m^2\eta)\right\}\int\exp\left\{\frac{-\phi_i}{2}\left[\mu_i^2(\eta+1)-2\mu_i(y_i+m\eta)\right]\right\}d\mu_i d\phi_i \\
&= \int\frac{\eta^{\frac{1}{2}}(\frac{ab}{2})^{\frac{a}{2}}\phi_i^{\frac{a}{2}}}{\Gamma(\frac{a}{2})2\pi}\exp\left\{\frac{-\phi_i}{2}\left(ab+\frac{\eta(y_i-m)^2}{\eta+1}\right)\right\}\int\exp\left\{\frac{-\phi_i(\eta+1)}{2}\left[\mu_i-\left(\frac{y_i+m\eta}{\eta+1}\right)\right]^2\right\}d\mu_i d\phi_i \\
&= \frac{\eta^{\frac{1}{2}}(\frac{ab}{2})^{\frac{a}{2}}}{\Gamma(\frac{a}{2})(\eta+1)^{\frac{1}{2}}\sqrt{2\pi}}\int\phi_i^{\frac{a+1}{2}-1}\exp\left\{\frac{-\phi_i}{2}\left(ab+\frac{\eta(y_i-m)^2}{\eta+1}\right)\right\}d\phi_i \\
&= \frac{\eta^{\frac{1}{2}}(\frac{ab}{2})^{\frac{a}{2}}}{\Gamma(\frac{a}{2})(\eta+1)^{\frac{1}{2}}\sqrt{2\pi}}\left(\frac{ab}{2}+\frac{\eta(y_i-m)^2}{2(\eta+1)}\right)^{-\frac{a+1}{2}}\Gamma(\tfrac{a+1}{2}), \\
&= \frac{\Gamma(\frac{a+1}{2})}{\Gamma(\frac{a}{2})}\left(\frac{\eta}{b(\eta+1)a\pi}\right)^{\frac{1}{2}}\left(1+\frac{\eta(y_i-m)^2}{b(\eta+1)a}\right)^{-\frac{a+1}{2}},
\end{aligned} \tag{4.4}$$

em que $\Gamma(.)$ representa a função gama.

A integral em $q_0$ fornece a densidade da distribuição t de Student com $a$ graus de liberdade e parâmetros $(m,\lambda)$, em que $\lambda = \frac{\eta}{b(\eta+1)}$.

Assim, as distribuições condicionais completas *a posteriori* para $\theta_i$, $i = 1,...,n$, são obtida da expressão,

$$p(\theta_i|\theta^{i-},y_i) = b\left(\alpha St(y_i|m,\lambda,a)Ng(\mu,\phi|\mu_0,\sigma_0^2,a_0,b_0) + \sum_{\substack{j=1\\ j\neq i}}^{n} N(y_i|\theta_j)\delta_{\theta_i}(\theta_j)\right), \tag{4.5}$$

na qual, $St(y_i|m,\lambda,a)$ representa a densidade da distribuição t de student, $N(y_i|\mu,\phi)$ representa a densidade da distribuição normal, ambas avaliadas no ponto $y_i$, e $Ng(\mu,\phi|\mu_0,\sigma_0^2,a_0,b_0)$ representa a densidade da distribuição normal-gama com parâmetros $\mu_0 = \frac{\eta m+y_i}{\eta+1}$, $\sigma_0^2 = \eta+1$, $a_0 = a+1$ e $b_0 = ab+\frac{\eta(y_i-m)^2}{(\eta+1)}$.

Na Equação (4.5), a densidade da distribuição t de Student representa a marginal dos dados, a distribuição normal-gama é a distribuição *a posteriori* para os parâmetros $(\mu,\phi)$ e a densidade da distribuição normal é a verossimilhança para cada ponto de dado da amostra.

## 4.2 Quantidades Preditivas

A densidade preditiva é de interesse em problemas de inferência pois nos permite obter uma estimativa de um valor futuro $(y_{n+1})$ de uma variável aleatória $Y$, dados $D_n =$

$(y_1, y_2, ..., y_n)$ observações da v.a. $Y$. Neste trabalho, além de simularmos valores da distribuição *a posteriori*, também temos interesse na estimativa de densidade, que em inferência bayesiana é obtida por meio da distribuição preditiva *a posteriori* que é dada pelo valor esperado em relação aos parâmetros, como

$$f(y_{n+1}|D_n) = \int f(y_{n+1}, \boldsymbol{\theta}|D_n)d\boldsymbol{\theta} = \int f(y_{n+1}|\boldsymbol{\theta})f(\boldsymbol{\theta}|D_n)d\boldsymbol{\theta} = E_{\boldsymbol{\theta}|y}[f(y_{n+1}|\boldsymbol{\theta})], \tag{4.6}$$

em que $D_n$ representa os dados observados.

No contexto analisado aqui, a densidade $f(y_{n+1}|\boldsymbol{\theta})$ é obtida da seguinte maneira,

$$\begin{aligned} f(y_{n+1}|\boldsymbol{\theta}) &= \int f(y_{n+1}|\theta_{n+1})f(\theta_{n+1}|\boldsymbol{\theta})d\theta_{n+1} \\ &= \int N(y_{n+1}|\theta_{n+1})f(\theta_{n+1}|\boldsymbol{\theta})d\theta_{n+1}, \end{aligned} \tag{4.7}$$

que é a esperança da densidade da distribuição normal, $N(y_{n+1}|\theta_{n+1})$. Tendo em vista a densidade em (3.17), então,

$$\begin{aligned} f(y_{n+1}|\boldsymbol{\theta}) &= \tfrac{\alpha}{\alpha+n}\int N(y_{n+1}|\theta_{n+1})G_0(\theta_{n+1})d\theta_{n+1} + \tfrac{1}{\alpha+n}N(y_{n+1}|\theta_{n+1})\sum_{i=1}^{n}\delta_{\theta_{n+1}}(\theta_i) \\ &= \tfrac{\alpha}{\alpha+n}\int N(y_{n+1}|\theta_{n+1})G_0(\theta_{n+1})d\theta_{n+1} + \tfrac{1}{\alpha+n}\sum_{i=1}^{n}N(y_{n+1}|\theta_{n+1}). \end{aligned} \tag{4.8}$$

De (4.4), a integral no primeiro termo desta equação é a densidade da distribuição t de Student, logo

$$f(y_{n+1}|\boldsymbol{\theta}) = \tfrac{\alpha}{\alpha+n}St(y_{n+1}|m, \lambda, a) + \tfrac{1}{\alpha+n}\sum_{i=1}^{n}N(y_{n+1}|\theta_i), \tag{4.9}$$

em que $St(.|m, \lambda, a)$ representa a densidade da distribuição t de Student com $a$ graus de liberdade.

Com isso, a distribuição preditiva é estimada como uma média sobre a distribuição preditiva condicional, ou seja,

$$\begin{aligned} f(y_{n+1}|D_n) &= E_{\boldsymbol{\theta}|y}[f(y_{n+1}|\boldsymbol{\theta})] \\ &\approx 1/T\sum_{t=1}^{T}f(y_{n+1}|\boldsymbol{\theta}^t), \end{aligned} \tag{4.10}$$

em que $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ e $\boldsymbol{\theta}^t$ são os valores imputados depois de $t$ iterações do algoritmo Gibbs sampling. Esta estimação é feita usando um método chamado de método de composição. Maiores detalhes pode ser visto em Tanner (1996) Seção 3.3

# Capítulo 5

# Performance do modelo DPMN-N

Neste Capítulo, verificamos aspectos relacionados à eficiência do modelo DPMN-N. Para isso, fizemos uma análise do desempenho do modelo no que diz respeito as condições em que é identificado corretamente o número de componentes, quando se tem uma mistura de distribuições. Vamos nos referir a esta análise como análise de precisão. Também fazemos uma análise de sensibilidade em relação ao parâmetro de locação da distribuição base do processo de Dirichlet. Vamos nos referir a esta análise como análise de sensibilidade. Antes de falarmos a respeito dessas análises, falaremos um pouco da estratégia computacional utilizada.

## 5.1 Estrategia computacional

Para proceder nossas análises, implementamos o algoritmo proposto por Escobar & West (1995) visto na Seção 3.3.1, em que a estimação do número de grupos na população ocorre de forma quase que automática, no modelo visto em (4.1). Com base no estudo de Bush & MacEachern (1996), adicionamos um passo a mais ao algoritmo de Escobar & West (1995) (como descrito em algoritmo 1 na página 30). Com base no trabalho de Bush & MacEachern (1996), a variável latente foi incluída no algoritmo de Escobar & West (1995) para ser possível atualizar os parâmetros do modelo, dada a identificação dos componentes para cada ciclo completo algoritmo. Esta atualização é feita por meio das densidades obtidas combinando a verossimilhança dos dados agrupados com a distribuição *a priori* para os parâmetros do

modelo vista em (4.3), com isso obtemos a distribuição *a posteriori* para os parâmetros em cada componente, ou seja,

$$\begin{cases} \sigma_j^{-2}|Y_{z_i=j} & \sim \; Ga\left(\frac{n_j+a}{2}, \frac{1}{2}\left(ab + \sum_{i=1}^{n}(y_{z_i=j} - \bar{y_j})^2 + \frac{\eta n_j (m-\bar{y_j})^2}{\eta+n_j}\right)\right) \\ \mu_j|\sigma_j^{-2}, Y_{z_i=j} & \sim \; N\left(\frac{\eta m + n_j \bar{y_j}}{\eta+n_j}, \left(\sigma_j^{-2}(\eta+n_j)\right)^{-1}\right), \end{cases} \tag{5.1}$$

em que $y_{z_i=j}$ é o conjunto de observações pertencentes à componente $j$,com $j = 1, ..., k$ sendo $k$ o número de componentes identificados pelo modelo, no passo anterior do algoritmo, e $\bar{y_j}$ é a média aritmética dos elementos dentro da componente $j$. Com a atualização dos parâmetros $\mu$ e $\sigma^2$ condicionada a identificação dos componentes, temos uma iteração completa do algoritmo utilizado.

