

Modelo de Mistura com Dependência Markoviana de Primeira Ordem

SILVANA APARECIDA MEIRA

UFSCar - São Carlos/SP

Outubro/2014

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

Modelo de Mistura com Dependência Markoviana de Primeira Ordem

SILVANA APARECIDA MEIRA

ORIENTADOR: PROF. DR. LUÍS APARECIDO MILAN

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

UFSCar - São Carlos/SP

Outubro/2014

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

M514mm Meira, Silvana Aparecida.
Modelo de mistura com dependência Markoviana de primeira ordem / Silvana Aparecida Meira. -- São Carlos : UFSCar, 2014.
78 f.

Tese (Doutorado) -- Universidade Federal de São Carlos, 2014.

1. Estatística. 2. Modelo de mistura. 3. Estimador de máxima verossimilhança. 4. Teoria bayesiana da estimação. 5. Algoritmos. 6. Basquetebol. I. Título.

CDD: 519.5 (20^a)



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) : Silvana Aparecida Meira

TESE DE DOUTORADO DEFENDIDA E APROVADA EM 12/09/2014 PELA
COMISSÃO JULGADORA:

Presidente

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

1º Examinador

Prof. Dr. Erlandson Ferreira Saraiva (UFMS)

2º Examinador

Prof. Dr. José Galvão Leite (DEs-UFSCar)

3º Examinador

Prof. Dr. Luis Gustavo Esteves (IME-USP)

4º Examinador

Profa. Dra. Vera Lúcia Damasceno Tomazella (DEs-UFSCar)

Agradecimentos

À Deus pela vida, força e coragem para complementar mais esta etapa em minha vida.

A minha família pelo apoio e incentivo.

Ao meu orientador Milan, pela dedicação, compreensão, apoio nos momentos difíceis, paciência e incentivo para a realização deste trabalho.

Ao prof. Galvão pela colaboração e sugestões e ao prof. Luis Gustavo pelas sugestões na qualificação.

Ao prof. Erlandson, prof. Galvão, prof. Luis Gustavo e a prof^a Vera pelas sugestões, comentários e críticas construtivas da tese.

Aos professores do departamento de estatística da UFSCar, que de forma direta ou indireta participaram da minha formação acadêmica.

Aos amigos que conheci na pós-graduação.

Aos funcionários do departamento de estatística da UFSCar, principalmente a Isabel pelas palavras de incentivo.

A todos que de maneira direta ou indireta possibilitaram completar mais esta etapa em minha vida.

Resumo

Nesse trabalho apresentamos o modelo de mistura com dependência markoviana de primeira ordem, MMM(1). A metodologia proposta corresponde a uma redefinição do modelo markoviano oculto (HMM) na qual utilizamos uma variável não observável como controladora da mistura. O modelo de mistura usual (sem dependência) é um caso particular do MMM(1).

A redefinição proposta permite uma adaptação de instrumentos usuais de estimação como por exemplo o algoritmo EM.

Apresentamos também os estimadores de máxima verossimilhança e bayesianos para os modelos MMM(1) e de mistura usual para os casos da distribuição normal e binomial.

Estudos de simulação demonstram a funcionalidade do modelo e estimadores propostos. Ao final apresentamos uma aplicação a um conjunto de dados reais apresentados na literatura para o caso binomial.

Abstract

We present the mixture model with first order dependence, MMM(1). This model corresponds to a redefinition of the hidden Markov model (HMM) where a non observable variable is used to control the mixture. The usual mixture model is a particular case of the MMM(1).

The proposed redefinition makes easier the application of usual estimation tools as the EM algorithm.

We present the maximum likelihood and Bayesian estimators for the normal and binomial cases of the MMM(1) and usual mixture models.

Simulation studies show the functionality of the proposed models and their estimators. And finally we present an application to a real data set for the binomial case.

Sumário

Agradecimentos	i
Resumo	ii
Abstract	iii
1 Introdução	1
2 Modelos de Mistura	6
2.1 Modelo de Mistura	7
2.1.1 Estimação via máxima verossimilhança	8
2.1.2 Algoritmo EM	10
2.2 Modelo de Mistura com Dependência Markoviana	12
2.2.1 Alguns resultados	14
2.2.2 Função de Verossimilhança	17
2.2.3 Algoritmo EM	20
3 Mistura de normais	23
3.1 MMM(1) de normais	26
3.2 Abordagem Bayesiana	28
3.2.1 Algoritmo MCMC	32

3.3	Abordagem Bayesiana para MMM(1)	33
3.3.1	Algoritmo MCMC para MMM(1)	37
4	Mistura de binomiais	39
4.1	MMM(1) de binomiais	41
4.2	Abordagem Bayesiana	43
4.3	Abordagem Bayesiana para MMM(1)	45
5	Aplicação de Mistura de Binomiais	51
5.1	Dados Vinnie Johnson	51
5.1.1	Estimação dos parâmetros dos modelos via Verossimilhança	52
5.1.2	Abordagem Bayesiana	55
5.2	Estudo de Simulação	59
5.2.1	Abordagem via Verossimilhança para MMM(1) usando algoritmo EM adaptado	59
5.2.2	Abordagem bayesiana para MMM(1) usando MCMC	60
6	Discussão	65
	Referências	67
A	Algoritmo em R para simulação	69
A.1	Simulação de mistura de binomial independentes e para MMM(1) com K componentes	69
B	Algoritmo em R para calculo de EM	71
B.1	Mistura de binomial independentes com K componentes	71
B.2	MMM(1) de distribuição binomial K componentes	73

C	Algoritmo em R para cálculo do MCMC	74
C.1	Mistura de binomial independentes com K componentes	74
C.2	Mistura de binomial dependentes com K componentes	76

Capítulo 1

Introdução

A primeira aplicação conhecida de mistura de distribuições consiste de uma soma ponderada de duas distribuições normais feita por Karl Pearson em 1890 (McLachlan & Peel, 2000). Podemos observar através da Figura (1.1) uma representação gráfica do modelo de mistura.

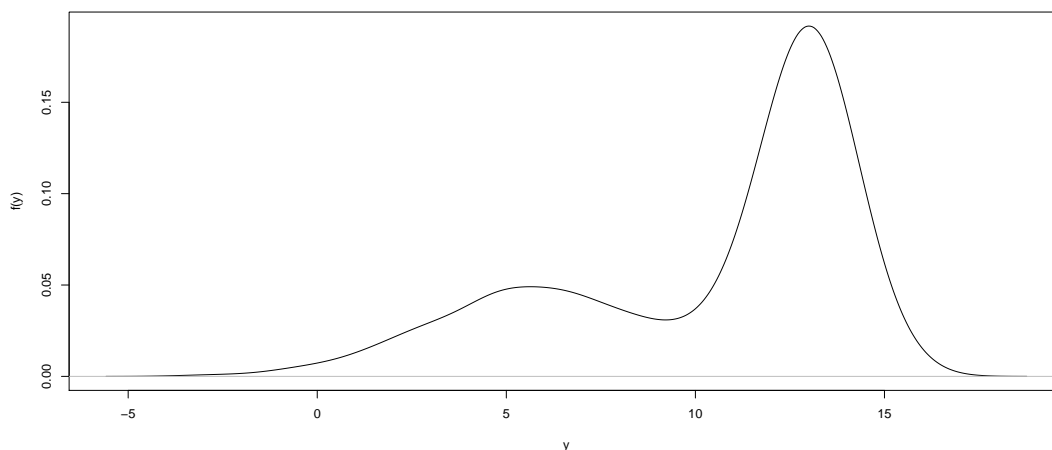


FIGURA 1.1: Mistura de distribuições

O modelo de mistura considerando o número de componentes, K , conhecido pode ser expresso por

$$\begin{aligned} f(y) &= w_1 f_1(y) + \dots + w_K f_K(y), \\ &= \sum_{k=1}^K w_k f_k(y), \end{aligned} \tag{1.1}$$

em que $w_k > 0$, $k = 1, \dots, K$ e $\sum_{k=1}^K w_k = 1$. Na forma original do modelo de mistura (1.1), os w 's são pesos ou ponderações e as $f(\cdot)$ são funções densidades que compõem a mistura de distribuições com suas componentes.

Diebold & Robert (1994) apresentam o modelo de mistura com variáveis latentes fazendo estimação bayesiana considerando K conhecido e aplicando a mistura de distribuições normais. Chib (1996) apresentam o fator de Bayes para o caso de mistura de distribuições normais. Escobar (1995) usam inferência bayesiana para estimação de mistura de distribuições normais usando mistura de processo de Diriclet. Choy & Chan (2008) Apresentam mistura de distribuição aplicada em distribuições escalas usando algoritmo MCMC eficiente para modelos estatísticos complicados. Mais referências sobre o modelo de mistura, tal como apresentado em (1.1), e suas aplicações podem ser encontradas em Titterington & Smith (1985) ou McLachlan & Peel (2000).

Os modelos que utilizam misturas de distribuições têm sido empregados em várias áreas do conhecimento, tais como na astronomia, biologia, genética, medicina, economia, engenharia, entre outras. Uma área onde a modelagem com mistura de distribuições apresenta resultados interessantes é na biologia molecular. Um exemplo desse tipo de aplicação pode ser visto em Coimbra (2005), que utiliza uma abordagem bayesiana para o modelo de mistura de distribuições multinomiais e variáveis latentes para identificar proteínas similares. Missão (2007) aplica o modelo de mistura a dados sobre o aproveitamento do jogador de basquetebol Vinnie Johnson.

Diversos métodos de estimação têm sido utilizados em modelos de mistura com enfoques clássico e bayesiano. Apresentamos a seguir algumas referências. Gelman *et al.* (1995), trabalha o modelo de mistura utilizando um algoritmo baseado em dados ampliados. Celeux & Soromenho (1999) Utilizam um critério de entropia para estimar o número de componentes de mistura do modelo de mistura. Missão (2007) utiliza máxima verossimilhança na estimação dos parâmetros e o fator de Bayes para selecionar o modelo de mistura mais adequado para um determinado conjunto de dados com o objetivo de encontrar o número de componentes do modelo. Roeder & Wasserman (1997) estimam o valor de K usando densidade preditiva considerando o modelo de mistura de normais. Richardson & Green (1997) apresenta uma análise bayesiana para o modelo de mistura com número de componentes desconhecido usando o método de Monte Carlo em

Cadeia de Markov (MCMC), introduzindo o algoritmo *reversible jump* para a estimação dos parâmetros do modelo de mistura conjuntamente com o número de componentes e aplicam ao acaso de mistura de distribuições normais. Stephens (2000) propõe um método alternativo usando um processo de nascimento e morte estacionário para o modelo de mistura essa metodologia: é uma versão contínua para o método MCMC *Reversible Jump*. Saraiva (2009) apresenta o algoritmo *split-merge* MCMC, que propõe novas componentes a partir da partição/junção dos dados e então propõe novos parâmetros. O método faz uma partição nos dados, *split*, e, condicional a esta partição propõe as novas componentes, e, pela proposta *merge*, diminui uma componente juntando componentes existentes.

Na literatura, o *modelo markoviano oculto*, é uma tradução para o português da expressão *Hidden Markov Models*, que explica também a sigla pela qual o modelo é conhecido, HMM pode ser visto em McLachlan & Peel (2000). Na sua forma original, o HMM tem recebido diversos tratamentos estatísticos.

Passamos a seguir a citar algumas referências sobre essas diversas abordagens. Baum *et al.* (1970) considera uma técnica iterativa para a estimação dos parâmetros de uma cadeia de Markov usando verossimilhança. Para a estimação do modelo markoviano oculto temos que usar um procedimento *forward-backward*, para estimar a sequência oculta usamos o algoritmo de Viterbi que otimiza a sequência e minimiza o número de caminhos possíveis, e para a estimação dos parâmetros e da sequência conjuntamente usamos o método de *Baum-Welch*. Estes procedimentos podem ser encontrados em Dugad & Desai (1996). Chib (1996) apresenta estimativas modais para calcular uma versão estocástica do algoritmo EM. Holst & Lindgren (1991) propõem um algoritmo recursivo para os parâmetros de modelos de mistura, onde as observações são governadas por uma cadeia de Markov oculta. Lember (2006) propões o algoritmo *Viterbi training* como uma alternativa computacional para o algoritmo *forward-backward* ou Baum-Welch. Koloydenko *et al.* (2007) provaram via teoria assintótica a propriedades *adjusted Viterbi training* para modelo de Markov oculto. Altman (2007) apresenta uma nova classe de modelos de mistura usando HMM para dados longitudinais. Schnatter (2008) usa a distribuição preditiva *a posteriori* para analisar a falta de ajuste do modelo em mistura markoviana. Dias *et al.* (2010) utilizam o HMM em pesquisas financeira usando dados de séries temporais.

Objetivo

Nosso interesse nesse trabalho está focado no modelo de mistura em que a sequência de variáveis que controla a componente da mistura no modelo é uma cadeia de Markov. Uma novidade sobre o HMM que defendemos nesse trabalho é o tratamento do mesmo como um modelo de mistura. Esse tratamento abrirá novas possibilidades para a abordagem estatística para o mesmo.

Neste trabalho, consideramos uma sequência de variáveis aleatórias discretas S_t , $t = 1, \dots, T$, que controlam qual componente da mistura está presente em cada ocorrência da sequência de variáveis aleatórias Y_t , $t = 1, \dots, T$. Naturalmente, as variáveis aleatórias S_t 's não são observáveis e a estrutura de dependência entre elas define modelos diversos tais como o modelo de mistura tradicional e o similar ao HMM.

Existem diversas situações em que o modelo de mistura com dependência markoviana de primeira ordem pode ser aplicado. São aplicações naturais a modelagem de cadeias de DNA, de sequência de jogos, de ações da bolsa de valores, etc.. Nestas situações, podemos supor uma sequência variáveis aleatórias não observáveis S_t , regida por uma cadeia de Markov, e que controla as componentes da mistura em cada ocorrência da variável observável Y_t . O modelo de mistura passa a ter componentes de mistura dada pela matriz de transição.

Para os modelos propostos, encontramos estimadores de máxima verossimilhança, usando uma adaptação do algoritmo EM, e bayesianos através do método MCMC. A estimação do número de componentes dos modelos é feita através de métodos de seleção de modelos tais como AIC (*Akaike Information Criteria*), BIC (*Bayesian Information Criteria*) e Fator de Bayes.

Desenvolvemos os estimadores para os parâmetros uma distribuição discreta, a binomial, e uma contínua, a normal. Os modelos propostos são aplicados a dados simulados e a dados reais.

Dentre as contribuições apresentadas nesse trabalho, podemos citar: formalização do modelo de mistura tradicional com uma variável de controle cuja distribuição substitui os tradicionais pesos; a formalização do modelo HMM na forma de um modelo de mistura cujas variáveis controladoras seguem o modelo da cadeia de Markov e tendo

como caso particular o modelo de mistura tradicional apresentado nessa nova forma; os estimadores de máxima verossimilhança acompanhados dos respectivos algoritmos EM devidamente modificados; os estimadores bayesianos para ambos os casos (modelos de mistura com dependência de primeira ordem e independente); os estimadores para casos específicos das distribuições binomial e normal; a exposição de limites dos métodos via simulação; e a aplicação dos métodos aos dados reais. Dentre as vantagens dos métodos propostos, podemos enfatizar os estimadores de máxima verossimilhança que substituem os complicados algoritmos *forward-backward*, Viterbi e Bawn-Welsh.

No Capítulo 2, apresentamos a definição de modelo de mistura sob uma nova formalização probabilística, onde a usual ponderação é substituída pela distribuição das variáveis aleatórias S_t . No Capítulo 3, aplicamos os modelos propostos para o caso específico da distribuição normal. No Capítulo 4, aplicamos os modelos propostos para o caso específico da distribuição binomial. No Capítulo 5, aplicamos os modelos a dados reais usando máxima verossimilhança e a abordagem bayesiana para a distribuição binomial. Finalmente, no Capítulo 6, apresentamos as considerações finais.

Capítulo 2

Modelos de Mistura

Neste capítulo, apresentamos o modelo de mistura considerando o número de componentes de mistura, K conhecido. A estimação dos parâmetros do modelo de mistura é feita via máxima verossimilhança usando uma adaptação do algoritmo Estimação e Maximização (EM) para o modelo apresentado.

Antes de apresentar a definição do modelo de mistura, introduzimos um exemplo simples, buscando estabelecer uma ideia inicial.

Exemplo 2.1 *Considere lançamento de uma moeda e dois dados, A e B , com probabilidades diferentes para as mesmas faces. Considere o experimento de lançar a moeda e verificar o resultado. Se o resultado for cara, lançamos a seguir o dado A , e se for coroa lançamos o dado B . Nosso interesse está no resultado do lançamento do dado.*

Representando o resultado do lançamento da moeda pela variável aleatória $S_t \in \{1, 2\}$, temos que $S_t = 1$ se o resultado for cara e $S_t = 2$ se o resultado for coroa.

O resultado do lançamento do dado é representado pela variável aleatória Y , dessa forma, Y assume o valor da face do dado, $Y = 1$ se obtemos a face 1 e assim por diante. O espaço de eventos é dado por $\Omega = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6)\}$. Não é difícil notar que ao condicionarmos em S , particionamos o espaço em duas partes, que corresponde aos eventos associados a cada um dos dados.

A distribuição da variável aleatória Y é então dada por

$$Pr(Y = y) = Pr(S_t = 1)Pr(Y = y|S_t = 1) + Pr(S_t = 2)Pr(Y = y|S_t = 2), \quad (2.1)$$

para $y \in \{1, 2, 3, 4, 5, 6\}$.

Em geral, nos problemas envolvendo o modelo de mistura, o resultado do lançamento da moeda não está disponível, somente o resultado do lançamento do dado. Não sabemos portanto qual dos dois dados foi lançado.

Dizemos que a distribuição de Y é uma mistura das duas distribuições condicionais da variável aleatória $Y|S_t = 1$ e $Y|S_t = 2$. Temos também que o número de componentes nesse caso é 2 pelo fato de termos dois dados disponíveis e, portanto, a distribuição de Y é a mistura das duas distribuições, cada uma associada a um dado.

□

2.1 Modelo de Mistura

Considere uma família de variáveis aleatórias discretas independentes $S = \{S_t; t = 1, \dots, T\}$, S_t com distribuição multinomial, $S_t \sim \text{mult}(1, \mathbf{p} = (p_1, \dots, p_K))$, com K resultados possíveis e $\sum_{k=1}^K p_k = 1$. Considere também uma família de variáveis aleatórias $Y = \{Y_t; t = 1, \dots, T\}$, sendo que a distribuição de cada variável aleatória Y_t é controlada pelo valor assumido pela variável aleatória S_t correspondente, ou seja, $Y_t|S_t = k, \theta$ tem distribuição condicional $Pr(Y_t = y_t|S_t = s_t, \theta)$ se Y_t for discreta e por $f(y_t|S_t = s_t, \theta)$ se Y_t for absolutamente contínua em que $\theta = (\theta_1, \dots, \theta_K)$ é o vetor de parâmetros com $\theta \in \mathfrak{R}^K$, e θ_k é o parâmetro associado a k -ésima componente da mistura para $k = 1, \dots, K$.

Complementando a notação, temos que \mathbf{s} é a sequência de valores assumidos pela família S , $\mathbf{s} = \{s_t; t = 1, \dots, T\}$, e \mathbf{y} é a sequência de valores assumidos pela família Y , $\mathbf{y} = \{y_t; t = 1, \dots, T\}$.

A distribuição marginal de Y_t no caso contínuo é dado por

$$f_{Y_t}(y|\theta) = \sum_{k=1}^K f_{Y_t}(y|S_t = k, \theta) Pr(S_t = k|\mathbf{p}) \quad (2.2)$$

em que $f_{Y_t}(y|S_t = k, \theta)$ é a função de densidade de probabilidade condicional. A distri-

buição marginal de Y_t no caso discreto é dado por

$$Pr(Y_t = y_t | \boldsymbol{\theta}) = \sum_{k=1}^K Pr(Y_t = y_t | S_t = k, \boldsymbol{\theta}) Pr(S_t = k | \mathbf{p}) \quad (2.3)$$

em que $Pr(Y_t = y_t | S_t = k, \boldsymbol{\theta})$ é a função de probabilidade condicional. Note que em ambos os casos a distribuição de Y_t é uma mistura de K componentes.

A Figura 2.1 mostra a relação entre as duas famílias de variáveis aleatórias Y_t e S_t . Como podemos notar, o grafo mostra a independência entre os S_t 's e entre os Y_t 's, pela não conexão entre os nós correspondentes. Mostra também cada S_t controlando sua correspondente Y_t , através das setas.

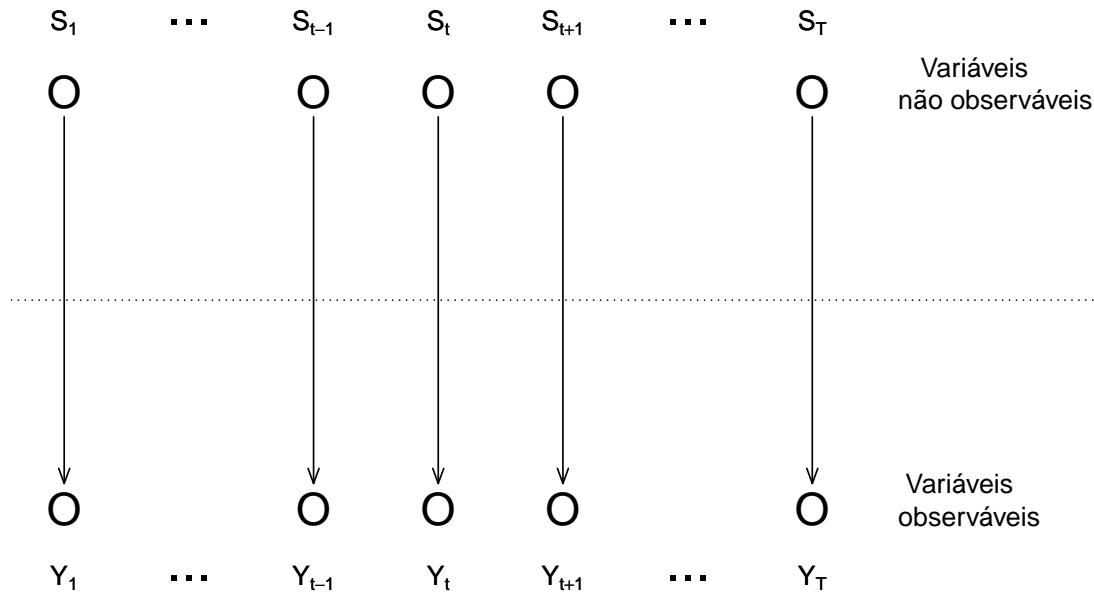


FIGURA 2.1: Grafo representando a relação de dependência entre as variáveis aleatórias no modelo de mistura com S_t 's independentes.

2.1.1 Estimação via máxima verossimilhança

Agora vamos construir a função de verossimilhança. Partimos da distribuição multinomial de S_t , estabelecida acima, e que pode ser expressa por

$$Pr(S_t = s_t | \mathbf{p}) = \prod_{k=1}^K p_k^{I_{st}(k)} \quad (2.4)$$

em que $I_{s_t}(k)$ é uma função indicadora que assume valor um se a condição ($s_t = k$) for verdadeira e zero caso contrário, e $\mathbf{p} = (p_1, \dots, p_K)$ é o vetor de parâmetros com $\mathbf{p} \in (0, 1)^K$ com a restrição $\sum_{k=1}^K p_k = 1$. Temos também que

$$Pr(Y_t = y_t | S_t = s_t, \boldsymbol{\theta}) = \prod_{k=1}^K Pr(Y_t = y_t | S_t = s_t, \theta_k)^{I_{s_t}(k)}. \quad (2.5)$$

A distribuição conjunta de Y_t e S_t é dada pelo produto de (2.4) e (2.5)

$$\begin{aligned} Pr(Y_t = y_t, S_t = s_t | \boldsymbol{\theta}, \mathbf{p}) &= Pr(S_t = s_t | \mathbf{p}) Pr(Y_t = y_t | S_t = s_t, \boldsymbol{\theta}, \mathbf{p}) \\ &= \prod_{k=1}^K p_k^{I_{s_t}(k)} Pr(Y_t = y_t | S_t = s_t, \theta_k)^{I_{s_t}(k)}. \end{aligned} \quad (2.6)$$

Logo, a distribuição conjunta de $\mathbf{Y} = (Y_1, \dots, Y_T)$ e $\mathbf{S} = (S_1, \dots, S_T)$ é dada por

$$\begin{aligned} Pr(\mathbf{Y} = \mathbf{y}, \mathbf{S} = \mathbf{s} | \boldsymbol{\theta}, \mathbf{p}) &= \prod_{t=1}^T \prod_{k=1}^K p_k^{I_{s_t}(k)} Pr(Y_t = y_t | S_t = s_t, \theta_k)^{I_{s_t}(k)} \\ &= \prod_{k=1}^K \prod_{t=1}^T p_k^{I_{s_t}(k)} Pr(Y_t = y_t | S_t = s_t, \theta_k)^{I_{s_t}(k)} \\ &= \prod_{k=1}^K p_k^{\sum_{t=1}^T I_{s_t}(k)} \prod_{t=1}^T Pr(Y_t = y_t | S_t = s_t, \theta_k)^{I_{s_t}(k)} \\ &= \prod_{k=1}^K p_k^{m_k} \prod_{t=1}^T Pr(Y_t = y_t | S_t = s_t, \theta_k)^{I_{s_t}(k)}, \end{aligned} \quad (2.7)$$

em que $m_k = \sum_{t=1}^T I_{s_t}(k)$ é o número de elementos da k -ésima componente de mistura.

