UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE CIÊNCIAS EXATAS E DA TERRA
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA
APLICADA E ESTATÍSTICA

ESTIMAÇÃO PARAMÉTRICA E NÃO-PARAMÉTRICA EM MODELOS DE MARKOV OCULTOS

FRANCISCO MOISÉS CÂNDIDO DE MEDEIROS
Orientador: Prof. Dr. André Gustavo Campos Pereira
Co-orientador: Prof. Dr. Paulo Sérgio Lucio

NATAL, FEVEREIRO DE 2010

UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE CIÊNCIAS EXATAS E DA TERRA
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA
APLICADA E ESTATÍSTICA

ESTIMAÇÃO PARAMÉTRICA E NÃO-PARAMÉTRICA EM MODELOS DE MARKOV OCULTOS

FRANCISCO MOISÉS CÂNDIDO DE MEDEIROS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Matemática Aplicada e Estatística da Universidade Federal do Rio Grande do Norte (PPGMAE-UFRN) como parte dos requisitos necessários para obtenção do título de Mestre em Matemática Aplicada e Estatística.

NATAL, FEVEREIRO DE 2010

"Sempre me pareceu estranho que todos aqueles que estudam seriamente esta ciência acabam tomados de uma espécie de paixão pela mesma. Em verdade, o que proporciona o máximo prazer não é o conhecimento e sim a aprendizagem, não é a posse mas a aquisição, não é a presença, mas o ato de atingir a meta."

Carl Friedrich Gauss

Agradecimentos

A Deus, acima de tudo.

Ao professor André Gustavo pela confiança depositada em mim e pela orientação. Sem ele esse trabalho não seria possível. Valeu cara.

À professora Viviane Simioli pela ajuda nos primeiros passos deste trabalho, pela atenção e disposição em ajudar nas horas de dúvidas e por participar da banca de qualificação contribuindo com suas valiosas sugestões.

Ao professor Paulo Lucio pela motivação e apoio na parte computacional.

Ao professor Pledson Guedes por participar da banca de qualificação e pelas correções e sugestões.

Ao professor Jaques Silveira pelos ensinamentos e conselhos valiosos.

Ao professor Damião Nóbrega pelo presente que recebi por tê-lo como meu professor, pelo excelente profissional e por me mostrar a beleza da Estatística.

À professora Dione Valença, a quem admiro como pessoa e como profissional, por acreditar no PPGMAE e por tratar todos com carinho e dedicação.

À minha mãe, meu pai, meus irmãos Mércia e Messias, minha querida avó Dona Maria e minha amada tia Maria de Fátima, que mesmo distantes sempre estiveram próximos a mim em orações e imensuravelmente me apoiaram na luta pelos meus objetivos, ensinando que uma das melhores coisas da vida é aprender. Obrigado por entenderem minha ausência e pelos poucos momentos felizes, porém intensos que passamos juntos durante esses seis anos. Sem eles eu não conseguiria chegar até aqui. Meu muito obrigado.

Às minhas amadas amigas e companheiras de jornada, desde a graduação, Maria Aparecida e Camila Nascimento, por compartilharmos inesquecíveis momentos de felicidade e tristeza, pelo companheirismo, conselhos e ajuda. Pelo cuidado em me ouvir, entender,

por acreditarem em mim quando ninguém mais acreditava, por todo amor e atenção, em fim, por existirem na minha vida.

À minha amiga Patrícia Borchardt por me aturar durante esses seis anos e me amar do jeito que sou. Obrigado pelo carinho, cuidado e atenção.

Aos meus colegas de turma, em especial: Enai, Tatiana Queiroz, João Batista e Daniel, pela companhia nos estudos e pelos vários momentos de descontração e sabedoria.

Aos companheiros Cecílio, Lenilson, Manasses, Renata Mendonça, Renata Fonseca, João Saturnino e Hermes, por compartilharem minhas alegrias e frustrações.

Aos demais colegas do programa: Renilma, Francinário, Kaline Andreza, Kaline Juliana, Kelly, Claúdia, Felipe, Julianne, Alysson, Marconio e Tatiana Farache, pelos ensinamentos.

Aos amigos que não fazem parte do PPGMAE mas compartilham da realização desse sonho:

À Liliana, Simone Pereira, Simone Santos, Úrsula e Vanessa por torcerem por mim.

A Kleber Klinger, Danny Stywart (Baraca obaca), Kalyne Lucena, Luana Paré e Raquel Dié, por conviverem com as minhas angústias, ansiedades, alegrias e tristezas.

Ao Departamento de Estatística na pessoa do professor André Pinho, que sempre me socorria nas dúvidas de inferência, e da professora Jeanete Alves, por me receberem maravilhosamente bem.

À Liandra e ao Russinho, funcionários e amigos do CCET, que sempre nos recebem com carinho e atenção.

A Rafael Pimenta pela disponibilidade em ajudar e resolver os nossos problemas.

Aos professores da banca de defesa pelas sugestões e correções.

À CAPES pelo apoio financeiro.

Resumo

Neste trabalho estudamos os modelos de Markov ocultos tanto em espaço de estados finito quanto em espaço de estados geral. No caso discreto, estudamos os algoritmos para frente e para trás para determinar a probabilidade da seqüência observada e, em seguida, estimamos os parâmetros do modelo via algoritmo EM. No caso geral, estudamos os estimadores do tipo núcleo e os utilizamos para conseguir uma seqüência de estimadores que converge na norma L_1 para a função densidade do processo observado.

Palavras chaves: Cadeia de Markov. Modelos de Markov oculto. Espaço de estados finito. Espaço de estados geral.

Áreas do conhecimento: Probabilidade e Estatística.

Abstract

In this work we study the Hidden Markov Models with finite as well as general state space. In the finite case, the forward and backward algorithms are considered and the probability of a given observed sequence is computed. Next, we use the EM algorithm to estimate the model parameters. In the general case, the kernel estimators are used and to built a sequence of estimators that converge in L_1-norm to the density function of the observable process.

Key words: Markov chain. Hidden Markov models. Finite state space. General state space.

Areas of knowledge: Probability and Statistics.

Sumário

Introdução

Quando estudamos cadeias de Markov, vemos que a característica marcante destes processos é que o próximo passo depende apenas de onde o processo se encontra no momento atual. Essa característica se encontra em muitos fenômenos da natureza ocasionando a otimização do processo. Por exemplo, em metalurgia o resfriamento do metal, via decréscimo programado da temperatura (agendamento de resfriamento), leva a estrutura interna dos átomos a uma configuração de energia mínima. Em biologia, a evolução sofrida pela população de geração em geração leva a evolução da espécie. Procurando emular tais situações conhecidas (meta-heurísticas) para solucionar problemas de otimização, algoritmos foram criados via cadeias de Markov. Podemos citar o *Simulated Annealing* (recozimento simulado) que procura utilizar o procedimento da metalurgia, criando uma cadeia de Markov não-homogênea, cuja energia a ser minimizada é a função que estamos querendo minimizar. Um outro algoritmo, que utiliza a meta-heurística da evolução das espécies, é o algoritmo genético. Neste algoritmo uma população inicial é gerada e depois procedimentos de seleção, cruzamento e mutação são efetuados nesta população a fim de fazê-la evoluir e chegar na população perfeita (formada só por pontos de ótimo). O processo de evolução de uma população para outra (depois dos três passos: seleção, mutação e cruzamento) também é representado por uma cadeia de Markov (homogênea).

Vemos desta forma que existem modelagens via cadeias de Markov utilizando diversas meta-heurísticas. Entretanto, existem outros tipos de problemas em que não temos a modelagem explícita via cadeias de Markov. Na verdade, sabemos que existe uma cadeia

de Markov controlando a dinâmica do processo, mas o resultado que observamos não é gerado apenas pela cadeia.

Por exemplo, suponhamos que uma pessoa seleciona, aleatoriamente, bolas de urnas que estão atrás de uma cortina. Uma urna é selecionada ao acaso, uma bola é retirada desta urna e mostrada para um observador que encontra-se do outro lado da cortina. Ele verifica qual é a cor da bola, no entanto, desconhece de qual urna a bola foi retirada. Considere também que existam apenas duas urnas, urna 1 e urna 2. Cada urna contém bolas de duas cores diferentes, digamos, vermelhas e brancas, e que a probabilidade de selecionarmos qualquer uma das urnas é a mesma.

Esse experimento pode ser modelado se considerarmos que a escolha da urna segue uma cadeia de Markov $\{X_t\}_{t\in\mathbb{N}}$ onde cada estado representa uma urna de onde a bola foi selecionada e que o processo observado $\{Y_t\}_{t\in\mathbb{N}}$ é gerado quando retiramos, ao acaso, as bolas das urnas. É importante lembrar que não sabemos de qual urna a bola foi selecionada (oculta), apenas conhecemos a cor da bola.

Tais modelos são denominados Modelos de Markov Ocultos (HMM, do inglês *Hidden Markov Models*). Nesse exemplo particular, o observado é a cor da bola (vermelha ou branca) e o oculto são os estados da cadeia de Markov (urna 1 ou urna 2).

Dado que a escolha das urnas é equiprovável,

$$P(X_{t+1} = j \,|\, X_t = i) = 0.5, \qquad \forall\ i, j\ \in \{\text{urna } 1, \text{urna } 2\} \quad \text{e} \quad t\ \in \mathbb{N}.$$

Neste caso, temos dois símbolos observáveis: bolas brancas ($Y_t = B$) e bolas vermelhas ($Y_t = V$). Suponhamos que as probabilidades de uma bola branca e uma bola vermelha serem selecionadas da urna 1 sejam 0.6 e 0.4, respectivamente, (ou seja, $P(Y_t = B \,|\, X_t = \text{urna } 1) = 0.6$ e $P(Y_t = V \,|\, X_t = \text{urna } 1) = 0.4$) e da urna 2 sejam 0.8 e 0.2, respectivamente, (ou seja, $P(Y_t = B \,|\, X_t = \text{urna } 2) = 0.8$ e $P(Y_t = V \,|\, X_t = \text{urna } 2) = 0.2$). Essas probabilidades são denominadas *distribuição de probabilidade dos símbolos observáveis*. Na Figura 1 representamos esse processo.

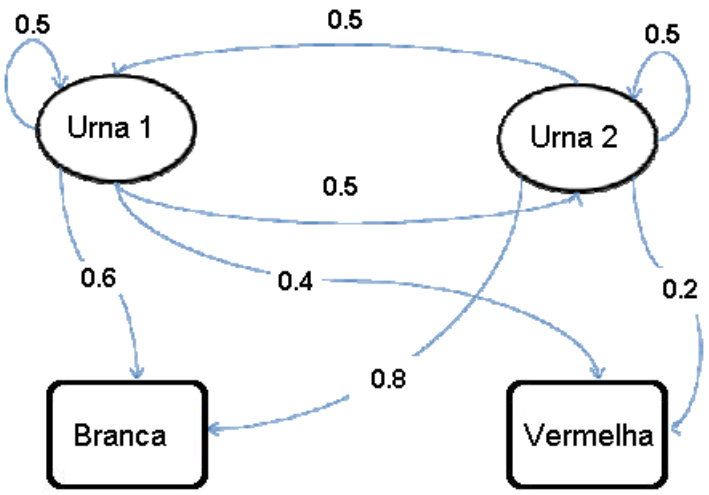

Figura 1: Estrutura do modelo.

Assim, associada a um HMM temos uma distribuição inicial para a cadeia de Markov (δ), uma matriz de transição da cadeia de Markov (Γ) e ainda uma matriz que indica as probabilidades de observarmos uma determinada saída, dado que a cadeia se encontrava em um estado específico (Π). Só para ilustrar, se supormos que o espaço de estados da cadeia de Markov é $E = \{1, 2, 3, ..., m\}$ e que o espaço de estados do processo observável é $S = \{y_1, y_2, ..., y_T\}$, então associado ao modelo teremos os seguintes dados: a distribuição inicial, definida por,

$$\delta_i = P(X_1 = i) \qquad \forall\, i \in E,$$

a matriz de probabilidades de transição da cadeia de Markov

$$\Gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1m} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m1} & \gamma_{m2} & \cdots & \gamma_{mm} \end{bmatrix},$$

sendo,

$$\gamma_{ji} = P(X_{t+1} = i \,|\, X_t = j) \qquad \forall\, t \in \mathbb{N} \quad \text{e} \quad i, j \in E,$$

e a matriz $T \times m$

$$\Pi = \begin{bmatrix} \pi_{y_{1,1}} & \pi_{y_{1,2}} & \cdots & \pi_{y_{1,m}} \\ \pi_{y_{2,1}} & \pi_{y_{2,2}} & \cdots & \pi_{y_{2,m}} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{y_{T,1}} & \pi_{y_{T,2}} & \cdots & \pi_{y_{T,m}} \end{bmatrix},$$

de probabilidades definidas por:

$$\pi_{y_t,j} = P(Y_t = y_t \,|\, X_t = j).$$

Portanto, diferentemente dos processos markovianos, onde cada estado corresponde a uma observação e a evolução do processo se dá pelas probabilidades de transição, nos HMMs uma observação pode estar associada a vários estados de uma cadeia de Markov, segundo uma distribuição de probabilidade, ou seja, cada observação pode ser gerada por uma distribuição de probabilidade de qualquer estado do processo, como mostramos na Figura 2.