Os comandos do software R, que foram usados para gerar os valores das amostras usadas para as simulações deste Capitulo se encontram no Apêndice juntamente com os códigos usados para fazer as estimativas do número de componentes na população e dos parâmetros do modelo de mistura.

## 5.2 Análise de sensibilidade

No contexto visto neste trabalho, uma parte importante no processo de estimação do número de componentes que compõem uma mistura de distribuições é a escolha dos parâmetros do processo de Dirichlet. Como pode ser visto na Seção 3.1, o DP tem dois parâmetros a serem considerados, um deles é o parâmetro de precisão $\alpha$ e o outro é a distribuição base, $G_0$, que também é a esperança do DP. Nesta análise, vamos verificar a performance do modelo sob diferentes escolhas para o parâmetro de locação da distribuição base do processo de Dirichlet, ou seja, sob diferentes escolhas do parâmetro de locação da distribuição $G_0$ vista na equação (4.3).

A fim de proceder a esta análise, ajustamos o modelo visto em (4.1) em 1000 conjuntos de amostras simuladas, utilizando o software R, todas com 100 elementos cada, ($n = 100$). As amostras foram simuladas da densidade de uma mistura de duas distribuições normais com médias $\mu_1$ e $\mu_2$ iguais a 40 e 60, respectivamente, e variâncias $\sigma_1^2$ e $\sigma_2^2$ iguais a 16 e 81,

respectivamente. As proporções da mistura são fixadas em $\omega_1 = \omega_2 = 1/2$. Assim,

$$y_i \sim \frac{1}{2}N\big(40, 4^2\big) + \frac{1}{2}N\big(60, 9^2\big), \tag{5.2}$$

em que $y_i$ para $i = 1, ..., 100$ representam os pontos simulados em cada amostra $\mathbf{y}$.

Para ajustar o modelo nestas amostras, a preocupação agora é quanto a escolha dos parâmetros da distribuição base do processo, $G_0$. Como visto na equação (4.3) adotamos uma distribuição normal-gama como distribuição base. Para obtermos variâncias grandes para as distribuições *a priori*, os parâmetros $\eta$ ,$a$ e $b$ são fixados com valores 1/100, 10 e 1/40, respectivamente. Assim,

$$G_0 \sim Ng\big(m, 100, 10/2, 10/80\big), \tag{5.3}$$

em que Ng(.) representa a distribuição normal-gama.

O parâmetro de precisão, $\alpha$, do processo de Dirichlet é fixado igual a 1. Como visto em Escobar (1994), o parâmetro de precisão $\alpha$ é a esperança do número de elementos distintos na amostra, $k$, em relação a distribuição *a priori* processo de Dirichlet. Em análise feita em Escobar (1994), vimos que para uma amostra de tamanho entre $50 - 200$, $\alpha = 1$ fornece um valor esperado de $k$ entre $4.5 - 5.88$.

Resta agora especificar o valores para o parâmetro de locação $m$ em (5.3). Ajustamos o modelo nas 1000 amostras simuladas, considerando valores de $m$ em $\{-20, 10, 40, 70, 100, 130\}$, como visto na primeira linha da Tabela 5.1. Esta tabela mostra os resultados das proporções de indicações para o número de componentes na mistura vista em (5.2), para o conjunto de 1000 amostras nos quais o modelo foi ajustado para cada um destes valores de $m$. Para cada simulação, foram feitas 2000 iterações completas do algoritmo, sendo descontado um período de aquecimento de 1000. Assim, foi considerada uma amostra de tamanho $L = 1000$ no processo de estimação em cada simulação. As estimativas para o número de componentes da mistura, foram obtidas utilizando a moda da distribuição atualizada, ou distribuição *a posteriori*.

Na Tabela 5.1, podemos observar que para valores de $m$ próximos dos valores de $\mu_1$ e $\mu_2$, o modelo não identifica corretamente o número de grupos, pois cada amostra foi simulada da mistura de duas distribuições, sendo assim o número correto de componentes é 2. Uma

explicação para essa baixa performance é que, uma nova componente surge com probabilidade proporcional a $q_0$, e no caso do modelo DPMN-N é $q_0$ a densidade da distribuição t de Student com parâmetro de locação $m$, se o valor deste parâmetro estiver próximo dos valores amostrados, esta probabilidade poderá aumentar substancialmente, fazendo com que muitos componentes sejam identificados pelo modelo.

Tabela 5.1: Resultados das proporções de indicação de número de grupos, que compõe a população, para diferentes valores de parâmetros de locação da distribuição base, na análise de 1000 conjuntos de dados simulados de (5.4).

| Nro de grupos ($k$) | $m = -20$ | $m = 10$ | $m = 40$ | $m = 70$ | $m = 100$ | $m = 130$ |
|---|---|---|---|---|---|---|
| $k = 1$ | $0,009$ | $0,06$ | $0$ | $0$ | $0,03$ | $0,02$ |
| $k = 2$ | $0,988$ | $0,927$ | $0$ | $0$ | $0,945$ | $0,975$ |
| $k = 3$ | $0,003$ | $0,013$ | $0$ | $0,006$ | $0,023$ | $0,005$ |
| $k > 3$ | $0$ | $0$ | $1$ (max. $k = 23$ ) | $0,994$ (max. $k = 16$ ) | $0,002$ | $0$ |

Podemos observar que o melhor resultado descrito na Tabela 5.1, é para $m = -20$, no caso dessas 1000 amostras analisadas, 998 fornecem maior probabilidade *a posteriori* para a existência de dois grupos na população, isso ocorre devido este valor estar distante dos valores dos dados simulados, fazendo com que a probabilidade de transição para um novo grupo, ocorra de modo mais próximo de uniforme, pois estas probabilidades são afetadas somente pela cauda da distribuição t de Student, cuja densidade é dada pela Equação (4.4).

Na quarta coluna da Tabela 5.1, observamos que com $m = 40$ o modelo identifica um número máximo de 23 grupos e para todas as amostras trabalhadas o modelo não identifica menos que 3 grupos, isso ocorre devido a média da primeira componente da mistura que deu origem aos dados, coincidir com o valor de $m$. Para a situação em que se tem $m = 70$, o modelo também identifica muitos grupos, isso se deve ao fato deste valor ser próximo da média da segunda componente da mistura, como pode ser visto em (5.4). Então, variando $m$ a análise fornece informações sobre o quão sensíveis os resultados para $k$ são em relação a escolha de $m$.

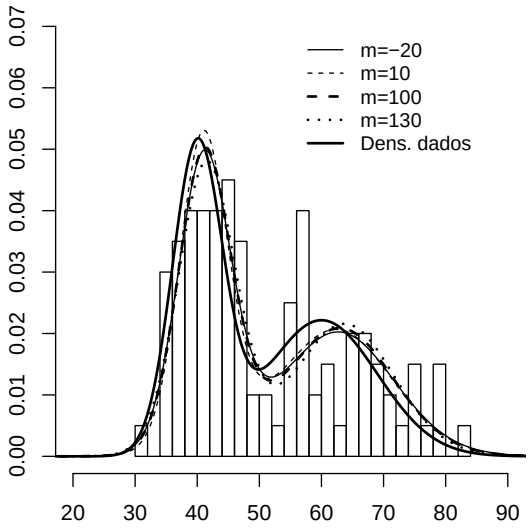


Figura 5.1: Estimativa de densidade obtida da distribuição preditiva *a posteriori* para $m \in \{-20, 10, 100, 130\}$.

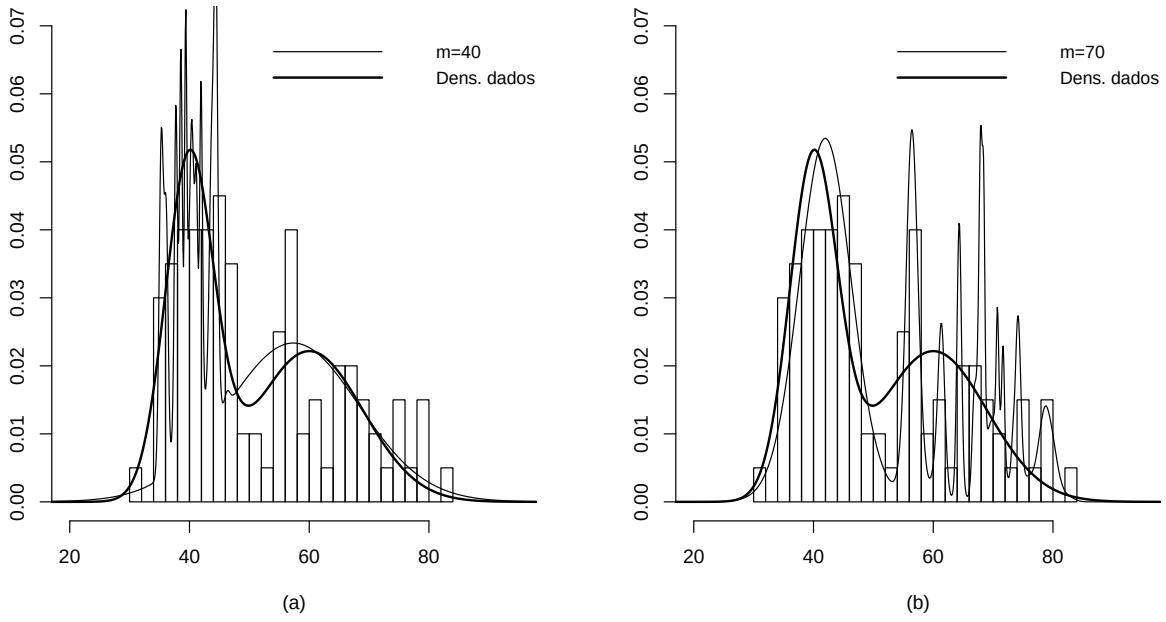


Figura 5.2: Estimativa de densidade obtida por meio da distribuição preditiva *a posteriori*, considerando $m = 40$ em (a) e $m = 70$ em (b).