A função de verossimilhança de $\mathbf{y} = (y_1, \dots, y_T)$ e $\mathbf{s} = (s_1, \dots, s_T)$ é portanto

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{p} | \mathbf{y}, \mathbf{s}) = \prod_{k=1}^K \left\{ p_k^{m_k} \left[\prod_{t=1}^T Pr(Y_t = y_t | S_t = s_t, \theta_k)^{I_{s_t}(k)} \right] \right\} \quad (2.8)$$

A obtenção dos estimadores de máxima verossimilhança utiliza a função log verossimilhança, $l(\boldsymbol{\theta}, \mathbf{p} | \mathbf{y}, \mathbf{s}) = \log \mathcal{L}(\boldsymbol{\theta}, \mathbf{p} | \mathbf{y}, \mathbf{s})$, e o multiplicador de Lagrange devido

a restrição $\sum_{k=1}^K p_k = 1$. Então,

$$\begin{aligned} l(\boldsymbol{\theta}, \mathbf{p} | \mathbf{y}, \mathbf{s}) &= \sum_{k=1}^K m_k \log p_k + \sum_{t=1}^T \sum_{k=1}^K I_{s_t}(k) \log Pr(Y_t = y_t | S_t = s_t, \theta_k) \\ &+ \lambda \left(1 - \sum_{k=1}^K p_k\right). \end{aligned} \quad (2.9)$$

Derivando (2.9) e igualando o resultado a zero temos o sistema de equações dado por

$$\left. \frac{\partial l(\boldsymbol{\theta}, \mathbf{p} | \mathbf{y}, \mathbf{s})}{\partial \theta_k} \right|_{\theta_k = \hat{\theta}_k} = \sum_{t=1}^T \frac{I_{s_t}(k)}{Pr(Y_t = y_t | S_t = s_t, \theta_k)} \left. \frac{\partial Pr(Y_t = y_t | S_t = s_t, \theta_k)}{\partial \theta_k} \right|_{\theta_k = \hat{\theta}_k} = 0, \quad (2.10)$$

$$\left. \frac{\partial l(\boldsymbol{\theta}, \mathbf{p} | \mathbf{y}, \mathbf{s})}{\partial p_k} \right|_{p_k = \hat{p}_k} = \left[m_k \frac{1}{p_k} - \lambda \right] \Big|_{p_k = \hat{p}_k} = 0 \quad (2.11)$$

em que $k = 1, \dots, K$ e $\lambda = \sum_{k=1}^K m_k = T$.

Os estimadores de máxima verossimilhança são

$$\hat{p}_k = \frac{m_k}{T}, \quad (2.12)$$

$$\hat{\theta}_k \text{ tal que } \left. \frac{\partial l(\boldsymbol{\theta}, \mathbf{p} | \mathbf{y}, \mathbf{s})}{\partial \theta_k} \right|_{\theta_k = \hat{\theta}_k} = 0 \quad (2.13)$$

para $k = 1, \dots, K$.

O problema com os estimadores de máxima verossimilhança mostrados acima é que a sequência \mathbf{S} não é observável e portanto não está disponível para ser utilizada em (2.12) e (2.13). Por essa razão, propomos, uma adaptação para o nosso modelo para o algoritmo EM.

2.1.2 Algoritmo EM

O algoritmo EM foi apresentado em Dempster *et al.* (1977) e a prova de convergência foi completada em Wu (1983).

A seguir apresentaremos algoritmo EM considerando a sequência \mathbf{S} não é observável (ou latente) e os dados \mathbf{Y} .

1. Passo E: calculamos

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)}) = E [\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{Y})|\mathbf{S}, \boldsymbol{\theta}^{(i)}]$$

ou seja, a esperança condicional da log-verossimilhança aumentada, supondo que dado \mathbf{Y} e o valor atual $\boldsymbol{\theta}^{(i)}$.

2. Passo M: escolher o valor $\boldsymbol{\theta}^{(i+1)}$ no espaço paramétrico que maximiza $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)})$.

3. Itere até a convergência, ou seja, até que $\|\boldsymbol{\theta}^{(i+1)} - \boldsymbol{\theta}^{(i)}\|$ ou $|Q(\boldsymbol{\theta}^{(i+1)}, \boldsymbol{\theta}^{(i)}) - Q(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(i)})|$ sejam suficientes pequenas.

Como a sequência \mathbf{S} não é observável recorreremos a um método iterativo de estimação baseado no algoritmo EM devidamente adaptado às necessidades do nosso modelo que podemos descrever da seguinte forma: Inicialmente atribuímos valores arbitrários a \mathbf{S} ; a seguir estimamos os parâmetros \hat{p}_k e $\hat{\theta}_k$ usando (2.12) e (2.13), respectivamente; depois, atualizamos os valores de \mathbf{s} usando a fórmula de Bayes, como mostrado com mais detalhes abaixo.

Partindo da distribuição conjunta para (\mathbf{Y}, \mathbf{S}) dada por (2.7) e usando o teorema de Bayes, obtemos

$$\begin{aligned} Pr(S_t = k|Y_t = y_t, \boldsymbol{\theta}, \mathbf{p}) &= \frac{Pr(S_t = k, Y_t = y_t, \boldsymbol{\theta}, \mathbf{p})}{Pr(Y_t = y_t, \boldsymbol{\theta}, \mathbf{p})} & (2.14) \\ &= \frac{Pr(S_t = k, Y_t = y_t, \boldsymbol{\theta}, \mathbf{p})}{\sum_{l=1}^K Pr(S_t = l, Y_t = y_t, \boldsymbol{\theta}, \mathbf{p})} \\ &= \frac{Pr(S_t = k|\mathbf{p})Pr(Y_t = y_t|S_t = s_t, \boldsymbol{\theta}, \mathbf{p})}{\sum_{l=1}^K Pr(s_t = l|\mathbf{p})Pr(Y_t = y_t|S_t = l, \boldsymbol{\theta}, \mathbf{p})} \\ &= \frac{p_k Pr(Y_t = y_t|S_t = k, \boldsymbol{\theta}, \mathbf{p})}{\sum_{l=1}^K p_l Pr(Y_t = y_t|S_t = l, \boldsymbol{\theta}, \mathbf{p})} \end{aligned}$$

em que $t = 1, \dots, T$ e $k = 1, \dots, K$.

A atualização dos valores de \mathbf{s} é feita utilizando

$$\hat{s}_t = \arg \max_k Pr(S_t = k|Y_t = y_t, \hat{\boldsymbol{\theta}}, \hat{\mathbf{p}}). \quad (2.15)$$

Temos então que descrever a nossa versão de algoritmo EM.

Algoritmo EM adaptado

- (1) Atribua valores a s_t^0 , $t = 1, \dots, T$, de forma que $s_t^0 \in \{1, \dots, K\}$, e faça $i = 1$;
- (2) Calcule $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$ e $\hat{\boldsymbol{p}} = (\hat{p}_1, \dots, \hat{p}_k)$ usando

$$\hat{\theta}_k^i \text{ tal que } \left. \frac{\partial l(\boldsymbol{\theta}, \hat{\boldsymbol{p}} | \mathbf{y}, \mathbf{s})}{\partial \theta_k} \right|_{\theta_k = \hat{\theta}_k} = 0; \quad \hat{p}_k^i = \frac{m_k}{T}$$

- (3) Atualize os valores de s^i utilizando (2.15) com os novos valores de $\hat{\boldsymbol{\theta}}^i$ e $\hat{\boldsymbol{p}}^i$,

$$\hat{s}_t^i = \arg \max_k (Pr(S_t = k | Y_t = y_t, \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^i, \boldsymbol{p} = \hat{\boldsymbol{p}}^i)),$$

para $t = 1, \dots, T$.

- (4) Repetir os passos (2) e (3) para $i = 2, 3, \dots$ até a convergência.

O critério de convergência utilizado é baseado na variação máxima para os estimadores dos parâmetros, $\hat{\boldsymbol{p}}$ e $\hat{\boldsymbol{\theta}}$, $\max(|\hat{\boldsymbol{\theta}}^i - \hat{\boldsymbol{\theta}}^{i-1}|) < tol_1$ e $\max(|\hat{\boldsymbol{p}}^i - \hat{\boldsymbol{p}}^{i-1}|) < tol_2$, e tol_i , $i = 1, 2$ são valores positivo pré estabelecido mediante o espaço dos parâmetros em questão.

Note que no algoritmo proposto não temos o usual $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)}) = E[l(\boldsymbol{\theta} | \mathbf{y}) | S, \boldsymbol{\theta}^{(i)}]$, ou seja, a esperança condicional da log verossimilhança aumentada, supondo que dado Y e o valor atual $\boldsymbol{\theta}^{(i)}$. Em contra partida, temos $\hat{s}_t = \arg \max_k Pr(S_t = k | Y_t = y_t, \boldsymbol{\theta}, \boldsymbol{p})$, ou seja, usamos o teorema de Bayes para atualizar os valores de s .

Além disso, observamos que o algoritmo EM no passo M escolhe os valores $\boldsymbol{\theta}^{(i+1)}$ no espaço paramétrico de forma a maximizar $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)})$. Enquanto que na nossa adaptação algoritmo EM escolhe os valores de $\boldsymbol{\theta}^{(i+1)}$ no espaço paramétrico de forma a maximizar a função verossimilhança, $\mathcal{L}(\boldsymbol{\theta}^{(i)}, \boldsymbol{p}^{(i)} | \mathbf{y}, \mathbf{s})$.

2.2 Modelo de Mistura com Dependência Markoviana

No modelo que tratamos agora, o Modelo de Mistura Markoviano com Dependência de Primeira Ordem, ou de forma abreviada MMM(1), a sequência de variáveis controladoras da mistura, $\boldsymbol{S} = (S_1, \dots, S_K)$, segue uma cadeia de Markov com dependência

de grau 1, diferindo da seção anterior, em que a sequência de variáveis controladoras da mistura tinha suas componentes S_t 's independentes e identicamente distribuídas.

Para ilustrar as relações entre as variáveis envolvidas no modelo MMM(1) apresentamos o grafo da Figura 2.2.

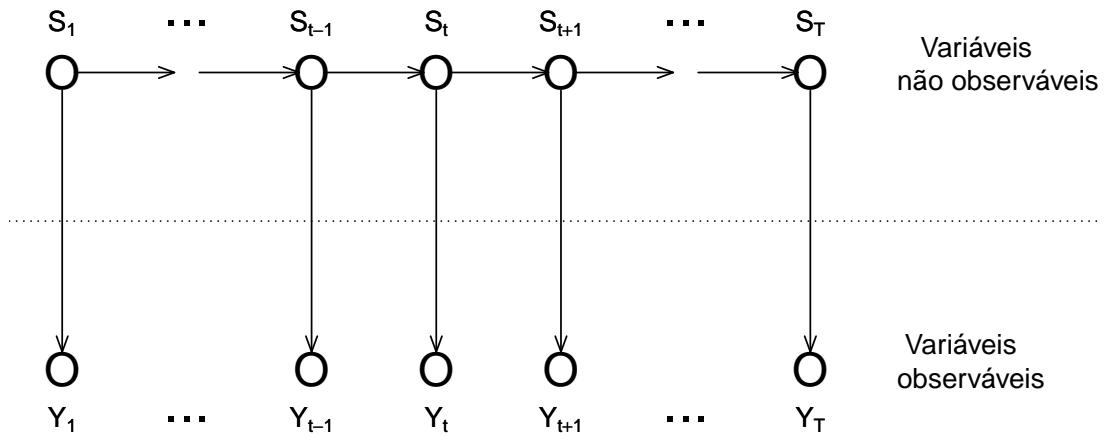


FIGURA 2.2: Modelo de mistura com dependência de grau 1 na variável não observável

Observamos que para simplificar a notação omitimos o condicionamento aos parâmetros \mathbf{P} , $\boldsymbol{\theta}$ e \mathbf{p}_0 nas distribuições de \mathbf{S} e \mathbf{Y} nas equações que seguem.

Considere a sequência de variáveis aleatórias $\mathbf{S} = (S_1, \dots, S_T)$, que assumem valores $S_t \in \{1, \dots, K\}$, é uma cadeia de Markov de primeira ordem, ou seja,

$$Pr(S_t = s_t | S_{t-1} = s_{t-1}, \dots, S_1 = s_1) = Pr(S_t = s_t | S_{t-1} = s_{t-1}). \quad (2.16)$$

Uma cadeia de Markov é caracterizada pela probabilidade do estado inicial da cadeia $\mathbf{p}_0 = (p_{01}, \dots, p_{0K})$ em que $P(S_1 = k) = p_{0k}$ para $k = 1, \dots, K$ e pelas probabilidades de transição $Pr(S_t = s_t | S_{t-1} = s_{t-1}) = p_{s_{t-1}s_t}$. A probabilidade de transição é a probabilidade de estar no estado k no tempo t dado que estava em j no tempo $t-1$. Esta pode ser expressa pela matriz de transição $\mathbf{P} = \{p_{jk}\}$,

$$\mathbf{P} = \begin{bmatrix} p_{11} & \dots & p_{1K} \\ p_{21} & \dots & p_{2K} \\ \vdots & & \\ p_{K1} & \dots & p_{KK} \end{bmatrix},$$

em que $p_{jk} \geq 0$, para todo j, k , e a soma das linhas deve ser igual a um, $\sum_{k=1}^K p_{jk} = 1$.

Uma característica interessante do modelo é que, embora as misturas \mathbf{Y} sejam controladas pela cadeia de Markov \mathbf{S} , as componentes da sequência Y são condicionalmente independentes, ou seja, $Y_u \perp Y_v | S = s$, para $u \neq v$.

Usando probabilidade condicional e considerando, sem perda de generalidade, que $u < v$ temos

$$\begin{aligned}
& Pr(Y_u = y_u, Y_v = y_v | \mathbf{S} = \mathbf{s}) = \\
& = Pr(Y_u = y_u, Y_v = y_v | S_1 = s_1, \dots, S_u = s_u, S_v = s_v, \dots, S_T = s_T) \\
& = \frac{Pr(Y_u = y_u, Y_v = y_v, S_1 = s_1, \dots, S_u = s_u, S_v = s_v, \dots, S_T = s_T)}{Pr(S_1 = s_1, \dots, S_u = s_u, S_v = s_v, \dots, S_T = s_T)} \\
& = \frac{Pr(S_1 = s_1)P(S_2 = s_2 | S_1 = s_1) \dots P(S_T = s_T | S_{T-1} = s_{T-1})}{Pr(S_1 = s_1)P(S_2 = s_2 | S_1 = s_1) \dots P(S_T = s_T | S_{T-1} = s_{T-1})} \\
& \times Pr(Y_u = y_u | S_1 = s_1, \dots, S_u = s_u, S_v = s_v, \dots, S_T = s_T) \\
& \times Pr(Y_v = y_v | S_1 = s_1, \dots, S_u = s_u, S_v = s_v, \dots, S_T = s_T, Y_u = y_u)
\end{aligned} \tag{2.17}$$

e usando o fato que Y_t tem dependência de Markov de primeira ordem em relação a S_t em (2.17), temos

$$Pr(Y_u = y_u, Y_v = y_v | S = s) = Pr(Y_u = y_u | S_u = s_u) Pr(Y_v = y_v | S_v = s_v) \tag{2.18}$$

2.2.1 Alguns resultados

Apresentamos, a seguir, alguns resultados relativos ao modelo proposto, MMM(1).

A distribuição de \mathbf{S} é dada por

$$\begin{aligned}
Pr(\mathbf{S} = \mathbf{s}) &= Pr(S_1 = s_1, \dots, S_T = s_T) \\
&= Pr(S_1 = s_1)Pr(S_2 = s_2|S_1 = s_1) \dots Pr(S_T = s_T|S_{T-1} = s_{T-1}) \\
&= Pr(S_1 = s_1) \prod_{t=2}^T Pr(S_t = s_t|S_{t-1} = s_{t-1}) \\
&= p_{0s_1} \prod_{t=2}^T p_{s_{t-1}s_t}.
\end{aligned} \tag{2.19}$$

Considerando apenas a primeira observação, temos que a distribuição conjunta para S_1 e Y_1 é

$$\begin{aligned}
Pr(Y_1 = y_1, S_1 = s_1) &= Pr(S_1 = s_1)Pr(Y_1 = y_1|S_1 = s_1) \\
&= p_{0s_1}Pr(Y_1 = y_1|S_1 = s_1),
\end{aligned} \tag{2.20}$$

e, aplicando o teorema da probabilidade total, temos

$$Pr(Y_1 = y_1) = \sum_{s_1=1}^K p_{0s_1}Pr(Y_1 = y_1|S_1 = s_1). \tag{2.21}$$

Considerando as duas primeiras observações temos

$$\begin{aligned}
&Pr(Y_1 = y_1, Y_2 = y_2, S_1 = s_1, S_2 = s_2) = \\
&= Pr(S_1 = s_1)Pr(Y_1 = y_1|S_1 = s_1)Pr(S_2 = s_2|S_1 = s_1)Pr(Y_2 = y_2|S_2 = s_2) \\
&= p_{0s_1}p_{s_1s_2}Pr(Y_1 = y_1|S_1 = s_1)Pr(Y_2 = y_2|S_2 = s_2) \\
&= p_{0s_1}p_{s_1s_2} \prod_{t=1}^2 Pr(Y_t = y_t|S_t = s_t)
\end{aligned} \tag{2.22}$$

e, aplicando a lei de probabilidade total, obtemos

$$Pr(Y_1 = y_1, Y_2 = y_2) = \sum_{s_1=1}^K \sum_{s_2=1}^K p_{0s_1}p_{s_1s_2} \prod_{t=1}^2 Pr(Y_t = y_t|S_t = s_t). \tag{2.23}$$

Para as três primeiras observações temos

$$Pr(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, S_1 = s_1, S_2 = s_2, S_3 = s_3) = \tag{2.24}$$

$$\begin{aligned}
&= Pr(S_1 = s_1)P(Y_1 = y_1|S_1 = s_1)Pr(S_2 = s_2|S_1 = s_1)Pr(Y_2 = y_2|S_2 = s_2) \\
&\quad \times Pr(S_3 = s_3|S_2 = s_2)Pr(Y_3 = y_3|S_3 = s_3) \\
&= p_{0s_1}p_{s_1s_2}p_{s_2s_3}Pr(Y_1 = y_1|S_1 = s_1)Pr(Y_2 = y_2|S_2 = s_2)Pr(Y_3 = y_3|S_3 = s_3) \\
&= p_{0s_1} \prod_{t=2}^3 p_{s_{t-1}s_t} \prod_{t=1}^3 Pr(Y_t = y_t|S_t = s_t). \tag{2.25}
\end{aligned}$$

Aplicando a lei de probabilidade total temos

$$Pr(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) = \sum_{s_1=1}^K \sum_{s_2=1}^K \sum_{s_3=1}^K p_{0s_1} \prod_{t=2}^3 p_{s_{t-1}s_t} \prod_{t=1}^3 Pr(Y_t = y_t|S_t = s_t). \tag{2.26}$$

Generalizando temos que a distribuição conjunta de (\mathbf{Y}, \mathbf{S}) é dada por

$$\begin{aligned}
&Pr(\mathbf{Y} = \mathbf{y}, \mathbf{S} = \mathbf{s}) = \\
&= Pr(Y_1 = y_1, \dots, Y_{T-1} = y_{T-1}, Y_T = y_T, S_1 = s_1, \dots, S_{T-1} = s_{T-1}, S_T = s_T) \\
&= Pr(S_1 = s_1)P(Y_1 = y_1|S_1 = s_1)Pr(S_2 = s_2|S_1 = s_1)Pr(Y_2 = y_2|S_2 = s_2) \\
&\quad \times Pr(S_3 = s_3|S_2 = s_2)Pr(Y_3 = y_3|S_3 = s_3) \dots Pr(S_{T-1} = s_{T-1}|S_{T-2} = s_{T-2}) \\
&\quad \times Pr(Y_{T-1} = y_{T-1}|S_{T-1} = s_{T-1})Pr(S_T = s_T|S_{T-1} = s_{T-1})Pr(Y_T = y_T|S_T = s_T) \\
&= p_{0s_1} \prod_{t=2}^T p_{s_{t-1}s_t} \prod_{t=1}^T Pr(Y_t = y_t|S_t = s_t) \tag{2.27}
\end{aligned}$$

em que $\mathbf{y} = (y_1, \dots, y_T)$ e $\mathbf{s} = (s_1, \dots, s_T)$.

Somando com relação aos possíveis valores de \mathbf{S} temos

$$\begin{aligned}
P(\mathbf{Y} = \mathbf{y}) &= \sum_{s_1=1}^K \dots \sum_{s_T=1}^K Pr(S_1 = s_1) Pr(Y_1 = y_1 | S_1 = s_1) \\
&\quad \times \prod_{t=2}^T Pr(S_t = s_t | S_{t-1} = s_{t-1}) Pr(Y_t = y_t | S_t = s_t) \\
&= \sum_{s_1=1}^K \dots \sum_{s_T=1}^K \prod_{t=1}^T p_{s_{t-1}s_t} Pr(Y_t = y_t | S_t = s_t)
\end{aligned} \tag{2.28}$$

se considerarmos que $s_0 = 0$.

2.2.2 Função de Verossimilhança

A função de verossimilhança conjunta de \mathbf{y} e \mathbf{s} para o modelo de mistura markoviano de primeira ordem, obtida de (2.27), pode ser escrita como

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}, P, \mathbf{p}_0 | \mathbf{y}, \mathbf{s}) &= Pr(\mathbf{S} = \mathbf{s} | \mathbf{P}, \mathbf{p}_0) Pr(\mathbf{Y} = \mathbf{y} | \mathbf{S} = \mathbf{s}, \boldsymbol{\theta}) \\
&= Pr(S_1 = s_1 | \mathbf{p}_0) \prod_{t=2}^T Pr(S_t = s_t | S_{t-1} = s_{t-1}, \mathbf{P}) \\
&\quad \times \prod_{t=1}^T Pr(Y_t = y_t | S_t = s_t, \boldsymbol{\theta})
\end{aligned} \tag{2.29}$$

em que $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ é o vetor de parâmetros das componentes e θ_k é o parâmetro associado a k -ésima componente da mistura; $\mathbf{p}_0 = (p_{01}, \dots, p_{0K})$ é o vetor de parâmetros relativos a S_1 (probabilidade inicial), com $0 \leq p_{0k} \leq 1$, para $k = 1, \dots, K$; e $\mathbf{P} = \{p_{jk}\}$ em que p_{jk} são as probabilidades de transição, sendo $\sum_{k=1}^K p_{jk} = 1$, para $j \in \{0, 1, \dots, K\}$.

Apresentamos, a seguir, algumas relações que serão utilizadas mais adiante. A distribuição inicial pode ser escrita como

$$Pr(S_1 = s_1 | \mathbf{p}_0) = \prod_{k=1}^K p_{0k}^{I_{s_1}(k)}, \tag{2.30}$$

e a distribuição condicional de $S_t | S_{t-1}$ como

$$\begin{aligned}
Pr(S_t = s_t | S_{t-1} = s_{t-1}, P, \mathbf{p}_0) &= \prod_{k=1}^K \prod_{j=1}^K p_{jk}^{I_{s_{t-1}, s_t}(j, k)} \\
&= \prod_{k=1}^K \prod_{j=1}^K p_{jk}^{I_{s_{t-1}}(j) I_{s_t}(k)},
\end{aligned} \tag{2.31}$$

sendo ambas com distribuição multinomial, ou seja, $S_t|S_{t-1} = j \sim \text{mult}(1, \mathbf{p} = (p_{j1}, \dots, p_{jK}))$ para $t = 2, \dots, T$, e $S_1 \sim \text{mult}(1, \mathbf{p} = (p_{01}, \dots, p_{0K}))$.