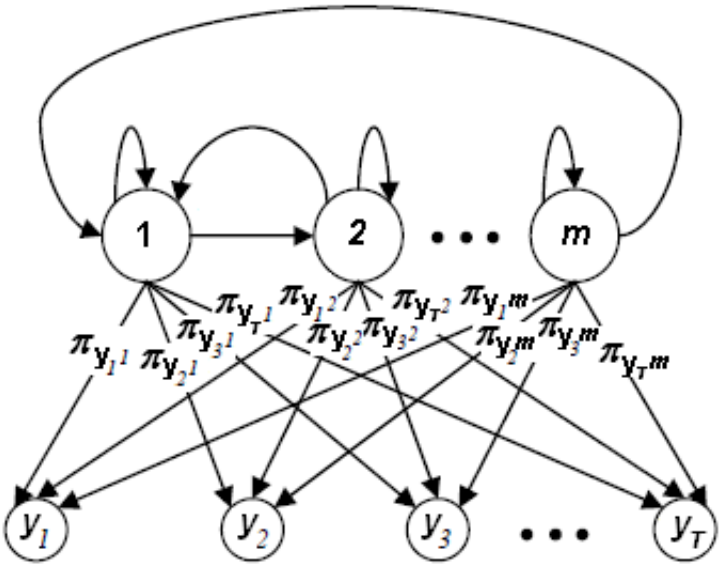

Figura 2: Estrutura de um modelo de Markov oculto a tempo discreto.

Observando a Figura 2 surgem algumas perguntas naturais:

1) Dado um modelo de Markov oculto, como calculamos a probabilidade da seqüência observada $\{y_1, \ldots, y_T\}$?

2) Dada a seqüência observada $\{y_1, \ldots, y_T\}$, quais os valores de δ, Γ e Π que fornecem a maior probabilidade dessa seqüência ter sido gerada por um modelo de Markov oculto?

Essas perguntas serão respondidas no Capítulo 1 onde assumimos um modelo de Markov oculto $\{X_t, Y_t\}_{t\in\mathbb{N}}$ com espaço de estados finito, onde $\{X_t\}_{t\in\mathbb{N}}$ é uma cadeia de Markov homogênea e $\{Y_t\}_{t\in\mathbb{N}}$ é um processo estocástico observado, gerado a partir de $\{X_t\}_{t\geq 0}$, segundo um conjunto de distribuições de probabilidades. Ainda neste capítulo introduzimos o algoritmo EM que será usado para responder ao segundo questionamento.

No Capítulo 2 tivemos como artigo base Dorea-Zhao (2002), onde trabalham com um modelo de Markov oculto $\{X_t, Y_t\}_{t\in\mathbb{N}}$ com espaço de estados geral dando condições sobre a cadeia de Markov homogênea $\{X_t\}_{t\in\mathbb{N}}$ e sobre o processo observado $\{Y_t\}_{t\in\mathbb{N}}$, de tal maneira a obter um estimador para a densidade de Y_t. Nesta direção foram introduzidos conceitos como estimador do tipo núcleo, um processo estocástico satisfazer a condição ϕ-*mixing* e possuir uma distribuição estacionária π, dentre outros.

Os HMMs têm se tornado, ao passar dos anos, uma importante ferramenta na modelagem de seqüências de variáveis aleatórias fracamente dependentes. Estes modelos estocásticos estão sendo aplicados em várias áreas do conhecimento, servindo de modelagem para problemas de processamento de voz (Rabiner, 1989), neurofisiologia (Fredkin-Rice, 1992), biologia (Leroux-Puterman, 1992), processamento de imagens (Huang et al., 2008), entre outras.

Considerando, então, a importância e utilidade dos HMMs em várias áreas de pesquisa, nosso objetivo principal é estudar esses modelos. Com esse estudo, buscamos introduzir os modelos markovianos ocultos aos leitores que busquem trabalhar com o tema.

Para termos idéia do desenvolvimento da teoria inferencial em HMMs, ela foi publicada pela primeira vez por Baum-Petrie (1966) no caso em que o modelo observável tinha espaço de estados finito. Neste trabalho, mostram a consistência e a normalidade assintótica dos estimadores de máxima verossimilhança (EMVs). Em Petrie (1969), as condições de consistência dos EMVs são enfraquecidas. Lindgren (1978) constroe estimadores consistentes e assintoticamente normais para os parâmetros, determinando as densidades condicionais de Y_n dado X_n, mas não considera a estimação das probabilidades de transição. Mais tarde, Leroux (1992) prova a consistência dos EMVs para HMMs gerais sob condições fracas. Em 1994, Rydén propõe uma nova classe de estimadores e prova a consistência e a normalidade assintótica sob algumas condições de regularidade. Depois, Bickel-Ritov (1996) constroem a normalidade assintótica local dos EMVs e em Bickel et al. (1998) a normalidade assintótica desses estimadores é provada.

Uma importante bibliografia no estudo inferencial em HMMs é o livro de Cappé et al. (2005). Trata-se de um estudo completo de inferência em HMMs gerais.

Capítulo 1

Modelos de Markov Ocultos em Espaço de Estados Discreto

Neste capítulo estudamos os modelos de Markov ocultos quando os espaços de estados envolvidos são finitos. Na seção 1.1 relembramos a definição e os elementos inerentes às cadeias de Markov. Na seção 1.2 introduzimos a definição dos modelos de Markov ocultos e derivamos algumas de suas propriedades, utilizando fortemente a suposição de independência condicional do modelo. Finalmente, na seção 1.3 apresentamos o algoritmo EM, as situações em que ele melhor se aplica, suas condições e o aplicamos na estimação dos parâmetros do modelo de Markov oculto, a saber as matrizes δ, Γ e Π.

1.1 Cadeia de Markov

Sejam $S = \{i_k, k = 0, \ldots, n\}$ um conjunto finito e $\delta = \{\delta_j, 0 \leq j \leq n\}$ uma distribuição de probabilidade, isto é, $\delta_j \geq 0$ para todo j e $\sum_{j=0}^{n} \delta_j = 1$.

Definição 1.1 *Uma seqüência de variáveis aleatórias (v.a.s) $\{X_n\}_{n\geq 0}$, assumindo valores em S é dita ser uma cadeia de Markov com distribuição inicial δ se,*

(i) $X_0 \sim \delta$, *ou seja,* $P(X_0 = i_k) = \delta_k$ *para todo* $k = 0, \ldots, n$.

(ii) $P(X_n = i_n \,|\, X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) = P(X_n = i_n \,|\, X_{n-1} = i_{n-1})$, *para todo* $i_0, \ldots, i_{n-1}, i_n \in S$ *e* $n = 0, 1, 2, \ldots$

Em outras palavras, dada uma distribuição inicial δ a sequência $\{X_n\}_{n\geq 0}$ é sem memória, isto é, o seu comportamento futuro (X_n) depende apenas do estado presente (X_{n-1}), independentemente do seu passado $(X_0, \ldots, X_{n-2})$.

O conjunto S é denominado *espaço de estados*. Quando S é um conjunto enumerável, dizemos que o processo é uma *cadeia* ou que o processo é *discreto*, caso S não seja enumerável dizemos que o processo tem *espaço de estados geral*.

Definição 1.2 *Considere* $\{X_n\}_{n\geq 0}$ *uma cadeia de Markov com distribuição inicial* δ. *O núcleo de transição (de um passo) da cadeia é definido como:*

$$\gamma_{ij}^{(n,n+1)} = P(X_{n+1} = j \,|\, X_n = i), \qquad \forall\, n \in \{0, 1, 2, \ldots\}$$

se $\gamma_{ij}^{(n,n+1)}$ *não depende de* n, *isto é,*

$$\gamma_{ij}^{(n,n+1)} = \gamma_{ij} = P(X_{n+1} = j \,|\, X_n = i), \qquad \forall\, n \in \{0, 1, 2, \ldots\}$$

dizemos que a cadeia é homogênea ou estacionária.

Considere $\{X_n\}_{n\in\mathbb{N}}$ uma cadeia de Markov homogênea com distribuição inicial δ e espaço de estados $S = \{1, 2, \ldots, k\}$. Para essa cadeia existem k^2 probabilidades de transição γ_{ij}, $i = 1, \ldots, k$ e $j = 1, \ldots, k$. Podemos organizar esses valores em uma matriz $\mathbf{\Gamma} = \{\gamma_{ij};\, i, j \in S\}$,

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1k} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{k1} & \gamma_{k2} & \cdots & \gamma_{kk} \end{bmatrix},$$

denominada *matriz de probabilidades de transição* ou simplesmente *matriz de transição*. Essa matriz nos dá todas as informações a respeito da dinâmica do processo entre os estados.

Note que a matriz de transição satisfaz as seguintes propriedades:

- $\gamma_{ij} \geq 0, \qquad i, j \in S.$
- $\sum_{j=1}^{k} \gamma_{ij} = 1, \qquad i \in S.$

Então, uma vez conhecida a distribuição inicial e a matriz de transição, o processo de Markov está totalmente caracterizado.

1.2 Cadeia de Markov oculta

Considere um experimento aleatório $\mathcal{E}$. Cada realização desse experimento gera uma seqüência de observações $\{y_1, ..., y_T\}$, que é considerada como uma realização de comprimento T de algum processo aleatório $\{Y_t : t \in \mathbb{N}\}$ com espaço de estados finito S. O processo $\{Y_t\}_{t\in\mathbb{N}}$ é gerado por dois mecanismos probabilísticos: em primeiro lugar, uma cadeia de Markov homogênea não observável $\{X_t : t \in \mathbb{N}\}$ com m estados que ocorrem em qualquer instante de tempo t e, em segundo lugar, um conjunto de distribuições de probabilidades, uma para cada estado, que produzem as observações a partir de um conjunto finito de T possibilidades.

Definição 1.3 *Uma cadeia de Markov oculta (HMM) a tempo discreto é um processo estocástico duplo* $\{(X_t, Y_t)\}_{t\in\mathbb{N}}$ *tal que:*

1. $\{X_t\}_{t\in\mathbb{N}}$ *é uma cadeia de Markov homogênea não observável, de espaço de estados finito* $E = \{1, 2, \ldots, m\}$.

2. $\{Y_t\}_{t\in\mathbb{N}}$ *é uma seqüência de variáveis aleatórias condicionalmente independentes com a distribuição de* Y_k *dependendo somente de* X_k*, ou seja*

$$\begin{aligned} P(\boldsymbol{Y}|\,\boldsymbol{X}) &= P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_T = y_T \,|\, X_1 = x_1, X_2 = x_2, \ldots, X_T = x_T) \\ &= \prod_{l=1}^{T} P(Y_l = y_l \,|\, X_1 = x_1, X_2 = x_2, \ldots, X_T = x_T) \\ &= \prod_{l=1}^{T} P(Y_l = y_l \,|\, X_l = x_l). \end{aligned}$$

O termo HMM é motivado por assumirmos que $\{X_t : t \in \mathbb{N}\}$ é uma cadeia de Markov não observável, ou seja, não sabemos de qual estado o processo observado foi gerado.