Os efeitos da variação de $m$ na inferência sobre a função de densidade estão representados graficamente nas figuras 5.1 e 5.2. Estas figuras contem os gráficos das distribuições preditivas para cada valor de $m$ em $\{-20, 10, 40, 70, 100, 130\}$. Na Figura 5.1 podemos observar que o método parece estimar bem a distribuição para $m$ assumindo valores em $\{-20, 10, 100, 130\}$, pois estas estimativas estão próximas da densidade obtida a partir da distribuição de onde os dados foram gerados. No entanto, o método não estima bem para $m = 40$, situação em que podemos observar multimodalidade no gráfico apresentado na Figura 5.2a), situação em que se tem $\mu_1 = m = 40$. Na figura 5.2b) a multimodalidade é menos acentuada, nesta situação temos $m = 70$ que é um valor próximo de $\mu_2 = 70$. Estes resultados gráficos já era de se esperar, tendo em vista os resultados apresentados na Tabela 5.1.

## 5.3 Análise de precisão

A análise feita nesta Seção diz respeito ao desempenho do modelo especificado em (4.1), na estimação do número de componentes quando se tem um modelo de mistura. O objetivo é verificar a performance do modelo quando ocorre mudanças no valor da média e do desvio padrão dos componentes de uma mistura. Para proceder com essa análise ajustamos o modelo DPMN-N em amostras, $\mathbf{x}$, de 100 elementos cada, simuladas da mistura de duas distribuições normais,

$$x_i \sim \frac{1}{2}N\big(\mu_1, \sigma_1\big) + \frac{1}{2}N\big(\mu_2^r, \sigma_2^s\big), \tag{5.4}$$

para $r, s \in \{1, ..., 5\}$, em que $x_i$ para $i = 1, ..., 100$ representam os pontos simulados em cada amostra $\mathbf{x}$.

Para cada par distinto de $(\mu_2^r, \sigma_2^s)$ em (5.4), simulamos 1000 amostras, com 100 elementos cada uma mantendo $\mu_1 = 40$ e $\sigma_1 = 4$, para todas as amostras. Os valores para $(\mu_2^r, \sigma_2^s)$, que foram usados estão descritos nas duas primeiras linhas da tabela 5.2 e foram obtidos de

modo que,

$$\begin{aligned} \mu_2^r &= \mu_2^{r-1} + \delta \\ \sigma_2^s &= \sigma_2^{s-1}.\gamma, \end{aligned}$$

para $r, s \in \{2, ..., 5\}$ e assumindo $\mu_2^1 = \mu_1 = 40$, $\sigma_2^1 = \sigma_1 = 4$, $\delta = 10$ e $\gamma = 1.5$. Com isso, geramos 1000 amostras da mistura em 5.4 para cada conjunto de parâmetros, $\{\mu_1, \sigma_1, \mu_2^r, \sigma_2^r\}$, diferentes. Como são 5 valores diferentes de médias e desvios padrões para a segunda componente, geramos um total de $1000 \times 5 \times 5$ amostras para esta análise. A proporção da mistura foi fixada em 1/2 para cada componente, em todas as gerações.

Tabela 5.2: Proporção de indicação para dois grupos em 1000 amostras simuladas a partir de uma mistura de duas distribuições normais

| | | $\mu_2$ | 40 | 50 | 60 | 70 | 80 |
|---|---|---|---|---|---|---|---|
| $\sigma_2$ | $\sigma_1$ | $\mu_1$ | 40 | 40 | 40 | 40 | 40 |
| 4 | 4 | | 0 | 0.154 | 1 | 1 | 1 |
| 6 | 4 | | 0 | 0.252 | 1 | 1 | 1 |
| 9 | 4 | | 0.023 | 0.699 | 0.988 | 0.997 | 1 |
| 13.50 | 4 | | 0.249 | 0.83 | 0.99 | 0.989 | 0.99 |
| 20.25 | 4 | | 0.606 | 0.869 | 0.962 | 0.969 | 0.979 |

O modelo visto em (4.1) foi ajustado com distribuição base normal gama inversa, ou seja, $G_o \sim Nig\big(m, (\eta\phi)^{-1}, a/2, ab/2\big)$, em que Nig(.) representa a distribuição normal gama inversa. Os valores dos parâmetros da distribuição base foram escolhidos de acordo com a análise feita na Seção 5.2 e assumem os valores

$$m = -20, \quad \eta = 1/100, \quad a = 10, \quad b = \tfrac{1}{40}.$$

O parâmetro de precisão foi mantido como na primeira análise, $\alpha = 1$.

Assim como na primeira análise, para cada simulação foram realizadas 2000 iterações completas do algoritmo, sendo descontado um período de aquecimento de 1000. Aqui também foi considerada uma sequência de gerações de tamanho $L = 1000$ no processo de esti-

mação em cada simulação e as estimativas para o número de componentes, foram obtidas utilizando a moda da distribuição atualizada.

Para cada grupo de 1000 sequências geradas da mesma distribuição, ou seja, da mistura em (5.4) com parâmetros iguais, calculamos a proporção de indicação para dois grupos, ou seja, $k = 2$, os valores destas proporções estão descritas na Tabela 5.2.

A primeira célula da Tabela 5.2, mostra o caso em que se tem médias e desvios padrões iguais para os dois componentes, neste caso o que se tem é apenas uma componente na mistura, isso justifica o fato de que 1000 em 1000 amostras, não forneceram maior probabilidade *a posteriori* para a existência de dois grupos na população. Partindo do modelo de uma componente, aumentando o desvio padrão, notamos que quando se tem pouca variação do desvio padrão mantendo as médias idênticas o modelo também não identifica componentes distintas. No entanto, a medida que a variabilidade dos dados dentro de cada componente aumenta, as proporções de indicação para dois componentes em cada amostra também aumenta. Com isso, existe evidências de que o modelo também identifica ( ou distingue) as componentes na mistura, com base na diferença da variabilidade dos dados em cada componente. Esta situação é melhor descrita no gráfico que pode ser visto na Figura 5.3(a). Neste gráfico, estão as curvas que representam os valores das proporções segundo as colunas da Tabela 5.2.

Com base nas proporções para indicação para dois componentes na população, exibidas na Tabela 5.2, observamos que o modelo identifica componentes pela variabilidade dos dados, como podemos observar nas duas primeiras colunas da Tabela 5.2. No entanto, como podemos observar nas três últimas colunas da Tabela 5.2, se as médias estiverem muito afastadas o aumento da variabilidade no segundo componente diminui a precisão do método em estimar o número de componentes da população. Observe que a proporção de indicação para $k = 2$ decresce a medida que o desvio padrão cresce, para o caso em que $\mu > 50$, como mostra a Figura 5.3(b). Ainda com base nos valores das proporções descritos na Tabela 5.2, observamos que a eficiência do método aumenta a medida que a distância entre as médias aumenta, situação que pode ser percebida graficamente na Figura 5.3(a).

Para ilustrar como se comportam os valores das proporção de indicação para $k = 2$, construimos o gráfico de superfície, como mostra a Figura 5.4. Este gráfico foi construído

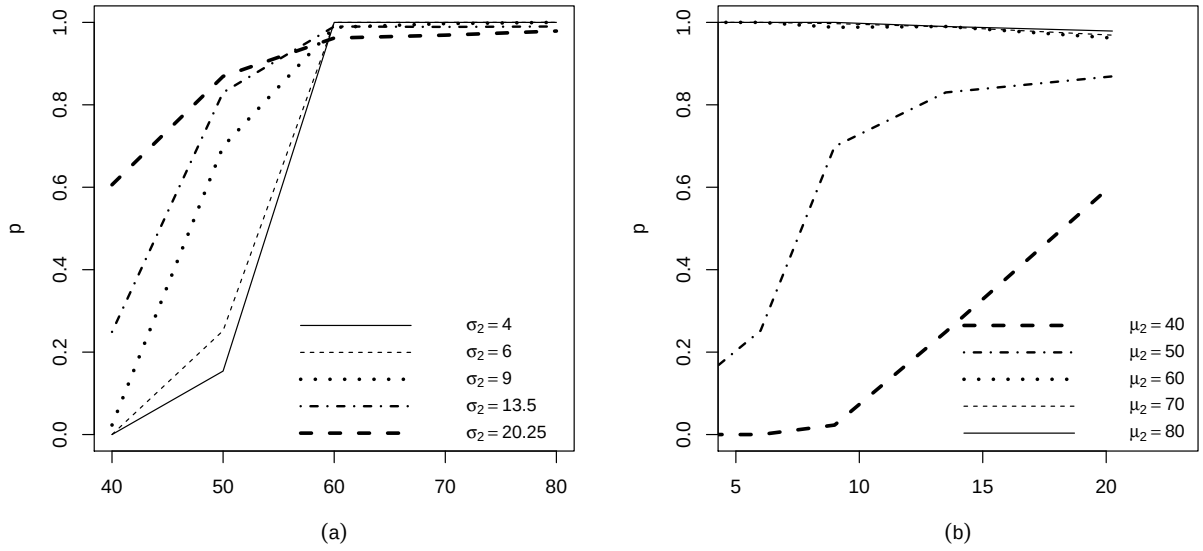


Figura 5.3: Gráficos para colunas e linhas da Tabela 5.2, respectivamente. Em (a) está o gráfico para a situação em que se toma valores fixos para o desvio padrão e a média é variada. Em (b) temos o gráfico dos valores das proporções tomando valores fixos das médias e variando o desvio padrão.

a partir da Tabela 5.2, para que possamos ter uma visão global do comportamento dos valores das proporções apresentados nessa tabela. Podemos notar, a partir desse gráfico, uma melhora da performance do modelo no que diz respeito a identificação de componentes distintas, quando se afasta as médias e os desvios padrões no modelo visto em 5.4, sendo que essa melhora é mais acentuada quando se afasta as médias.