A distribuição conjunta de \mathbf{S} é obtida fazendo o produtório

$$\begin{aligned}
Pr(\mathbf{S} = \mathbf{s}) &= Pr(S_1 = s_1 | \mathbf{p}_0) \prod_{t=2}^T Pr(S_t = s_t | S_{t-1} = s_{t-1}, \mathbf{P}, \mathbf{p}_0) \\
&= \left[\prod_{k=1}^K p_{0k}^{I_{s_1}(k)} \right] \prod_{t=2}^T \prod_{k=1}^K \prod_{j=1}^K p_{jk}^{I_{s_{t-1}, s_t}(j, k)} \\
&= \left[\prod_{k=1}^K p_{0k}^{I_{s_1}(k)} \right] \prod_{k=1}^K \prod_{j=1}^K p_{jk}^{\sum_{t=2}^T I_{s_{t-1}}(j) I_{s_t}(k)}, \\
&= \left[\prod_{k=1}^K p_{0k}^{m_{0k}} \right] \prod_{k=1}^K \prod_{j=1}^K p_{jk}^{m_{jk}}, \tag{2.32}
\end{aligned}$$

considerando que $m_{0k} = I_{s_1}(k)$ e que m_{jk} é o número de transições de j para k na sequência s , ou seja, $m_{jk} = \sum_{t=2}^T I_{s_{t-1}, s_t}(j, k) = \sum_{t=2}^T I_{s_{t-1}}(j) I_{s_t}(k)$.

Temos também que

$$Pr(Y_t = y_t | S_t = s_t) = \prod_{k=1}^K Pr(Y_t = y_t | S_t = s_t)^{I_{s_t}(k)}.$$

Portanto, a distribuição conjunta de \mathbf{Y} condicional a \mathbf{S} é dada pelo produto

$$\begin{aligned}
Pr(\mathbf{Y} = \mathbf{y} | \mathbf{S} = \mathbf{s}) &= \prod_{t=1}^T Pr(Y_t = y_t | S_t = s_t) \\
&= \prod_{t=1}^T \prod_{k=1}^K Pr(Y_t = y_t | S_t = s_t)^{I_{s_t}(k)} \\
&= \prod_{k=1}^K \prod_{t=1}^T Pr(Y_t = y_t | S_t = s_t)^{I_{s_t}(k)}. \tag{2.33}
\end{aligned}$$

e a distribuição conjunta de (\mathbf{Y}, \mathbf{S}) é dada por

$$\begin{aligned}
Pr(\mathbf{Y} = \mathbf{y}, \mathbf{S} = \mathbf{s}) &= Pr(\mathbf{S} = \mathbf{s}) Pr(\mathbf{Y} = \mathbf{y} | \mathbf{S} = \mathbf{s}) \\
&= \left[\prod_{k=1}^K p_{0k}^{m_{0k}} \right] \left[\prod_{k=1}^K \prod_{j=1}^K p_{jk}^{m_{jk}} \right] \left[\prod_{k=1}^K \prod_{t=1}^T Pr(Y_t = y_t | S_t = s_t)^{I_{s_t}(k)} \right]. \tag{2.34}
\end{aligned}$$

A partir dessas relações, podemos reescrever a função de verossimilhança (2.29) na forma

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}|\mathbf{y}, \mathbf{s}) &= \left[\prod_{k=1}^K p_{0k}^{m_{0k}} \right] \left[\prod_{k=1}^K \prod_{j=1}^K p_{jk}^{m_{jk}} \right] \\ &\times \left\{ \prod_{k=1}^K \prod_{t=1}^T [Pr(Y_t = y_t | S_t = s_t, \theta_k)]^{I_{s_t}(k)} \right\} \\ &= \prod_{k=1}^K \left\{ p_{0k}^{m_{0k}} \prod_{j=1}^K p_{jk}^{m_{jk}} \left[\prod_{t=1}^T [Pr(Y_t = y_t | S_t = s_t, \theta_k)]^{I_{s_t}(k)} \right] \right\} \end{aligned} \quad (2.35)$$

em que $\sum_{k=1}^K p_{jk} = 1$ para $j = 0, \dots, K$.

A função log-verossimilhança, $l(\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}|\mathbf{y}, \mathbf{s}) = \log \mathcal{L}(\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}|\mathbf{y}, \mathbf{s})$, de (2.35) juntamente com o multiplicador de Lagrange é dado por

$$\begin{aligned} l(\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}|\mathbf{y}, \mathbf{s}) &= \sum_{k=1}^K m_{0k} \log p_{0k} + \sum_{k=1}^K \sum_{t=1}^T I(s_t = k) \log Pr(Y_t = y_t | S_t = s_t, \theta_k) \\ &+ \sum_{k=1}^K \sum_{j=1}^K m_{jk} \log p_{jk} + \lambda_0 (1 - \sum_{k=1}^K p_{0k}) + \lambda_1 (1 - \sum_{k=1}^K p_{1k}) \\ &+ \lambda_2 (1 - \sum_{k=1}^K p_{2k}) + \dots + \lambda_j (1 - \sum_{k=1}^K p_{jk}) + \dots + \lambda_K (1 - \sum_{k=1}^K p_{Kk}). \end{aligned} \quad (2.36)$$

Derivando (2.36) e igualando a derivada a zero, obtemos os estimadores de máxima verossimilhança

$$\begin{aligned} \left. \frac{\partial l(\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}|\mathbf{y}, \mathbf{s})}{\partial \theta_k} \right|_{\theta_k = \hat{\theta}_k} &= \sum_{t=1}^T \frac{I(s_t = k)}{Pr(Y_t = y_t | S_t = s_t, \theta_k)} \left. \frac{\partial Pr(Y_t = y_t | S_t = s_t, \theta_k)}{\partial \theta_k} \right|_{\theta_k = \hat{\theta}_k} = 0 \\ \left. \frac{\partial l(\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}|\mathbf{y}, \mathbf{s})}{\partial p_{0k}} \right|_{p_{0k} = \hat{p}_{0k}} &= \left[m_{0k} \frac{1}{p_{0k}} - \lambda_0 \right] \Big|_{p_{0k} = \hat{p}_{0k}} = 0 \\ \left. \frac{\partial l(\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}|\mathbf{y}, \mathbf{s})}{\partial p_{jk}} \right|_{p_{jk} = \hat{p}_{jk}} &= \left[m_{jk} \frac{1}{p_{jk}} - \lambda_j \right] \Big|_{p_{jk} = \hat{p}_{jk}} = 0 \end{aligned}$$

em que $\lambda_0 = \sum_{k=1}^K m_{0k}$ e $\lambda_j = \sum_{k=1}^K m_{jk}$ com $j = 1, \dots, K$.

Os estimadores de máxima verossimilhança são

$$\hat{\theta}_k \text{ tal que } \left. \frac{\partial l(\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P} | \mathbf{y}, \mathbf{s})}{\partial \theta_k} \right|_{\theta_k = \hat{\theta}_k} = 0, \quad (2.37)$$

$$\hat{p}_{0k} = \frac{m_{0k}}{\sum_{l=1}^K m_{0l}}, \quad (2.38)$$

$$\hat{p}_{jk} = \frac{m_{jk}}{\sum_{l=1}^K m_{jl}} \quad (2.39)$$

para $j, k = 1, \dots, K$.

Podemos reescrever os estimadores de máxima verossimilhança dados em (2.37), (2.38) e (2.39) da seguinte forma

$$\hat{\mathbf{p}}_0 = \arg \max_{\mathbf{p}_0} l(\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P} | \mathbf{y}, \mathbf{s}), \quad (2.40)$$

$$\hat{\mathbf{P}} = \arg \max_{\mathbf{P}} l(\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P} | \mathbf{y}, \mathbf{s}), \quad (2.41)$$

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P} | \mathbf{y}, \mathbf{s}) \quad (2.42)$$

em que $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_K)$, $\hat{\mathbf{p}}_0 = (\hat{p}_{01}, \dots, \hat{p}_{0K})$ e $\hat{\mathbf{P}} = (\hat{p}_{j1}, \dots, \hat{p}_{jK})$ para $j = 1, \dots, K$.

Da mesma forma que na Seção 2.2, em que é tratado o modelo de mistura sem dependência entre as variáveis S_t 's, aqui também o problema com os estimadores de máxima verossimilhança mostrados acima é que a sequência \mathbf{S} não é observável e, portanto, não está disponível para ser utilizada em (2.13). Da mesma forma, utilizaremos uma modificação no algoritmo EM para o modelo proposto.

2.2.3 Algoritmo EM

Como no caso da sequência \mathbf{S} com variáveis independentes, Seção 2.1.2, aqui também temos a sequência \mathbf{S} não é observável. Por essa razão, recorreremos à mesma abordagem utilizada anteriormente, o método iterativo de estimação baseado no algoritmo EM devidamente adaptado às necessidades do nosso modelo atual em que \mathbf{S} é composta por variáveis com dependência markoviana de primeira ordem.

Inicialmente, atribuímos valores arbitrários a \mathbf{S} ; a seguir, estimamos os parâmetros $\hat{\mathbf{p}}_0$, $\hat{\mathbf{P}}$ e $\hat{\boldsymbol{\theta}}_k$ usando (2.40), (2.41) e (2.42), respectivamente; depois, atualizamos os valores de s usando a fórmula de Bayes, como mostrado abaixo. Os dois últimos passos são

repetidos até a convergência, convergência que é garantida pelos resultados de Dempster *et al.* (1977) e Wu (1983).

Partimos da distribuição conjunta para (\mathbf{Y}, \mathbf{S}) dada por (2.34) e usando o teorema de Bayes, obtemos

$$\begin{aligned}
Pr(S_1 = k | Y_1 = y_1, \theta, p_0) &= \frac{Pr(S_1 = k, Y_1 = y_1, \theta, p_0, P)}{Pr(Y_1 = y_1, \theta, p_0, P)} \\
&= \frac{Pr(S_1 = k, Y_1 = y_1, \theta, p_0, P)}{\sum_{l=1}^K Pr(Y_1 = y_1, S_1 = l, \theta, p_0, P)} \\
&= \frac{Pr(S_1 = k | p_0, P) Pr(Y_1 = y_1 | S_1 = k, \theta, p_0, P)}{\sum_{l=1}^K Pr(S_1 = l | p_0, P) Pr(Y_1 = y_1 | S_1 = l, \theta, p_0, P)} \\
&= \frac{p_{0k} Pr(Y_1 = y_1 | S_1 = k, \theta, p_0, P)}{\sum_{l=1}^K p_{0l} Pr(Y_1 = y_1 | S_1 = l, \theta, p_0, P)} \tag{2.43}
\end{aligned}$$

em que $k = 1, \dots, K$. Para $t = 2, \dots, T$ temos

$$\begin{aligned}
Pr(S_t = k | S_{t-1} = j, Y_t = y_t, \theta, p_0, P) &= \frac{Pr(S_t = k, S_{t-1} = j, Y_t = y_t, \theta, p_0, P)}{Pr(Y_t = y_t, S_{t-1} = j, \theta, p_0, P)} \\
&= \frac{Pr(S_t = k, S_{t-1} = j, Y_t = y_t, \theta, p_0, P)}{\sum_{l=1}^K Pr(Y_t = y_t, S_t = l, S_{t-1} = j, \theta, p_0, P)} \\
&= \frac{Pr(S_t = k | S_{t-1} = j, p_0, P) Pr(Y_t = y_t | S_t = k, \theta, p_0, P)}{\sum_{l=1}^K Pr(S_t = l | S_{t-1} = j, p_0, P) Pr(Y_t = y_t | S_t = l, \theta, p_0, P)} \\
&= \frac{p_{jk} Pr(Y_t = y_t | S_t = k, \theta, p_0, P)}{\sum_{l=1}^K p_{jl} Pr(Y_t = y_t | S_t = l, \theta, p_0, P)} \tag{2.44}
\end{aligned}$$

Na estimação dos parâmetros do modelo MMM(1), utilizamos o algoritmo apresentado a seguir.

Algoritmo EMM para MMM(1)

- (1) Atribua valores iniciais para S_t , $\mathbf{s}^0 \leftarrow \mathbf{s}$ e faça $i = 1$.
- (2) Calcule os valores dos parâmetros $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{p}}_0$ e $\hat{\mathbf{P}}$ usando

$$\begin{aligned}
\hat{\boldsymbol{\theta}}^i &= \arg \max_{\boldsymbol{\theta}} (l(\boldsymbol{\theta}, \hat{\mathbf{p}}_0, \hat{\mathbf{P}}; \mathbf{y}, \mathbf{s}^{i-1})), \\
\hat{\mathbf{P}}^i &= \arg \max_{\mathbf{P}} (l(\hat{\boldsymbol{\theta}}, \hat{\mathbf{p}}_0, \mathbf{P}; \mathbf{y}, \mathbf{s}^{i-1})), \\
\hat{\mathbf{p}}_0^i &= \arg \max_{p_0} (l(\hat{\boldsymbol{\theta}}, p_0, \hat{\mathbf{P}}; \mathbf{y}, \mathbf{s}^{i-1})).
\end{aligned}$$

(3) Calcule os valores de s_1^i usando

$$s_1^i = \arg \max_k (P(S_1 = k | Y_1 = y_1, \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^i, \mathbf{p}_0 = \hat{\mathbf{p}}_0^i, \mathbf{P} = \hat{\mathbf{P}}^i)),$$

e de s_t^i , para $t = 2, \dots, T$, usando

$$s_t^i = \arg \max_k (S_t = k | s_{t-1}^i = j, Y_t = y_t, \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^i, \mathbf{p}_0 = \hat{\mathbf{p}}_0^i, \mathbf{P} = \hat{\mathbf{P}}^i).$$

(4) Repita os passos (2) e (3) para $i = 2, 3, \dots$ até obter a convergência.

O critério de convergência utilizado é baseado na variação máxima para os parâmetros $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{p}}_0$ e $\hat{\mathbf{P}}$, $\max(|\hat{\boldsymbol{\theta}}^i - \hat{\boldsymbol{\theta}}^{i-1}|) < tol_1$ e $\max(|\hat{\mathbf{p}}_0^i - \hat{\mathbf{p}}_0^{i-1}|) < tol_2$ $\max(|\hat{\mathbf{P}}^i - \hat{\mathbf{P}}^{i-1}|) < tol_3$, em que tol_l $l = 1, 2, 3$ são valores positivo pré estabelecido conforme o espaço das variáveis.

Vale notar que o algoritmo apresentado acima difere radicalmente dos procedimentos de estimação utilizados em HMM, tais como os algoritmos de Viterbi, Baum-Welch e *forward-backward*.

Capítulo 3

Mistura de normais

Neste capítulo, apresentamos uma aplicação do modelo de mistura apresentado no capítulo 2 utilizando a distribuição normal para o caso em que temos independência na variável não observável e para o MMM(1). O modelo de mistura de distribuições normais será apresentado primeiramente utilizando a estimação dos parâmetros via verossimilhança e usando algoritmo EMM para a estimação dos parâmetros, e consideramos os critérios AIC (*Akaike Information Criterion*) e BIC (*Bayesian Information Criterion*) para seleção do modelo. Num segundo momento, utilizamos a abordagem bayesiana e algoritmo MCMC para estimação dos parâmetros.

O modelo de mistura de distribuições normais em que $Y_t|S_t = k \sim N(\mu_k, \sigma_k^2)$, $t = 1, \dots, T$, e $k = 1, \dots, K$ em que μ_k é o parâmetro de locação e σ_k^2 é o parâmetro de escala da k -ésima distribuição normal, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ e $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_K^2)$ são os vetores de parâmetros. A sequência de variáveis aleatórias não observáveis S é de variáveis aleatórias independentes e identicamente distribuídas e a variável aleatória indicadora que associa Y_t a k -ésima componente de mistura, $I_{s_t}(k)$, $t = 1, \dots, T$, $k = 1, \dots, K$.

Considerando os dados $\mathbf{y} = (y_1, \dots, y_T)$ uma realização de $\mathbf{Y} = (Y_1, \dots, Y_T)$, $\mathbf{s} = (s_1, \dots, s_T)$, com $s_t \in \{1, \dots, K\}$, uma realização de $\mathbf{S} = (S_1, \dots, S_T)$ e a função de verossimilhança conjunta dada (2.8) em que a distribuição de probabilidade $Pr(Y_t = y_t|S_t = k, \mu_k, \sigma_k^2)$ é substituída pela função densidade $f_{Y_t}(y_t|S_t = k, \mu_k, \sigma_k^2)$

de uma distribuição normal. Temos que a função de máxima verossimilhança de \mathbf{y} e \mathbf{s} é

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}|\mathbf{y}, \mathbf{s}) &= \prod_{k=1}^K \left\{ p_k^{m_k} \left[\prod_{t=1}^T \left(\frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\left\{ -\frac{(y_t - \mu_k)^2}{2\sigma_k^2} \right\}} I_{(-\infty, +\infty)}(y_t) \right)^{I_{s_t}(k)} \right] \right\} \\ &= \prod_{k=1}^K \left\{ p_k^{m_k} (2\pi)^{-\frac{m_k}{2}} (\sigma_k^2)^{-\frac{m_k}{2}} e^{\left\{ -\frac{\sum_{t=1}^T I_{s_t}(k)(y_t - \mu_k)^2}{2\sigma_k^2} \right\}} \right. \\ &\quad \left. \times \prod_{t=1}^T I_{(-\infty, +\infty)}(y_t)^{I_{s_t}(k)} \right\} \end{aligned} \quad (3.1)$$

em que $m_k = \sum_{t=1}^T I_{s_t}(k)$ é o número de elementos da k -ésima componente de mistura, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ com $\mu_k \in (-\infty, +\infty)$, $\mathbf{p} = (p_1, \dots, p_K)$ com $\sum_{k=1}^K p_k = 1$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_K^2)$, $\sigma_k^2 > 0$.

Usando os multiplicadores de Lagrange, devido a restrição $\sum_{k=1}^K p_k = 1$ e a função log verossimilhança, $l(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}|\mathbf{y}, \mathbf{s}) = \log \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}|\mathbf{y}, \mathbf{s})$, temos

$$\begin{aligned} l(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}|\mathbf{y}, \mathbf{s}) &= \sum_{k=1}^K m_k \log p_k - \frac{1}{2} \sum_{k=1}^K m_k \log 2\pi - \frac{1}{2} \sum_{k=1}^K m_k \log \sigma_k^2 + \lambda \left(1 - \sum_{k=1}^K p_k \right) \\ &\quad - \sum_{k=1}^K \frac{\sum_{t=1}^T I_{s_t}(k)(y_t - \mu_k)^2}{2\sigma_k^2}, \end{aligned} \quad (3.2)$$

em que $m_k = \sum_{t=1}^T I_{s_t}(k)$.

Derivando a função log verossimilhança em relação aos parâmetros e igualando a zero temos o seguinte sistema de equações

$$\left. \frac{\partial l(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}|\mathbf{y}, \mathbf{s})}{\partial \mu_k} \right|_{\mu_k = \hat{\mu}_k} = \left[\sum_{t=1}^T \frac{2}{2\sigma_k^2} I_{s_t}(k)(y_t - \mu_k) \right] \Big|_{\mu_k = \hat{\mu}_k} = 0, \quad (3.3)$$

$$\left. \frac{\partial l(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}|\mathbf{y}, \mathbf{s})}{\partial \sigma_k^2} \right|_{\sigma_k^2 = \hat{\sigma}_k^2} = \left[-\frac{m_k}{2\sigma_k^2} + \frac{\sum_{t=1}^T I_{s_t}(k)(y_t - \mu_k)^2}{2\sigma_k^4} \right] \Big|_{\sigma_k^2 = \hat{\sigma}_k^2} = 0, \quad (3.4)$$

$$\left. \frac{\partial l(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}|\mathbf{y}, \mathbf{s})}{\partial p_k} \right|_{p_k = \hat{p}_k} = \left[\frac{m_k}{p_k} - \lambda \right] \Big|_{p_k = \hat{p}_k} = 0 \quad (3.5)$$

em que $\lambda = \sum_{k=1}^K m_k = T$.

Os estimadores de máxima verossimilhança são

$$\hat{\mu}_k = \frac{\sum_{t=1}^T I_{s_t}(k)y_t}{m_k}, \quad (3.6)$$

$$\hat{\sigma}_k^2 = \frac{\sum_{t=1}^T I_{s_t}(k)(y_t - \hat{\mu}_k)^2}{m_k}, \quad (3.7)$$

$$\hat{p}_k = \frac{m_k}{T} \quad (3.8)$$

para $k = 1, \dots, K$.

Temos que a variável S é não observável. Por essa razão, não conseguimos calcular os valores estimadores de máxima verossimilhança diretamente dados pelas equações (3.6), (3.7) e (3.8). Devido a esse fato, temos que utilizar o algoritmo EM adaptado para obter os valores das estimativas.

Consideramos dois critérios para selecionar o melhor modelo, o *AIC* e o *BIC*. O critério AIC (*Akaike Information Criterion*) proposto por (Akaike, 1974) é um indicativo da qualidade do ajuste do modelo dada por

$$AIC(k) = -2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}, \hat{\mathbf{p}} | \mathbf{y}, \mathbf{s}) + \gamma_k \quad (3.9)$$

em que γ_k é o número de parâmetros considerados no modelo para $k = 1, \dots, K$.

O *BIC* (*Bayesian Information Criterion*) é um critério de seleção de modelo proposto por Schwarz (1978), o qual pondera entre o máximo da função log verossimilhança e o número de parâmetros do modelo. O valor do *BIC* é dado por

$$BIC(k) = -2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}, \hat{\mathbf{p}} | \mathbf{y}, \mathbf{s}) + (\gamma_k) \log(T), \quad (3.10)$$

em que $\log \mathcal{L}(\hat{\boldsymbol{\theta}}, \hat{\mathbf{p}} | \mathbf{y}, \mathbf{s})$ é o valor da função log verossimilhança avaliada no estimador de máxima verossimilhança, γ_k é o número de parâmetros considerados no modelo com $k = 1, \dots, K$ componentes de mistura e T é o tamanho da amostra.

O AIC para o modelo de mistura de normais em que $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}^2)$, $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_K)$ e $\hat{\boldsymbol{\sigma}}^2 = (\hat{\sigma}_1^2, \dots, \hat{\sigma}_K^2)$ é dado por

$$\begin{aligned} AIC(k) = & -2 \left[\sum_{k=1}^K m_k \log \hat{p}_k - \frac{1}{2} \sum_{k=1}^K m_k \log 2\pi - \sum_{k=1}^K \frac{\sum_{t=1}^T I_{s_t}(k)(y_t - \hat{\mu}_k)^2}{2\hat{\sigma}_k^2} \right. \\ & \left. - \frac{1}{2} \sum_{k=1}^K m_k \log \hat{\sigma}_k^2 \right] + \gamma_k. \end{aligned} \quad (3.11)$$

com γ_k dado por

$$\gamma_k = 3k - 1 \quad (3.12)$$

para $k = 1, \dots, K$ e o *BIC* (*Bayesian Information Criterion*) é

$$\begin{aligned} BIC(k) &= -2 \left[\sum_{k=1}^K m_k \log \hat{p}_k - \frac{1}{2} \sum_{k=1}^K m_k \log 2\pi - \sum_{k=1}^K \frac{\sum_{t=1}^T I_{st}(k) (y_t - \hat{\mu}_k)^2}{2\hat{\sigma}_k^2} \right. \\ &\quad \left. - \frac{1}{2} \sum_{k=1}^K m_k \log \hat{\sigma}_k^2 \right] + (\gamma_k) \log(T). \end{aligned} \quad (3.13)$$

3.1 MMM(1) de normais

O modelo de mistura com dependência markoviana de primeira ordem de distribuições normais em que $Y_t | S_t = k \sim N(\mu_k, \sigma_k^2)$, $t = 1, \dots, T$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ e $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_K^2)$ são os vetores de parâmetros. A sequência de variáveis aleatórias não observáveis \mathbf{S} é uma cadeia de Markov de primeira ordem que assume valores no espaço de estados $E = \{1, \dots, K\}$, probabilidade inicial p_0 e cuja matriz de transição é dada por $\mathbf{P} = \{p_{jk}\}$ com $j, k \in \{1, \dots, K\}$.