Como falamos na introdução, para caracterizarmos uma cadeia de Markov basta que conheçamos a distribuição inicial e a matriz de probabilidades de transição do processo. Já nos HMMs além da distribuição inicial (denotada por δ) definida por,

$$\delta_i = P(X_1 = i), \qquad \forall\, i \in E,$$

e da matriz de probabilidades de transição da cadeia de Markov (denotada por Γ)

$$\Gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1m} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m1} & \gamma_{m2} & \cdots & \gamma_{mm} \end{bmatrix},$$

sendo

$$\gamma_{ji} = P(X_{t+1} = i \,|\, X_t = j), \qquad \forall\, t \in \mathbb{N} \quad \text{e} \quad i, j \in E,$$

é necessário que conheçamos uma matriz de probabilidades $T \times m$ (denotada por Π)

$$\Pi = \begin{bmatrix} \pi_{y_{1,1}} & \pi_{y_{1,2}} & \cdots & \pi_{y_{1,m}} \\ \pi_{y_{2,1}} & \pi_{y_{2,2}} & \cdots & \pi_{y_{2,m}} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{y_{T,1}} & \pi_{y_{T,2}} & \cdots & \pi_{y_{T,m}} \end{bmatrix},$$

sendo

$$\pi_{y_t,j} = P(Y_t = y_t \,|\, X_t = j),$$

com

$$\sum_{t=1}^{T} \pi_{y_t,j} = 1, \qquad \forall\, j \in E.$$

No exemplo apresentado na introdução, se a distribuição inicial para a cadeia de Markov for

$$\delta = \begin{bmatrix} 1 \\ \\ 0 \end{bmatrix},$$

os outros dados retirados do modelo são

$$\Gamma = \begin{bmatrix} 0.5 & 0.5 \\ \\ 0.5 & 0.5 \end{bmatrix} \qquad \text{e} \qquad \Pi = \begin{bmatrix} 0.6 & 0.2 \\ \\ 0.4 & 0.8 \end{bmatrix}.$$

Para respondermos ao primeiro questionamento feito na introdução, precisaremos de uma série de propriedades dos modelos de Markov ocultos que detalharemos a seguir. Todas estas propriedades referem-se ao cenário da definição, isto é, o processo $\{X_t\}_{t\in\mathbb{N}}$ é uma cadeia de Markov homogênea com espaço de estados finito E e $\{Y_t : t \in \mathbb{N}\}$ tem, para todo T, a seguinte propriedade: condicionado a $\mathbf{X} = \{X_t : t = 1, \ldots, T\}$, as variáveis aleatórias $Y_1, \ldots, Y_T$ são mutuamente independentes e a distribuição condicional de Y_t só depende de X_t e não das variáveis X_k, $k \neq t$.

Por simplicidade e para não carregar a notação, o evento $\{X_t = x_t\}$ será denotado por X_t. No entanto, quando necessário, usaremos a notação $\{X_t = x_t\}$. De maneira análoga, o evento $\{Y_t = y_t\}$ será denotado por Y_t e quando necessário faremos uso da notação $\{Y_t = y_t\}$.

Proposição 1.1 *Para todo inteiro t e l tais que $1 \leq t \leq l \leq T$:*

$$P(Y_l, \ldots, Y_T \,|\, X_t, \ldots, X_T) = P(Y_l, \ldots, Y_T \,|\, X_l, \ldots, X_T).$$

Demonstração. Para $t = 1$ e $t = l$ não há o que demonstrar. Suponhamos que $1 < t < l$. Assim,

$$\begin{aligned} P(Y_l, \ldots, Y_T \,|\, X_t, \ldots, X_T) &= \frac{1}{P(X_t, \ldots, X_T)} P(Y_l, \ldots, Y_T, X_t, \ldots, X_T) \\ &= \frac{1}{P(X_t, \ldots, X_T)} \sum_{x_1, \ldots, x_{t-1} \in E} P(Y_l, \ldots, Y_T, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}, X_t, \ldots, X_T) \\ &= \frac{1}{P(X_t, \ldots, X_T)} \sum_{x_1, \ldots, x_{t-1} \in E} P(Y_l, \ldots, Y_T \,|\, X_1, \ldots, X_T) P(X_1, \ldots, X_T), \end{aligned}$$

segue da Definição 1.3 que,

$$P(Y_l, \ldots, Y_T \,|\, X_1, \ldots, X_T) = P(Y_l \,|\, X_l) \cdots P(Y_T \,|\, X_T),$$

então,

$$\begin{aligned} P(Y_l, \ldots, Y_T \,|\, X_t, \ldots, X_T) &= \frac{P(Y_l \,|\, X_l) \cdots P(Y_T \,|\, X_T)}{P(X_t, \ldots, X_T)} \sum_{x_1, \ldots, x_{t-1} \in E} P(X_1, \ldots, X_T) \\ &= \frac{P(Y_l, \ldots, Y_T \,|\, X_l, \ldots, X_T)}{P(X_t, \ldots, X_T)} \underbrace{\sum_{x_1, \ldots, x_{t-1} \in E} P(X_1, \ldots, X_T)}_{P(X_t, \ldots, X_T)} \\ &= \frac{P(Y_l, \ldots, Y_T \,|\, X_l, \ldots, X_T)}{P(X_t, \ldots, X_T)} P(X_t, \ldots, X_T) \\ &= P(Y_l, \ldots, Y_T \,|\, X_l, \ldots, X_T). \end{aligned}$$

Proposição 1.2 *Para $t = 1, 2, \ldots, T-1$:*

$$P(Y_{t+1}, \ldots, Y_T \,|\, X_1, \ldots, X_t) = P(Y_{t+1}, \ldots, Y_T \,|\, X_t).$$

Demonstração. Aplicando a Proposição 1.1 temos:

$$\begin{aligned} P(Y_{t+1}, \ldots, Y_T \,|\, X_1, \ldots, X_t) &= \frac{1}{P(X_1, \ldots, X_t)} P(Y_{t+1}, \ldots, Y_T, X_1, \ldots, X_t) \\ &= \frac{1}{P(X_1, \ldots, X_t)} \sum_{x_{t+1}, \ldots, x_T \in E} P(Y_{t+1}, \ldots, Y_T, X_1, \ldots, X_t, X_{t+1}, \ldots, X_T) \\ &= \frac{1}{P(X_1, \ldots, X_t)} \sum_{x_{t+1}, \ldots, x_T \in E} P(Y_{t+1}, \ldots, Y_T \,|\, X_1, \ldots, X_T) P(X_1, \ldots, X_T) \\ &= \frac{1}{P(X_1, \ldots, X_t)} \sum_{x_{t+1}, \ldots, x_T \in E} P(Y_{t+1}, \ldots, Y_T \,|\, X_{t+1}, \ldots, X_T) P(X_1, \ldots, X_T) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{P(X_1,\ldots,X_t)} \sum_{x_{t+1},\ldots,x_T \in E} P(Y_{t+1},\ldots,Y_T \,|\, X_t,\ldots,X_T) P(X_1,\ldots,X_T) \\
&= \sum_{x_{t+1},\ldots,x_T \in E} P(Y_{t+1},\ldots,Y_T \,|\, X_t,\ldots,X_T) P(X_{t+1},\ldots,X_T \,|\, X_1,\ldots,X_t),
\end{aligned}
$$

pela propriedade de Markov,

$$
\begin{aligned}
P(Y_{t+1},\ldots,Y_T \,|\, X_1,\ldots,X_t) &= \sum_{x_{t+1},\ldots,x_T \in E} \frac{P(Y_{t+1},\ldots,Y_T,X_t,\ldots,X_T)}{P(X_t,\ldots,X_T)} \frac{P(X_t,\ldots,X_T)}{P(X_t)} \\
&= \frac{P(Y_{t+1},\ldots,Y_T,X_t)}{P(X_t)} \\
&= P(Y_{t+1},\ldots,Y_T \,|\, X_t).
\end{aligned}
$$

Proposição 1.3 *Para $t = 1, 2, \ldots, T$:*

$$P(Y_1,\ldots,Y_t \,|\, X_1,\ldots,X_T) = P(Y_1,\ldots,Y_t \,|\, X_1,\ldots,X_t).$$

Demonstração. Segue da Definição 1.3 que,

$$P(Y_1,\ldots,Y_t \,|\, X_1,\ldots,X_T) = P(Y_1 \,|\, X_1) \cdots P(Y_t \,|\, X_t)$$

e

$$P(Y_1,\ldots,Y_t \,|\, X_1,\ldots,X_t) = P(Y_1 \,|\, X_1) \cdots P(Y_t \,|\, X_t),$$

então

$$P(Y_1,\ldots,Y_t \,|\, X_1,\ldots,X_T) = P(Y_1,\ldots,Y_t \,|\, X_1,\ldots,X_t).$$

Proposição 1.4 *Para $t = 1, 2, \ldots, T$:*

$$P(Y_1,\ldots,Y_T \,|\, X_t) = P(Y_1,\ldots,Y_t \,|\, X_t) P(Y_{t+1},\ldots,Y_T \,|\, X_t).$$

Demonstração. Usando a independência condicional de $Y_1,\ldots,Y_T$ dado $\mathbf{X} = \{X_1,\ldots,X_T\}$, podemos escrever

$$P(Y_1,\ldots,Y_T \,|\, X_t) = \frac{1}{P(X_t)} \sum\nolimits^{(1)} \sum\nolimits^{(2)} P(\mathbf{X}) \left[P(Y_1,\ldots,Y_t \,|\, \mathbf{X}) P(Y_{t+1},\ldots,Y_T \,|\, \mathbf{X}) \right]$$

onde $\sum^{(1)}$ e $\sum^{(2)}$ denotam o somatório em $x_1, \ldots, x_{t-1} \in E$ e $x_{t+1}, \ldots, x_T \in E$ respectivamente. Aplicando as Proposições 1.3 e 1.2, o lado direito da igualdade pode ser escrito como,

$$\begin{aligned}
& \frac{1}{P(X_t)} \sum^{(1)} \sum^{(2)} P(\mathbf{X}) \left[P(Y_1, \ldots, Y_t \,|\, X_1, \ldots, X_t) P(Y_{t+1}, \ldots, Y_T \,|\, \mathbf{X}) \right] \\
= & \frac{1}{P(X_t)} \sum^{(1)} \sum^{(2)} P(\mathbf{X}) \left[\frac{P(Y_1, \ldots, Y_t, X_1, \ldots, X_t)}{P(X_1, \ldots, X_t)} P(Y_{t+1}, \ldots, Y_T \,|\, \mathbf{X}) \right] \\
= & \frac{1}{P(X_t)} \left[\sum^{(1)} \frac{P(Y_1, \ldots, Y_t, X_1, \ldots, X_t)}{P(X_1, \ldots, X_t)} P(X_1, \ldots, X_t) \sum^{(2)} P(Y_{t+1}, \ldots, Y_T \,|\, \mathbf{X}) \right] \\
= & \frac{1}{P(X_t)} P(Y_1, \ldots, Y_t, X_t) P(Y_{t+1}, \ldots, Y_T \,|\, X_t) = P(Y_1, \ldots, Y_t \,|\, X_t) P(Y_{t+1}, \ldots, Y_T \,|\, X_t),
\end{aligned}$$

então,

$$P(Y_1, \ldots, Y_T \,|\, X_t) = P(Y_1, \ldots, Y_t \,|\, X_t) P(Y_{t+1}, \ldots, Y_T \,|\, X_t).$$

Proposição 1.5 *Para $t = 1, 2, \ldots, T$:*

$$P(Y_t, \ldots, Y_T \,|\, X_t) = P(Y_t \,|\, X_t) P(Y_{t+1}, \ldots, Y_T \,|\, X_t).$$

Demonstração. Pela Proposição 1.4,

$$P(Y_1, \ldots, Y_T \,|\, X_t) = P(Y_1, \ldots, Y_t \,|\, X_t) P(Y_{t+1}, \ldots, Y_T \,|\, X_t).$$

Somando ambos os lados com respeito a $y_1, \ldots, y_{t-1} \in S$ temos,

$$\sum_{y_1, \ldots, y_{t-1} \in S} P(Y_1, \ldots, Y_T \,|\, X_t) = \sum_{y_1, \ldots, y_{t-1} \in S} P(Y_1, \ldots, Y_t \,|\, X_t) P(Y_{t+1}, \ldots, Y_T \,|\, X_t),$$

então

$$P(Y_t, \ldots, Y_T \,|\, X_t) = P(Y_t \,|\, X_t) P(Y_{t+1}, \ldots, Y_T \,|\, X_t).$$

Proposição 1.6 *Para $t = 1, 2, \ldots, T$:*

$$P(Y_1, \ldots, Y_T \,|\, X_t, X_{t+1}) = P(Y_1, \ldots, Y_t \,|\, X_t) P(Y_{t+1}, \ldots, Y_T \,|\, X_{t+1}).$$

A demonstração desta propriedade é análoga à demostração da Proposição 1.4.