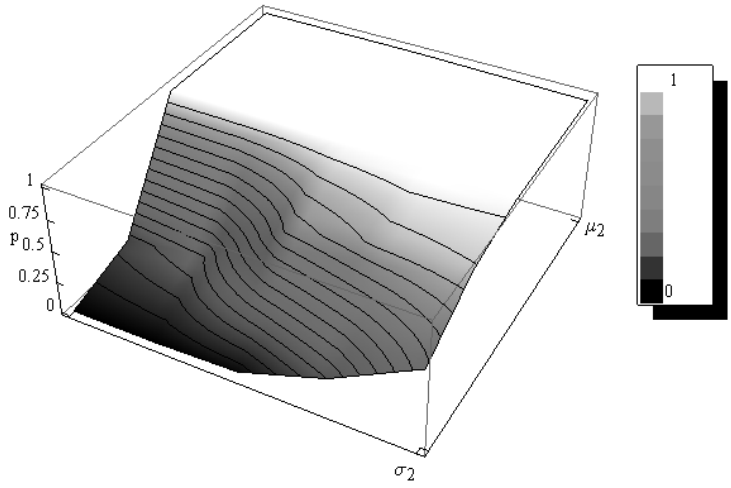


Figura 5.4: Gráfico de superfície exibindo os resultados das proporções exibidos na Tabela (5.2).

# Capítulo 6

# Aplicação

Neste Capítulo, aplicamos o método estudado anteriormente em três conjuntos de dados reais, que são considerados aqui como oriundos de misturas de distribuições normais. O primeiro conjunto trata-se de dados de acidez de lagos. O segundo diz respeito a medida de comprimento de peixes (em centímetros), usada para identificar grupos etários na população. O último conjunto de dados, são dados de tons musicais. Este último exemplo é uma aplicação em modelos de regressão com uma covariável, que aqui é considerada como variável. Os comandos do software R, usados para fazer as estimativas apresentadas neste trabalho se encontram no apêndice que se encontra no final do texto.

## 6.1 Dados de acidez

O conjunto de dados analisado nesta Seção, foi obtido a parti do site

http://www.inside-r.org/packages/cran/gamlss.data/docs/acidity

e contém índices de acidez de uma amostra de 155 lagos do estado de Wisconsin nos Estados Unidos. Este índice descreve a capacidade do lago em absorver a acidez. Valores baixos destes índices, podem levar a uma perda de recursos biológicos, pois é um indicativo de que o lago esteja acidificado. Estes dados foram analisados como uma mistura de distribuições normais na escala log, por Crawford *et al.* (1992), Crawford (1994) e também por Richardson & Green (1997), sendo que o último usa *reversible-jump* MCMC em sua análise.

Ajustamos o modelo descrito no Capítulo 4, neste conjunto de dados, considerando para a distribuição base valores para

$$m = -5,\ \eta = 1/100,\ a = 10,\ b = \frac{1}{40} \text{ e } \alpha = 1.$$

As probabilidades *a posteriori* para o número de componentes ($k$) da mistura, estão descritas na Tabela 6.1, onde observamos que com probabilidade *a posteriori* próximo de 1, é identificada duas componentes na população. Também observamos que com probabilidade *a posteriori* próximo de zero, são identificados três componentes e para os demais valores de $k$, as probabilidades *a posteriori* são zero, ou seja, o modelo não identifica mais que três componentes na população de onde foram extraídos os dados.

Tabela 6.1: Probabilidades *a posteriori* para número de grupos.

| k | probabilidades *a posteriori* . |
|---|---|
| 1 | 0 |
| 2 | 0.985 |
| 3 | 0.015 |

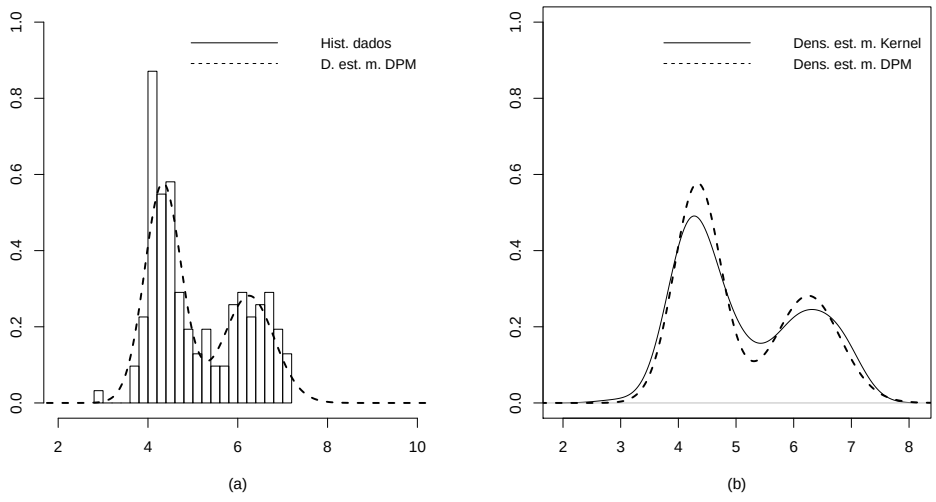


Figura 6.1: Em (a) está o gráfico da densidade estimada a partir da distribuição preditiva *a posteriori* , sobre o histograma dos dados de índices de acidez e em (b) a densidade estimada sobre a densidade dos dados de índices de acidez.

Tabela 6.2: Estimativas para os parâmetros obtidas a partir dos dados de índices de acidez em lagos, considerando $k = 2$.

| | Est. pontual | Int. HPD com 95% |
|---|---|---|
| $\omega_1$ | 0.599 | $(0.512, 0.675)$ |
| $\omega_2$ | 0.401 | $(0.324, 0.488)$ |
| $\mu_1$ | 4.34 | $(4.24, 4.45)$ |
| $\mu_2$ | 6.25 | $(6.07, 6.42)$ |
| $\sigma_1$ | 0.41 | $(0.34, 0.49)$ |
| $\sigma_2$ | 0.56 | $(0.46, 0.70)$ |

A Figura 6.1(a), mostra o gráfico da estimativa de densidade sobre o histograma dos dados e a Figura 6.1(b) mostra a curva da densidade dos dados, obtida utilizando o método do Kernel. O método do Kernel é um método não paramétrico para estimação de curvas de densidades, onde cada observação é ponderada pela distância em relação a um valor central. Este método é discutido em Silverman (1986) e Venables & Ripley (2002) e foi utilizado aqui por meio do software R com o comando *density*. Além da curva da estimativa de densidade obtida pelo método do Kernel, a Figura 6.1(b) também apresenta a curva da densidade estimada a partir do método apresentado neste trabalho. Nesta figura, observamos que as duas curvas estão próximas, como os métodos de estimação são diferentes, essa proximidade pode ser uma evidência de que se tem boas estimativas para a densidade dos dados.

Em cada iteração completa do algoritmo, foram obtidas as estimativas para os parâmetros. Estas estimativas estão apresentadas na Tabela 6.2, de modo que a segunda coluna mostra as estimativas pontuais para as proporções da mistura ($\omega$), para as médias ($\mu$) e para os desvios padrões ($\sigma$). Na última coluna estão os intervalos HPD com 95% de credibilidade para estes parâmetros. Todos os intervalos HPD (HPD, do inglês, Highest Posterior Density, significa maior densidade posterior), apresentados na Tabela 6.2, foram obtidos por meio do pacote MCMCpack de Martin *et al.* (2011). A análise feita em Richardson & Green (1997), a distribuição *a posteriori* para $k$ favorece $3-5$ componentes, sendo que a maior probabilidade *a posteriori* é para $k = 3$, em nossa análise a probabilidade *a posteriori* para $k = 3$ foi baixa em relação a probabilidade *a posteriori* para $k = 2$.