A função de máxima verossimilhança de \mathbf{y} e \mathbf{s} dada em (2.35) e aplicada a mistura de distribuições normais é

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}_0, \mathbf{P} | \mathbf{y}, \mathbf{s}) &= \prod_{k=1}^K \left\{ p_{0k}^{m_{0k}} \prod_{j=1}^K p_{jk}^{m_{jk}} \prod_{t=1}^T \left(\frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\left\{ -\frac{(y_t - \mu_k)^2}{2\sigma_k^2} \right\}} I_{(-\infty, +\infty)}(y_t) \right)^{I_{st}(k)} \right\} \\ &= \prod_{k=1}^K \left\{ p_{0k}^{m_{0k}} \left[\prod_{j=1}^K p_{jk}^{m_{jk}} \right] \left[(2\pi\sigma_k^2)^{-\frac{m_k}{2}} e^{\left\{ -\frac{\sum_{t=1}^T I_{st}(k) (y_t - \mu_k)^2}{2\sigma_k^2} \right\}} \right. \right. \\ &\quad \left. \left. \times \prod_{t=1}^T I_{(-\infty, +\infty)}(y_t)^{I_{st}(k)} \right] \right\} \end{aligned} \quad (3.14)$$

em que $m_k = \sum_{t=1}^T I_{st}(k)$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ com $\mu_k \in (-\infty, +\infty)$, $\mathbf{p} = (p_1, \dots, p_K)$ com $\sum_{k=1}^K p_k = 1$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_K^2)$, $\sigma_k^2 > 0$.

Usando os multiplicadores de Lagrange devido as restrições $\sum_{k=1}^K p_{0k} = 1$, $\sum_{k=1}^K p_{jk} = 1$ para $j = 1, \dots, K$ e a função log verossimilhança, $l(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p} | \mathbf{y}, \mathbf{s}) =$

$\log \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p} | \mathbf{y}, \mathbf{s})$, temos

$$\begin{aligned}
l(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}_0, \mathbf{P} | \mathbf{y}, \mathbf{s}) &= \sum_{k=1}^K m_{0k} \log p_{0k} + \sum_{k=1}^K \sum_{j=1}^K m_{jk} \log p_{jk} - \frac{1}{2} \sum_{k=1}^K m_k \log 2\pi - \\
&- \frac{1}{2} \sum_{k=1}^K m_k \log \sigma_k^2 - \sum_{k=1}^K \frac{\sum_{t=1}^T I_{st}(k) (y_t - \mu_k)^2}{2\sigma_k^2} + \\
&+ \lambda_0 \left(1 - \sum_{k=1}^K p_{0k}\right) + \lambda_1 \left(1 - \sum_{k=1}^K p_{1k}\right) + \dots + \lambda_K \left(1 - \sum_{k=1}^K p_{Kk}\right).
\end{aligned} \tag{3.15}$$

Derivando a função log verossimilhança em relação aos parâmetros e igualando a zero temos

$$\begin{aligned}
\left. \frac{\partial l(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}_0, \mathbf{P} | \mathbf{y}, \mathbf{s})}{\partial \mu_k} \right|_{\mu_k = \hat{\mu}_k} &= \left[\sum_{t=1}^T \frac{2}{2\sigma_k^2} I_{st}(k) (y_t - \mu_k) \right] \Bigg|_{\mu_k = \hat{\mu}_k} = 0, \\
\left. \frac{\partial l(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}_0, \mathbf{P} | \mathbf{y}, \mathbf{s})}{\partial \sigma_k^2} \right|_{\sigma_k^2 = \hat{\sigma}_k^2} &= \left[-\frac{m_k}{2\sigma_k^2} + \frac{\sum_{t=1}^T I_{st}(k) (y_t - \mu_k)^2}{2\sigma_k^4} \right] \Bigg|_{\sigma_k^2 = \hat{\sigma}_k^2} = 0, \\
\left. \frac{\partial l(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}_0, \mathbf{P} | \mathbf{y}, \mathbf{s})}{\partial p_{0k}} \right|_{p_{0k} = \hat{p}_{0k}} &= \left[\frac{m_{0k}}{p_{0k}} - \lambda_0 \right] \Bigg|_{p_{0k} = \hat{p}_{0k}} = 0, \\
\left. \frac{\partial l(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}_0, \mathbf{P} | \mathbf{y}, \mathbf{s})}{\partial p_{jk}} \right|_{p_{jk} = \hat{p}_{jk}} &= \left[m_{jk} \frac{1}{p_{jk}} - \lambda_j \right] \Bigg|_{p_{jk} = \hat{p}_{jk}} = 0
\end{aligned}$$

em que $\lambda_0 = \sum_{k=1}^K m_{0k}$ e $\lambda_j = \sum_{k=1}^K m_{jk}$ para $j = 1, \dots, K$.

Os estimadores de máxima verossimilhança são

$$\hat{\mu}_k = \frac{\sum_{t=1}^T I_{st}(k) y_t}{m_k}, \tag{3.16}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{t=1}^T I_{st}(k) (y_t - \hat{\mu}_k)^2}{m_k}, \tag{3.17}$$

$$\hat{p}_{0k} = \frac{m_{0k}}{\sum_{k=1}^K m_{0k}}, \tag{3.18}$$

$$\hat{p}_{jk} = \frac{m_{jk}}{\sum_{k=1}^K m_{jk}} \tag{3.19}$$

em que $j, k = 1, \dots, K$.

Para o cálculo dos estimadores de máxima verossimilhança do MMM(1) de distribuições normais, utilizamos o algoritmo EM adaptado.

O AIC para o MMM(1) de distribuições normais é

$$\begin{aligned} AIC(k) &= -2 \left[\sum_{k=1}^K m_{0k} \log p_{0k} + \sum_{k=1}^K \sum_{j=1}^K m_{jk} \log p_{jk} - \frac{1}{2} \sum_{k=1}^K m_k \log 2\pi - \frac{1}{2} \sum_{k=1}^K m_k \log \sigma_k^2 \right. \\ &\quad \left. - \sum_{k=1}^K \frac{\sum_{t=1}^T I_{st}(k)(y_t - \mu_k)^2}{2\sigma_k^2} \right] + \gamma_k. \end{aligned} \quad (3.20)$$

em γ_k é dado por

$$\gamma_k = 3k + k(k - 1) - 1 \quad (3.21)$$

para $k = 1, \dots, K$ e o *BIC* (*Bayesian Information Criterion*) é

$$\begin{aligned} BIC(k) &= -2 \left[\sum_{k=1}^K m_{0k} \log p_{0k} + \sum_{k=1}^K \sum_{j=1}^K m_{jk} \log p_{jk} - \frac{1}{2} \sum_{k=1}^K m_k \log 2\pi - \frac{1}{2} \sum_{k=1}^K m_k \log \sigma_k^2 \right. \\ &\quad \left. - \sum_{k=1}^K \frac{\sum_{t=1}^T I_{st}(k)(y_t - \mu_k)^2}{2\sigma_k^2} \right] + (\gamma_k) \log(T). \end{aligned} \quad (3.22)$$

3.2 Abordagem Bayesiana

Nesta seção, desenvolvemos uma abordagem bayesiana para o modelo de mistura em que a variável aleatória não observável é independente.

A distribuição *a priori* conjunta para os parâmetros do modelo de mistura com independência na variável não observável é estabelecida considerando independência entre os parâmetros associados a \mathbf{Y} , $\boldsymbol{\theta}$, e os parâmetros associados a \mathbf{S} e \mathbf{p} . Temos, portanto,

$$\pi(\boldsymbol{\theta}, \mathbf{p}) = \pi(\boldsymbol{\theta})\pi(\mathbf{p}) \quad (3.23)$$

Considerando a função de verossimilhança dada em (2.8), a distribuição *a priori* conjunta (3.23) temos que a distribuição *a posteriori* conjunta é

$$\pi(\boldsymbol{\theta}, \mathbf{p} | \mathbf{y}, \mathbf{s}) = \frac{\mathcal{L}(\boldsymbol{\theta}, \mathbf{p} | \mathbf{y}, \mathbf{s})\pi(\boldsymbol{\theta})\pi(\mathbf{p})}{\int_{\Theta} \int_{(0,1)} \mathcal{L}(\boldsymbol{\theta}, \mathbf{p} | \mathbf{y}, \mathbf{s})\pi(\boldsymbol{\theta})\pi(\mathbf{p})d\boldsymbol{\theta}d\mathbf{p}}. \quad (3.24)$$

Considerando a aplicação do modelo de mistura de distribuições normais e a função de verossimilhança dada pela equação (3.1), podemos reescrever a função de

verossimilhança da seguinte maneira

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p} | \mathbf{y}, \mathbf{s}) \propto \prod_{k=1}^K p_k^{m_k} \left(\frac{1}{\sigma_k^2} \right)^{\frac{m_k}{2}} e^{-\frac{(H_k^2)(m_k-1) + m_k(\bar{y}_k - \mu_k)^2}{2\sigma_k^2}} \prod_{t=1}^T I_{(-\infty, +\infty)}(y_t)^{I_{s_t}(k)} \quad (3.25)$$

em que $m_k = \sum_{t=1}^T I_{s_t}(k)$, $\bar{y}_k = \frac{\sum_{t=1}^T I_{s_t}(k)y_t}{m_k}$ e $H_k^2 = \frac{\sum_{t=1}^T I_{s_t}(k)(y_t - \bar{y}_k)^2}{m_k - 1}$.

Para as componentes de mistura, $\mathbf{p} = (p_1, \dots, p_K)$, assumimos uma distribuição *a priori* Dirichlet com parâmetros $\boldsymbol{\tau} = (\tau_1, \dots, \tau_K)$ dada por

$$\pi(\mathbf{p} | \boldsymbol{\tau}) = \frac{\Gamma(\sum_{k=1}^K \tau_k)}{\prod_{k=1}^K \Gamma(\tau_k)} \prod_{k=1}^K p_k^{\tau_k - 1} I_{\{C_K\}}(p) \quad (3.26)$$

em que $C_K = \left\{ \mathbf{p} = (p_1, \dots, p_K) : 0 \leq p_k \leq 1, k = 1, \dots, K, \sum_{k=1}^K p_k = 1 \right\}$ é o simplex de dimensão K , $\boldsymbol{\tau} = (\tau_1, \dots, \tau_K)$ com $\tau_k > 0$ e conhecido para $k = 1, \dots, K$.

A distribuição *a priori* para μ_k dado σ_k^2 é uma distribuição *a priori* normal com média α_k conhecida e variância condicionada σ_k^2 , $N(\alpha_k, \sigma_k^2)$, dada por

$$\pi(\mu_k | \alpha_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\left\{ -\frac{(\mu_k - \alpha_k)^2}{2\sigma_k^2} \right\}} I_{(-\infty, \infty)}(\mu_k) \quad (3.27)$$

em que $-\infty < \alpha_k < \infty$ e $\sigma_k^2 > 0$ $k = 1, \dots, K$ são conhecidos.

A distribuição *a priori* conjunta para $\boldsymbol{\mu}$ dado $\boldsymbol{\sigma}^2$ é

$$\pi(\boldsymbol{\mu} | \boldsymbol{\alpha}, \boldsymbol{\sigma}^2) = \prod_{k=1}^K \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\left\{ -\frac{(\mu_k - \alpha_k)^2}{2\sigma_k^2} \right\}} I_{(-\infty, \infty)}(\mu_k) \quad (3.28)$$

em que $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$ são conhecidos e $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_K^2)$ condicionado.

A distribuição *a priori* para σ_k^2 é uma distribuição gama inversa (GI) com parâmetros λ_k e β_k , $GI(\lambda_k, \beta_k)$, dada por

$$\pi(\sigma_k^2 | \lambda_k, \beta_k) = \frac{(\beta_k)^{\lambda_k}}{\Gamma(\lambda_k)} \left(\frac{1}{\sigma_k^2} \right)^{\lambda_k + 1} e^{\left\{ -\frac{\beta_k}{\sigma_k^2} \right\}} I_{(0, \infty)}(\sigma_k^2), \quad (3.29)$$

em que $\beta_k > 0$ e $\lambda_k > 0$ $k = 1, \dots, K$ são conhecidos. A distribuição *a priori* conjunta para $\boldsymbol{\sigma}^2$ é

$$\pi(\boldsymbol{\sigma}^2 | \boldsymbol{\lambda}, \boldsymbol{\beta}) = \prod_{k=1}^K \frac{(\beta_k)^{\lambda_k}}{\Gamma(\lambda_k)} \left(\frac{1}{\sigma_k^2} \right)^{\lambda_k + 1} e^{\left\{ -\frac{\beta_k}{\sigma_k^2} \right\}} I_{(0, \infty)}(\sigma_k^2) \quad (3.30)$$

em que $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)$ e $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$ são conhecidos.

A densidade da distribuição *a posteriori* conjunta resulta da função de verossimilhança (3.25) e da distribuição *a priori* para os parâmetros \mathbf{p} , $\boldsymbol{\mu}$ e $\boldsymbol{\sigma}^2$, dadas por (3.26), (3.27) e (3.30). Assim, obtemos

$$\begin{aligned}
\pi(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p} | \mathbf{y}, \mathbf{s}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &\propto \prod_{k=1}^K \left\{ p_k^{m_k} \left(\frac{1}{\sigma_k^2} \right)^{\frac{m_k}{2}} e^{-\frac{H_k^2(m_k-1)}{2\sigma_k^2}} e^{-\frac{m_k(\bar{y}_k - \mu_k)^2}{2\sigma_k^2}} \prod_{t=1}^T I_{(-\infty, +\infty)}(y_t)^{I_{st}(k)} \right. \\
&\times p_k^{\tau_k-1} I_{\{C_K\}}(p) \left(\frac{1}{\sigma_k^2} \right)^{\frac{1}{2}} e^{\left\{ -\frac{(\mu_k - \alpha_k)^2}{2\sigma_k^2} \right\}} I_{(-\infty, \infty)}(\mu_k) \\
&\times \left(\frac{1}{\sigma_k^2} \right)^{\lambda_k+1} e^{\left\{ -\frac{\beta_k^2}{\sigma_k^2} \right\}} I_{(0, \infty)}(\sigma_k^2) \\
&\propto \prod_{k=1}^K \left\{ p_k^{m_k + \tau_k - 1} \left(\frac{1}{\sigma_k^2} \right)^{\frac{m_k+1}{2} + \lambda_k + 1} e^{\left\{ -\frac{H_k^2(m_k-1)}{2\sigma_k^2} - \frac{\beta_k^2}{\sigma_k^2} \right\}} \right. \\
&\times \left. e^{\left\{ -\frac{m_k(\bar{y}_k - \mu_k)^2 + (\mu_k - \alpha_k)^2}{2\sigma_k^2} \right\}} \right\}. \tag{3.31}
\end{aligned}$$

Desenvolvendo a soma de quadrados, $\frac{m_k(\bar{y}_k - \mu_k)^2 + (\mu_k - \alpha_k)^2}{2\sigma_k^2}$, temos

$$\begin{aligned}
\frac{m_k(\bar{y}_k - \mu_k)^2 + (\mu_k - \alpha_k)^2}{2\sigma_k^2} &= \frac{[m_k \bar{y}_k^2 - 2m_k \bar{y}_k \mu_k + m_k \mu_k^2 + \mu_k^2 - 2\mu_k \alpha_k + \alpha^2]}{2\sigma_k^2} \\
&= \frac{[\mu_k^2(m_k + 1) - 2\mu_k(m_k \bar{y}_k + \alpha_k) + m_k \bar{y}_k^2 + \alpha^2]}{2\sigma_k^2} \\
&= \frac{(m_k + 1) \left[\mu_k^2 - 2\mu_k \left(\frac{m_k \bar{y}_k + \alpha_k}{m_k + 1} \right) + \frac{m_k \bar{y}_k^2}{m_k + 1} + \frac{\alpha^2}{m_k + 1} \right]}{2\sigma_k^2} \\
&= \frac{\left[\mu_k^2 - 2\mu_k \left(\frac{m_k \bar{y}_k + \alpha_k}{m_k + 1} \right) + \frac{m_k \bar{y}_k^2}{m_k + 1} + \frac{\alpha^2}{m_k + 1} \right]}{2 \frac{\sigma_k^2}{(m_k + 1)}}. \tag{3.32}
\end{aligned}$$

Usando (3.32) em (3.31), complementando os quadrados e eliminando os termos constantes, podemos reescrever (3.31) na forma

$$\begin{aligned}
\pi(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p} | \mathbf{y}, \mathbf{s}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &\propto \prod_{k=1}^K \left[p_k^{m_k + \tau_k - 1} \left(\frac{1}{\sigma_k^2} \right)^{\frac{m_k+1}{2} + \lambda_k + 1} e^{\left\{ -\frac{H_k^2(m_k-1)}{2\sigma_k^2} - \frac{\beta_k^2}{\sigma_k^2} \right\}} \right. \\
&\times \left. e^{\left\{ -\frac{\left(\mu_k - \frac{m_k \bar{y}_k + \alpha_k}{m_k + 1} \right)^2}{2 \frac{\sigma_k^2}{(m_k + 1)}} \right\}} \right] \tag{3.33}
\end{aligned}$$

Observamos, que dado \mathbf{y} os parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\sigma}^2$ são independente do parâmetro \mathbf{p} associado a \mathbf{s} na distribuição *a posteriori* dada pela equação(3.33).

Os estimadores bayesianos dos parâmetros são obtidos através do método MCMC usando as distribuições condicionais. As distribuições condicionais para os parâmetros são dadas por

$$\mathbf{p}|\boldsymbol{\lambda}, \mathbf{y}, \mathbf{s} \sim \text{Dirichlet}(m_1 + \tau_1, \dots, m_K + \tau_K), \quad (3.34)$$

$$\sigma_k^2|\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{y}, \mathbf{s} \sim \text{GI}\left(\frac{m_k + 1}{2} + \lambda_k; \beta_k + \frac{S_k^2(m_k - 1)}{2}\right) \quad (3.35)$$

em que $k = 1, \dots, K$ e

$$\mu_k|\boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \mathbf{y}, \mathbf{s} \sim N\left(\frac{m_k \bar{y}_k + \alpha_k}{m_k + 1}, \frac{\sigma_k^2}{m_k + 1}\right) \quad (3.36)$$

em que $k = 1, \dots, K$.

A probabilidade condicional de S_t dado Y_t para $t = 1, \dots, T$ é

$$S_t|\mathbf{y}, \boldsymbol{\theta}, \mathbf{p} \sim \text{mult}(1, q_{t1}, \dots, q_{tK}). \quad (3.37)$$

Usando a distribuição conjunta e o teorema de Bayes, temos

$$\begin{aligned} q_{tk} &= \text{Pr}(S_t = k|Y_t = y_t, \boldsymbol{\theta}, \mathbf{p}) \\ &= \frac{\text{Pr}(S_t = k, Y_t = y_t, \boldsymbol{\theta}, \mathbf{p})}{\text{Pr}(Y_t = y_t, \boldsymbol{\theta}, \mathbf{p})} \\ &= \frac{\text{Pr}(S_t = k, Y_t = y_t, \boldsymbol{\theta}, \mathbf{p})}{\sum_{l=1}^K \text{Pr}(Y_t = y_t, S_t = l, \boldsymbol{\theta}, \mathbf{p})} \\ &= \frac{\text{Pr}(S_t = k|p)\text{Pr}(Y_t = y_t|S_t = k, \boldsymbol{\theta}, \mathbf{p})}{\sum_{j=1}^K \text{Pr}(S_t = j|p)\text{Pr}(Y_t = y_t|S_t = j, \boldsymbol{\theta}, \mathbf{p})} \\ &= \frac{p_k \text{Pr}(Y_t = y_t|S_t = k, \boldsymbol{\theta}, \mathbf{p})}{\sum_{j=1}^K p_j \text{Pr}(Y_t = y_t|S_t = j, \boldsymbol{\theta}, \mathbf{p})}. \end{aligned} \quad (3.38)$$

Se Y_t tem distribuição contínua, então, reescrevemos (3.38) da seguinte forma

$$q_{tk} = \frac{p_k f_{Y_t}(y_t|S_t = k, \boldsymbol{\theta}, \mathbf{p})}{\sum_{j=1}^K p_j f_{Y_t}(y_t|S_t = j, \boldsymbol{\theta}, \mathbf{p})}. \quad (3.39)$$

em que $t = 1, \dots, T$ e $k = 1, \dots, K$.

Aplicando (3.39) ao modelo de mistura de distribuições normais e considerando $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ com $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ e $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_K^2)$, assim, a estimação de S_t é dada por

$$S_t | \mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p} \sim \text{mult}(1, q_{t1}, \dots, q_{tK}) \quad (3.40)$$

em que

$$q_{tk} = \frac{p_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(y_t - \mu_k)^2}{2\sigma_k^2}}}{\sum_{j=1}^K p_j \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(y_t - \mu_j)^2}{2\sigma_j^2}}} \quad (3.41)$$

em que $t = 1, \dots, T$ e $k = 1, \dots, K$.

A sequência dos valores estimados da variável não observável é

$$s_t = \max_k Pr(S_t = k | Y_t = y_t, \boldsymbol{\theta}, \mathbf{p}) \quad (3.42)$$

em que $t = 1, \dots, T$ e $k = 1, \dots, K$. As distribuições condicionais são conhecidas, utilizamos o algoritmo *Gibbs sampling* para gerar as amostras dos parâmetros e fazer a estimação dos mesmos.

3.2.1 Algoritmo MCMC

1. Faça $i = 0$ e atribua valores iniciais aos parâmetros $\boldsymbol{\mu}^{(i)} = (\mu_1^{(i)}, \dots, \mu_K^{(i)})$, $\mathbf{p}^{(i)} = (p_1^{(i)}, \dots, p_K^{(i)})$ e $\boldsymbol{\sigma}^{2(i)} = (\sigma_1^{2(i)}, \dots, \sigma_K^{2(i)})$.

2. Construa a sequência $s_t^{(i+1)}$

$$s_t^{(i+1)} = \max_k Pr(S_t = k | Y_t = y_t, \boldsymbol{\theta}^{(i)}, \mathbf{p}^{(i)}) \quad (3.43)$$

em que $k = 1, \dots, K$ e $t = 1, \dots, T$.

3. Gere valores das distribuições condicionais

$$p^{(i+1)} | \boldsymbol{\tau}, \mathbf{s} \sim \text{Dirichlet}(\tau_1 + m_k^{(i+1)}, \dots, \tau_K + m_k^{(i+1)}), \quad (3.44)$$

em que $\tau_k > 0$ são conhecidos para $k = 1, \dots, K$,

$$\sigma_k^{2(i+1)} | \boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{y}, \mathbf{s} \sim GI \left(\frac{m_k^{(i+1)}}{2} + \lambda_k; \beta_k + \frac{H_k^{2(i+1)} (m_k^{(i+1)} + 1)}{2} \right) \quad (3.45)$$

em que λ_k, β_k conhecidos e positivos, e

$$\mu_k^{(i+1)} | \boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \mathbf{y}, \mathbf{s} \sim N \left(\frac{m_k^{(i+1)} \bar{y}_k^{(i+1)} + \alpha_k}{m_k^{(i+1)} + 1}; \frac{\sigma_k^{2(i+1)}}{m_k^{(i+1)} + 1} \right) \quad (3.46)$$

em que $k = 1, \dots, K$.

4. Faça $i = i + 1$ e repita os passos 2 e 3 um número grande de vezes até a convergência. Descarte os valores iniciais.

3.3 Abordagem Bayesiana para MMM(1)

Considerando a aplicação do modelo de mistura de distribuições normais e a função de verossimilhança dada pela equação (3.14), podemos reescrever a função de verossimilhança da seguinte forma

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}_0, \mathbf{P} | \mathbf{y}, \mathbf{s}) &= \prod_{k=1}^K \left\{ p_{0k}^{m_{0k}} \left[\prod_{j=1}^K p_{jk}^{m_{jk}} \right] \left[(2\pi\sigma_k^2)^{-\frac{m_k}{2}} e^{-\frac{(H_k^2)(m_k-1) + m_k(\bar{y}_k - \mu_k)^2}{2\sigma_k^2}} \right] \right. \\ &\quad \left. \times \prod_{t=1}^T I_{(-\infty, \infty)}(y_t)^{I_{st}(k)} \right\} \end{aligned} \quad (3.47)$$

em que $m_k = \sum_{t=1}^T I_{st}(k)$, $\bar{y}_k = \frac{\sum_{t=1}^T I_{st}(k)y_t}{m_k}$ e $H_k^2 = \frac{\sum_{t=1}^T I_{st}(k)(y_t - \bar{y}_k)^2}{m_k - 1}$.

A distribuição *a priori* para μ_k dada σ_k^2 é uma distribuição normal dada por

$$\pi(\mu_k | \alpha_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\left\{ -\frac{(\mu_k - \alpha_k)^2}{2\sigma_k^2} \right\}} I_{(-\infty, \infty)}(\mu_k) \quad (3.48)$$

em que $-\infty < \alpha_k < \infty$ $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$ são conhecidos e $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_K^2)$ condicionados $\sigma_k^2 > 0$, para $k = 1, \dots, K$.