Proposição 1.7 *Para todo inteiro t e l tais que $1 \leq t \leq l \leq T$:*

$$P(Y_l, \ldots, Y_T \,|\, X_t, \ldots, X_l) = P(Y_l, \ldots, Y_T \,|\, X_l).$$

Demonstração. Para $t = l$ não há o que demonstrar. Suponhamos que $1 \leq t < l$ assim,

$$\begin{aligned} P(Y_l, \ldots, Y_T \,|\, X_t, \ldots, X_l) &= \frac{1}{P(X_t, \ldots, X_l)} P(Y_l, \ldots, Y_T, X_t, \ldots, X_l) \\ &= \frac{1}{P(X_t, \ldots, X_l)} \sum\nolimits^{(2)} \sum\nolimits^{(1)} P(Y_l, \ldots, Y_T \,|\, \mathbf{X}) P(\mathbf{X}) \end{aligned}$$

onde $\sum^{(1)}$ denota o somatório em $x_1, \ldots, x_{t-1} \in E$ e $\sum^{(2)}$ denota o somatório em $x_{l+1}, \ldots, x_T \in E$. Fazendo $t = 1$ na Proposição 1.1 temos,

$$P(Y_l, \ldots, Y_T \,|\, \mathbf{X}) = P(Y_l, \ldots, Y_T \,|\, X_l, \ldots, X_T),$$

logo,

$$\begin{aligned} & P(Y_l, \ldots, Y_T \,|\, X_t, \ldots, X_l) = \frac{1}{P(X_t, \ldots, X_l)} \sum\nolimits^{(2)} \sum\nolimits^{(1)} P(Y_l, \ldots, Y_T \,|\, X_l, \ldots, X_T) P(\mathbf{X}) \\ = \; & \frac{1}{P(X_t, \ldots, X_l)} \left[\sum\nolimits^{(2)} P(Y_l, \ldots, Y_T \,|\, X_l, \ldots, X_T) \sum\nolimits^{(1)} P(X_1, \ldots, X_{t-1}, \ldots, X_T) \right] \\ = \; & \frac{1}{P(X_t, \ldots, X_l)} \sum\nolimits^{(2)} P(Y_l, \ldots, Y_T \,|\, X_l, \ldots, X_T) P(X_t, \ldots, X_l, X_{l+1}, \ldots, X_T) \\ = \; & \frac{1}{P(X_t, \ldots, X_l)} \sum\nolimits^{(2)} P(Y_l, \ldots, Y_T \,|\, X_l, \ldots, X_T) P(X_{l+1}, \ldots, X_T \,|\, X_t, \ldots, X_l) P(X_t, \ldots, X_l), \end{aligned}$$

pela propriedade de Markov,

$$\begin{aligned} P(Y_l, \ldots, Y_T \,|\, X_t, \ldots, X_l) &= \sum\nolimits^{(2)} P(Y_l, \ldots, Y_T \,|\, X_l, \ldots, X_T) P(X_{l+1}, \ldots, X_T \,|\, X_l) \\ &= \frac{1}{P(X_l)} \sum\nolimits^{(2)} P(Y_l, \ldots, Y_T \,|\, X_l, \ldots, X_T) P(X_l, X_{l+1}, \ldots, X_T) \\ &= \frac{1}{P(X_l)} \sum\nolimits^{(2)} P(Y_l, \ldots, Y_T, X_l, \ldots, X_T) \\ &= \frac{1}{P(X_l)} P(Y_l, \ldots, Y_T, X_l) = P(Y_l, \ldots, Y_T \,|\, X_l). \end{aligned}$$

Agora estamos com todas as ferramentas necessárias para respondermos a primeira pergunta levantada na introdução, ou seja, calcular a probabilidade da seqüência observada de um dado modelo de Markov oculto a tempo discreto. Os resultados seguintes

nos fornecem duas formas de encontrarmos essa probabilidade. O primeiro consiste no cálculo das probabilidades diretas e o segundo no cálculo das probabilidades reversas. Ambos conhecidos na literatura como algoritmo *forward* (para frente) e *backward* (para trás), respectivamente. A função de verossimilhança dos dados observados é usualmente avaliada por esses procedimentos.

Definimos as *probabilidades diretas* e as *probabilidades reversas* para todos os estados da cadeia de Markov e todo $t = 1, \ldots, T$ como segue,

Definição 1.4 *Seja* $\mathbf{Y} = (Y_1, \ldots, Y_t, \ldots, Y_T)$ *a seqüência de observações do modelo* $\lambda = (\delta, \Gamma, \Pi)$. *Para todo* $t \in \{1, \ldots, T\}$ *e* $i \in E = \{1, \ldots, m\}$ *definimos,*

$$\alpha_t(i) = P(Y_1 = y_1, \ldots, Y_t = y_t, X_t = i) \tag{1.1}$$

e

$$\beta_t(i) = P(Y_{t+1} = y_{t+1}, \ldots, Y_T = y_T \,|\, X_t = i). \tag{1.2}$$

Quando $t = T$ convencionamos que $\beta_T(i) = 1$ para todo $i \in E$.

Proposição 1.8 *Seja* $\mathbf{Y} = (Y_1, \ldots, Y_T)$ *a seqüência de observações do modelo* $\lambda = (\delta, \Gamma, \Pi)$ *então*

$$P(\mathbf{Y} \,|\, \lambda) = \sum_{i=1}^{m} \alpha_t(i)\beta_t(i), \qquad 1 \leq t \leq T. \tag{1.3}$$

Demonstração. Da Definição 1.4 e da Proposição 1.4, para $t = 1, 2, \ldots, T$ e $i \in E$ temos,

$$\begin{aligned}
\alpha_t(i)\beta_t(i) &= P(Y_1, \ldots, Y_t, X_t = i)P(Y_{t+1}, \ldots, Y_T \,|\, X_t = i) \\
&= P(Y_1, \ldots, Y_t \,|\, X_t = i)P(X_t = i)P(Y_{t+1}, \ldots, Y_T \,|\, X_t = i) \\
&= P(Y_1, \ldots, Y_t \,|\, X_t = i)P(Y_{t+1}, \ldots, Y_T \,|\, X_t = i)P(X_t = i) \\
&= P(Y_1, \ldots, Y_T \,|\, X_t = i)P(X_t = i) \\
&= P(Y_1, \ldots, Y_T, X_t = i).
\end{aligned} \tag{1.4}$$

Então

$$\begin{aligned} P(\mathbf{Y} \mid \lambda) &= P(Y_1 = y_1, \ldots, Y_T = y_T) \\ &= P\left(Y_1, \ldots, Y_T, \bigcup_{i=1}^{m} \{X_t = i\}\right) \\ &= \sum_{i=1}^{m} P(Y_1, \ldots, Y_T, X_t = i) \\ &= \sum_{i=1}^{m} \alpha_t(i)\beta_t(i). \end{aligned}$$

Esse resultado nos mostra uma maneira eficiente de calcularmos a probabilidade da seqüência de observações, para isso basta que as probabilidades diretas e as probabilidades reversas, em um determinado instante de tempo t, sejam conhecidas. Como a equação (1.3) é válida para todo $t \in \{1, \ldots, T\}$, em particular, para $t = T$ temos,

$$P(\mathbf{Y} \mid \lambda) = \sum_{i=1}^{m} \alpha_T(i)\beta_T(i) = \sum_{i=1}^{m} \alpha_T(i), \tag{1.5}$$

uma vez que $\beta_T(i) = 1$ para todo $i \in E$.

Portanto, basta determinarmos $\alpha_T(i)$, $i \in E$. O resultado seguinte nos dá uma maneira recursiva de calcularmos essas probabilidades.

Teorema 1.1 (Cálculo das probabilidades diretas) *Seja* $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_T)$ *a seqüência observável do modelo* $\lambda = (\delta, \Gamma, \Pi)$. *Então*

$$\alpha_1(i) = \pi_{y_1,i}\delta_i, \qquad 1 \leq i \leq m$$

e para $1 \leq j \leq m$ *e* $1 \leq t \leq T - 1$,

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{m} \alpha_t(i)\gamma_{ij}\right) \pi_{y_{t+1},i}.$$

Demonstração. Fazendo $t = 1$ em (1.1), para todo $1 \leq i \leq m$ obtemos,

$$\alpha_1(i) = P(Y_1 = y_1, X_1 = i) = P(Y_1 = y_1 \mid X_1 = i)P(X_1 = i) = \pi_{y_1,i}\delta_i.$$

Os demais valores são determinados de forma recursiva, válidos para $1 \leq t \leq T-1$. Usando a Proposição 1.6,

$$\begin{aligned}
\alpha_{t+1}(j) &= \sum_{i=1}^{m} P(Y_1, \ldots, Y_{t+1}, X_t = i, X_{t+1} = j) \\
&= \sum_{i=1}^{m} P(Y_1, \ldots, Y_{t+1} \,|\, X_t = i, X_{t+1} = j) P(X_t = i, X_{t+1} = j) \\
&= \sum_{i=1}^{m} P(Y_1, \ldots, Y_t \,|\, X_t = i) P(Y_{t+1} \,|\, X_{t+1} = j) P(X_t = i, X_{t+1} = j) \\
&= \sum_{i=1}^{m} P(Y_1, \ldots, Y_t \,|\, X_t = i) P(Y_{t+1} \,|\, X_{t+1} = j) P(X_t = i) \underbrace{P(X_{t+1} = j \,|\, X_t = i)}_{\gamma_{ij}} \\
&= \sum_{i=1}^{m} P(Y_1, \ldots, Y_t \,|\, X_t = i) \underbrace{P(Y_{t+1} \,|\, X_{t+1} = j)}_{\pi_{y_{t+1},j}} P(X_t = i) \gamma_{ij} \\
&= \sum_{i=1}^{m} P(Y_1, \ldots, Y_t \,|\, X_t = i) P(X_t = i) \pi_{y_{t+1},j} \gamma_{ij} \\
&= \pi_{y_{t+1},j} \sum_{i=1}^{m} P(Y_1, \ldots, Y_t, X_t = i) \gamma_{ij} \\
&= \left(\sum_{i=1}^{m} \alpha_t(i) \gamma_{ij} \right) \pi_{y_{t+1},j}.
\end{aligned}$$

A estrutura do algoritmo *forward* pode ser resumida da seguinte forma:

1° Passo: Inicialização:

$$\alpha_1(i) = \pi_{y_1,i} \delta_i, \qquad 1 \leq i \leq m.$$

2° Passo: Indução:

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{m} \alpha_t(i) \gamma_{ij} \right) \pi_{y_{t+1},j}, \qquad 1 \leq j \leq m \quad \text{e} \quad 1 \leq t \leq T-1.$$

3° Passo: Finalização:

$$P(\mathbf{Y} \,|\, \lambda) = \sum_{j=1}^{m} \alpha_T(j).$$

No primeiro passo, a probabilidade direta é iniciada como a probabilidade conjunta do estado i e da observação inicial Y_1. O passo de indução, a etapa mais importante do cálculo, é ilustrado na Figura 1.1. Esta figura mostra que o estado j é alcançado no instante $t+1$ a partir de qualquer um dos m estados possíveis no tempo t. Como $\alpha_t(i)$ é a probabilidade de $Y_1, \ldots, Y_t$ ser observado e o estado ser i no tempo t, o produto $\alpha_t(i)\gamma_{ij}$ é a probabilidade de $Y_1, \ldots, Y_t$ ser observado e alcançarmos o estado j no tempo $t+1$ estando no estado i no passo anterior. Somando estes produtos para todos os m estados possíveis no tempo t obtém-se a probabilidade de estarmos em j no tempo $t+1$ com todas as observações parciais anteriores.

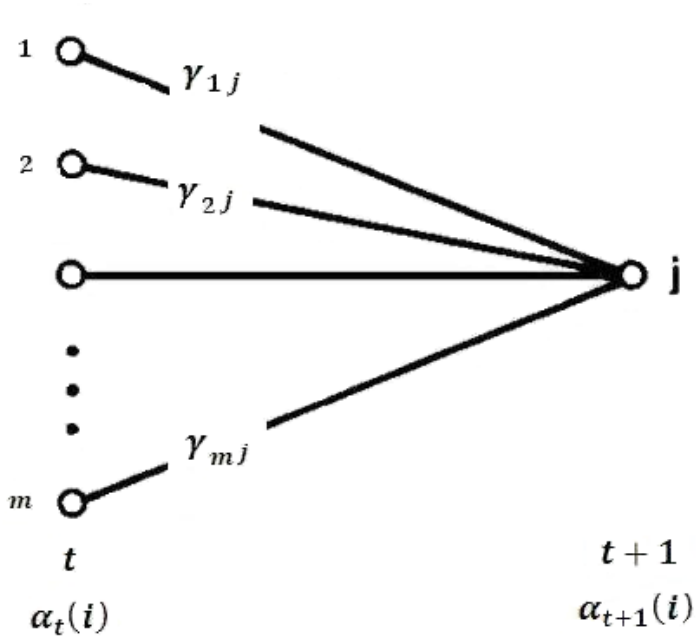

Figura 1.1: Ilustração do número de operações necessárias para o cálculo das probabilidades diretas.