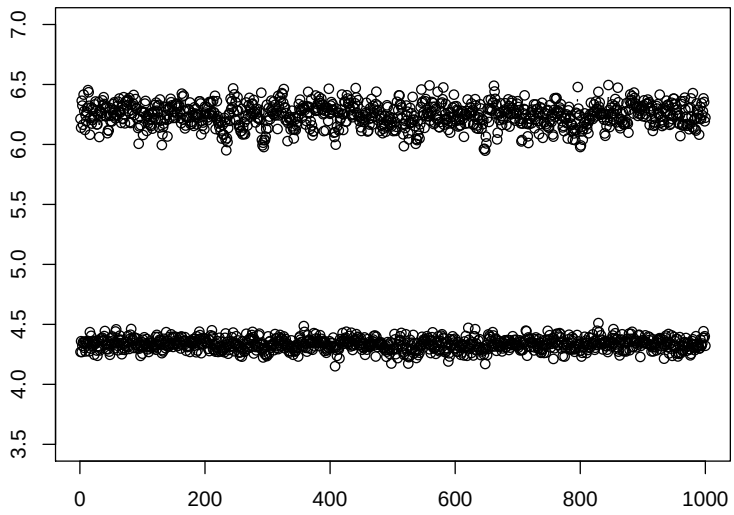


Figura 6.2: Gráfico de frequência de 1000 estimativas de médias obtidas considerando $k = 2$, para os dados de índices de acidez em lagos

A Figura 6.2 mostra o gráfico de sequência das distribuições *a posteriori* para as médias das distribuições em cada componente, para uma amostra *a posteriori* de tamanho $L = 1000$ já descontado o período de aquecimento. Verificamos que existem evidências de que a cadeia atingiu convergência para distribuição de equilíbrio, considerando $k = 2$ componentes. Com isso, temos que existe evidências favoráveis de que podemos classificar os lagos em dois grupos. Para o primeiro grupo temos um índice de acidez médio estimado de 4.34 e para o segundo temos uma acidez média estimada de 6.25.

## 6.2 Dados de comprimento de peixes

Os dados de comprimento de peixes foi obtido por intermédio do site

http://svitsrv25.epfl.ch/R-doc/library/bayesmix/html/fish.html

e pode ser visto em Titterington *et al.* (1985). Este conjunto contém dados referente ao comprimentos de 256 peixes da família dos Lutjanídeos. Como a desova dos peixes ocorre numa determina época do ano, a heterogeneidade nos dados surge devido aos grupos etários, deste modo um peixe que se originou devido a desova do ano corrente pertence ao grupo dos peixes menores e mais jovens, enquanto que um peixe que surgiu devido a desova do ano

Tabela 6.3: Probabilidades *a posteriori* para número de grupos.

| Grupos | probabilidades *a posteriori* . |
|---|---|
| $k < 4$ | 0 |
| $k = 4$ | 0.999 |
| $k = 5$ | 0.001 |
| $k > 5$ | 0 |

anterior pertence a um grupo com tamanho médio um pouco maior e assim por diante. Para este conjunto de dados Titterington *et al.* (1985) assumem que para um certo grupo de idade o comprimento de um peixe segue uma distribuição normal, alem disso a idade do peixe é desconhecida. Com isso, um peixe escolhido de forma aleatória pertence ao componente $j$ com probabilidade proporcional a $\omega_j$, com $j = 1, ..., k$ representando todos os componentes não vazios na mistura. As densidades dos componentes descrevem as distribuições de comprimentos dos peixes de diferentes idades e os pesos da mistura, $\omega_j$, indicam a distribuição etária dos peixes na população total.

O modelo descrito no Capítulo 4, também foi ajustado neste conjunto de dados, assumindo

$$m = -6,\ \eta = 1/100,\ a = 10,\ b = \frac{1}{40} \text{ e } \alpha = 1,$$

para a distribuição base $G_0 \sim Ng\big(m, \eta, a, b\big)$. Consideramos uma amostra *a posteriori* de tamanho $L = 1000$ já descontado o período de aquecimento de 1000 iterações completas. Para obter as estimativas para o número de componentes não vazias, usamos a moda da distribuição *a posteriori* . Em cada ciclo completo do algoritmo, obtivemos uma estimativa para os parâmetros, obtendo assim, uma amostra para cada parâmetro do modelo de mistura.

As probabilidades *a posteriori* para o número de componentes, $k$, está descrito na Tabela 6.3, onde observamos que com probabilidade *a posteriori* próximo de 1, são identificados quatro componentes na população, com probabilidade *a posteriori* próximo de zero, são identificados cinco grupos na população e para os demais valores de $k$ a probabilidade *a posteriori* é sempre zero. Com isso, vemos que existe evidencias favoráveis para a existência de quatro componentes na população.

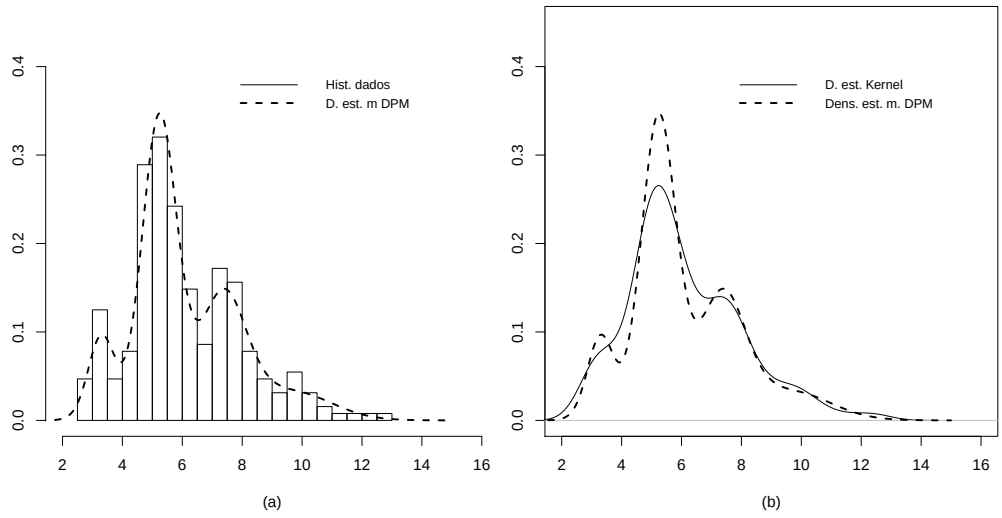


Figura 6.3: Gráfico da densidade dos dados estimados a partir da distribuição preditiva *a posteriori* sobre o histograma e a densidade dos dados de comprimeto de peixes.

A Figura 6.3(a) mostra a densidade da distribuição preditiva *a posteriori* sobreposta ao histograma dos dados observados e a Figura 6.3(b) mostra o gráfico da densidade estimada sobreposta à densidade estimada pelo método do Kernel. Observamos que as curvas de densidades estão próximas, o que indica que os dados são oriundos de uma densidade que está próxima destas estimativas.

Os parâmetros estimados condicionados a existência de quatro grupos na população foram, número de grupos ($k$), médias ($\mu$), desvios padrões ($\sigma$) e pesos da mistura ($\omega$). As médias das distribuições *a posteriori* para os parâmetros são usadas para obter as estimativas pontuais dos parâmetros condicionadas ao número de grupos e é mostrada na Tabela 6.4, juntamente com intervalos HPD com 95%.

Este conjunto de dados foi analisado por Frühwirth-Schnatter (2006) usando *reversible-jump* MCMC de Richardson & Green (1997), em que foi considerado quatro componentes *a priori* para a mistura. Os resultados obtidos em Frühwirth-Schnatter (2006) estão apresentados, neste trabalho, na tabela 6.4 ao lado dos resultados que obtivemos usando o modelo DPMN-N. Nesta tabela, podemos observar que as estimativas fornecidas pelo *reversible-jump* estão próximas dos resultados obtidos no nosso trabalho.

Tabela 6.4: Estimativas para os parâmetros dos dados de comprimento de peixes, obtida com o ajuste do modelo DPMN-N juntamente com as estimativas obtidas por Frühwirth-Schnatter (2006) usando *reversible-jump*.

| | Estimativas obtidas com DPMN-N | | Estimativas obtidas com o *reversible-jump* | |
|---|---|---|---|---|
| Parâmetros | Média *a posteriori* | HPD com 95% | Média *a posteriori* | HPD com 95% |
| $\omega_1$ | 0.12 | $(0.08, 0.16)$ | 0.115 | $(0.076, 0.164)$ |
| $\omega_2$ | 0.47 | $(0.35, 0.55)$ | 0.469 | $(0.336, 0.569)$ |
| $\omega_3$ | 0.28 | $(0.19, 0.43)$ | 0.222 | $(0.102, 0.377)$ |
| $\omega_4$ | 0.13 | $(0.04, 0.22)$ | 0.195 | $(0.084, 0.368)$ |
| $\mu_1$ | 3.33 | $(3.09, 3.57)$ | 3.30 | $(3.11, 3.58)$ |
| $\mu_2$ | 5.22 | $(5.08, 5.37)$ | 5.23 | $(5.06, 5.39)$ |
| $\mu_3$ | 7.23 | $(6.68, 7.71)$ | 7.29 | $(6.64, 7.63)$ |
| $\mu_4$ | 9.55 | $(8.43, 10.73)$ | 8.78 | $(7.36, 10.21)$ |
| $\sigma_1$ | 0.46 | $(0.33, 0.63)$ | 0.399 | $(0.25, 0.645)$ |
| $\sigma_2$ | 0.56 | $(0.45, 0.66)$ | 0.559 | $(0.425, 0.706)$ |
| $\sigma_3$ | 0.83 | $(0.56, 1.18)$ | 0.699 | $(0.349, 1.186)$ |
| $\sigma_4$ | 1.34 | $(0.95, 1.79)$ | 1.68 | $(1.01, 2.29)$ |

Fonte: Frühwirth-Schnatter (2006)

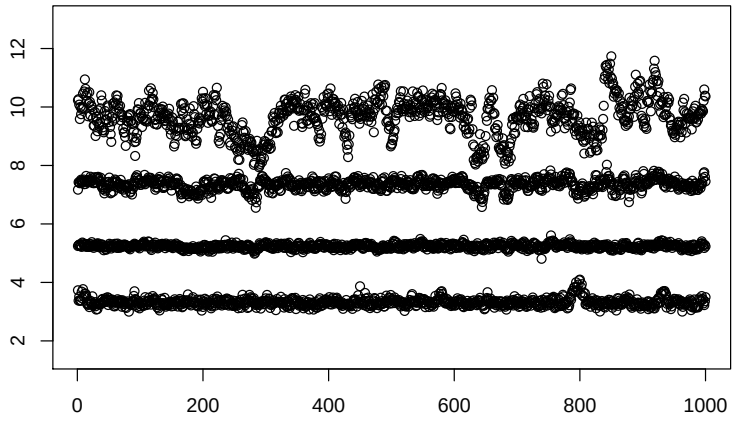


Figura 6.4: Gráfico de frequência das estimativas das médias das distribuições dentro dos grupos.