A distribuição *a priori* conjunta para $\boldsymbol{\mu}$ dado $\boldsymbol{\sigma}^2$ é uma distribuição normal dada por

$$\pi(\boldsymbol{\mu} | \boldsymbol{\alpha}, \boldsymbol{\sigma}^2) = \prod_{k=1}^K \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\left\{ -\frac{(\mu_k - \alpha_k)^2}{2\sigma_k^2} \right\}} I_{(-\infty, \infty)}(\mu_k) \quad (3.49)$$

em que $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$ são conhecidos e $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_K^2)$ condicionados.

A distribuição *a priori* para $\sigma_k^2 | \lambda_k, \beta_k$ é a gama inversa (GI), $GI(\lambda_k, \beta_k)$, e é dada por

$$\pi(\sigma_k^2 | \lambda_k, \beta_k) = \frac{(\beta_k)^{\lambda_k}}{\Gamma(\lambda_k)} \left(\frac{1}{\sigma_k^2} \right)^{\lambda_k + 1} e^{\left\{ -\frac{\beta_k}{\sigma_k^2} \right\}} I_{(0, \infty)}(\sigma_k^2) \quad (3.50)$$

em que $\beta_k > 0$ e $\lambda_k > 0$ $k = 1, \dots, K$ são conhecidos. A distribuição *a priori* conjunta para σ^2 é

$$\pi(\sigma^2 | \boldsymbol{\lambda}, \boldsymbol{\beta}) = \prod_{k=1}^K \frac{(\beta_k)^{\lambda_k}}{\Gamma(\lambda_k)} \left(\frac{1}{\sigma_k^2} \right)^{\lambda_k+1} e^{\left\{ -\frac{\beta_k^2}{\sigma_k^2} \right\}} I_{(0, \infty)}(\sigma^2) \quad (3.51)$$

em que $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)$ e $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$ são conhecidos.

A distribuição *a priori* para \mathbf{p}_0 a distribuição Dirichlet com parâmetros $\boldsymbol{\tau} = (\tau_1, \dots, \tau_K)$ dada por

$$\pi(\mathbf{p}_0 | \boldsymbol{\tau}) = \frac{\Gamma(\sum_{k=1}^K \tau_k)}{\Gamma(\tau_1) \dots \Gamma(\tau_K)} \prod_{k=1}^K p_{0k}^{\tau_k-1} I_{\{C_k\}}(p_{01}, \dots, p_{0K}) \quad (3.52)$$

em que $C_K = \left\{ \mathbf{p}_0 = (p_{01}, \dots, p_{0K}) : 0 < p_{0k} < 1, k = 1, \dots, K, \sum_{k=1}^K p_{0k} = 1 \right\}$ é o simplex, $\boldsymbol{\tau} = (\tau_1, \dots, \tau_K)$, $\tau_k > 0$ e conhecido com $k = 1, \dots, K$.

A distribuição *a priori* para cada linha da matriz de transição, \mathbf{p}_j , é uma distribuição Dirichlet com parâmetros $\boldsymbol{\gamma}_j = (\gamma_{j1}, \dots, \gamma_{jK})$ dada

$$\pi(\mathbf{p}_j | \boldsymbol{\gamma}_j) = \frac{\Gamma(\sum_{k=1}^K \gamma_{jk})}{\Gamma(\gamma_{j1}) \dots \Gamma(\gamma_{jK})} \prod_{k=1}^K p_{jk}^{\gamma_{jk}-1} I_{\{C_{jK}\}}(p_{j1}, \dots, p_{jK}) \quad (3.53)$$

em que o simplex é $C_{jK} = \left\{ \mathbf{p}_j = (p_{j1}, \dots, p_{jK}) : 0 < p_{jk} < 1, k = 1, \dots, K, \sum_{k=1}^K p_{jk} = 1 \right\}$, $\boldsymbol{\gamma}_j = (\gamma_{j1}, \dots, \gamma_{jK})$ com $\gamma_{jk} > 0$ e conhecido para $j, k = 1, \dots, K$.

A distribuição *a priori* conjunta para a matriz de transição, \mathbf{P} , é estabelecida de modo que \mathbf{p}_j 's são independente para cada linha de \mathbf{P} , $\mathbf{p}_j \sim Dir(\gamma_{j1}, \dots, \gamma_{jK})$, assim temos

$$\begin{aligned} \pi(\mathbf{P} | \boldsymbol{\gamma}) &= \prod_{j=1}^K \pi(\mathbf{p}_j | \boldsymbol{\gamma}_j) \\ &= \prod_{j=1}^K \frac{\Gamma(\sum_{k=1}^K \gamma_{jk})}{\Gamma(\gamma_{j1}) \dots \Gamma(\gamma_{jK})} \prod_{k=1}^K p_{jk}^{\gamma_{jk}-1} I_{\{C_{jK}\}}(p_{j1}, \dots, p_{jK}). \end{aligned} \quad (3.54)$$

A distribuição *a posteriori* conjunta para $(\mathbf{p}_0, \mathbf{P}, \boldsymbol{\theta})$ é obtida multiplicando a função de verossimilhança (3.47) pelas distribuições *a priori* (3.49), (3.51), (3.52) e (3.54)

e é dada por

$$\begin{aligned} \pi(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}_0, \mathbf{P} | \mathbf{y}, \mathbf{s}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &\propto \prod_{k=1}^K \left\{ p_{0k}^{m_{0k} + \tau_k - 1} \left[\prod_{j=1}^K p_{jk}^{m_{jk} + \gamma_{jk} - 1} \right] \left(\frac{1}{\sigma_k^2} \right)^{\frac{m_k + 1}{2} + \lambda_k + 1} \right. \\ &\times \left. \exp \left\{ -\frac{H_k^2(m_k - 1)}{2\sigma_k^2} - \frac{\beta_k^2}{\sigma_k^2} \right\} \exp \left\{ -\frac{\left(\mu_k - \frac{m_k \bar{y}_k + \alpha_k}{m_k + 1} \right)^2}{2 \frac{\sigma_k^2}{(m_k + 1)}} \right\} \right\}. \end{aligned} \quad (3.55)$$

Temos, que dado \mathbf{y} os parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\sigma}^2$ são independente dos parâmetros \mathbf{p}_0 e \mathbf{P} associados a \mathbf{s} na distribuição *a posteriori* dada pela equação(3.55).

Os estimadores bayesianos dos parâmetros são obtidos através do método MCMC usando as distribuições condicionais. As distribuições condicionais para os parâmetros são dadas por

$$\mathbf{p}_0 | \boldsymbol{\lambda}, \mathbf{y}, \mathbf{s} \sim \text{Dirichlet}(\tau_1 + m_{01}, \dots, \tau_K + m_{0K}), \quad (3.56)$$

$$\mathbf{p}_j | \boldsymbol{\gamma}_j, \mathbf{y}, \mathbf{s} \sim \text{Dirichlet}(\gamma_{j1} + m_{j1}, \dots, \gamma_{jK} + m_{jK}) \quad (3.57)$$

em que $j = 1, \dots, K$,

$$\sigma_k^2 | \boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{y}, \mathbf{s} \sim GI \left(\frac{m_k + 1}{2} + \lambda_k; \beta_k + \frac{H_k^2(m_k - 1)}{2} \right) \quad (3.58)$$

em que $k = 1, \dots, K$ e

$$\mu_k | \boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \mathbf{y}, \mathbf{s} \sim N \left(\frac{m_k \bar{y}_k + \alpha_k}{m_k + 1}, \frac{\sigma_k^2}{m_k + 1} \right) \quad (3.59)$$

em que $k = 1, \dots, K$.

Considere a probabilidade condicional de S_1 dado Y_1 em que

$$S_1 | Y_1 = y_1, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P} \sim \text{mult}(1, (q_{11}, \dots, q_{1K})) \quad (3.60)$$

em que utilizando a distribuição conjunta e o teorema de Bayes temos

$$\begin{aligned} q_{1k} = Pr(S_1 = k | Y_1 = y_1, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}) &= \frac{Pr(S_1 = k, Y_1 = y_1, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}{Pr(Y_1 = y_1, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})} \\ &= \frac{Pr(S_1 = k, Y_1 = y_1, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}{\sum_{l=1}^K Pr(Y_1 = y_1, S_1 = l, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})} \\ &= \frac{p_{0k} P(Y_1 = y_1 | S_1 = k, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}{\sum_{l=1}^K p_{0l} P(Y_1 = y_1 | S_1 = l, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}. \end{aligned} \quad (3.61)$$

em que $k = 1, \dots, K$.

No caso que Y tem distribuição contínua a equação (3.61) passa a ser

$$Pr(S_1 = k | Y_1 = y_1, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}) = \frac{p_{0k} f_{Y_1}(y_1 | S_1 = k, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}{\sum_{l=1}^K p_{0l} f_{Y_1}(y_1 | S_1 = l, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})} \quad (3.62)$$

Considerando j conhecido e $t = 2, \dots, T$ temos a probabilidade condicional de S_t dado Y_t em que

$$S_t | Y_t = y_t, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P} \sim \text{mult}(1, (q_{tj1}, \dots, q_{tjK})) \quad (3.63)$$

em que

$$q_{tjk} = Pr(S_t = k | Y_t = y_t, S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}) = \frac{Pr(Y_t = y_t, S_t = k, S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}{Pr(Y_t = y_t, S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}$$

$$\begin{aligned} bm_{\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}} &= \frac{Pr(Y_t = y_t, S_t = k, S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}{\sum_{l=1}^K Pr(Y_t = y_t, S_t = l, S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})} \\ &= \frac{p_{jk} Pr(Y_t = y_t | S_t = k, S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}{\sum_{l=1}^K p_{jl} Pr(Y_t = y_t | S_t = l, S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}. \end{aligned} \quad (3.64)$$

e para Y com distribuição contínua temos que

$$q_{tjk} = Pr(S_t = k | Y_t = y_t, S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}) = \frac{p_{jk} f_{Y_t}(y_t | S_t = k, S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}{\sum_{l=1}^K p_{jl} f_{Y_t}(y_t | S_t = l, S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}. \quad (3.65)$$

Considerando distribuição normal com parâmetros $\boldsymbol{\theta} = (\mu, \sigma)$ com $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ e $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_K^2)$ temos para $t = 1$ e $k = 1, \dots, K$ que

$$S_1 | Y_1 = y_1, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}_0, \mathbf{P} \sim \text{mult}(1, (q_{11}, \dots, q_{1K})) \quad (3.66)$$

em que

$$q_{1k} = Pr(S_1 = k | Y_1 = y_1, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}_0, \mathbf{P}) = \frac{p_{0k} \frac{2}{\sqrt{2\pi\sigma_k^2}} e^{\left\{-\frac{(y_1 - \mu_k)^2}{2\sigma_k^2}\right\}}}{\sum_{l=1}^K p_{0l} \frac{1}{\sqrt{2\pi\sigma_l^2}} e^{\left\{-\frac{(y_1 - \mu_l)^2}{2\sigma_l^2}\right\}}} \quad (3.67)$$

e para j dado e $t = 2, \dots, T$ temos

$$S_t | Y_t = y_t, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{p}_0, \mathbf{P} \sim \text{mult}(1, (q_{tj1}, \dots, q_{tjK})) \quad (3.68)$$

em que

$$q_{tjk} = Pr(S_1 = k | Y_1 = y_1, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p}_0, \mathbf{P}) = \frac{p_{jk} \frac{2}{\sqrt{2\pi\sigma_k^2}} e^{\left\{-\frac{(y_t - \mu_k)^2}{2\sigma_k^2}\right\}}}{\sum_{l=1}^K p_{jl} \frac{1}{\sqrt{2\pi\sigma_l^2}} e^{\left\{-\frac{(y_t - \mu_l)^2}{2\sigma_l^2}\right\}}} \quad (3.69)$$

em que $j, k = 1, \dots, K$.

A sequência estimada de s é dada por

$$s_1 = \max_k Pr(S_1 = k | Y_1 = y_1, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}) \quad (3.70)$$

$$s_t = \max_k Pr(S_t = k | Y_t = y_t, S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}) \quad (3.71)$$

em que $j, k = 1, \dots, K$ e $t = 2, \dots, T$.

Considerando que o número da componentes de mistura, K , é conhecido a estimação dos parâmetros é realizada via algoritmo *Gibbs sampling*.

3.3.1 Algoritmo MCMC para MMM(1)

1. Faça $i = 0$ e atribua valores iniciais para os parâmetros $\boldsymbol{\theta}^{(i)} = (\theta_1^{(i)}, \dots, \theta_K^{(i)})$, $\mathbf{p}_0^{(i)} = (p_{01}^{(i)}, \dots, p_{0K}^{(i)})$ e

$$\mathbf{P}^{(i)} = \begin{bmatrix} p_{11}^{(i)} & \dots & p_{1K}^{(i)} \\ p_{21}^{(i)} & \dots & p_{2K}^{(i)} \\ \vdots & & \\ p_{K1}^{(i)} & \dots & p_{KK}^{(i)} \end{bmatrix},$$

2. Construa a sequência $\mathbf{s}^{(i+1)}$ com base em \mathbf{y} , $\boldsymbol{\theta}^{(i)}$, $\mathbf{p}_0^{(i)}$, $\mathbf{P}^{(i)}$, usando

$$s_1^{(i+1)} = \max_k Pr(S_1 = k | Y_1 = y_1, \boldsymbol{\theta}^{(i)}, \mathbf{p}_0^{(i)}, \mathbf{P}^{(i)}) \quad (3.72)$$

$$s_t^{(i+1)} = \max_k Pr(S_t = k | Y_t = y_t, s_{t-1}^{(i+1)} = j, \boldsymbol{\theta}^{(i)}, \mathbf{p}_0^{(i)}, \mathbf{P}^{(i)}) \quad (3.73)$$

em que $j, k = 1, \dots, K$ e $t = 2, \dots, T$

3. Gere novos valores dos parâmetros das distribuições condicionais

$$\mathbf{p}_0^{(i+1)} | \boldsymbol{\lambda}, \mathbf{s} \sim \text{Dirichlet}(\tau_1 + m_{01}^{(i+1)}, \dots, \tau_K + m_{0K}^{(i+1)}), \quad (3.74)$$

em que $\lambda_k > 0$ e conhecido com $k = 1, \dots, K$,

$$\mathbf{p}_j^{(i+1)} | \boldsymbol{\gamma}_j, \mathbf{s} \sim \text{Dirichlet}(\gamma_{j1} + m_{j1}^{(i+1)}, \dots, \gamma_{jK} + m_{jK}^{(i+1)}) \quad (3.75)$$

em que $j, k = 1, \dots, K$ e $\gamma_{jk} > 0$ e conhecido,

$$\sigma_k^{2(i+1)} | \boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{y}, \mathbf{s} \sim GI \left(\frac{m_k^{(i+1)} + 1}{2} + \lambda_k; \beta_k + \frac{H_k^{2(i+1)} (m_k^{(i+1)} - 1)}{2} \right) \quad (3.76)$$

em que $k = 1, \dots, K$ e

$$\mu_k^{(i+1)} | \boldsymbol{\sigma}, \boldsymbol{\alpha}, \mathbf{y}, \mathbf{s} \sim N \left(\frac{m_k^{(i+1)} \bar{y}_k^{(i+1)} + \alpha_k}{m_k^{(i+1)} + 1}; \frac{\sigma_k^2}{m_k^{(i+1)} + 1} \right) \quad (3.77)$$

4. Faça $i = i + 1$ e repeta os passos 2 e 3 um número grande de vezes até atingir a convergência. Descarte os valores iniciais.

Capítulo 4

Mistura de binomiais

Nesse capítulo, num primeiro momento, particularizamos os resultados estabelecidos no capítulo 2 para o caso em que Y_t , $t = 1, \dots, T$, tem distribuição binomial. Num segundo momento, apresentamos o modelo de mistura de distribuição binomial bayesiano estabelecendo *prioris* para os parâmetros.

Para aplicação consideramos o modelo de mistura de distribuições binomiais em que $Y_t|S_t = k \sim \text{bin}(n_t, \theta_k)$ em que θ_k é o parâmetro associado à k -ésima componente da mistura, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ é o vetor de parâmetros e S_t a variável controladora da componente a que pertence cada observação. Da mesma forma que no caso normal, as variáveis aleatórias não observáveis S_t , para $t = 1, \dots, T$, são inicialmente consideradas independentes e identicamente distribuídas e, num segundo momento, são modeladas por uma cadeia de Markov.

Considerando os dados $\mathbf{y} = (y_1, \dots, y_T)$ uma realização de $\mathbf{Y} = (Y_1, \dots, Y_T)$, $\mathbf{s} = (s_1, \dots, s_T)$ com $s_t \in \{1, \dots, K\}$, uma realização de $\mathbf{S} = (S_1, \dots, S_T)$ e a função de verossimilhança conjunta (2.8) temos que a função de verossimilhança de \mathbf{y} e \mathbf{s} é

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \mathbf{p}|\mathbf{y}, \mathbf{n}, \mathbf{s}) &= \prod_{k=1}^K \left\{ p_k^{m_k} \left[\prod_{t=1}^T \left(\binom{n_t}{y_t} \theta_k^{y_t} (1 - \theta_k)^{(n_t - y_t)} I_{\{0, \dots, n_t\}}(y_t) \right)^{I_{s_t}(k)} \right] \right\} \\ &= \prod_{k=1}^K \left\{ p_k^{m_k} \left[\theta_k^{\sum_{t=1}^T I_{s_t}(k) y_t} (1 - \theta_k)^{\sum_{t=1}^T I_{s_t}(k) (n_t - y_t)} \right. \right. \\ &\quad \left. \left. \times \prod_{t=1}^T \binom{n_t}{y_t} I_{\{0, \dots, n_t\}}(y_t)^{I_{s_t}(k)} \right] \right\} \end{aligned} \quad (4.1)$$

em que $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ com $\theta_k \in (0, 1)$, $\mathbf{p} = (p_1, \dots, p_K)$ com $\sum_{k=1}^K p_k = 1$,

$\mathbf{n} = (n_1, \dots, n_T)$, $n_t \in N - \{0\}$.

Considerando os multiplicadores de Lagrange devido a restrição $\sum_{k=1}^K p_k = 1$ e a função log verossimilhança de (4.1), $l(\boldsymbol{\theta}, \mathbf{p}|\mathbf{y}, \mathbf{n}, \mathbf{s}) = \log \mathcal{L}(\boldsymbol{\theta}, \mathbf{p}|\mathbf{y}, \mathbf{n}, \mathbf{s})$, temos

$$\begin{aligned} l(\boldsymbol{\theta}, \mathbf{p}|\mathbf{y}, \mathbf{n}, \mathbf{s}) &= \sum_{k=1}^K m_k \log p_k + \sum_{k=1}^K \sum_{t=1}^T I_{s_t}(k) \log \binom{n_t}{y_t} + \sum_{k=1}^K \sum_{t=1}^T I_{s_t}(k) y_t \log \theta_k \\ &+ \sum_{k=1}^K \sum_{t=1}^T I_{s_t}(k) (n_t - y_t) \log(1 - \theta_k) + \lambda \left(1 - \sum_{k=1}^K p_k\right). \end{aligned} \quad (4.2)$$

em que $m_k = \sum_{t=1}^T I_{s_t}(k)$.

Derivando (4.1) e igualando o resultado a zero, temos o sistema de equações dado por

$$\begin{aligned} \left. \frac{\partial l(\boldsymbol{\theta}, \mathbf{p}|\mathbf{y}, \mathbf{n}, \mathbf{s})}{\partial \theta_k} \right|_{\theta_k = \hat{\theta}_k} &= \left[\sum_{t=1}^T I_{s_t}(k) y_t \frac{1}{\theta_k} - \sum_{t=1}^T I_{s_t}(k) (n_t - y_t) \frac{1}{1 - \theta_k} \right] \Bigg|_{\theta_k = \hat{\theta}_k} = 0 \\ \left. \frac{\partial l(\boldsymbol{\theta}, \mathbf{p}|\mathbf{y}, \mathbf{s})}{\partial p_k} \right|_{p_k = \hat{p}_k} &= \left[m_k \frac{1}{p_k} - \lambda \right] \Bigg|_{p_k = \hat{p}_k} = 0 \end{aligned}$$

em que $\lambda = \sum_{k=1}^K m_k = T$.

Os estimadores de máxima verossimilhança são

$$\hat{\theta}_k = \frac{\sum_{t=1}^T I_{s_t}(k) y_t}{\sum_{t=1}^T I_{s_t}(k) n_t}, \quad (4.3)$$

$$\hat{p}_k = \frac{m_k}{T} \quad (4.4)$$

para $k = 1, \dots, K$, $T = \sum_{k=1}^K m_k$.

Após a obtenção dos estimadores dos parâmetros do modelo de mistura de distribuições binomiais, (4.3) e (4.4), aplicamos o algoritmo de estimação e maximização (EM) adaptado para obter as estimativas dos parâmetros.

O AIC aplicado no modelo de mistura de binomiais dado por (4.1) é

$$\begin{aligned} AIC(k) &= -2 \left[\sum_{k=1}^K m_k \log \hat{p}_k + \sum_{k=1}^K \sum_{t=1}^T I_{s_t}(k) \log \binom{n_t}{y_t} + \sum_{k=1}^K \sum_{t=1}^T I_{s_t}(k) y_t \log \hat{\theta}_k \right. \\ &\left. + \sum_{k=1}^K \sum_{t=1}^T I_{s_t}(k) (n_t - y_t) \log(1 - \hat{\theta}_k) \right] + \gamma_k \end{aligned} \quad (4.5)$$

em que γ_k é dado por

$$\gamma_k = 2k - 1$$

para $k = 1, \dots, K$ e o *BIC* (*Bayesian Information Criterion*) é

$$\begin{aligned} BIC(k) &= -2 \left[\sum_{k=1}^K m_k \log \hat{p}_k + \sum_{k=1}^K \sum_{t=1}^T I_{s_t}(k) \log \binom{n_t}{y_t} + \sum_{k=1}^K \sum_{t=1}^T I_{s_t}(k) y_t \log \hat{\theta}_k \right. \\ &\quad \left. + \sum_{k=1}^K \sum_{t=1}^T I_{s_t}(k) (n_t - y_t) \log(1 - \hat{\theta}_k) \right] + (\gamma_k) \log(T). \end{aligned} \quad (4.6)$$

4.1 MMM(1) de binomiais

Consideramos agora o modelo de mistura de distribuições binomiais, $Y_t | S_t = k \sim \text{bin}(n_t, \theta_k)$, em que θ_k é o parâmetro associado à k -ésima componente da mistura e $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ é o vetor de parâmetros. A sequência de variáveis aleatórias não observáveis S_t é uma cadeia de Markov de primeira ordem que assume valores no espaço de estados $E = \{1, \dots, K\}$. A cadeia de Markov é caracterizada pela distribuição inicial \mathbf{p}_0 e pela matriz de transição $\mathbf{P} = \{p_{jk}\}$, com $j, k = 1, \dots, K$.

Considerando os dados $\mathbf{y} = (y_1, \dots, y_T)$ uma realização de $\mathbf{Y} = (Y_1, \dots, Y_T)$ em que $Y_t \sim \text{bin}(n_t, \theta_k)$, com $\theta_k \in (0, 1)$, $\mathbf{s} = (s_1, \dots, s_T)$ e $s_t \in \{1, \dots, K\}$ uma realização de $\mathbf{S} = (S_1, \dots, S_T)$.

A função de verossimilhança para o MMM(1) é dada por

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \mathbf{P}, \mathbf{p}_0 | \mathbf{y}, \mathbf{n}, \mathbf{s}) &= \prod_{k=1}^K \left\{ p_{0k}^{m_{0k}} \left[\prod_{j=1}^K p_{jk}^{m_{jk}} \right] \left[\prod_{t=1}^T \left[\binom{n_t}{y_t} \theta_k^{y_t} (1 - \theta_k)^{(n_t - y_t)} \right. \right. \right. \\ &\quad \left. \left. \left. \times I_{\{0, \dots, n_t\}}(y_t) \right]^{I_{s_t}(k)} \right] \right\} \\ &= \prod_{k=1}^K \left\{ p_{0k}^{m_{0k}} \left[\prod_{j=1}^K p_{jk}^{m_{jk}} \right] \left[\theta_k^{\sum_{t=1}^T I_{s_t}(k) y_t} (1 - \theta_k)^{\sum_{t=1}^T I_{s_t}(k) (n_t - y_t)} \right. \right. \\ &\quad \left. \left. \times \prod_{t=1}^T \binom{n_t}{y_t}^{I_{s_t}(k)} I_{\{0, \dots, n_t\}}(y_t)^{I_{s_t}(k)} \right] \right\} \\ &\propto \prod_{k=1}^K \left\{ p_{0k}^{m_{0k}} \left[\prod_{j=1}^K p_{jk}^{m_{jk}} \right] \left[\theta_k^{\sum_{t=1}^T I_{s_t}(k) y_t} (1 - \theta_k)^{\sum_{t=1}^T I_{s_t}(k) (n_t - y_t)} \right] \right\} \end{aligned} \quad (4.7)$$

em que $\mathbf{n} = (n_1, \dots, n_T)$, $n_t \in N - \{0\}$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ com $\theta_k \in (0, 1)$, $\mathbf{p}_0 = (p_{01}, \dots, p_{0K})$ em que $\sum_{k=1}^K p_{0k} = 1$ e $\mathbf{P} = \{p_{jk}\}$ com $\sum_{k=1}^K p_{jk} = 1$ para $j =$

$1, \dots, K$.