Tendo calculado esses valores, para obter-se $\alpha_{t+1}(j)$ basta multiplicarmos por $\pi_{y_{t+1},j}$. O cálculo é feito para todos os estados j, $1 \leq j \leq m$ e para todo $1 \leq t \leq T-1$. O cálculo das probabilidades diretas é ilustrado na Figura 1.2:

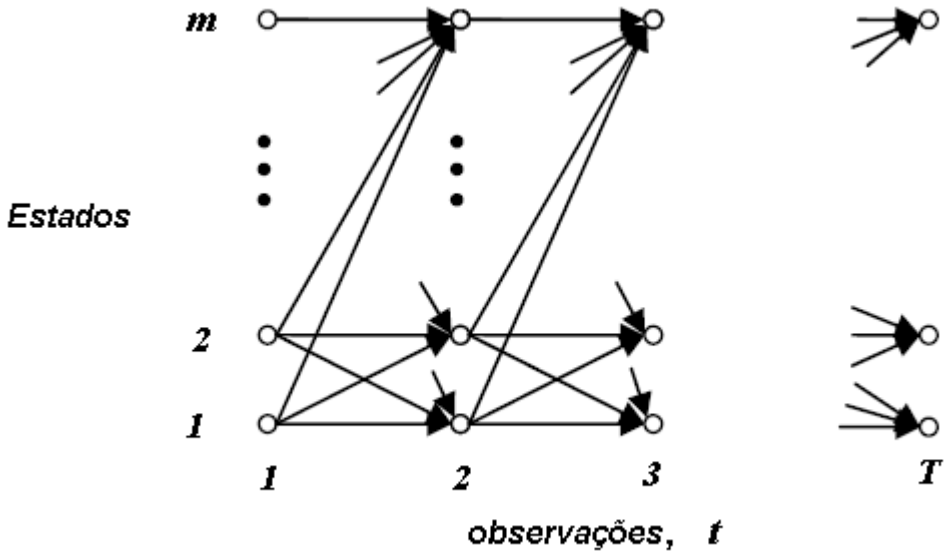

Figura 1.2: Procedimento do cálculo de α_{t+1}.

Uma outra maneira de obtermos a probabilidade da seqüência de observações é a partir do cálculo das probabilidades reversas.

Proposição 1.9 *Seja* $\mathbf{Y} = (Y_1, \ldots, Y_T)$ *a seqüência observável do modelo* $\lambda = (\delta, \Gamma, \Pi)$. *Então,*

$$P(\mathbf{Y} \mid \lambda) = \sum_{j=1}^{m} \pi_{y_1,j} \beta_1(j) \delta_j.$$

Demonstração. Fazendo $t = 1$ na Proposição 1.5 vemos que,

$$\begin{aligned} P(\mathbf{Y} \mid \lambda) &= P(Y_1 = y_1, \ldots, Y_T = y_T) = \sum_{j=1}^{m} P(Y_1, \ldots, Y_T, X_1 = j) \\ &= \sum_{j=1}^{m} P(Y_1, \ldots, Y_T \mid X_1 = j) P(X_1 = j) \\ &= \sum_{j=1}^{m} \underbrace{P(Y_1 \mid X_1 = j)}_{\pi_{y_1,j}} \underbrace{P(Y_2, \ldots, Y_T \mid X_1 = j)}_{\beta_1(j)} \underbrace{P(X_1 = j)}_{\delta_j} \\ &= \sum_{j=1}^{m} \pi_{y_1,j} \beta_1(j) \delta_j. \end{aligned}$$

Portanto, conhecendo δ_j, $\pi_{y_1,j}$ e $\beta_1(j)$, para todo $j \in E$, determinamos $P(\mathbf{Y} \mid \lambda)$. Como δ_j e $\pi_{y_1,j}$ são dados, basta calcularmos $\beta_1(j)$ para todo $j \in E$. O Teorema 1.2 nos mostra uma maneira recursiva para o cálculo das probabilidades reversas.

Teorema 1.2 (Cálculo das probabilidades reversas) *Seja* $\mathbf{Y} = (Y_1, \ldots, Y_T)$ *a seqüência observável do modelo* $\lambda = (\delta, \Gamma, \Pi)$. *Então para* $1 \leq i \leq m$ *e* $1 \leq t \leq T-1$,

$$\beta_t(i) = \sum_{j=1}^{m} \pi_{y_{t+1},j} \beta_{t+1}(j) \gamma_{ij}.$$

Demonstração. Usando as Proposições 1.7 (com $l = t+1$) e 1.5 temos,

$$\begin{aligned}
\beta_t(i) &= \sum_{j=1}^{m} P(Y_{t+1}, \ldots, Y_T, X_{t+1} = j \mid X_t = i) \\
&= \sum_{j=1}^{m} \frac{P(Y_{t+1}, \ldots, Y_T, X_t = i, X_{t+1} = j)}{P(X_t = i)} \\
&= \sum_{j=1}^{m} \frac{P(Y_{t+1}, \ldots, Y_T \mid X_t = i, X_{t+1} = j) P(X_t = i, X_{t+1} = j)}{P(X_t = i)} \\
&= \sum_{j=1}^{m} \frac{P(Y_{t+1}, \ldots, Y_T \mid X_{t+1} = j) P(X_t = i, X_{t+1} = j)}{P(X_t = i)} \\
&= \sum_{j=1}^{m} \frac{P(Y_{t+1}, \ldots, Y_T \mid X_{t+1} = j) P(X_{t+1} = j \mid X_t = i) P(X_t = i)}{P(X_t = i)} \\
&= \sum_{j=1}^{m} \underbrace{P(Y_{t+1} \mid X_{t+1} = j)}_{\pi_{y_{t+1},j}} \underbrace{P(Y_{t+2}, \ldots, Y_T \mid X_{t+1} = j)}_{\beta_{t+1}(j)} \gamma_{ij} \\
&= \sum_{j=1}^{m} \pi_{y_{t+1},j} \beta_{t+1}(j) \gamma_{ij}.
\end{aligned}$$

A estrutura do algoritmo *backward* pode ser resumida da seguinte forma:

1° Passo: Inicialização:

$$\beta_T(i) = 1, \qquad 1 \leq i \leq m.$$

2° Passo: Indução:

$$\beta_t(i) = \sum_{j=1}^{m} \pi_{y_{t+1},j} \beta_{t+1}(j) \gamma_{ij}, \quad 1 \leq i \leq m \quad \text{e} \quad 1 \leq t \leq T-1.$$

3° Passo: Finalização:

$$P(\mathbf{Y} \mid \lambda) = \sum_{j=1}^{m} \pi_{y_1,j} \beta_1(j) \delta_j.$$

A inicialização do procedimento é dada por $\beta_T(i) = 1$ para todos os valores de $i \in E$. No passo de indução, dada a seqüência de observações $Y_{t+1}, \ldots, Y_T$, o estado i é alcançado considerando que o sistema modelado esteja em qualquer um dos j estados possíveis. Este procedimento é realizado levando em conta as probabilidades de transição do estado i para o estado j (γ_{ij}, $1 \leq j \leq m$), a observação Y_{t+1} do estado j ($\pi_{y_{t+1},j}$), e a seqüência de observações restantes a partir do estado j ($\beta_{t+1}(j)$). O cálculo de $\beta_t(i)$ é realizado para todo $1 \leq t \leq T-1$ e $1 \leq i \leq m$. A ilustração do cálculo das probabilidades reversas é mostrada na Figura 1.3:

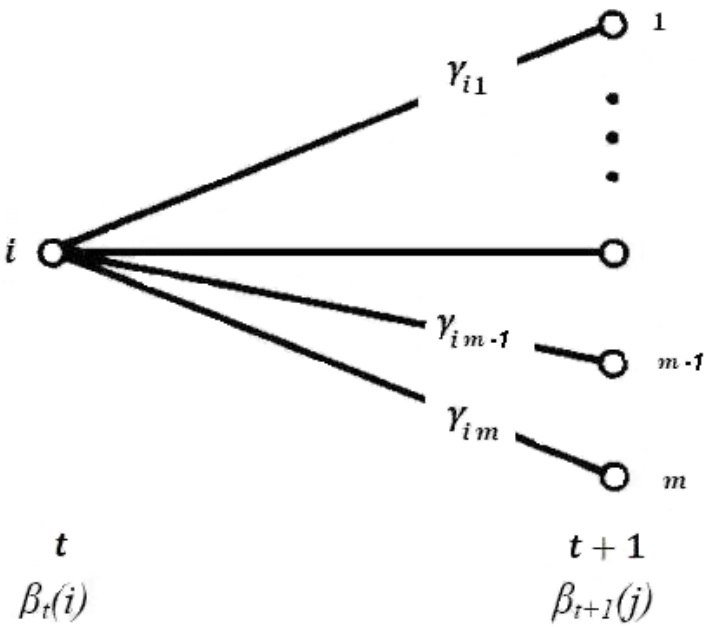

Figura 1.3: Ilustração do número de operações necessárias para o cálculo das probabilidades reversas.

Os dois algoritmos acima nos fornecem resultados iguais. Portanto, para calcularmos a probabilidade da seqüência de observações basta utilizarmos um dos procedimentos (*forward* ou *backward*). Enquanto o algoritmo *forward* calcula a $P(\mathbf{Y} \mid \lambda)$ utilizando as

probabilidades diretas ($\alpha_t(i)$), o algoritmo *backward* calcula a mesma quantidade utilizando as probabilidades reversas ($\beta_t(i)$). Daí a terminologia dos algoritmos.

Os resultados obtidos também podem ser representados na forma matricial. As matrizes:

$$A = (\alpha_{ti})_{T\times m} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{T1} & \alpha_{T2} & \cdots & \alpha_{Tm} \end{bmatrix} \quad \text{e} \quad B = (\beta_{ti})_{T\times m} = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{T1} & \beta_{T2} & \cdots & \beta_{Tm} \end{bmatrix},$$

são denominadas matrizes das probabilidades diretas e reversas respectivamente.

Se definirmos os vetores,

$$\alpha_t = (\alpha_{t1}, \alpha_{t2}, \ldots, \alpha_{tm}) \qquad \text{e} \qquad \beta_t = (\beta_{t1}, \beta_{t2}, \ldots, \beta_{tm}), \tag{1.6}$$

podemos reescrever (1.3) como,

$$P(\mathbf{Y} \mid \lambda) = \alpha_t \left[\beta_t\right]^t. \tag{1.7}$$

Observemos que (1.6) são linhas das matrizes A e B respectivamente. Em outras palavras, basta escolhermos uma das t linhas da matriz A e multiplicarmos pela transposta da linha t da matriz B. Para fixarmos as idéias, vamos considerar o seguinte exemplo:

Exemplo 1.1 *Seja $\lambda = (\delta, \Gamma, \Pi)$ um modelo de Markov oculto com espaço de estados $\{1, 2, 3\}$ e a seguinte seqüência observável $\mathbf{Y} = (a\, a\, b\, b)$ com*

$$\delta = \begin{bmatrix} 0.6 \\ 0.4 \\ 0 \end{bmatrix}, \qquad \Gamma = \begin{bmatrix} 0.3 & 0.5 & 0.2 \\ 0 & 0.3 & 0.7 \\ 0 & 0 & 1 \end{bmatrix} \qquad \text{e} \qquad \Pi = \begin{bmatrix} 1 & 0.5 & 0 \\ 0 & 0.5 & 1 \end{bmatrix}.$$

1. *Utilizando os algoritmos forward e backward como calculamos as probabilidades diretas e reversas?*

2. Qual a probabilidade da seqüência de observações?

Solução:

Cálculo das probabilidades diretas

1° Passo: Inicialização:

$$\alpha_1(1) = P(Y_1 = y_1, X_1 = 1) = P(Y_1 = y_1 \,|\, X_1 = 1)P(X_1 = 1) = \pi_{y_1,1}\delta_1 = 1 \cdot 0.6 = 0.6.$$

$$\alpha_1(2) = P(Y_1 = y_1, X_1 = 2) = P(Y_1 = y_1 \,|\, X_1 = 2)P(X_1 = 2) = \pi_{y_1,2}\delta_2 = 0.5 \cdot 0.4 = 0.2.$$

$$\alpha_1(3) = P(Y_1 = y_1, X_1 = 3) = P(Y_1 = y_1 \,|\, X_1 = 3)P(X_1 = 3) = \pi_{y_1,3}\delta_3 = 0 \cdot 0 = 0.$$