O gráfico de frequência das médias *a posteriori* para 1000 iterações, já descontado o período de aquecimento, é mostrado na Figura 6.4, onde verificamos que existe evidência de que a cadeia convergiu para a distribuição de equilibro, pelo menos para as três primeiras médias, considerando a ordenação do menor para o maior valor. Com este resultado, juntamente com os resultados apresentados na Tabela 6.4 e na figura 6.3, são bastante co-

erentes e fornecem evidências de que o modelo fornece resultados interessantes no que diz respeito a estimativa do número de componentes na mistura e também de estimativas para os parâmetros do modelo de mistura.

Os resultados apresentados na Tabela 6.4 indicam que exitem 4 grupos etários na população de peixes, sendo que os mais jovens constituem o grupo menor com uma proporção média de $0,12$ em relação ao total da amostra. Como o segundo grupo é bem maior, sendo constituído de quase a metade da quantidade total estimada (proporção média de $0,47$ em relação ao total da amostra), podemos concluir que o primeiro grupo ainda não está completo, devendo existir peixes tão pequenos que não foram considerados na amostra. Deste modo, se a idade máxima para esta espécie for de 4 a 5 anos, é compreensivo que o terceiro grupo seja menor que o segundo, pois no decorrer do tempo a população mais velha tende a diminuir devido a ataques predatórios e outros fatores. Se a expectativa de vida é de 5 anos, por volta dos 4 anos um peixe já estará atingindo seu tamanho máximo e chegando ao seu limite de tempo de vida.

## 6.3 Aplicação em mistura de modelos de regressão

Nesta Seção, o modelo descrito no Capítulo 4, é ilustrado pelos dados de percepção de tons que foram obtidos por intermédio do site

http://www.inside-r.org/packages/cran/fpc/docs/tonedata.

Este conjunto é resultado de uma experiência de Cohen (1980) e trata-se de dados de percepção de tons musicais, em que a taxa (ou razão) de sintonização do tom foi analisada por um músico treinado. Os resultados de 150 ensaios realizados com um único músico estão exibidos na figura 6.5, onde podemos observar que pelo menos duas retas são evidentes. Em situações como esta, uma mistura de modelos de regressão é apropriado, desde que não se tenha informação sobre a associação dos pontos para cada reta. Para fazermos uma breve revisão da mistura de modelos de regressão, vamos considerar o caso em que se tem duas componentes neste modelo,assim a mistura de modelo de regressão é dado por,

$$y_i = \begin{cases} \alpha_1 + \beta_1 x_i + \epsilon_{1i} & \text{com probabilidade } \omega \\ \alpha_2 + \beta_2 x_i + \epsilon_{2i} & \text{com probabilidade } 1-\omega \end{cases} \tag{6.1}$$

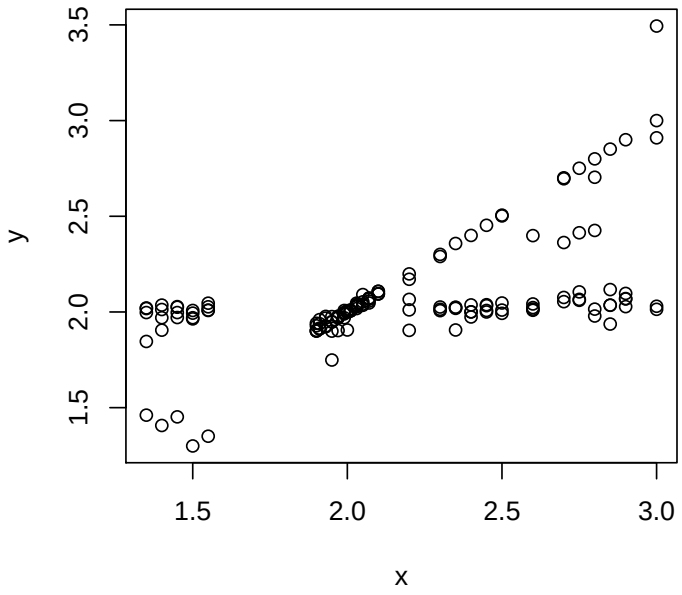


Figura 6.5: Gráfico de dispersão do valor sintonizado (variável resposta y ) versus a taxa de sintonização (variável explicativa x).

em que $\epsilon_{ji} \sim N(0, \sigma_j^2)$ para $j = 1, 2$ e $i = 1, ..., n$ são considerados independentes.

Vamos considerar aqui, uma matriz $X$ cuja dimensão é $n \times 2$ sendo que a primeira coluna é formada apenas por 1's e a segunda é formada por $(x_1, x_2, ..., x_n)$. Vamos supor também que $x_i$, para $1 \leq i \leq n$, são tais que $X'X$ é positiva definida para todo $n \geq 1$. Condicionando em $x_i$, os $y_i$'s tem função de densidade de probabilidade dada por,

$$f(y|x) = \omega f_1(y|x) + (1 - \omega) f_2(y|x) \tag{6.2}$$

em que $f_j(y|x)$ é a densidade da distribuição normal com média $\mu_j = \alpha_j + x_i\beta_j$ e variância $\sigma_j^2$, $j = 1, 2$. Vamos assumir que $\mu_1 \geq \mu_2$ para se fazer a distinção das densidades $f_1$ e $f_2$.

Agora faremos uma generalização da mistura de modelos de regressão vista até agora, para o caso em que se tem um número $k \geq 1$ de componentes não vazias, em que $k$ é

desconhecido. Para isso basta tomar $j = 1, ..., k$ em vez de $j = 1, 2$. Assim teremos

$$f(y|x) = \sum_{j=1}^{k} \omega_j f_j(y|x) \tag{6.3}$$

em que $\sum_{j=1}^{k} \omega_j$=1 com $0 \leq \omega_j \leq 1$, vamos considerar que cada $f_j(y|x)$ representa a densidade da distribuição normal com média $\mu_j = \alpha_j + x_i\beta_j$ e variância $\sigma_j^2$, $j = 1, ..., k$. Com isso, vamos considerar que nos dados de percepção de tons, o número de componentes é desconhecido, para aplicarmos o método estudado no Capítulo 4. A intenção aqui é obter uma estimativa para o número de componentes da mistura de modelos de regressão e com isso classificar os dados de percepção de tons, procurando associar cada par ordenado $(x_i, y_i)$ a sua componente.

Vamos considerar para os dados de percepção de tons, erros de medida nas variáveis explicativas. Com isso vamos assumir para este conjunto o modelo dado por,

$$\begin{aligned} Y_i|x_i, \alpha_i, \beta_i, \sigma_{i,y} &\sim N(\alpha_i + x_i\beta_i, \sigma_{i,y}^2) \\ X_i|\mu_{i,x}, \sigma_{i,x} &\sim N(\mu_i, \sigma_{i,x}^2) \\ (\mu_i, \sigma_{i,x}, \alpha_i, \beta_i, \sigma_{i,y})|G &\sim G \\ G|\alpha, G_0 &\sim DP(\alpha, G_0). \end{aligned} \tag{6.4}$$

Aqui, cada $X_i$ é assumido ter distribuição normal com vetor de parâmetro $\theta_{x,i} = (\mu_i, \sigma_{i,x})$ e $Y_i|X_i$ também tem distribuição normal com vetor de parâmetros $\theta_{i,y} = (\alpha_i + x_i\beta_i, \sigma_{i,y}^2)$. Desta forma, assumimos que $\theta_i = (\theta_{i,x}, \theta_{i,y})$ segue uma distribuição $G$, com $G$ seguindo uma distribuição que é um processo de Dirichlet, cuja distribuição base é dada por,

$$G_0 \rightarrow \begin{cases} \phi_{i,y} &\sim Ga(\frac{a_y}{2}, \frac{a_y b_y}{2}), \\ \alpha_i &\sim N(m_0, (\phi_{i,y} c_0)^{-1}, ) \\ \beta_i &\sim N(m_1, (\phi_{i,y} c_1)^{-1}), \\ \phi_{i,x} &\sim Ga(\frac{a_x}{2}, \frac{a_x b_x}{2}), \\ \mu_{i,x} &\sim N(m_x, (\phi_{i,x} c_x)^{-1}), \end{cases} \tag{6.5}$$

com $c_x, a_x, a_y, b_x$, $b_y$, $c_0$ e $c_1$ sendo valores reais positivos, $m_0 = 0$ e $m_x$, $m_1 \in \mathbb{R}$. Além disso, temos que $\phi_{i,y} = \sigma_{i,y}^{-2}$ e $\phi_{i,x} = \sigma_{i,x}^{-2}$.