Usando os multiplicadores de Lagrange, devido as restrições $\sum_{k=1}^K p_{0k} = 1$ e $\sum_{k=1}^K p_{jk} = 1$, a função log-verossimilhança pode ser reescrita como

$$\begin{aligned}
l(\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P} | \mathbf{y}, \mathbf{s}, \mathbf{n}) &= \sum_{k=1}^K m_{0k} \log p_{0k} + \sum_{k=1}^K \sum_{j=1}^K m_{jk} \log p_{jk} + \sum_{k=1}^K \sum_{t=1}^T I_{s_t}(k) y_t \log \theta_k \\
&+ \sum_{k=1}^K \sum_{t=1}^T I_{s_t}(k) (n_t - y_t) \log(1 - \theta_k) + \lambda_0 \left(1 - \sum_{k=1}^K p_{0k}\right) \\
&+ \lambda_1 \left(1 - \sum_{k=1}^K p_{1k}\right) + \dots + \lambda_K \left(1 - \sum_{k=1}^K p_{Kk}\right). \tag{4.8}
\end{aligned}$$

Derivando a equação (4.8) e igualando a zero temos

$$\begin{aligned}
\left. \frac{\partial l(\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P} | \mathbf{y}, \mathbf{s}, \mathbf{n})}{\partial \theta_k} \right|_{\theta_k = \hat{\theta}_k} &= \left[\sum_{t=1}^T I_{s_t}(k) y_t \frac{1}{\theta_k} - \sum_{t=1}^T I_{s_t}(k) (n_t - y_t) \frac{1}{1 - \theta_k} \right] \Bigg|_{\theta_k = \hat{\theta}_k} = 0 \\
\left. \frac{\partial l(\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P} | \mathbf{y}, \mathbf{s}, \mathbf{n})}{\partial p_{0k}} \right|_{p_{0k} = \hat{p}_{0k}} &= \left[m_{0k} \frac{1}{p_{0k}} - \lambda_1 \right] \Bigg|_{p_{0k} = \hat{p}_{0k}} = 0 \\
\left. \frac{\partial l(\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P} | \mathbf{y}, \mathbf{s}, \mathbf{n})}{\partial p_{jk}} \right|_{p_{jk} = \hat{p}_{jk}} &= \left[m_{jk} \frac{1}{p_{jk}} - \lambda_j \right] \Bigg|_{p_{jk} = \hat{p}_{jk}} = 0
\end{aligned}$$

em que $\lambda_0 = \sum_{k=1}^K m_{0k}$ e $\lambda_j = \sum_{k=1}^K m_{jk}$ para $j = 1, \dots, K$.

Os estimadores de máxima verossimilhança são

$$\hat{\theta}_k = \frac{\sum_{t=1}^T I(s_t = k) y_t}{\sum_{t=1}^T I_{s_t}(k) n_t}, \tag{4.9}$$

$$\hat{p}_{0k} = \frac{m_{0k}}{\sum_{l=1}^K m_{0l}}, \tag{4.10}$$

$$\hat{p}_{jk} = \frac{m_{jk}}{\sum_{l=1}^K m_{jl}} \tag{4.11}$$

em que $j, k = 1, \dots, K$.

Após a obtenção dos estimadores dos parâmetros do MMM(1) de distribuições binomiais, aplicamos o algoritmo de estimação e maximização (EM) adaptado até a convergência das estimativas dos parâmetros.

Aplicando o AIC ao modelo de mistura de distribuições binomiais e usando a

função log verossimilhança (4.4) no ponto de máximo temos

$$\begin{aligned} AIC(k) = & -2 \left[\sum_{k=1}^K m_{0k} \log \hat{p}_{0k} + \sum_{k=1}^K \sum_{t=1}^T I_{st}(k) \log \binom{n_t}{y_t} + \sum_{k=1}^K \sum_{j=1}^K m_{jk} \log \hat{p}_{jk} \right. \\ & \left. + \sum_{k=1}^K \sum_{t=1}^T I_{st}(k) y_t \log \hat{\theta}_k + \sum_{k=1}^K \sum_{t=1}^T I_{st}(k) (n_t - y_t) \log(1 - \hat{\theta}_k) \right] + \gamma_k \end{aligned}$$

em que γ_k é dado por

$$\gamma_k = 2k + k(k - 1) - 1$$

para $k = 1, \dots, K$ e o BIC é dado por

$$\begin{aligned} BIC = & -2 \left[\sum_{k=1}^K m_{0k} \log \hat{p}_{0k} + \sum_{k=1}^K \sum_{t=1}^T I_{st}(k) \log \binom{n_t}{y_t} + \sum_{k=1}^K \sum_{j=1}^K m_{jk} \log \hat{p}_{jk} \right. \\ & \left. + \sum_{k=1}^K \sum_{t=1}^T I_{st}(k) y_t \log \hat{\theta}_k + \sum_{k=1}^K \sum_{t=1}^T I_{st}(k) (n_t - y_t) \log(1 - \hat{\theta}_k) \right] + (\gamma_k) \log(T). \end{aligned}$$

4.2 Abordagem Bayesiana

Consideramos, inicialmente, o modelo de mistura de distribuições binomiais da variável observável \mathbf{Y} para o caso em que as S_t 's são independentes e identicamente distribuídos $Y_t | S_t = k \sim \text{bin}(n_t, \theta_k)$ para $t = 1, \dots, T$ e a função de verossimilhança é dada na equação (4.1).

A distribuição conjunta *a priori* para $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ é um produto de distribuições beta com parâmetros (a_k, b_k) , dada por

$$\pi(\boldsymbol{\theta} | \mathbf{a}, \mathbf{b}) = \prod_{k=1}^K \frac{\Gamma(a_k + b_k)}{\Gamma(a_k)\Gamma(b_k)} \theta_k^{a_k-1} (1 - \theta_k)^{b_k-1} I_{(0,1)}(\theta_k) \quad (4.12)$$

em que $a_k, b_k > 0$ são conhecidos, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$, $\mathbf{a} = (a_1, \dots, a_K)$ e $\mathbf{b} = (b_1, \dots, b_K)$.

Para as componentes de mistura, $\mathbf{p} = (p_1, \dots, p_K)$, assumimos uma distribuição *a priori* Dirichlet com parâmetros $\boldsymbol{\tau} = (\tau_1, \dots, \tau_K)$ dada por

$$\pi(\mathbf{p} | \boldsymbol{\tau}) = \frac{\Gamma(\sum_{k=1}^K \tau_k)}{\prod_{k=1}^K \Gamma(\tau_k)} \prod_{k=1}^K p_k^{\tau_k-1} I_{\{C_K\}}(p) \quad (4.13)$$

em que $C_K = \left\{ \mathbf{p} = (p_1, \dots, p_K) : 0 \leq p_k \leq 1, k = 1, \dots, K, \sum_{k=1}^K p_k = 1 \right\}$ é o simplex, $\boldsymbol{\tau} = (\tau_1, \dots, \tau_K)$ com $\tau_k > 0$ é conhecido para $k = 1, \dots, K$.

Considerando a função de verossimilhança (4.1), e as distribuições *a priori* (4.12) e (4.13), a distribuição *a posteriori* conjunta de $\boldsymbol{\theta}$ e \mathbf{p} é dada por

$$\begin{aligned} \pi(\boldsymbol{\theta}, \mathbf{p} | \mathbf{y}, \mathbf{n}, \mathbf{s}, \mathbf{a}, \mathbf{b}, \boldsymbol{\tau}) &\propto \prod_{k=1}^K \left\{ p_k^{\sum_{t=1}^T I_{s_t}(k) + \tau_k - 1} I_{\{C_K\}}(p_1, \dots, p_K) \right. \\ &\times \theta_k^{\sum_{t=1}^T I_{s_t}(k) y_t + a_k - 1} (1 - \theta_k)^{\sum_{t=1}^T I_{s_t}(k) (n_t - y_t) + b_k - 1} I_{(0,1)}(\theta_k) \\ &\times \left. \left[\prod_{t=1}^T I_{\{0, \dots, n_t\}}(y_t)^{I(s_t=k)} \right] \right\} \end{aligned} \quad (4.14)$$

em que $\boldsymbol{\tau} = (\tau_1, \dots, \tau_K)$, $\mathbf{a} = (a_1, \dots, a_K)$, $\mathbf{b} = (b_1, \dots, b_K)$, com $\tau_k, a_k, b_k > 0$ conhecidos para $k = 1, \dots, K$.

As estimativas dos parâmetros são obtidas via MCMC utilizando as distribuições condicionais.

A distribuição condicional de $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ é um produto de distribuições beta. A distribuição condicional de cada θ_k é dada por

$$\theta_k | \mathbf{a}, \mathbf{b}, \mathbf{y}, \mathbf{n}, \mathbf{s} \sim \prod_{k=1}^K \text{beta}(a_k + \sum_{t=1}^T I_{s_t}(k) y_t; b_k + \sum_{t=1}^T I_{s_t}(k) (n_t - y_t)) \quad (4.15)$$

para $k = 1, \dots, K$.

A distribuição condicional para \mathbf{p} é uma distribuição Dirichlet

$$\mathbf{p} | \boldsymbol{\tau}, \mathbf{s} \sim \text{Dirichlet}(\tau_1 + \sum_{t=1}^T I(s_t = 1), \dots, \tau_K + \sum_{t=1}^T I_{s_t}(k)). \quad (4.16)$$

A probabilidade condicional de S_t dado Y_t para $t = 1, \dots, T$ é

$$S_t | \mathbf{y}, \boldsymbol{\theta}, \mathbf{p} \sim \text{mult}(1, (q_{t1}, \dots, q_{tK})), \quad (4.17)$$

em que \mathbf{Y} tem distribuição binomial. Assim, reescrevemos a equação (3.38) da seguinte forma

$$q_{tk} = \frac{p_k \binom{n_t}{y_t} \theta_k^{y_t} (1 - \theta_k)^{(n_t - y_t)}}{\sum_{j=1}^K p_j \binom{n_t}{y_t} \theta_j^{y_t} (1 - \theta_j)^{(n_t - y_t)}} \quad (4.18)$$

A sequência dos valores estimados da variável não observável é

$$s_t = \max_k \text{Pr}(S_t = k | Y_t = y_t, \boldsymbol{\theta}, \mathbf{p}) \quad (4.19)$$

em que $t = 1, \dots, T$ e $k = 1, \dots, K$.

Como as distribuições condicionais são conhecidas, utilizamos o algoritmo *Gibbs sampling* para gerar as amostras dos parâmetros e fazer a estimação dos mesmos.

Algoritmo

1. Faça $i = 0$ e atribua valores iniciais aos parâmetros $\boldsymbol{\theta}^{(i)} = (\theta_1^{(i)}, \dots, \theta_K^{(i)})$, $\mathbf{p}^{(i)} = (p_1^{(i)}, \dots, p_K^{(i)})$ e
2. Construa a sequência $s_t^{(i+1)}$

$$s_t^{(i+1)} = \max_k Pr(S_t = k | Y_t = y_t, \boldsymbol{\theta}^{(i)}, \mathbf{p}^{(i)}) \quad (4.20)$$

em que $k = 1, \dots, K$ e $t = 1, \dots, T$.

3. Gere valores das distribuições condicionais

$$p^{(i+1)} | \boldsymbol{\tau}, \mathbf{s} \sim Dirichlet(\tau_1 + \sum_{t=1}^T I_{s_t}(k)^{(i+1)}, \dots, \tau_K + \sum_{t=1}^T I_{s_t}(k)^{(i+1)}), \quad (4.21)$$

em que $\tau_k > 0$ são conhecidos para $k = 1, \dots, K$,

$$\theta_k^{(i+1)} | \mathbf{a}, \mathbf{b}, \mathbf{y}, \mathbf{n}, \mathbf{s} \sim beta(a_k + \sum_{t=1}^T y_t I_{s_t}(k)^{(i+1)}; b_k + \sum_{t=1}^T I_{s_t}(k)^{(i+1)}(n_t - y_t)) \quad (4.22)$$

em que a_k, b_k são conhecidos e positivos.

4. Faça $i = i + 1$ e repita os passos 2 e 3 um número grande de vezes até que haja convergência. Descarte os valores iniciais.

4.3 Abordagem Bayesiana para MMM(1)

Considere que cada variável aleatória Y_t tem distribuição binomial, $Y_t | S_t = k \sim bin(n_t, \theta_k)$ com $t = 1, \dots, T$, $k = 1, \dots, K$.

A função de verossimilhança para \mathbf{P} , \mathbf{p}_0 e $\boldsymbol{\theta}$ é

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \mathbf{P}, \mathbf{p}_0 | \mathbf{y}, \mathbf{n}, \mathbf{s}) &= \prod_{k=1}^K \left\{ p_{0k}^{m_{0k}} \left[\prod_{j=1}^K p_{jk}^{m_{jk}} \right] \left[\theta_k^{\sum_{t=1}^T I_{s_t}(k) y_t} (1 - \theta_k)^{\sum_{t=1}^T I_{s_t}(k) (n_t - y_t)} \right] \right. \\ &\quad \left. \times \left[\prod_{t=1}^T \binom{n_t}{y_t}^{I_{s_t}(k)} I_{\{0, \dots, n_t\}}(y_t)^{I_{s_t}(k)} \right] \right\} \quad (4.23) \end{aligned}$$

em que $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ com $\theta_k \in (0, 1)$, $\mathbf{n} = (n_1, \dots, n_T)$ com $n_t \in N - \{0\}$, $\mathbf{P} = \{p_{jk}\}$, para $j = 1, \dots, K$ e $k = 1, \dots, K$, com $\sum_{k=1}^K p_{jk} = 1$ e $\mathbf{p}_0 = (p_{01}, \dots, p_{0K})$ em que $\sum_{k=1}^K p_{0k} = 1$.

A distribuição *a priori* para θ_k é uma distribuição beta com parâmetros (α_k, β_k) dada por

$$\pi(\theta_k | \alpha_k, \beta_k) = \frac{\Gamma(\alpha_k)\Gamma(\beta_k)}{\Gamma(\alpha_k + \beta_k)} \theta_k^{\alpha_k-1} (1 - \theta_k)^{\beta_k-1} I_{(0,1)}(\theta_k) \quad (4.24)$$

em que $\alpha_k, \beta_k > 0$ para $k = 1, \dots, K$. A distribuição *a priori* conjunta para $\boldsymbol{\theta}$ supondo $\theta_1, \dots, \theta_K$ independentes, é dada por

$$\pi(\boldsymbol{\theta} | \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{k=1}^K \frac{\Gamma(\alpha_k)\Gamma(\beta_k)}{\Gamma(\alpha_k + \beta_k)} \theta_k^{\alpha_k-1} (1 - \theta_k)^{\beta_k-1} I_{(0,1)}(\theta_k) \quad (4.25)$$

em que $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$ e $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)$.

A distribuição *a priori* para \mathbf{p}_0 é uma distribuição Dirichlet com parâmetros $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$ dada por

$$\pi(\mathbf{p}_0 | \boldsymbol{\lambda}) = \frac{\Gamma(\sum_{k=1}^K \lambda_k)}{\Gamma(\lambda_1) \dots \Gamma(\lambda_K)} \prod_{k=1}^K p_{0k}^{\lambda_k-1} I_{\{C_k\}}(p_{01}, \dots, p_{0K}) \quad (4.26)$$

em que $C_K = \left\{ \mathbf{p}_0 = (p_{01}, \dots, p_{0K}) : 0 < p_{0k} < 1, k = 1, \dots, K, \sum_{k=1}^K p_{0k} = 1 \right\}$ é o simplex, $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$, $\lambda_k > 0$ e conhecido com $k = 1, \dots, K$.

A distribuição *a priori* para cada linha da matriz de transição, \mathbf{p}_j , é uma distribuição Dirichlet com parâmetros $\boldsymbol{\gamma}_j = (\gamma_{j1}, \dots, \gamma_{jK})$ dada

$$\pi(\mathbf{p}_j | \boldsymbol{\gamma}_j) = \frac{\Gamma(\sum_{k=1}^K \gamma_{jk})}{\Gamma(\gamma_{j1}) \dots \Gamma(\gamma_{jK})} \prod_{k=1}^K p_{jk}^{\gamma_{jk}-1} I_{\{C_{jK}\}}(p_{j1}, \dots, p_{jK}) \quad (4.27)$$

em que o simplex é $C_{jK} = \left\{ \mathbf{p}_j = (p_{j1}, \dots, p_{jK}) : 0 < p_{jk} < 1, k = 1, \dots, K, \sum_{k=1}^K p_{jk} = 1 \right\}$, $\boldsymbol{\gamma}_j = (\gamma_{j1}, \dots, \gamma_{jK})$ com $\gamma_{jk} > 0$ e conhecido para $j, k = 1, \dots, K$.

A distribuição *a priori* conjunta para a matriz de transição, \mathbf{P} , é estabelecida de forma independente para cada linha de \mathbf{P} , $\mathbf{p}_j \sim Dir(\gamma_{j1}, \dots, \gamma_{jK})$, assim temos

$$\begin{aligned} \pi(\mathbf{P} | \boldsymbol{\gamma}) &= \prod_{j=1}^K \pi(\mathbf{p}_j | \boldsymbol{\gamma}_j) \\ &= \prod_{j=1}^K \frac{\Gamma(\sum_{k=1}^K \gamma_{jk})}{\Gamma(\gamma_{j1}) \dots \Gamma(\gamma_{jK})} \prod_{k=1}^K p_{jk}^{\gamma_{jk}-1} I_{\{C_{jK}\}}(p_{j1}, \dots, p_{jK}). \end{aligned} \quad (4.28)$$

A distribuição *a posteriori* conjunta para $(\mathbf{p}_0, \mathbf{P}, \boldsymbol{\theta})$ é obtida multiplicando a função de verossimilhança (4.23) pelas distribuições *a priori* (4.25) (4.26) e (4.28),

$$\begin{aligned}
\pi(\boldsymbol{\theta}, \mathbf{P}, \mathbf{p}_0 | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \mathbf{y}, \mathbf{n}, \mathbf{s}) &\propto \left[\prod_{k=1}^K p_{0k}^{m_{0k}} \right] \left[\prod_{k=1}^K \prod_{j=1}^K p_{jk}^{m_{jk}} \right] \\
&\times \left[\prod_{k=1}^K \theta_k^{\sum_{t=1}^T I_{st}(k) y_t} (1 - \theta_k)^{\sum_{t=1}^T I_{st}(k) (n_t - y_t)} \right] \\
&\times \left[\prod_{t=1}^T \binom{n_t}{y_t}^{I_{st}(k)} I_{\{0, \dots, n_t\}}(y_t)^{I_{st}(k)} \right] \\
&\times \left[\prod_{k=1}^K \frac{\Gamma(\alpha_k) \Gamma(\beta_k)}{\Gamma(\alpha_k + \beta_k)} \theta_k^{\alpha_k - 1} (1 - \theta_k)^{\beta_k - 1} I_{(0,1)}(\theta_k) \right] \\
&\times \left[\prod_{k=1}^K p_{0k}^{\lambda_k - 1} \frac{\Gamma(\sum_{k=1}^K \lambda_k)}{\Gamma(\lambda_1) \dots \Gamma(\lambda_K)} I_{\{C_k\}}(p_{01}, \dots, p_{0K}) \right] \\
&\times \left[\prod_{j=1}^K \frac{\Gamma(\sum_{k=1}^K \gamma_{jk})}{\Gamma(\gamma_{j1}) \dots \Gamma(\gamma_{jK})} \prod_{k=1}^K p_{jk}^{\gamma_{jk} - 1} I_{\{C_{jK}\}}(p_{j1}, \dots, p_{jK}) \right] \\
&\propto \prod_{k=1}^K \left\{ \left[p_{0k}^{m_{0k} + \lambda_k - 1} \right] I_{\{C_k\}}(p_{01}, \dots, p_{0K}) \right. \\
&\times \left[\prod_{j=1}^K p_{jk}^{m_{jk} + \gamma_{jk} - 1} \right] \left[\prod_{j=1}^K I_{\{C_{jK}\}}(p_{j1}, \dots, p_{jK}) \right] \\
&\times \left[\theta_k^{\sum_{t=1}^T I_{st}(k) y_t + \alpha_k - 1} (1 - \theta_k)^{\sum_{t=1}^T I_{st}(k) (n_t - y_t) + \beta_k - 1} \right] I_{(0,1)}(\theta_k) \\
&\times \left. \left[\prod_{t=1}^T I_{\{0, \dots, n_t\}}(y_t)^{I_{st}(k)} \right] \right\}, \tag{4.29}
\end{aligned}$$

em que $\lambda_k > 0$, $\gamma_{jk} > 0$, $\alpha_k, \beta_k > 0$, conhecidos, com $j, k = 1, \dots, K$.

Temos que os parâmetros \mathbf{p}_0 e \mathbf{P} associado a \mathbf{s} são independentes do parâmetro $\boldsymbol{\theta}$ associado a \mathbf{y} . Os estimadores bayesianos dos parâmetros são obtidos através do método MCMC usando as distribuições condicionais. As distribuições condicionais conhecidas para os parâmetros são dadas por

$$\theta_k | \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{y}, \mathbf{n}, \mathbf{s} \sim \text{beta}(\alpha_k + \sum_{t=1}^T I_{st}(k) y_t; \beta_k + \sum_{t=1}^T I_{st}(k) (n_t - y_t)) \tag{4.30}$$

em que $k = 1, \dots, K$,

$$\mathbf{p}_0 | \boldsymbol{\lambda}, \mathbf{y}, \mathbf{s} \sim \text{Dirichlet}(\lambda_1 + m_{01}, \dots, \lambda_K + m_{0K}), \tag{4.31}$$

$$\mathbf{p}_j | \boldsymbol{\gamma}_j, \mathbf{y}, \mathbf{s} \sim \text{Dirichlet}(\gamma_{j1} + m_{j1}, \dots, \gamma_{jK} + m_{jK}) \tag{4.32}$$

em que $j = 1, \dots, K$.

Considere a probabilidade condicional de \mathbf{S} dado \mathbf{Y} para $t = 1$ temos

$$S_1|Y_1 = y_1, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P} \sim \text{mult}(1, (q_{11}, \dots, q_{1K})) \quad (4.33)$$

em que

$$\begin{aligned} q_{1k} &= Pr(S_1 = k|Y_1 = y_1, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}) \\ &= \frac{Pr(S_1 = k, Y_1 = y_1|\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}{Pr(Y_1 = y_1, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})} \\ &= \frac{Pr(S_1 = k, Y_1 = y_1|\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}{\sum_{l=1}^K Pr(Y_1 = y_1, S_1 = l, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})} \\ &= \frac{Pr(S_1 = k|\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})P(Y_1 = y_1|S_1 = k, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}{\sum_{l=1}^K Pr(S_1 = l|\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})P(Y_1 = y_1|S_1 = l, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})} \\ &= \frac{p_{0k} \binom{n_1}{y_1} \theta_k^{y_1} (1 - \theta_k)^{(n_1 - y_1)}}{\sum_{l=1}^K p_{0l} \binom{n_1}{y_1} \theta_l^{y_1} (1 - \theta_l)^{(n_1 - y_1)}} \end{aligned} \quad (4.34)$$

com $k = 1, \dots, K$, para j conhecido e $t = 2, \dots, T$ temos

$$S_t|Y_t = y_t, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P} \sim \text{mult}(1, (q_{tj1}, \dots, q_{tjK})) \quad (4.35)$$

$$\begin{aligned} q_{tjk} &= Pr(S_t = k|Y_t = y_t, S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}) \\ &= \frac{Pr(Y_t = y_t, S_t = k, S_{t-1} = j|\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}{Pr(Y_t = y_t, S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})} \\ &= \frac{Pr(Y_t = y_t, S_t = k, S_{t-1} = j|\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}{\sum_{l=1}^K Pr(Y_t = y_t, S_t = l, S_{t-1} = j|\boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})} \\ &= \frac{Pr(S_t = k|S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})Pr(Y_t = y_t|S_t = k, S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})}{\sum_{l=1}^K Pr(S_t = l|S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})Pr(Y_t = y_t|S_t = l, S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P})} \\ &= \frac{p_{jk} \binom{n_t}{y_t} \theta_k^{y_t} (1 - \theta_k)^{(n_t - y_t)}}{\sum_{l=1}^K p_{jl} \binom{n_t}{y_t} \theta_l^{y_t} (1 - \theta_l)^{(n_t - y_t)}} \end{aligned} \quad (4.36)$$

em que $j, k = 1, \dots, K$.