2° Passo: Indução:

$$\alpha_2(1) = \left(\sum_{i=1}^{3} \alpha_1(i)\gamma_{i1}\right)\pi_{y_2,1} = (0.6 \cdot 0.3 + 0.2 \cdot 0 + 0 \cdot 0) \cdot 1 = 0.18.$$

$$\alpha_2(2) = \left(\sum_{i=1}^{3} \alpha_1(i)\gamma_{i2}\right)\pi_{y_2,2} = (0.6 \cdot 0.5 + 0.2 \cdot 0.3 + 0 \cdot 0) \cdot 0.5 = 0.18.$$

$$\alpha_2(3) = \left(\sum_{i=1}^{3} \alpha_1(i)\gamma_{i3}\right)\pi_{y_2,3} = (0.6 \cdot 0.2 + 0.2 \cdot 0.7 + 0 \cdot 1) \cdot 0 = 0.$$

$$\alpha_3(1) = \left(\sum_{i=1}^{3} \alpha_2(i)\gamma_{i1}\right)\pi_{y_3,1} = (0.18 \cdot 0.3 + 0.18 \cdot 0 + 0 \cdot 0) \cdot 0 = 0.$$

$$\alpha_3(2) = \left(\sum_{i=1}^{3} \alpha_2(i)\gamma_{i2}\right)\pi_{y_3,2} = (0.18 \cdot 0.5 + 0.18 \cdot 0.3 + 0 \cdot 0) \cdot 0.5 = 0.072.$$

$$\alpha_3(3) = \left(\sum_{i=1}^{3} \alpha_2(i)\gamma_{i3}\right)\pi_{y_3,3} = (0.18 \cdot 0.2 + 0.18 \cdot 0.7 + 0 \cdot 1) \cdot 1 = 0.162.$$

$$\alpha_4(1) = \left(\sum_{i=1}^{3} \alpha_3(i)\gamma_{i1}\right)\pi_{y_4,1} = (0 \cdot 0.3 + 0.072 \cdot 0 + 0.162 \cdot 0) \cdot 0 = 0.$$

$$\alpha_4(2) = \left(\sum_{i=1}^{3} \alpha_3(i)\gamma_{i2}\right)\pi_{y_4,2} = (0 \cdot 0.5 + 0.072 \cdot 0.3 + 0.162 \cdot 0) \cdot 0.5 = 0.0108.$$

$$\alpha_4(3) = \left(\sum_{i=1}^{3} \alpha_3(i)\gamma_{i3}\right)\pi_{y_4,3} = (0 \cdot 0.2 + 0.072 \cdot 0.7 + 0.162 \cdot 1) \cdot 1 = 0.2124.$$

Cálculo das probabilidades reversas

1° Passo: Inicialização:

$$\beta_4(1) = 1, \quad \beta_4(2) = 1, \quad \beta_4(3) = 1.$$

2° Passo: Indução:

$$\begin{aligned}
\beta_3(1) &= \sum_{j=1}^{3} \pi_{y_4,j}\beta_4(j)\gamma_{1j} = 0 \cdot 1 \cdot 0.3 + 0.5 \cdot 1 \cdot 0.5 + 1 \cdot 1 \cdot 0.2 = 0.45. \\
\beta_3(2) &= \sum_{j=1}^{3} \pi_{y_4,j}\beta_4(j)\gamma_{2j} = 0 \cdot 1 \cdot 0 + 0.5 \cdot 1 \cdot 0.3 + 1 \cdot 1 \cdot 0.7 = 0.85. \\
\beta_3(3) &= \sum_{j=1}^{3} \pi_{y_4,j}\beta_4(j)\gamma_{3j} = 0 \cdot 1 \cdot 0 + 0.5 \cdot 1 \cdot 0 + 1 \cdot 1 \cdot 1 = 1. \\
\beta_2(1) &= \sum_{j=1}^{3} \pi_{y_3,j}\beta_3(j)\gamma_{1j} = 0 \cdot 0.45 \cdot 0.3 + 0.5 \cdot 0.85 \cdot 0.5 + 1 \cdot 1 \cdot 0.2 = 0.4125. \\
\beta_2(2) &= \sum_{j=1}^{3} \pi_{y_3,j}\beta_3(j)\gamma_{2j} = 0 \cdot 0.45 \cdot 0 + 0.5 \cdot 0.85 \cdot 0.3 + 1 \cdot 1 \cdot 0.7 = 0.8275. \\
\beta_2(3) &= \sum_{j=1}^{3} \pi_{y_3,j}\beta_3(j)\gamma_{3j} = 0 \cdot 0.45 \cdot 0 + 0.5 \cdot 0.85 \cdot 0 + 1 \cdot 1 \cdot 1 = 1. \\
\beta_1(1) &= \sum_{j=1}^{3} \pi_{y_2,j}\beta_2(j)\gamma_{1j} = 1 \cdot 0.4125 \cdot 0.3 + 0.5 \cdot 0.8275 \cdot 0.5 + 0 \cdot 1 \cdot 0.2 = 0.330625. \\
\beta_1(2) &= \sum_{j=1}^{3} \pi_{y_2,j}\beta_2(j)\gamma_{2j} = 1 \cdot 0.4125 \cdot 0 + 0.5 \cdot 0.8275 \cdot 0.3 + 0 \cdot 1 \cdot 0.7 = 0.124125. \\
\beta_1(3) &= \sum_{j=1}^{3} \pi_{y_2,j}\beta_2(j)\gamma_{3j} = 1 \cdot 0.4125 \cdot 0 + 0.5 \cdot 0.8275 \cdot 0 + 0 \cdot 1 \cdot 1 = 0.
\end{aligned}$$

Portanto, obtemos as matrizes formadas pelas probabilidades diretas e reversas, respectivamente:

$$A = \begin{bmatrix} 0.6 & 0.2 & 0 \\ 0.18 & 0.18 & 0 \\ 0 & 0.072 & 0.162 \\ 0 & 0.0108 & 0.2124 \end{bmatrix} \quad \text{e} \quad B = \begin{bmatrix} 0.330625 & 0.124125 & 0 \\ 0.4125 & 0.8275 & 1 \\ 0.45 & 0.85 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

O cálculo da probabilidade da seqüência de observações dado pelos procedimentos *forward* e *backward* é, respectivamente,

$$P(\mathbf{Y} \mid \lambda) = \sum_{i=1}^{3} \alpha_T(i) = 0 + 0.0108 + 0.2124 = 0.2232$$

e

$$P(\mathbf{Y} \mid \lambda) = \sum_{i=1}^{3} \pi_{y_1,j} \beta_1(j) \delta_j = 1 \times 0.330625 \times 0.6 + 0.5 \times 0.124125 \times 0.4 + 0 \times 0 \times 0 = 0.2232.$$

Também poderíamos calcular $P(\mathbf{Y} \mid \lambda)$ escolhendo qualquer linha da matriz A, e de forma correspondente, da matriz B e aplicarmos (1.7). Por exemplo, vamos considerar a terceira linha,

$$\alpha_3 = (0, 0.072, 0.162) \qquad \text{e} \qquad \beta_3 = (0.45, 0.85, 1),$$

logo,

$$P(\mathbf{Y} \mid \lambda) = \alpha_3 \left[\beta_3\right]^t = 0 \times 0.45 + 0.072 \times 0.85 + 0.162 \times 1 = 0.2232.$$

Com esse exemplo observamos alguns aspectos importantes. Primeiro, ambos os procedimentos geram resultados iguais. Segundo, o cálculo manual das probabilidades é enfadonho, por esse motivo sugerimos o uso de softwares. Neste caso particular, implementamos no *software* R, versão 2.7.0, algoritmos específicos. No entanto, já existem pacotes, por exemplo `HiddenMarkov`, com as funções `forward` e `backward` implementadas. Vejamos como utilizá-lo para resolver, computacionalmente, o exemplo que acabamos de fazer.

```
# Estados da Cadeia: E={1,2,3}
# Observáveis: {a,b}
# Sequência de Observações: {aabb}
# Cálculo das probabilidades diretas e reversas
# Distribuição inicial da Cadeia de Markov
(delta <- matrix(c(0.6,0.4, 0), nrow = 3, ncol = 1, byrow = TRUE))
```

```
     [,1]
[1,]  0.6
[2,]  0.4
[3,]  0.0
  # Matriz de Transição da Cadeia de Markov:
  (gama <- matrix(c(0.3,0.5,0.2, 0,0.3,0.7, 0,0,1), nrow = 3,
   ncol = 3, byrow = TRUE))
     [,1] [,2] [,3]
[1,]  0.3  0.5  0.2
[2,]  0.0  0.3  0.7
[3,]  0.0  0.0  1.0
  # Matriz de Probabilidades do Observado dado o Oculto: P{Y(t)|X(t)}
  (pi <- matrix(c(1,0.5,0, 0,0.5,1), nrow = 2, ncol = 3, byrow = TRUE))
     [,1] [,2] [,3]
[1,]    1  0.5    0
[2,]    0  0.5    1
 # Dado os estados ocultos, temos dois possíveis resultados para o
 observado {a,b}. Podemos supor que a distribuição do observado
 dado o oculto é uma binomial, ou seja,
 # P{obs | estado oculto = 1} = Binomial(n=1,p=1)
 # P{obs | estado oculto = 2} = Binomial(n=1,p=0.5)
 # P{obs | estado oculto = 3} = Binomial(n=1,p=0)
  obs <- c(1,1,0,0) # sequência observável
  pn <- list(size=rpois(4, 10))
  (x <- dthmm(obs, gama, delta, "binom", list(prob=c(1, 0.5, 0)),
  pn, discrete=TRUE))
$x
[1] 1 1 0 0
$Pi
```

```
     [,1] [,2] [,3]
[1,]  0.3  0.5  0.2
[2,]  0.0  0.3  0.7
[3,]  0.0  0.0  1.0
$delta
     [,1]
[1,]  0.6
[2,]  0.4
[3,]  0.0
$distn
[1] "binom"
$pm
$pm$prob
[1] 1.0 0.5 0.0
$pn
$pn$size
[1]  8 16 11 10
$discrete
[1] TRUE
$nonstat
[1] TRUE
attr(,"class")
[1] "dthmm"
  (yfb <- \textit{forward}back(x$x, gama, delta, "binom", list(c(1,1,1),
  prob=c(1, 0.5, 0))))
```

Matriz das probabilidades diretas

```
  exp(yfb$logalpha)
     [,1]   [,2]   [,3]
```

```
[1,] 0.60 0.2000 0.0000
[2,] 0.18 0.1800 0.0000
[3,] 0.00 0.0720 0.1620
[4,] 0.00 0.0108 0.2124
```

Matriz das probabilidades reversas

```
  exp(yfb$logbeta)
        [,1]     [,2] [,3]
[1,] 0.330625 0.124125    0
[2,] 0.412500 0.827500    1
[3,] 0.450000 0.850000    1
[4,] 1.000000 1.000000    1
```

Cálculo da probabilidade da sequência de observações

```
  exp(yfb$LL)
[1] 0.2232
```

Um outro problema também levantado na introdução do trabalho está relacionado com os parâmetros do modelo. Estamos interessados nos valores de $\lambda = (\delta, \Gamma, \Pi)$ para os quais a probabilidade da seqüência de observações seja máxima. Em outras palavras, quais são os parâmetros do modelo tais que $P(\mathbf{Y} \mid \lambda)$ seja máxima?

Nas seções seguintes iremos abordar esse problema.

1.3 Algoritmo EM

Queremos determinar os estimadores de máxima verossimilhança de um modelo de Markov oculto a tempo discreto. Para isso, utilizamos a metodologia do algoritmo EM na

obtenção dos estimadores. Na seção 1.3.1 apresentamos o algoritmo EM e os resultados que garantem a sua convergência e na seção 1.3.2 estudamos essa metodologia aplicada a cadeias de Markov ocultas.