A verossimilhança dos dados é escrita da forma,

$$\begin{aligned} f(x_i, y_i) &= f(y_i|x_i, \theta_{y,i}) f(x_i|\theta_{x,i}) \\ &= N(y_i|\theta_{y,i}) N(x_i|\theta_{x,i}) \end{aligned} \tag{6.6}$$

com $N(.|\theta)$ representando a densidade da distribuição normal com vetor de parâmetros $\theta$. As distribuições *a posteriori* para os parâmetros $\alpha_i, \beta_i, \sigma^2_{i,y}, \mu_i$ e $\sigma^2_{i,x}$, são obtidas combinando a verossimilhança dos dados, dada pela equação (6.6), com a distribuição base $G_0$, dada pela equação (6.5), de forma análoga ao Capítulo 4.

A maginal dos dados, $q_0$, é obtida por integração, em duas etapas, ou seja,

$$\begin{aligned} q_{0,y} &= \int Ga(\phi_{i,y}|\tfrac{a_y}{2}, \tfrac{a_y b_y}{2}) \int N(\beta_i|m_1, (\phi_{i,y}c_0)^{-1},) \int N(\alpha_i|m_0, (\phi_{i,y}c_0)^{-1},) N(y_i|\theta_{y,i}) d\theta_{y,i} \\ &= \frac{c_0 c_1}{\Gamma(\frac{a_y}{2})\sqrt{2\pi(1+c_0+c_0x_i^2)}} \int \phi_{i,y}^{\frac{a_y+1}{2}-1} \exp\{-\phi_{i,y}\tfrac{Q}{2}\} d\phi_{i,y} \end{aligned} \tag{6.7}$$

em que

$$Q = a_y b_y + \frac{c_0(y_i^2 - 2y_i m_0 + m_1^2) + m_1^2 - \frac{(m_1(1+c_0)+(y_i x_i - m_0 x_i))^2}{1+c_0+c_0x_i^2})}{1+c_0} \tag{6.8}$$

Fazendo $m = 0$ em (6.8), esta expressão se reduz a,

$$Q = a_y b_y + \frac{c_0(y_i - m_1 x_i)^2}{1 + c_0 + c_0 x_i^2} > 0. \tag{6.9}$$

Assim, obtemos o resultado da Expressão (6.7), utilizando o núcleo da distribuição gama, com isso,

$$q_{0,y} = \frac{\Gamma(\frac{a_y+1}{2})}{\Gamma(\frac{a_y}{2})} \frac{c_0 c_1 (\frac{Q}{2})^{-\frac{a_y+1}{2}}}{\sqrt{2\pi(1+c_0+c_0x_i^2)}} \tag{6.10}$$

Pelo que vimos no Capítulo (4),

$$\begin{aligned} q_{0,x} &= \int Ga(\phi_{i,x}|\tfrac{a_y}{2}, \tfrac{a_y b_y}{2}) \int N(\mu_{i,x}|\mu, (\phi_{i,x}c)^{-1}) N(y_i|\theta_{y,i}) d\theta_{x,i}. \\ &= \frac{\Gamma(\frac{a_x+1}{2})}{\Gamma(\frac{a_x}{2})} \left(\frac{c_x}{b_x(c_x+1)a_x\pi}\right)^{\frac{1}{2}} \left(1 + \frac{c_x(y_i-m_x)^2}{b_x(c_x+1)a_x}\right)^{-\frac{a_x+1}{2}}. \end{aligned} \tag{6.11}$$

Com isso, obtemos,

$$q_0 = (q_{0,y}).(q_{0,x}) \tag{6.12}$$

Esta expressão, juntamente com o parâmetro $\alpha$, fornece a probabilidade de que seja criado um novo componente em cada ciclo do algoritmo.

Para implementar o modelo (6.4) para os dados de percepção de tons, consideramos $m_x = -6,\ c_x = 1/1000,\ a_x = 2/3,\ a_y = 1/3,\ b_x = 1/2,\ b_y,\ c_0 = 1/1000,\ c_1 = 1/1000,\ m_0 = 0$ e $m_1 = 5$, para a distribuição base $G_0$ dada pela Equação (6.5). Após um período de aquecimento de 5000 iterações, uma amostra de tamanho $L = 1000$ foi selecionada para proceder a estimação dos indicadores de componentes. A estimação de cada indicador foi feita usando a moda da distribuição *a posteriori* , ou seja, se um ponto $(x, y)$ foi alocado maior quantidade de vezes no componente 1, então ele será alocado no componente 1 e assim por diante.

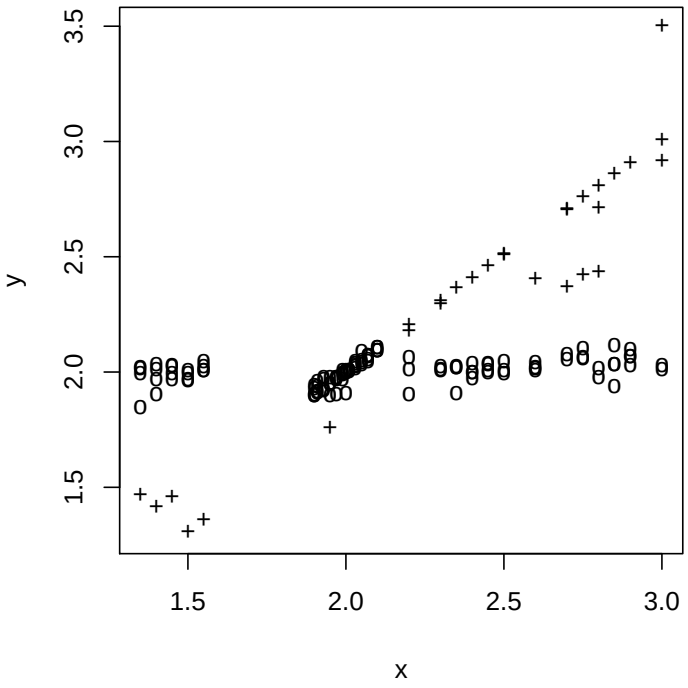


Figura 6.6: Gráfico de dispersão do valor sintonizado (variável resposta y ) versus a taxa de sintonização (variável explicativa x), após a classificação dos dados em dois componentes, sendo que os elementos do componente 1 são os pontos representados por '+' e os elementos do componente 2 são os pontos representados por 'o'.

Os componentes identificados pelo modelo foram distinguidos pela ordenação das médias da distribuição de $y|x$, conforme especificado no inicio desta Seção. O resultado desta classificação está exposto na Figura 6.6, onde podemos observar dois componentes distintos na amostra; a primeira componente é formada pelos pontos representados pelo simbolo '+'

e o segundo componente é formando pelos pontos representados por 'o'. A estimativa dos parâmetros do modelo requer uma análise multivariada, pois teríamos que proceder as estimativas dos parâmetros dados os componentes distintos identificados pelo modelo. Esta análise foge do escopo do nosso trabalho e não será apresentada aqui, ficando como proposta de trabalho futuro. O conjunto de dados apresentado nesta Seção também foi analisado por de Veaux (1989) e em sua análise, de Veaux (1989) estima dois componentes para esta mistura de modelos de regressão.

Para mais informações a respeito da aplicação do modelo DPM em modelos de regressão, podem ser consultados, por exemplo, Shahbaba & Neal (2009) e Hannah *et al.* (2011).

# Capítulo 7

# Discussão

Apresentamos, neste trabalho, um estudo de desempenho para os modelos DPM relacionados a dois aspectos práticos da estimativa de mistura de distribuições normais. A primeira discussão considera a influência dos parâmetros da distribuição base normal-gama sobre a estimativa do número de componentes na mistura. O segundo aspecto refere-se às diferenças entre as médias e variâncias das distribuições dos componentes e da sua influência na estimativa do número de componentes na população.

Com base nos valores das proporções visto na Tabela 5.1, concluímos que os estudos de simulação discutidos no Capítulo 5.2 indicam que o parâmetro $m$ da distribuição de base normal-gama-invertida é relevante para a precisão do método, quando consideramos que os dados observados seguem uma mistura de distribuições normais.

A primeira colula da Tabela 5.2 mostra que o modelo DPMN-N é sensível à variabilidade dos dados, ou seja, o modelo distingue componentes se suas variâncias são diferentes, mesmo quando se tem médias iguais. Esse aspecto favorece a identificação de componentes distintos na amostra, pois ajuda a identificar a multimodalidade nos dados quando se tem valores de médias próximos entre si. É claro que, quando se tem valores de médias bem afastados, uma alta variabilidade dos dados em cada componente desfavorece o método, pois, neste caso, variâncias pequenas fazem com que os dados que compartilham a mesma média tenham valores próximos entre si, e afastados em relação aos demais valores de dados.

O Capítulo 6.3 mostra uma aplicação do modelo DPMN-N em mistura de modelos de regressão, em que o modelo identificou dois componentes na amostra e classificou os dados de

forma coerente, como mostra a Figura 6.6. A estimação dos parâmetros, nesta aplicação, foi omitida. No entanto, é possível estimar os parâmetros dada a classificação dos dados. Como proposta para trabalhos futuros, vamos deixar a estimação dos parâmetros em modelos deste tipo, além da atribuição de uma distribuição ao parâmetro $m$ da distribuição base normal-gama, que aqui foi fixado em alguns valores, e também fornecer uma distribuição ao parâmetro $\alpha$ do processo de Dirichlet, que foi fixado igual a 1. Assim, iremos atribuir mais um nível ao modelo DPMN-N e analisar o seu desempenho nestas circunstância.