A sequência estimada de s é dada por

$$s_1 = \max_k Pr(S_1 = k | Y_1 = y_1, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}) \quad (4.37)$$

$$s_t = \max_k Pr(S_t = k | Y_t = y_t, S_{t-1} = j, \boldsymbol{\theta}, \mathbf{p}_0, \mathbf{P}) \quad (4.38)$$

em que $j, k = 1, \dots, K$ e $t = 2, \dots, T$. Considerando que o número da componentes de mistura, K , é conhecido, a estimação dos parâmetros é realizada via algoritmo *Gibbs sampling*.

Algoritmo MCMC

1. Faça $i = 0$ e atribua valores iniciais para os parâmetros $\boldsymbol{\theta}^{(i)} = (\theta_1^{(i)}, \dots, \theta_K^{(i)})$, $\mathbf{p}_0^{(i)} = (p_{01}^{(i)}, \dots, p_{0K}^{(i)})$ e

$$\mathbf{P}^{(i)} = \begin{bmatrix} p_{11}^{(i)} & \cdots & p_{1K}^{(i)} \\ p_{21}^{(i)} & \cdots & p_{2K}^{(i)} \\ \vdots & & \\ p_{K1}^{(i)} & \cdots & p_{KK}^{(i)} \end{bmatrix},$$

2. Construa a sequência $S^{(i+1)}$ com base em \mathbf{y} , $\boldsymbol{\theta}^{(i)}$, $\mathbf{p}_0^{(i)}$, $\mathbf{P}^{(i)}$, usando

$$s_1^{(i+1)} = \max_k Pr(S_1 = k | Y_1 = y_1, \boldsymbol{\theta}^{(i)}, \mathbf{p}_0^{(i)}, \mathbf{P}^{(i)}) \quad (4.39)$$

$$s_t^{(i+1)} = \max_k Pr(S_t = k | Y_t = y_t, s_{t-1} = j, \boldsymbol{\theta}^{(i)}, \mathbf{p}_0^{(i)}, \mathbf{P}^{(i)}) \quad (4.40)$$

3. Gere novos valores dos parâmetros das distribuições condicionais

$$\mathbf{p}_0^{(i+1)} | \boldsymbol{\lambda}, \mathbf{s} \sim Dirichlet(\lambda_1 + m_{01}^{(i+1)}, \dots, \lambda_K + m_{0K}^{(i+1)}), \quad (4.41)$$

em que $\lambda_k > 0$ e conhecido com $k = 1, \dots, K$,

$$\mathbf{p}_j^{(i+1)} | \boldsymbol{\gamma}_j, \mathbf{s} \sim Dirichlet(\gamma_{j1} + m_{j1}^{(i+1)}, \dots, \gamma_{jK} + m_{jK}^{(i+1)}) \quad (4.42)$$

em que $j, k = 1, \dots, K$ e $\gamma_{jk} > 0$ e conhecido,

$$\theta_k^{(i+1)} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{y}, \mathbf{n}, \mathbf{s} \sim beta(\alpha_k + \sum_{t=1}^T y_t I_{s_t}(k)^{(i+1)}; \beta_k + \sum_{t=1}^T I_{s_t}(k)^{(i+1)}(n_t - y_t)) \quad (4.43)$$

para $\alpha_k, \beta_k > 0$ conhecidos e $k = 1, \dots, K$.

4. Faça $i = i + 1$ e repita os passos 2 e 3 um número grande de vezes até atingir a convergência. Descarte os valores iniciais.

A não identificabilidade dos rótulos que identificam as componentes é um problema comum aos modelos de mistura devido a permutabilidade que pode ser visto em Stephens (2000). A falta de identificabilidade afeta a estimação dos parâmetros do modelo, esse problema pode ser resolvido estabelecendo uma regra para fixação dos rótulos, um exemplo: é estabelecer que as médias menores tenham os menores índices.

Uma outra questão que pode ser levantada é que o número de componentes K geralmente é desconhecido e os métodos tratados até aqui o consideram conhecido. Aprofundar esse aspecto da modelagem está fora do escopo dessa pesquisa, por essa razão optamos por uma solução simples que é utilizar metodologias para seleção de modelos para estimar o número de componentes de mistura K . Assim, utilizamos o AIC e BIC com os estimadores de máxima verossimilhança e o Fator de Bayes através em conjunto com do método MCMC para os estimadores bayesianos. Uma alternativa interessante é a utilização de modelagens transdimensionais, tais como o *reversible jump*, que é deixada para algum momento no futuro.

Capítulo 5

Aplicação de Mistura de Binomiais

Apresentamos, a seguir, a aplicação dos modelos de mistura de binomiais tratados nos capítulos anteriores a um conjunto de dados reais e a dados simulados.

5.1 Dados Vinnie Johnson

Para ilustrar a aplicação dos métodos para ajuste de modelos de mistura, utilizamos os dados referentes ao aproveitamento do jogador de basquetebol Vinnie Johnson no período de 1985 a 1989, com um total de 380 observações. Nesse caso, o interesse é acompanhar o aproveitamento do jogador ao longo de sua carreira. O aproveitamento corresponde à proporção esperada de arremessos convertidos pelo jogador. Caso a modelagem fosse bem sucedida, isso permitiria traçar um perfil da carreira e, eventualmente verificar a existência (ou não) de fases boas (ou ruins). Como detalhado no decorrer do capítulo, o ajuste do modelo MMM(1) a esse conjunto de dados não foi bem sucedido, o que nos ofereceu uma excelente oportunidade para fazer o diagnóstico do ocorrido, usando para isso a ferramenta simulação.

Os dados estão disponíveis em <http://lib.stat.cmu.edu/datasets/Vinnie.Johnson>. Esse conjunto de dados foi previamente analisado por Kass & Raftery (1995) para verificar se o jogador de basquetebol Vinnie Johnson tem mão quente em relação aos acertos, usando o fator de Bayes e distribuição binomial chegou à conclusão que ele não tem mão quente. Missão (2007) analisa esse conjunto de dados com o objetivo de verificar se a

probabilidade de acertos é constante ou se existe variação de jogo para jogo, utilizou uma abordagem bayesiana e o fator de Bayes chegando à mesma conclusão de Kass & Raftery (1995).

Na Figura (5.1) apresentamos a proporção do número de acertos em cada jogo do jogador Vinnie Johnson ao longo de sua carreira.

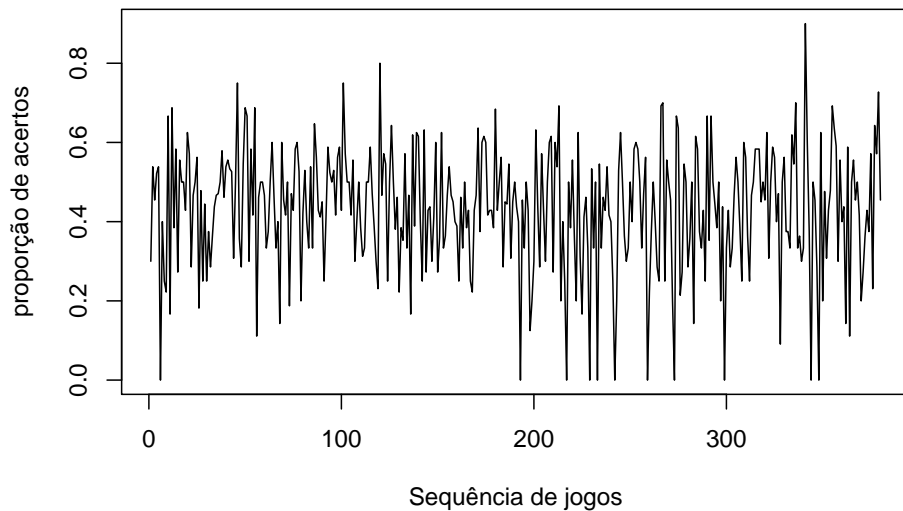


FIGURA 5.1: Sequência dos jogos de basquete.

A Figura (5.2) apresenta a densidade estimada da proporção de acertos nos jogos feita pelo método de estimação de densidades. Nela podemos observar a forma da distribuição dos dados. Nesse caso a estimação da função densidade de probabilidade estimada é feita utilizando a rotina, disponível no software R, através de um estimador tipo *kernel* a instrução *density(.)*. Para mais detalhes sobre o estimador tipo *kernel* veja (Silverman, 1986).

5.1.1 Estimação dos parâmetros dos modelos via Verossimilhança

Ajustamos o modelo de mistura de distribuições binomiais em que a variável aleatória não observável é dependente e para simplificar a notação das tabelas chamamos os modelos estimados de M_k , $k = 1, \dots, 6$.

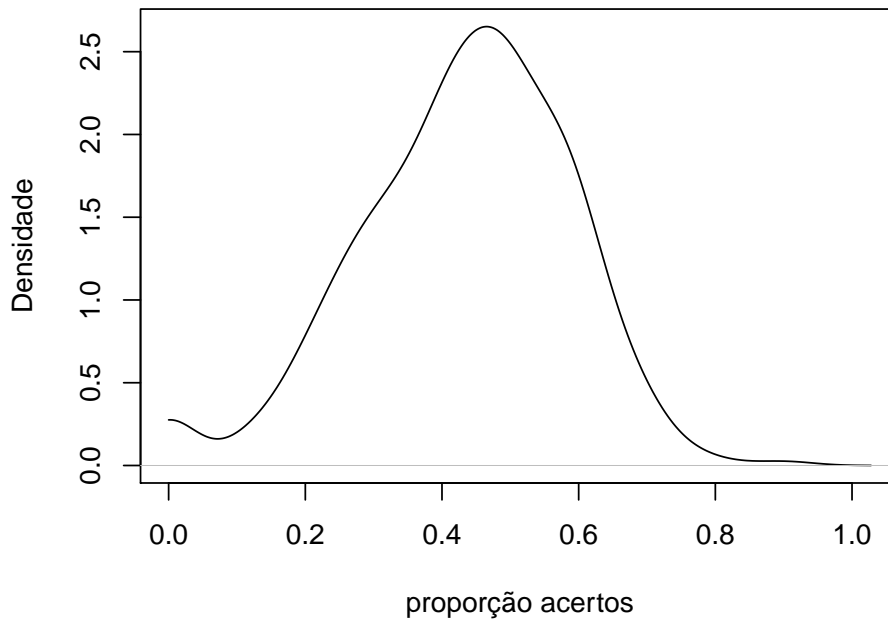


FIGURA 5.2: Gráfico da densidade estimada da proporção de acertos nos jogos

Na Tabela 5.1, apresentamos os valores dos parâmetros obtidos via máxima verossimilhança, os modelos estimados juntamente com os critérios de seleção de modelos AIC e BIC do modelo de mistura de distribuições binomiais considerando que a variável não observável tem dependência markoviana de primeira ordem.

Na tabela, observamos que o modelo M_5 , que é uma mistura de cinco distribuições binomiais, é o modelo que apresentou menor AIC e BIC, indicando que este é o modelo que melhor se ajusta aos dados. A matriz de transição estimada para a variável não observável para os modelos de mistura de distribuições binomiais para $k = 2$ é \hat{P}_2 e, assim, sucessivamente.

$$\hat{P}_2 = \begin{bmatrix} 0,5129534 & 0,4870466 \\ 0,5107527 & 0,4892473 \end{bmatrix},$$

$$\hat{P}_3 = \begin{bmatrix} 0,3303571 & 0,2857143 & 0,3839286 \\ 0,3047619 & 0,2000000 & 0,4952381 \\ 0,2654321 & 0,3148148 & 0,4197531 \end{bmatrix},$$

TABELA 5.1: Valores dos parâmetros dos modelos e critérios de seleção de modelo

Modelos	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	AIC	BIC
M_1	0,4566						6627,910	6630,850
M_2	0,5473	0,3308					6602,316	6767,803
M_3	0,5981	0,2549	0,4421				6449,8879	6457,759
M_4	0,6141	0,3742	0,2105	0,4876			6405,662	6452,994
M_5	0,4392	0,5271	0,3347	0,1840	0,6321		6374,123	6452,926
M_6	0,3869	0,6320	0,4513	0,1292	0,2899	0,5271	6421,8844	6587,371

$$\hat{P}_4 = \begin{bmatrix} 0,2666667 & 0,3000000 & 0,1777778 & 0,2555556 \\ 0,2162162 & 0,2792793 & 0,1981982 & 0,3063063 \\ 0,2753623 & 0,2463768 & 0,1739130 & 0,3043478 \\ 0,2110092 & 0,3211009 & 0,1743119 & 0,2935780 \end{bmatrix},$$

$$\hat{P}_5 = \begin{bmatrix} 0,2417582 & 0,2307692 & 0,1868132 & 0,2197802 & 0,1208791 \\ 0,2643678 & 0,2413793 & 0,2298851 & 0,1149425 & 0,1494253 \\ 0,2345679 & 0,2716049 & 0,2469136 & 0,1111111 & 0,1358025 \\ 0,2777778 & 0,1666667 & 0,1296296 & 0,1481481 & 0,2777778 \\ 0,1969697 & 0,2121212 & 0,2424242 & 0,1060606 & 0,2424242 \end{bmatrix},$$

$$\hat{P}_6 = \begin{bmatrix} 0,07017544 & 0,1754386 & 0,1754386 & 0,08771930 & 0,2982456 & 0,1929825 \\ 0,15151515 & 0,2424242 & 0,1363636 & 0,07575758 & 0,1818182 & 0,2121212 \\ 0,12500000 & 0,0937500 & 0,1562500 & 0,09375000 & 0,2656250 & 0,2656250 \\ 0,09375000 & 0,2500000 & 0,2187500 & 0,09375000 & 0,1250000 & 0,2187500 \\ 0,19178082 & 0,1780822 & 0,2054795 & 0,06849315 & 0,1232877 & 0,2328767 \\ 0,20689655 & 0,1494253 & 0,1609195 & 0,09195402 & 0,1494253 & 0,2413793 \end{bmatrix},$$

Ajustamos o modelo de mistura de distribuições binomiais para os dados considerando independência na variável não observável. Os estimadores dos parâmetros das distribuições binomiais juntamente com os critérios de seleção de modelos encontram-se na Tabela 5.2. Observamos na tabela que ao considerarmos independência entre as variáveis

TABELA 5.2: Modelos de mistura de binomiais estimados

Modelos	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	AIC	BIC
M_1	0,4566						6627,910	6630,850
M_2	0,5473	0,3308					6945,698	6963,339
M_3	0,5992	0,2549	0,4429				7191,760	7235,862
M_4	0,6141	0,3742	0,2105	0,4876			7427,859	7510,184
M_5	0,4392	0,5271	0,3347	0,1840	0,6321		7633,484	7765,7926
M_6	0,4513	0,6320	0,3869	0,1292	0,2899	0,5271	7808,821	8002,872

aleatórias não observáveis, o modelo que melhor explica os dados é o M_1 , ou seja, não identificamos a presença de mistura.

Note, que nas Tabelas (5.1) e (5.2) os valores de θ são iguais devido estimadores de máxima verossimilhança serem iguais tanto no modelo de mistura quanto no MMM(1). Além disso, a seleção das distribuições binomiais é feita da mesma forma em ambos os modelos. O que diferencia um modelo do outro é o fato de \mathbf{S} ser independente ou dependente.

O modelo ajustado, o MMM(1), tal qual os modelos de mistura em geral, apresenta a falta de identificabilidade de seus parâmetros nos valores assumidos pela variável controladora da mistura, ou seja, é possível um rearranjo dos rótulos dos elementos da sequência sem alterar os critérios de ajuste do modelo.

5.1.2 Abordagem Bayesiana

Nessa seção, aplicamos a metodologia desenvolvida no capítulo anterior, Seções 4.2 e 4.3, para modelar os dados referentes ao aproveitamento do jogador de basquete Vinnie Johnson já descritos no início desse capítulo.

Consideramos as distribuições *a priori* beta(1,1) para os parâmetros θ_k com $k = 1, \dots, K$, e para p_0 e P distribuições *a priori* Dirichlet com todos os parâmetros iguais a 1. No caso em que S é independente utilizamos os mesmos valores dos parâmetros.

Na aplicação do MCMC para o ajuste do modelo, foram gerados 101000 valores, descartamos os 1000 primeiros valores de cada sequência e, usando um salto de 100,

selecionamos uma amostra de tamanho 1000 para a estimação dos parâmetros.

Inicialmente ajustamos o caso em que a variável aleatória não observável S é independente. Chamamos de M_k , com $k = 1, \dots, K$, o modelo com K componentes de mistura. Também consideramos o modelo de mistura com dependência markoviana na variável não observável para os dados do jogador.

Na Tabela 5.3, apresentamos o modelo de mistura para uma até cinco componentes de mistura, M_1 a M_5 . Para cada um dos modelos apresentamos as médias e os intervalos de credibilidade (IC), dados pelos quantis 0,025 e 0,975 dos valores simulados. A proporção de acertos estimada está próxima de 0,45.

TABELA 5.3: Resumos dos parâmetro θ dos modelos de mistura

Parâmetros	Independente		Dependente	
	Média	IC(95%)	Média	IC(95%)
M_1 θ	0,4565	(0,4425;0,4704)		
M_2	θ_1	0,4551	(0,3873;0,5101)	0,4371 (0,1579;0,6944)
	θ_2	0,4535	(0,3811;0,5073)	0,4491 (0,2186;0,6681)
M_3	θ_1	0,4552	(0,3669;0,5284)	0,4480 (0,2033;0,6363)
	θ_2	0,4533	(0,3571;0,5217)	0,4462 (0,2199;0,6507)
	θ_3	0,4512	(0,0315;0,7312)	0,4484 (0,2349;0,5754)
M_4	θ_1	0,4517	(0,3398;0,5402)	0,4509 (0,2623;0,5988)
	θ_2	0,4539	(0,3509;0,5402)	0,4494 (0,2528;0,5853)
	θ_3	0,4513	(0,3123;0,5301)	0,4486 (0,2692;0,5801)
	θ_4	0,4507	(0,3392;0,5303)	0,4493 (0,2262;0,5880)
M_5	θ_1	0,4492	(0,3119;0,5503)	0,4494 (0,2761;0,5707)
	θ_2	0,4508	(0,3086;0,5483)	0,4478 (0,2724;0,5528)
	θ_3	0,4499	(0,3001;0,5443)	0,4530 (0,3187;0,5548)
	θ_4	0,4466	(0,3007;0,5494)	0,4506 (0,3102;0,5517)
	θ_5	0,4509	(0,3062;0,5499)	0,4487 (0,0029;0,8279)

Na Tabela (5.4), apresentamos o fator de Bayes para o modelo de mistura para o caso em que as variáveis não observáveis são independentes e para o caso de dependência markoviana de primeira ordem.

No caso do modelo de mistura com variável não observável independente temos

que os valores do fator de Bayes ficaram todos próximos de um, assim, podemos considerar o modelo 1, M_1 , que corresponde ao modelo sem mistura e chegamos a mesma conclusão que em Missão (2007).

Para o modelo de mistura com dependência markoviana de primeira ordem temos que o fator de Bayes maior foi $F_{15} = 26,80553$ o que significa que o modelo com 5 componentes, M_5 , se ajusta melhor aos dados em relação ao modelo com uma componente, M_1 .

TABELA 5.4: Fatores de Bayes para a comparação de modelo.

Fator Bayes	Independente	Dependente
F_{12}	1,0025	4,547348
F_{13}	1,1684	9,931581
F_{14}	1,1994	17,38471
F_{15}	1,2282	26,80553
F_{23}	1,1656	2,184038
F_{24}	1,1964	3,823044
F_{25}	1,2251	5,894761
F_{34}	1,0265	1,750448
F_{35}	1,0511	2,69902
F_{45}	1,0240	1,541902

As matrizes de transição das cadeias de Markov estimadas, para os modelos M_2 a M_5 , são

$$\hat{P}_2 = \begin{bmatrix} 0,6133 & 0,3867 \\ 0,3193 & 0,6807 \end{bmatrix},$$

$$\hat{P}_3 = \begin{bmatrix} 0,4263 & 0,2862 & 0,2875 \\ 0,3011 & 0,3860 & 0,3129 \\ 0,2908 & 0,2634 & 0,4458 \end{bmatrix},$$

$$\hat{P}_4 = \begin{bmatrix} 0,2846 & 0,2454 & 0,2337 & 0,2363 \\ 0,2458 & 0,2952 & 0,2189 & 0,2401 \\ 0,2230 & 0,2164 & 0,3058 & 0,2548 \\ 0,2319 & 0,2214 & 0,2499 & 0,2968 \end{bmatrix},$$

$$\hat{P}_5 = \begin{bmatrix} 0,2277 & 0,1859 & 0,2039 & 0,2013 & 0,1812 \\ 0,1935 & 0,2364 & 0,1942 & 0,1878 & 0,1881 \\ 0,1899 & 0,1785 & 0,2273 & 0,2142 & 0,1900 \\ 0,1918 & 0,1823 & 0,2201 & 0,2272 & 0,1786 \\ 0,1907 & 0,1826 & 0,1955 & 0,1951 & 0,2361 \end{bmatrix},$$

respectivamente.

O fator de Bayes, Tabela 5.4, indica que o modelo com cinco componentes, o M_5 , é o que melhor se ajusta aos dados. Poderíamos encerrar a análise aqui, mas vamos dar um passo adiante e realizar uma análise crítica do modelo ajustado.

Numa análise crítica dos modelos ajustados, observamos através do fator de Bayes, Tabela 5.4, que o modelo com cinco componentes, o M_5 , é o que melhor se ajusta aos dados. Entretanto, quando observamos os valores dos θ_k 's na Tabela 5.3, não conseguimos identificar claramente tais componentes já que todas as estimativas dos θ_k 's ficam em torno de 0,45 (São muito similares!) não caracterizando componentes diferentes.

A hipótese de que os valores dos θ_k 's são realmente similares não deve ser descartada, mas vamos procurar fazer um diagnóstico mais apurado da questão.

Apresentamos, na Figura 5.3, um histograma que mostra as frequências de ocorrência dos valores n_t , $t = 1, \dots, T$. Como podemos observar, os n_t 's estão abaixo de 25 em sua maioria, 352 destes valores são menores que 20. Isso poderia explicar em parte a dificuldade em estimar os valores dos parâmetros do modelo tratada acima, casos esses n_t 's se enquadrem no que seriam amostras pequenas para a estimação pretendida. Para o conjunto de dados temos uma distribuição $bin(n_t, \theta_k)$ com o número de arremessos, n_t 's. Através de um estudo de simulação poderemos compreender melhor o que está acontecendo com modelo.

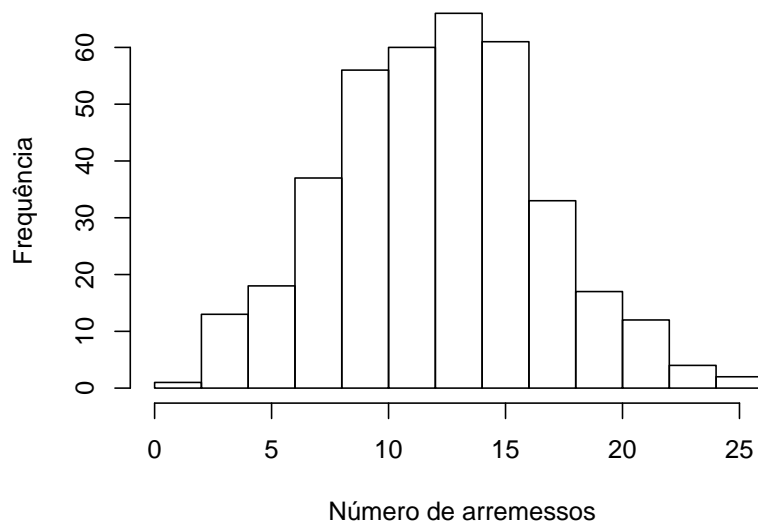


FIGURA 5.3: Número de Arremessos em cada jogo de basquetebol.