1.3.1 O algoritmo EM

Considere uma família $\{f(x\,;\,\theta)\}_{\theta\in\Theta}$ de funções não negativas e integráveis. Esta família é indexada pelo parâmetro $\theta \in \Theta \subseteq \mathbb{R}^d$. Nosso objetivo é maximizar a função

$$L(\theta) = \int f(x\,;\,\theta)dx, \tag{1.8}$$

com respeito ao parâmetro θ. No nosso caso consideramos f a densidade conjunta de duas variáveis aleatórias Y (observável) e X (não observável). A variável aleatória X corresponde aos dados ocultos, f à função de verossimilhança dos dados completos e L à densidade de Y.

Suponhamos que $L(\theta)$ é positiva para todo $\theta \in \Theta$, então a maximização de $L(\theta)$ é equivalente a maximizar

$$l(\theta) = \log L(\theta). \tag{1.9}$$

Também associamos à função $f(\cdot\,;\,\theta)$ a densidade de probabilidade $p(\cdot\,;\,\theta)$ definida por,

$$p(x\,;\,\theta) = \frac{f(x\,;\,\theta)}{L(\theta)}. \tag{1.10}$$

O algoritmo EM (do inglês, *expectation-maximization*) é um dos métodos mais populares em problemas de otimização em que se busca um ótimo local. O artigo de Dempster et al. (1977) é referência neste método.

A idéia desenvolvida por Dempster et al. (1977), introduz uma função auxiliar, conforme a Definição 1.5,

Definição 1.5 *A quantidade intermediária do EM é uma família* $\{\mathcal{Q}(\cdot\,;\,\theta^{'})\}_{\theta^{'}\in\Theta}$ *de funções a valores reais em* Θ*, indexada por* $\theta^{'}$ *definida por,*

$$\mathcal{Q}(\theta\,;\,\theta^{'}) = \int \log f(x\,;\,\theta) p(x\,;\,\theta^{'}) dx. \tag{1.11}$$

A quantidade intermediária $\mathcal{Q}(\cdot\,;\,\theta^{'})$ pode ser interpretada como a esperança da variável aleatória $\log f(X\,;\,\theta)$ quando X tem uma função densidade $p(x\,;\,\theta^{'})$ indexada por um valor $\theta^{'}$ do parâmetro. Usando (1.9) e (1.10) podemos reescrever (1.11) como

$$\mathcal{Q}(\theta\,;\,\theta^{'}) = l(\theta) - \mathcal{H}(\theta\,;\,\theta^{'}), \tag{1.12}$$

sendo

$$\mathcal{H}(\theta\,;\,\theta^{'}) = -\int \log p(x\,;\,\theta) p(x\,;\,\theta^{'}) dx.$$

Condição 1.1 *Suponhamos que:*

(a) O conjunto Θ *é um subconjunto aberto de* $\mathbb{R}^d$*.*

(b) Para qualquer $\theta \in \Theta$*,* $L(\theta)$ *é uma função limitada.*

(c) Para qualquer $(\theta, \theta^{'}) \in \Theta \times \Theta$*,* $\int |\nabla \log p(x;\theta)|\, p(x;\theta^{'}) dx$ *é limitada.*

O resultado que justifica a construção do algoritmo EM é o seguinte:

Proposição 1.10 *Sob a condição 1.1, para qualquer* $(\theta\,,\,\theta^{'}) \in \Theta \times \Theta$*,*

$$l(\theta) - l(\theta^{'}) \geq \mathcal{Q}(\theta\,;\,\theta^{'}) - \mathcal{Q}(\theta^{'}\,;\,\theta^{'})$$

e se,

a) $\theta \mapsto L(\theta)$ *é continuamente diferenciável em* Θ*.*

b) Para qualquer $\theta^{'} \in \Theta$*,* $\theta \mapsto \mathcal{H}(\theta\,;\,\theta^{'})$ *é continuamente diferenciável em* Θ*.*

Então para qualquer $\theta^{'} \in \Theta$, $\theta \mapsto \mathcal{Q}(\theta\,;\,\theta^{'})$ *é continuamente diferenciável em* Θ *e*

$$\nabla l(\theta^{'}) = \nabla \mathcal{Q}(\theta\,;\,\theta^{'})|_{\theta=\theta'}$$

onde ∇ *denota o gradiente.*

O algoritmo proposto por Dempster et al. (1977) consiste na construção de uma seqüência de estimadores a partir de um valor inicial θ_0. Cada iteração é dividida em dois passos

Passo E: Determinar a função $\mathcal{Q}(\theta\,;\,\theta^{'})$;

Passo M: Escolher θ^{i+1} para ser o valor de $\theta \in \Theta$ que maximiza $\mathcal{Q}(\theta\,;\,\theta^{i})$.

De maneira simples, o algoritmo EM utiliza a quantidade intermediária como uma função suporte para a maximização de $l(\theta)$. Ambas não são necessariamente comparadas, mas de (1.12), podemos notar que qualquer valor de $\theta \in \Theta$ tal que,

$$\mathcal{Q}(\theta\,;\,\theta^{'}) \geq \mathcal{Q}(\theta^{'}\,;\,\theta^{'}) \quad \Rightarrow \quad l(\theta) \geq l(\theta^{'}) \tag{1.13}$$

A Proposição 1.10 garante dois resultados importantes na construção do algoritmo EM. Primeiro, para qualquer seqüência $\{\theta^{i}\}_{i\geq 0}$ tal que $\mathcal{Q}(\theta^{i+1}\,;\,\theta^{i}) \geq \mathcal{Q}(\theta^{i}\,;\,\theta^{i})$, a seqüência $\{l(\theta^{i})\}_{i\geq 0}$ é não decrescente. Então o EM é um algoritmo de otimização monótono. Segundo, se o processo de iteração parar no ponto θ^{*}, então $\mathcal{Q}(\theta\,;\,\theta^{*})$ tem um máximo em θ^{*} (caso contrário, ainda será possível encontrar um outro θ^{*}), e então θ^{*} é tal que $\nabla_{\theta} l(\theta_{*}) = 0$, e este valor será o ponto de estacionariedade da função de verossimilhança.

1.3.2 Ajustando o EM para HMMs

Agora retomamos ao nosso objetivo principal e discutiremos a aplicação do algoritmo EM para o caso específico em HMMs.

Os HMMs correspondem a uma categoria de modelos de dados incompletos conhecida como *modelos de dados omissos*. Nesses modelos, os dados observáveis $\mathbf{Y}$ são um subconjunto dos *dados completos* $(\mathbf{Y}, \mathbf{X})$. Aqui, assumimos que a distribuição conjunta de $\mathbf{X}$ e $\mathbf{Y}$, para um dado valor do parâmetro θ, admite uma função densidade $f(\mathbf{x}, \mathbf{y}\,;\,\theta)$. É importante lembrar que f é uma função densidade somente quando a consideramos como função de $\mathbf{x}$ e $\mathbf{y}$. Para um dado valor fixado $\mathbf{y}$, f é uma função positiva e integrável. Na verdade, a função de verossimilhança da seqüência observada, que é definida como a função densidade de $\mathbf{Y}$, é obtida pela marginal,

$$L(\mathbf{y}\,;\,\theta) = \int f(\mathbf{x}, \mathbf{y}\,;\,\theta)dx. \tag{1.14}$$

Para um dado valor de $\mathbf{y}$ este é um caso particular de (1.8). Em modelos de dados omissos, a família de funções densidades $\{p(\cdot\,;\,\theta)\}_{\theta\in\Theta}$, definida em (1.10), é interpretada como,

$$p(\mathbf{x}\,|\,\mathbf{y}\,;\,\theta) = \frac{f(\mathbf{x}, \mathbf{y}\,;\,\theta)}{\int f(\mathbf{x}, \mathbf{y}\,;\,\theta)dx},$$

que é função densidade condicional de $\mathbf{X}$ dado $\mathbf{Y}$.

1.4 Aplicando o EM em HMMs

Dada a amostra observada $\mathbf{y} = \{y_1, \ldots, y_T\}$, utilizamos o algoritmo EM para estimar os parâmetros $\lambda = (\delta, \Gamma, \Pi)$ de uma cadeia de Markov oculta. Esta estimação é tal que a cadeia de Markov oculta que tenha estes parâmetros é a que apresenta a maior probabilidade de ter gerado a amostra inicial.

Precisaremos definir algumas ferramentas auxiliares e para isso vamos relembrar o modelo. O modelo de Markov oculto é representado por um processo estocástico duplo $\{Y_t, X_t\}_{t\geq 0}$. O processo $\{Y_t\}_{t\in\mathbb{N}}$ é gerado por uma cadeia de Markov homogênea não observável com espaço de estados $E = \{1, \ldots, m\}$ e por um conjunto de distribuições de

probabilidades, uma para cada estado da cadeia que produzem as observações a partir de um conjunto de T possibilidades. Supomos ainda que a distribuição de Y_k só depende do correspondente X_k.

Definição 1.6 *Seja $\lambda = (\delta, \Gamma, \Pi)$ um modelo de Markov oculto. Definimos a função de verossimilhança da seqüência observada como,*

$$L(\lambda; \mathbf{Y}) = P(\mathbf{Y} \mid \lambda) = P(Y_1 = y_1, \ldots, Y_T = y_T \mid \lambda). \tag{1.15}$$

Queremos determinar $\widehat{\lambda} = (\widehat{\delta}, \widehat{\Gamma}, \widehat{\Pi})$ tal que $L(\lambda; \mathbf{Y})$ seja máxima. Inicialmente vamos definir:

Definição 1.7 *Seja $\mathbf{Y} = (Y_1, \ldots, Y_T)$ uma seqüência observável do modelo $\lambda = (\delta, \Gamma, \Pi)$. Para todo $i \in E$ definimos*

$$e_t(i) = P(X_t = i \mid \mathbf{Y}, \lambda), \qquad 1 \leq t \leq T \tag{1.16}$$

e

$$a_t(jk) = P(X_t = j, X_{t+1} = k \mid \mathbf{Y}, \lambda), \qquad 1 \leq t \leq T - 1. \tag{1.17}$$

A Proposição seguinte nos mostra uma maneira de calcularmos $e_t(i)$ e $a_t(jk)$ por meio das probabilidades diretas e reversas.

Proposição 1.11 *Considere $e_t(i)$ e $a_t(jk)$ definidos em (1.16) e (1.17). Então, para todo $i, j \in E$,*

$$e_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{m} \alpha_t(i)\beta_t(i)}, \qquad 1 \leq t \leq T \tag{1.18}$$

e

$$a_t(jk) = \frac{\alpha_t(j)\gamma_{jk}\pi_{y_{t+1},k}\beta_{t+1}(k)}{\sum_{j=1}^{m}\sum_{k=1}^{m} \alpha_t(j)\gamma_{jk}\pi_{y_{t+1},k}\beta_{t+1}(k)}, \qquad 1 \leq t \leq T - 1. \tag{1.19}$$

Demonstração. Da Proposição 1.3 e da equação (1.4) obtemos (1.18). De fato,

$$e_t(i) = P(X_t = i \,|\, \mathbf{Y}, \lambda) = \frac{P(\mathbf{Y}, X_t = i \,|\, \lambda)}{P(\mathbf{Y} \,|\, \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{m} \alpha_t(i)\beta_t(i)}.$$

Para demonstrarmos (1.19) escrevemos,

$$\begin{aligned} a_t(jk) &= P(X_t = j, X_{t+1} = k \,|\, \mathbf{Y}, \lambda) && (1.20)\\ &= \frac{P(\mathbf{Y}, X_t = j, X_{t+1} = k \,|\, \lambda)}{P(\mathbf{Y} \,|\, \lambda)} \\ &= \frac{P(\mathbf{Y}, X_t = j, X_{t+1} = k \,|\, \lambda)}{\sum_{j=1}^{m}\sum_{k=1}^{m} P(\mathbf{Y}, X_t = j, X_{t+1} = k \,|\, \lambda)} && (1.21) \end{aligned}$$