# Referências


Antoniak, C. E. . (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Statistics*, **2**(6), 1152–1174.

Bernardo, J. & Smith, A. (1994). *Bayesian Theory*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Wiley.

Bickel, P. & Doksum, K. (2001). *Mathematical statistics: basic ideas and selected topics*. Number v. 1 in Mathematical Statistics: Basic Ideas and Selected Topics. Prentice Hall.

Blackwell, D. (1973). Discreteness of Ferguson Selections. *The Annals of Statistics*, **1**(2), 356–358.

Blackwell, D. & MacQueen, J. B. (1973). Ferguson distributions via Polya urn schemes. *Annals of Statistics*, **1**(2), 353–355.

Blei, D. M., Griffiths, T. L. & Jordan, M. I. (2010). The nested chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. ACM*, **57**(2), 7:1–7:30.

Bush, C. A. & MacEachern, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika*, **83**(2), 275–285.

Chib, S. & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *American Statistician*, **49**(4), 327–335.

Cohen, E. A. (1980). *Inharmonic tone perception*. Mestrado, Stanford University.

Crawford, S. L. (1994). An application of the Laplace method to finite mixture distributions. *Journal of the American Statistical Association*, **89**(425), 259–267.

Crawford, S. L., DeGroot, M. H., Kadane, J. B. & Small, M. J. (1992). Modeling lake-chemistry distributions: approximate Bayesian methods for estimating a finite-mixture model. *Technometrics*, **34**(4), 441–453.

de Veaux, R. D. (1989). Mixtures of linear regressions. *Comput. Stat. Data Anal.*, **8**(3), 227–245.

Escobar, M. D. (1994). Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, **89**(425), 268–277.

Escobar, M. D. & West, M. (1994). Hierarchical priors and mixture models with application in regression and density estimation. *Statistics*, pages 363–386.

Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**(430), 577–588.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**(2), 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, **2**(4), 615–629.

Ferreira Da Silva, A. R. (2007). A Dirichlet process mixture model for brain mri tissue classification. *Medical Image Analysis*, **11**(2), 169–182.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer series in statistics. Springer, Dordrecht.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**(410), 398–409.

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

Ghosh, J. & Ramamoorthi, R. (2003). *Bayesian nonparametrics*. Springer Series in Statistics. Springer-Verlag.

Hannah, L. A., Blei, D. M. & Powell, W. B. (2011). Dirichlet process mixtures of generalized linear models. *J. Mach. Learn. Res.*, **12**, 1923–1953.

Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**(453), 161–173.

MacEachern, S. N. & Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal Of Computational And Graphical Statistics*, **7**(2), 223–238.

Martin, A. D., Quinn, K. M. & Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*, **42**(9), 22.

Müller, P., Erkanli, A. & West, M. I. K. E. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, **83**(1), 67–79.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal Of Computational And Graphical Statistics*, **9**(2), 249.

Quintana, F. A. & Müller, P. (2004). Nonparametric bayesian data analysis. *Statistical Science*, **19**(1), 95–110.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, **59**(4), 731–792.

Shahbaba, B. & Neal, R. (2009). Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research*, **10**(0707), 1829–1850.

Sheng, Q., Thijs, G., Moreau, Y. & Moor, B. D. (2005). Applications of Gibbs sampling in bioinformatics. *Optimization Methods and Software*, **00**(00), 1–15.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall. ISBN 9780412246203.

Tanner, M. A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Springer Series in Statistics. Springer.

Teh, Y. W. (2003). Dirichlet process. *The annals of applied statistics*, **4**(2), 1–11.

Titterington, D., Smith, A. & Makov, U. (1985). *Statistical analysis of finite mixture distributions*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Venables, W. & Ripley, B. (2002). *Modern Applied Statistics with S*. Statistics and Computing. Springer. ISBN 9780387954578.

Wilks, S. (1962). *Mathematical Statistics: Samuel S. Wilks*. Wiley series in probability and mathematical statistics. John Wiley.

# Apendice

Código do software R, que foram usados para gerar os dados usados para as análises feitas no Capítulo 5.

```
n=100; k.s=2
set.seed(03062005) #semente
mu.s=c(40,60) #vetor de médias
sigma.s=c(4,9) #vetor de desvios padrões
id=sample(1:k.s, size=n, replace=TRUE, prob=rep(1,k.s)/k.s )
x=rnorm(n,mu.s[id],sigma.s[id])
```

Segue os comandos do software R usados para obtenção de estimativas de indicadores de componentes e os demais parâmetros usando o modelo DPMN-N. Este código foi usado para os dados de acidez. O código para as demais simulações, com dados simulados e com os dados de comprimento de peixes, são similares a este e foram omitidos.

```
library(MASS)
library(coda)
library(lattice)library(MCMCpack)
library(msm)
library(mixAK)#=================================================================;
library("bayesmix")
data(Acidity)
x=Acidity
#=================================================================;
n.s=1000  # número de iterações sem periodo de burnig
bn=1000   # burning
sa=10     # salto
#=================================================================;
#chutes
n=length(x); mu=runif(n,min(x),max(x)); sig=1/runif(n,5,5.5);
g=seq(1:n); mi=mu; phi=sig; gr=g; f=rep(1,n); ka=n; sq=seq(1:(n+1))
#=================================================================;
#Hiper-parâmetros;
```

```
alfa=1; m=-5; c0 =1/100; n0=10; sigma0=1/40;
lambda=c0/((c0+1)*sigma0) # parametro da t de student
k=(gamma((n0+1)/2)/gamma(n0/2))*(1/n0*pi)^(1/2)
#============================================================;
#definindo matrizes
NR=0; KF=numeric(); KAF=numeric()
#============================================================;
#Função para cálculo da d. preditiva
yp=seq(0,15,0.01)
pre=function(yp,alfa,k,lambda,m,n0){
ee=numeric()
for(r in 1:length(yp))
ee[r]= 1/(alfa+n)*(alfa*k*(1+(lambda)*1/n0*(yp[r]-m)^2)^(-(n0+1)/2)+
sum(f[1:ka]*dnorm(yp[r],mi[1:ka],1/sqrt(phi[1:ka]))))
ee}
#============================================================================;
cat( file="P", append=FALSE);
cat( file="MU", append=FALSE);
cat( file="SIGMA", append=FALSE);
cat( file="G", append=FALSE);
cat( file="W", append=FALSE);   gg=numeric()
  KA=numeric(50); KAA=numeric(50);   Y=numeric() ; YY=numeric()
 for(z in 1:n.s){
for(i in 1:n) {
f[gr==g[i]]=f[gr==g[i]]-1
aux  =f[1:ka]*dnorm(x[i],mi[1:ka],1/sqrt(phi[1:ka]))
aux[ka+1] = alfa*k*(1+(lambda)*1/n0*(x[i]-m)^2)^(-(n0+1)/2)
isa = sample(1:(ka+1),size=1,prob=aux/sum(aux))
if(isa==(ka+1)){
sig[i]=rgamma(1,(n0+1)/2,rate=(n0*sigma0/2+c0*((x[i]-m)^2)/(c0+1)*2))
mu[i] = rnorm(1,(c0*m+x[i])/(1+c0),sqrt(1/sig[i]*(1+c0)))
g[i]= max(g)+1
phi=unique(sig)
mi=unique(mu)
gr=unique(g) }else{
mu[i]=mi[isa]
sig[i]=phi[isa]
mi=unique(mu)
phi=unique(sig)
g[i]=gr[isa]
gr=unique(g)}
ka=length(gr)
for(l in 1:ka){f[l]=length(which(g==gr[l])); f=f[1:ka]}}
for (j in 1:ka){
 y  = x[g==gr[j]]
```

```
 nj = length(y)
phi[j] =
rgamma(1,(nj+n0)/2,(n0*sigma0/2+sum((y-mean(y))^2)/2+
(c0*nj*(m-mean(y))^2)/(c0+nj)*2))
mi[j]  =
rnorm(1,(c0*m+nj*mean(y))/(c0+nj),sqrt(1/(phi[j]*(c0+nj))))
mu[g==gr[j]]=mi[j]
sig[g==gr[j]]=phi[j] }
Y=pre(yp,alfa,k,lambda,m,n0)
or=order(mi[1:ka], decreasing=TRUE)
mi=mi[or]
phi=phi[or]
f=f[or]
zz=sort(unique(g))
for(l in 1:ka) {g[g==zz[l]]=l}
gr=unique(g)
gr=gr[or]
#==============atualização de w
 delta=rep(1,ka)
 n.z=numeric(ka)
 for(l in 1:ka){
 n.z[l]=length(which(g==l))  }
w=rdirichlet(1,n.z+delta)
cat(c(Y),"\n",file="P", append=TRUE)
cat(c(mi,rep(0,(100-ka)) ),"\n",file="MU", append=TRUE)
cat(c(1/sqrt(phi),rep(0,(100-ka))) ,"\n",file="SIGMA", append=TRUE)
cat(g,"\n",file="G", append=TRUE)
cat(c(w,rep(0,(100-ka)) ),"\n",file="W", append=TRUE)}
#fim das 1000 iterações para a amostra
#=============================================================================
PP=read.table("P")   #Amostra estimada
M=read.table("MU")   #médias
V=read.table("SIGMA")#Variancias
Z=read.table("G")     # indicadores de grupo
W=read.table("W")     # pesos
```