5.2 Estudo de Simulação

5.2.1 Abordagem via Verossimilhança para MMM(1) usando algoritmo EM adaptado

Realizamos um estudo de simulação utilizando uma amostra de tamanho 1000 para o modelo de mistura de distribuições binomiais com dependência markoviana na variável não observável considerando diferentes os valores de $n_t \in \{10, 20, 30, 50, 100\}$ e de $\theta = (\theta_1, \theta_2, \theta_3)$ para a mistura de distribuição binomiais, $\text{bin}(n_t, \theta_k)$ com $K = 1, 2, 3$. Fixamos inicialmente $\theta_2 = 0,5$ e variámos $\theta_1 = \theta_2 - \Delta$ e $\theta_3 = \theta_2 + \Delta$ com $\Delta \in \{0,05; 0,1; 0,2; 0,3; 0,4\}$ e construímos um caso com dois valores de θ próximos e um distante, ou seja, $\theta_1 = 0,1$, $\theta_2 = 0,45$ e $\theta_3 = 0,55$. Na tabela (5.5), observar que conforme os θ 's estão próximos temos que os valores das estimativas se aproximam do verdadeiro valor quando n cresce. Conforme os valores dos θ 's se afastam, podemos estimá-los com valores de n menores.

Na Figura (5.4), observamos que conforme os valores de θ_1 e θ_3 aproximam de $\theta_2 = 0,5$ a estimação dos parâmetros melhoram conforme aumenta os valores de n_t . No gráfico F da figura (5.4), podemos observar o que ocorre quando temos dois valores de θ 's

próximos e um distante dos demais, as estimativas aproximam da verdadeira conforme aumentamos o valor de n_t para os θ 's próximos.

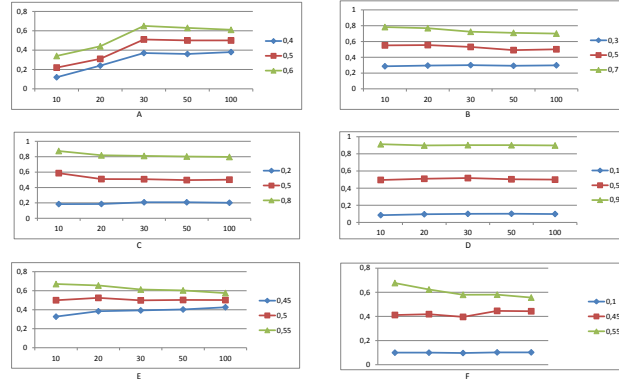


FIGURA 5.4: Valores estimados de θ por n_t

5.2.2 Abordagem bayesiana para MMM(1) usando MCMC

Realizamos um estudo de simulação fixando o número de componentes de mistura, $K = 3$, para observar a resposta do modelo em relação a variações em n_t , $t = 1, \dots, T$, em que a distribuição dos dados é binomial com parâmetros (n_t, θ_k) , $y_t \sim \text{bin}(n_t, \theta_k)$, $k = 1, \dots, K$. Escolhemos quatro valores de $n_t \in \{20, 30, 50, 100\}$ para ver o comportamento do modelo de mistura quando temos três componentes de mistura e fixamos uma distribuição binomial com probabilidade de sucesso igual a 0.5 e variamos as outras duas probabilidades. Para cada valor de n_t , variamos o valor de $\theta = (\theta_1, \theta_2, \theta_3)$ da seguinte forma: $\theta_2 = 0,5$, $\theta_1 = \theta_2 - \Delta$ e $\theta_3 = \theta_2 + \Delta$, com $\Delta \in \{0,1; 0,2; 0,3; 0,4\}$.

Para o ajuste do modelo de mistura considerando $K = 3$ componentes, simulamos 11000 iterações do MCMC, descartamos os 1000 valores iniciais e, usando um salto de tamanho 20, obtemos uma sequência de 500 valores para cada parâmetro e elemento de S .

Os resultados da simulação são apresentados na Tabela 5.6. Observamos que para $\Delta = 0,1$, somente com $n_t = 100$, $t = 1, \dots, T$, os valores estimados de θ_1 e θ_3 se aproximam dos verdadeiros. Para $\Delta = 0,2$, somente com $n_t \geq 30$ temos que todas as estimativas ficaram próximas de seus valores verdadeiros.

TABELA 5.5: Valores estimados para θ fixado e diferentes valores de n

valor real de θ	$n = 10$	$n = 20$	$n = 30$	$n = 50$	$n = 100$
0,4	0,124064	0,242718	0,377451	0,364859	0,388425
0,5	0,228771	0,313916	0,518039	0,506162	0,501714
0,6	0,345775	0,443366	0,659731	0,636265	0,610191
0,3	0,286223	0,295089	0,300738	0,292270	0,298622
0,5	0,550951	0,554948	0,531538	0,490867	0,501568
0,7	0,783175	0,768996	0,723860	0,709975	0,700739
0,2	0,184674	0,186013	0,208000	0,207963	0,201885
0,5	0,586595	0,509649	0,507835	0,496610	0,501649
0,8	0,874520	0,818261	0,810244	0,803006	0,797047
0,1	0,084776	0,095213	0,100104	0,101931	0,097993
0,5	0,495417	0,509239	0,517554	0,503441	0,499200
0,9	0,912736	0,898030	0,902099	0,901552	0,898309
0,45	0,327321	0,383901	0,392553	0,402678	0,426183
0,5	0,500000	0,524775	0,498527	0,502717	0,501452
0,55	0,670333	0,655282	0,612770	0,603601	0,575479
0,1	0,100000	0,100000	0,096726	0,102298	0,102384
0,45	0,411749	0,418683	0,395895	0,446000	0,442895
0,55	0,676632	0,622959	0,579072	0,580280	0,556318

Conforme aumentamos os valores de K , temos que a medida que a distância entre os valores de θ_k , $k = 1, \dots, K$, diminui isso aumenta a dificuldade de detecção das componentes envolvidas, o que é natural, e aumenta exigência (n_t maior) para que o modelo seja bem sucedido em identificar as componentes. O resultado se observarmos apenas os θ_k 's é que as estimativas se aproximam dos valores centrais, 0, 5. Observamos também que os intervalos de credibilidade aumentam a medida que os valores de θ_k 's estão próximos e os valores de $n_t \leq 30$.

TABELA 5.6: Estimativas e intervalos de credibilidade para diferentes θ e n .

	θ	Estimativas(IC(95%))	θ	Estimativas(IC(95%))
$n = 20$	0,4	0,4946(0,3812;0,6393)	0,3	0,3327(0,2680;0,5486)
	0,5	0,5035(0,3828;0,6416)	0,5	0,4934(0,2680;0,5486)
	0,6	0,5105(0,3792;0,6560)	0,7	0,6338(0,3335;0,7131)
	0,2	0,2012(0,1867;0,2157)	0,1	0,0989(0,0897;0,1085)
	0,5	0,5160(0,4777;0,5534)	0,5	0,4945(0,4723;0,5163)
	0,8	0,8149(0,7848;0,8243)	0,9	0,9005(0,8900;0,9114)
$n = 30$	0,4	0,4263(0,3548;0,5918)	0,3	0,3092(0,2930;0,5486)
	0,5	0,4920(0,3672;0,6113)	0,5	0,5246(0,4898;0,5582)
	0,6	0,5382(0,3703;0,6197)	0,7	0,7285(0,6995;0,7605)
	0,2	0,2061(0,1944;0,2182)	0,1	0,1054(0,0978;0,1133)
	0,5	0,4856(0,4615;0,5089)	0,5	0,5104(0,4961;0,5253)
	0,8	0,8076(0,7962;0,8195)	0,9	0,8031(0,8949;0,9109)
$n = 50$	0,4	0,4520(0,3931;0,6003)	0,3	0,3005(0,2879;0,3117)
	0,5	0,5080(0,4041;0,6073)	0,5	0,4805(0,4577;0,5042)
	0,6	0,5482(0,4108;0,6153)	0,7	0,7030(0,6889;0,7165)
	0,2	0,2025(0,1938;0,2115)	0,1	0,0982(0,0931;0,1034)
	0,5	0,4919(0,4801;0,5040)	0,5	0,5012(0,4879;0,5153)
	0,8	0,8028(0,7937;0,8117)	0,9	0,9012(0,8950;0,9076)
$n = 100$	0,4	0,4023(0,3780;0,4581)	0,3	0,2970(0,2903;0,3041)
	0,5	0,4908(0,4001;0,6016)	0,5	0,4997(0,4895;0,5094)
	0,6	0,5852(0,4620;0,6105)	0,7	0,7023(0,6951;0,7097)
	0,2	0,2012(0,1959;0,2069)	0,1	0,0992(0,0953;0,1032)
	0,5	0,4963(0,4886;0,5045)	0,5	0,4954(0,4874;0,5049)
	0,8	0,8015(0,7948;0,8082)	0,9	0,8964(0,8917;0,9009)

Na Figura (5.5), temos no gráfico A as estimativas de $\theta_2 = 0,5$ com $\Delta = 0,1$ no qual podemos observar que para $n_t = 20$ os valores ficam muito próximo de 0,5 com isso não conseguimos observar a diferença entre as estimativas, no entanto, quando os valores de n_t vai aumentando observamos a separação das estimativas. Conforme os valores de Δ aumenta as estimativas ficam melhores.

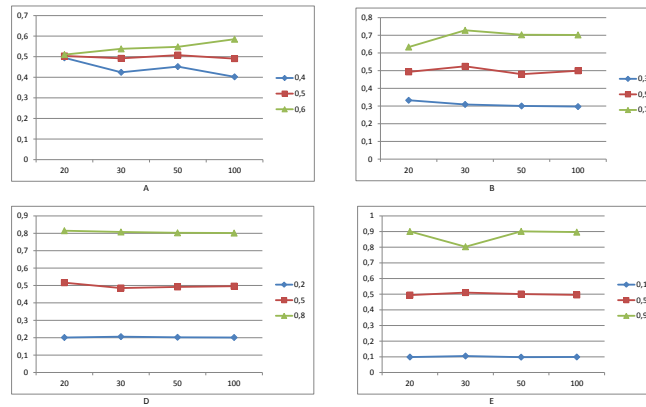


FIGURA 5.5: Valores estimados de θ por n_t .

Na Tabela 5.7, estimamos 1000 vezes o modelo para calcular a cobertura. Para cada uma das estimações consideramos 11000 iterações usando MCMC no qual descartamos os 1000 valores iniciais, com o restante selecionamos uma amostra de tamanho 500 usando um salto de 20. Considerando n_t fixo e diferentes valores de Δ , obtivemos as estimativas de θ : temos que para $\Delta \leq 0,2$, o modelo apresentou uma cobertura mínima de 774 e uma máxima de 995; para $\Delta \geq 0,3$, tivemos uma cobertura de 100%.

TABELA 5.7: Cobertura de $\theta = (\theta_1, \theta_2, \theta_3)$ para $K = 3$

Valor n	(0, 1; 0, 5, ; 0, 9)	(0, 2; 0, 5; 0, 8)	(0, 3; 0, 5; 0, 7)	(0, 4; 0, 5; 0, 6)
20	1000	1000	774	995
30	1000	1000	998	850
50	1000	1000	777	869
100	1000	1000	973	965

Os resultados obtidos na simulação indicam que a capacidade de estimar os parâmetros está associada aos valores das diferenças entre os θ_k 's. Quanto menor a diferença entre os θ_k 's maior a dificuldade para identificar as componentes envolvidas.

A dificuldade de convergência para valores próximos a 0.5, ou seja, nos casos em que $\Delta = \{0, 1; 0, 2\}$ pode ser explicada pela dificuldade do modelo em identificar as componentes a que pertence cada observação. Com os erros de classificação os valores das médias tendem a ir para o centro entre as médias das componentes reais.

Na tabela (5.7), observamos que para $\Delta = 0, 1$ e $\Delta = 0, 2$ é necessário novos estudos de simulação nesta região juntamente com proposta de solução para problema que será desenvolvido em trabalhos futuros.

Como podemos observar na Tabela 5.7, a probabilidade de cobertura dos intervalos de credibilidade também depende da distância entre os θ_k 's. Mas esse tema sai fora do escopo dessa monografia e somente será explorado com mais detalhes em trabalho futuro.

Capítulo 6

Discussão

Neste trabalho, propomos uma nova abordagem para modelos de mistura. Nessa nova abordagem, temos uma sequência de variáveis aleatórias não observáveis S_t que controla a mistura, no sentido que essa variável diz a que componente cada variável aleatória observável Y_t pertence, ou, equivalentemente, qual é sua distribuição condicional.

A forma com que estabelecemos a relação entre as variáveis aleatórias S_t estabelece o tipo de modelo de mistura obtido. Se as variáveis da sequência são independentes, temos o modelo de mistura tradicional, mas com uma abordagem probabilística em lugar de pesos e ponderações. Se S_t é uma cadeia de Markov, temos o modelo MMM(1), um modelo similar ao HMM. Essa nova apresentação possibilita a utilização adaptada do algoritmo EM. Na sua versão original o HMM necessita de algoritmos como *forward-backward*, Baum-Welch e Viterbi.

Métodos de seleção de modelos são utilizados para escolher entre os modelos propostos e também para selecionar o número de componentes de mistura. Na abordagem de máxima verossimilhança, usamos os critérios AIC e BIC e na abordagem bayesiana utilizamos o Fator de Bayes. Não é parte dos objetivos desse trabalho exaurir o tema seleção de modelo: os métodos escolhidos visam apenas a utilização de um método para este fim.

O modelo de mistura com dependência markoviana de primeira ordem foi apresentado para mistura de distribuição normal e binomial. Aplicamos o MMM(1) em um conjunto de dados reais em que a distribuição dos dados é binomial. No modelo

de mistura em que a variável aleatória não observável é independente, o melhor modelo, indicado pelos critérios AIC e BIC, é o modelo com uma componente ($K = 1$). Já no modelo com dependência de primeira ordem, o modelo mais adequado é o modelo com cinco componentes ($K = 5$).

Através de simulações determinamos limitações do método.

Como propostas futuras temos: o desenvolvimento futuro da pesquisa temos o aprofundamento do tema seleção de modelos, o que inclui a metodologia *reversible jump*, aplicada aos modelos descritos; determinação de incertezas na estimação de parâmetros estimados via máxima verossimilhança, possivelmente através de bootstrap; e alguns resultados teóricos com ou via simulação para comparação do EM.

Referências

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, pages 716–723.

Altman, R. M. (2007). Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *Journal of American Statistical Association*, **109**, 201–210.

Baum, L. E., Petrie, T., Soule, G. & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic function of Markov chain. *The Annals of mathematical Statistics*, **41**(1), 164–171.

Celeux, G. & Soromenho, G. (1999). An entropy criterion for assessing the number of clusters. *Journal of Classification*, (13), 195–212.

Chib, S. (1996). Calculating posterior distribution and modal estimates in Markov mixture models. *Journal Econometrics*, **75**, 79–97.

Choy, S. T. B. & Chan, J. S. K. (2008). Scale mixture distributions in statistical modelling. *Aust. N. Z. J. Stat.*, **50**(2), 135–146.

Coimbra, R. G. (2005). *Modelo de mistura de multinomiais aplicado à identificação de proteínas similares*. Dissertação de mestrado, Departamento de Estatística – Universidade Federal de São Carlos, São Carlos.

Dempster, A. P., Laird, N. M. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38.

Dias, J. G., Vermunt, J. K. & Ramos, S. (2010). *Advantage in data analysis, data handling and business intelligence*. Springer, Berlin.

Diebold, J. & Robert, C. P. (1994). Estimation of finite mixture distribution through Bayesian sampling. *Journal Royal Statistics Society, serie B*, **56**(2), 363–375.

Dugad, R. & Desai, U. B. (1996). A tutorial on hidden Markov models. *Technical Report No.: SPANN-96.11*, pages 1–16.

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

- Gelman, A. B., Carlin, J. S., Stern, H. S. & Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman Hall, London.
- Holst, V. & Lindgren, G. (1991). Recursive estimation in mixture models with Markov regime. *IEEE transactions on information theory*, **37**(6).
- Kass, R. & Raftery, A. (1995). Bayes factor. *Journal American Statistics Association*, **63**, 773–795.
- Koloydenko, A., Kaarik, M. & Lember, J. (2007). On adjusted Viterbi training . *Aca Appli. Math.*, (96), 309–326.
- Lember, J. (2006). *Adjusted Viterbi training for hidden Markov models*. Technical report 07-01, School of Mathematical sciences, Nottingham University, U.K.
- McLachlan, G. & Peel, D. (2000). *Finite mixture models*. Wiley, Canada.
- Missão, E. C. M. (2007). *Uma revisão do fator de Bayes com aplicações à modelos de misturas*. Dissertação de mestrado, Departamento de Estatística – Universidade Federal de São Carlos, São Carlos.
- Richardson, S. & Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Serie B*, **59**(4), 731–792.
- Roeder, K. & Wasserman, L. (1997). Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**(439), 894–920.
- Saraiva, E. F. (2009). *Modelo de mistura com componentes desconhecidos: estimação via método split-merge*. Tese de doutorado, Departamento de Estatística – Universidade Federal de São Carlos, São Carlos.
- Schnatter, S. F. (2008). *Finite mixture and Markov switching models*. Springer, Berlin.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**(2), 461–464.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall, London.
- Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components - Un alternative to reversible jump methods. *The Annals of Statistics*, **28**(1), 40–74.
- Titterington, D. M. & Smith, A. M. U. (1985). *Statistical analysis of finite mixture distributions*. Wiley, New York.
- Wu, C. F. T. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, **11**(1), 95–103.

Apêndice A

Algoritmo em R para simulação

A.1 Simulação de mistura de binomial independentes e para MMM(1) com K componentes

Definição das variáveis

y : dados gerados;

n : número de observações a gerar;

w : Probabilidade inicial de cada componente;

pa : Parâmetros das distribuições (uma distrib para cada componente). No caso de binomiais, cada linha corresponde a uma componente, e cada coluna a um parâmetro (col 1 p e col 2 n);

k : número de componetes de mistura;

$(U \geq sw[k])$ este é um comando lógico que retorna True ou False, mas quando colocado numa soma ele retorna 1 se True e 0 se False .

Mt : Matriz de transição da cadeia de Markov para variável não observável.

Função para geração de modelo de mistura de binomiais com S independente

```

MistBinomialIndep = function(n, pa, w) {
  K = nrow(pa)
  U = runif(n)
  sw = cumsum(w)
  id = rep(1, n); for(kin1 : K)id = id + (U >= sw[k])
  y = rbinom(n, pa[id, 2], pa[id, 1])}

```

Função para simulação de mistura de binomial dependente com K componentes

```

MistBinDepMarkov = function(n, pa, w, Mt) {
  if(abs(1 - sum(w)) > 1e - 10)stop("Valores de w devem
  ser probabilidades")
  K = nrow(pa)
  for(kin1 : K)if(abs(1 - sum(Mt[k, ])) > 1e - 10)stop("Mt deve
  ser Mat Transição")
  U = runif(n)
  sw = cumsum(w)
  id = rep(1, n);
  for(kin1 : K)id[1] = id[1] + (U[1] >= sw[k])
  Ms = Mt;
  for(kin1 : K)Ms[k, ] = cumsum(Mt[k, ])
  for(iin2 : n)for(kin1 : K)id[i] = id[i] + (U[i] >= Ms[id[i - 1], k])
  y = rbinom(n, pa[id, 2], pa[id, 1])}

```


Apêndice B

Algoritmo em R para calculo de EM

B.1 Mistura de binomial independentes com K componentes

Definição das variáveis

y : dados;

M : tamanho da amostra da distribuição binomial;

k : número de componetes de mistura;

T : número de observações;

$theta$: parâmetro da distribuição binomial;

S : variável não observável;

it : contador;

Valores iniciais

$y = c()$

$K = const$

$T = length(y)$

$M = c()$

```

i = floor(T/K)

id = rep(c(1, ..., K), rep(T, K))

n = n = length(id)

S = c(rep(1 : K, rep(i, K)), rep(K, T - i * K))

Tol = 1e - 2

theta = rep(0, K)

Di = 1e - 2

pm = matrix(0, T, K)

mt = matrix(0, K, K)

it = 0;

```

Programa

```

while(Di >= Tol)

theta0 = theta

for(k in 1 : K) {id = S == k;

theta[k] = sum(id * y)/sum(id * M)

m[k] = sum(S == k)}

for(k in 1 : K) {pm[, k] = dbinom(y, M, theta[k])}

sm = apply(pm, 1, sum)

pm = pm/sm

sx = apply(pm, 1, max)

S = (sx == pm)% * %(1 : K)

Di = sum(abs(theta0 - theta))

it = it + 1;

cat('it =', it, 'theta =', theta)}

```

B.2 MMM(1) de distribuição binomial K componentes

Função para calcular matriz de transição

```
tra = function(K, S) {
  m = matrix(0, K, K);
  n = length(S);
  for(i in 1 : K) {for(j in 1 : K) {m[i, j] = sum(S[1 : (n - 1)] == i & S[2 : n] == j)}}
  m[i, ] = m[i, ] / sum(m[i, ])
  m}

```

Programa

```
while(Di >= Tol) {theta0 = theta
for(k in 1 : K) {id = S == k
theta[k] = sum(id * y) / sum(id * M)}
S[1] = which.max(dbinom(y[1], M[1], theta));
mt = tra(K, S)
for(i in 2 : T) {S[i] = which.max(mt[S[i - 1], S[i]] * dbinom(y[i], M[i], theta));}
Di = sum(abs(theta0 - theta));
it = it + 1;
cat('n it =', it, ' theta =', theta)}

```

Apêndice C

Algoritmo em R para cálculo do MCMC

C.1 Mistura de binomial independentes com K componentes

Definição das variáveis

y : dados;

M : tamanho da amostra da distribuição binomial;

k : número de componentes de mistura;

T : número de observações;

θ : parâmetro da distribuição binomial;

S : variável não observável;

N : número de iterações;

p_0 : distribuição inicial;

α : parâmetro da distribuição beta;

β : parâmetro da distribuição beta;

a : parâmetro da distribuição Dirichlet para p ;

gama: parâmetro da distribuição Dirichlet de P

a: parâmetro da distribuição Dirichlet de p_0

Valores iniciais

set.seed()

y = c()

M = c()

T = length(y)

K = const

N = const

theta = c()

p0 = c()

a = c()

alfa = c()

beta = c()

it = 0

S1 = c()

v0 = matrix(0, N, K)

b = c()

p = c()

vthetas = matrix(0, N, K)

S = matrix(0, N, T)

Cálculo da distribuições condicionais

```

for(i in 1 : N) {
  for(lin1 : T) {pp1 = p0;
    v = (pp1 * dbinom(y[l], M[l], theta))/(sum(pp1 * dbinom(y[l], M[l], theta)));
    q = which.max(rmultinom(1, 1, prob = c(v)));
    j = q;
    S1[l] = j; }
  S[i,] = S1
  p0 = rdirichlet(1, c(a[1] + sum(S1 == 1), a[2] + sum(S1 == 2)))
  v0[i,] = p0;
  for(kin1 : K) {b[k] = rbeta(1, alfa[k] + sum(y * (S1 == k)), beta[k] + sum((S1 == k) * (M - y)))
  theta = b;
  vthetas[i,] = cbind(theta[1], theta[2]);
  it = it + 1}

```

C.2 Mistura de binomial dependentes com K componentes

Valores iniciais

```

set.seed()

y = c()

M = c()

T = length(y)

N = const

theta = c()

w1 = c()

pp = matrix(c(), nrow = K)

```

$p0 = c()$

$a = c()$

$alfa = c()$

$beta = c()$

$gama = matrix(c(), nrow = K)$

$it = 0$

$S1 = c()$

$v0 = matrix(0, N, K)$

$b = c()$

$p = c()$ $vthetas = matrix(0, N, K)$

$S = matrix(0, N, T)$

$P = matrix(0, K, K)$

$P1 = matrix(0, N, 3 * K)$

Cálculo das iterações do MCMC

```

for(i in 1 : N) {
  v = (w1 * dbinom(y[1], M, theta))/(sum(w1 * dbinom(y[1], M, theta)))
  q = which(rmultinom(1, 1, prob = c(v)) > 0)
  j = q
  S1[1] = j
  for(l in 2 : T) {pp1 = pp[j,]
    v = (pp1 * dbinom(y[l], M, theta))/(sum(pp1 * dbinom(y[l], M, theta)))
    q = which(rmultinom(1, 1, prob = c(v)) > 0)
    j = q
    S1[l] = j; }
  S[i,] = S1
  p0 = rdirichlet(1, c(a[1] + (S[1] == 1), a[2] + (S[1] == 2), a[3] + (S[1] == 3)))
  v0[i,] = p0
  z = matrix(0, K, K)
  for(k in 1 : K) {for(j in 1 : K) {z[k, j] = sum(S[1 : (T - 1)] == k & S[2 : T] == j)}}
  for(k in 1 : K) {P[k,] = rdirichlet(1, c(gama[k,] * z[k, ]))}
  P1[i,] = cbind(P[1,], P[2,], P[3,])
  pp = P
  for(k in 1 : K) {b[k] = rbeta(1, alfa[k] + sum(y * (S == k)), beta[k] + sum((S == k) * (M[k] - y
  theta = b
  vthetas[i,] = cbind(b[1], b[2], b[3])
  it = it + 1}

```