Aplicando as Proposições 1.5 e 1.6 podemos reescrever:

$$\begin{aligned} & P(\mathbf{Y}, X_t = j, X_{t+1} = k \,|\, \lambda) = P(Y_1, Y_2, \ldots, Y_t, Y_{t+1}, \ldots, Y_T, X_t = j, X_{t+1} = k \,|\, \lambda) \\ = \;& P(Y_1, Y_2, \ldots, Y_t, Y_{t+1}, \ldots, Y_T \,|\, X_t = j, X_{t+1} = k, \lambda) P(X_t = j, X_{t+1} = k \,|\, \lambda) \\ = \;& P(Y_1, Y_2, \ldots, Y_t \,|\, X_t = j, \lambda) P(Y_{t+1}, Y_{t+2} \ldots, Y_T \,|\, X_{t+1} = k, \lambda) P(X_{t+1} = k, X_t = j, \lambda) \\ = \;& \underbrace{P(Y_1, Y_2, \ldots, Y_t, X_t = j \,|\, \lambda)}_{\alpha_t(j)} \underbrace{P(Y_{t+1} \,|\, X_{t+1} = k, \lambda)}_{\pi_{y_{t+1},k}} \underbrace{P(Y_{t+2}, \ldots, Y_T \,|\, X_{t+1} = k, \lambda)}_{\beta_{t+1}(k)} \\ & \underbrace{P(X_{t+1} = k \,|\, X_t = j, \lambda)}_{\gamma_{jk}} \\ = \;& \alpha_t(j)\pi_{y_{t+1},k}\beta_{t+1}(k)\gamma_{jk}. && (1.22) \end{aligned}$$

Substituindo (1.22) em (1.20) temos,

$$a_t(jk) = \frac{\alpha_t(j)\gamma_{jk}\pi_{y_{t+1},k}\beta_{t+1}(k)}{\sum_{j=1}^{m}\sum_{k=1}^{m} \alpha_t(j)\gamma_{jk}\pi_{y_{t+1},k}\beta_{t+1}(k)}.$$

A Figura 1.4 ilustra graficamente como funciona o cálculo de uma transição usando as variáveis *forward* e *backward*. Para se ter uma transição do estado j no instante de tempo t para o estado k no tempo $t+1$, é necessário calcular a probabilidade de $Y_1, \ldots, Y_t$ ser observado e o estado ser j no tempo t ($\alpha_t(j)$). A seguir, calcula-se a probabilidade de transição de j para k (γ_{jk}) e finalmente a probabilidade de se emitir a observação y_{t+1} dado o estado k ($\pi_{y_{t+1},k}$). O termo $\beta_{t+1}(k)$ significa obter a seqüência parcial a partir do

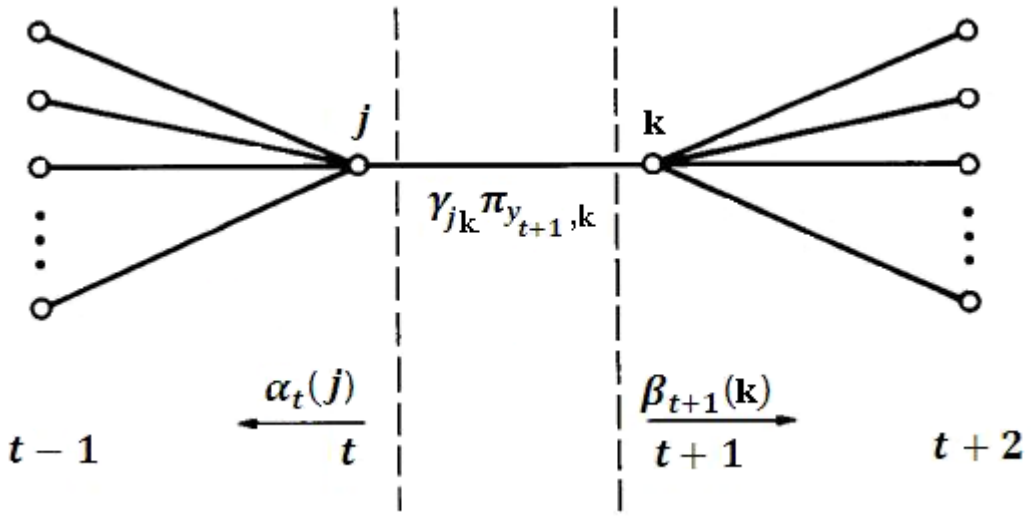

Figura 1.4: Cálculo da transição.

instante de tempo $t+1$ até o instante T dado que o processo de Markov encontra-se no estado j no instante de tempo t.

Uma vez definidas essas probabilidades, vamos estabelecer quais são os estimadores de máxima verossimilhança do modelo.

Teorema 1.3 *Seja $\lambda = (\delta, \Gamma, \Pi)$ um modelo de Markov oculto. Os estimadores de máxima verossimilhança de λ são dados por,*

$$\widehat{\delta}_i = e_1(i), \quad \widehat{\gamma}_{ij} = \frac{\sum_{t=1}^{T-1} a_t(ij)}{\sum_{t=1}^{T-1} e_t(i)} \quad e \quad \widehat{\pi}_{k,j} = \frac{\sum_{\{t;y_t=k\}}^{T} e_t(j)}{\sum_{t=1}^{T} e_t(j)}. \tag{1.23}$$

Demonstração. Defina,

$$\mathcal{Q}(\widehat{\lambda}, \lambda) = E\left\{\left[\log P(\mathbf{Y}, \mathbf{X} \,|\, \widehat{\lambda})\right] \,|\, \mathbf{Y}, \lambda\right\}. \tag{1.24}$$

Como $\{X_t; t \in \mathbb{N}\}$ é um processo estocástico assumindo valores em um conjunto finito, podemos escrever (1.24) como,

$$\mathcal{Q}(\widehat{\lambda}, \lambda) = \sum_{\mathbf{i} \in \aleph} \log P(\mathbf{Y}, \mathbf{X} = \mathbf{i} \,|\, \widehat{\lambda}) P(\mathbf{X} = \mathbf{i} \,|\, \mathbf{Y}, \lambda),$$

sendo $\aleph$ o conjunto de todas as seqüências de estados de comprimento T. Logo,

$$\mathcal{Q}(\widehat{\lambda}, \lambda) = \sum_{\mathbf{i} \in \aleph} \left[\frac{P(\mathbf{Y}, \mathbf{X} = \mathbf{i} \,|\, \lambda)}{P(\mathbf{Y} \,|\, \lambda)} \right] \log P(\mathbf{Y}, \mathbf{X} = \mathbf{i} \,|\, \widehat{\lambda}), \tag{1.25}$$

e como

$$\begin{aligned}
P(\mathbf{Y}, \mathbf{X} = \mathbf{i} \,|\, \lambda) &= P(Y_1, Y_2, \ldots, Y_T \,|\, X_1, X_2, \ldots, X_T, \lambda) P(X_1, X_2, \ldots, X_T \,|\, \lambda) \\
&= P(Y_1 \,|\, X_1, \lambda) \ldots P(Y_{T-1} \,|\, X_{T-1}, \lambda) P(Y_T \,|\, X_T, \lambda) P(X_1, X_2, \ldots, X_T \,|\, \lambda) \\
&= \pi_{y_1, i_1} \ldots \pi_{y_{T-1}, i_{T-1}} \pi_{y_T, i_T} P(X_1, X_2, \ldots, X_T \,|\, \lambda) \\
&= \left[\prod_{t=1}^{T} \pi_{y_t, i_t} \right] P(X_T \,|\, X_1, X_2, \ldots, X_{T-1}, \lambda) P(X_1, X_2, \ldots, X_{T-1} \,|\, \lambda) \\
&= \left[\prod_{t=1}^{T} \pi_{y_t, i_t} \right] P(X_T \,|\, X_{T-1}, \lambda) P(X_{T-1} \,|\, X_1, \ldots, X_{T-2} \,|\, \lambda) \\
&\quad P(X_1, X_2, \ldots, X_{T-2} \,|\, \lambda) \\
&= \left[\prod_{t=1}^{T} \pi_{y_t, i_t} \right] \left[\gamma_{i_{T-1} i_T} \gamma_{i_{T-2} i_{T-1}} \ldots \gamma_{i_2 i_1} P(X_1 \,|\, \lambda) \right] \\
&= \delta_{i_1} \left[\prod_{t=1}^{T} \pi_{y_t, i_t} \right] \left[\gamma_{i_{T-1} i_T} \gamma_{i_{T-2} i_{T-1}} \ldots \gamma_{i_2 i_1} \right] \\
&= \delta_{i_1} \left[\prod_{t=1}^{T} \pi_{y_t, i_t} \right] \left[\prod_{t=1}^{T-1} \gamma_{i_t i_{t+1}} \right],
\end{aligned}$$

então,

$$\log P(\mathbf{Y}, \mathbf{X} = \mathbf{i} \,|\, \widehat{\lambda}) = \log \widehat{\delta}_{i_1} + \sum_{t=1}^{T-1} \log \widehat{\gamma}_{i_t i_{t+1}} + \sum_{t=1}^{T} \log \widehat{\pi}_{y_t, i_t}.$$

Substituindo em (1.25), temos

$$\begin{aligned}
\mathcal{Q}(\widehat{\lambda}, \lambda) &= \sum_{i \in \aleph} \frac{P(\mathbf{Y}, \mathbf{X} = \mathbf{i} \,|\, \lambda)}{P(\mathbf{Y} \,|\, \lambda)} \left[\log \widehat{\delta}_{i_1} + \sum_{t=1}^{T-1} \log \widehat{\gamma}_{i_t i_{t+1}} + \sum_{t=1}^{T} \log \widehat{\pi}_{y_t, i_t} \right] \\
&= \sum_{i \in \aleph} \frac{P(\mathbf{Y}, \mathbf{X} = \mathbf{i} \,|\, \lambda)}{P(\mathbf{Y} \,|\, \lambda)} \log \widehat{\delta}_{i_1} + \sum_{i \in \aleph} \sum_{t=1}^{T-1} \frac{P(\mathbf{Y}, \mathbf{X} = \mathbf{i} \,|\, \lambda)}{P(\mathbf{Y} \,|\, \lambda)} \log \widehat{\gamma}_{i_t i_{t+1}} + \\
&+ \sum_{i \in \aleph} \sum_{t=1}^{T} \frac{P(\mathbf{Y}, \mathbf{X} = \mathbf{i} \,|\, \lambda)}{P(\mathbf{Y} \,|\, \lambda)} \log \widehat{\pi}_{y_t, i_t} \\
&= \mathcal{Q}_\delta(\widehat{\delta}, \lambda) + \mathcal{Q}_\gamma(\widehat{\gamma}, \lambda) + \mathcal{Q}_\pi(\widehat{\pi}, \lambda),
\end{aligned} \tag{1.26}$$

sendo,

$$\begin{aligned}\mathcal{Q}_{\delta}(\widehat{\delta},\lambda) &= \sum_{\mathbf{i}\in\aleph}\frac{P(\mathbf{Y},\mathbf{X}=\mathbf{i}\,|\,\lambda)}{P(\mathbf{Y}\,|\,\lambda)}\log\widehat{\delta}_i \\ &= \sum_{i=1}^{m}\left[\frac{P(\mathbf{Y},X_1=i\,|\,\lambda)}{P(\mathbf{Y}\,|\,\lambda)}\right]\log\widehat{\delta}_i, \end{aligned} \tag{1.27}$$

$$\begin{aligned}\mathcal{Q}_{\gamma}(\widehat{\gamma},\lambda) &= \sum_{\mathbf{i}\in\aleph}\sum_{t=1}^{T-1}\frac{P(\mathbf{Y},\mathbf{X}=\mathbf{i}\,|\,\lambda)}{P(\mathbf{Y}\,|\,\lambda)}\log\widehat{\gamma}_{i_t i_{t+1}} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{t=1}^{T-1}\left[\frac{P(X_t=i,X_{t+1}=j,\mathbf{Y}\,|\,\lambda)}{P(\mathbf{Y}\,|\,\lambda)}\right]\log\widehat{\gamma}_{ij} \end{aligned} \tag{1.28}$$

e

$$\begin{aligned}\mathcal{Q}_{\pi}(\widehat{\pi},\lambda) &= \sum_{\mathbf{i}\in\aleph}\sum_{t=1}^{T}\frac{P(\mathbf{Y},\mathbf{X}=\mathbf{i}\,|\,\lambda)}{P(\mathbf{Y}\,|\,\lambda)}\log\widehat{\pi}_{y_t,i_t} \\ &= \sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{\{t;y_t=k\}}^{T}\left[\frac{P(\mathbf{Y},X_t=j\,|\,\lambda)}{P(\mathbf{Y}\,|\,\lambda)}\right]\log\widehat{\pi}_{k,j}. \end{aligned} \tag{1.29}$$

Os parâmetros que desejamos maximizar estão divididos em uma soma de três termos independentes, então podemos maximizar cada termo individualmente. Para isso utilizaremos os multiplicadores de Lagrange sujeito às restrições

$$\sum_{i=1}^{m}\widehat{\delta}_i = 1, \qquad \sum_{j=1}^{m}\widehat{\gamma}_{ij} = 1, \qquad \sum_{k=1}^{n}\widehat{\pi}_{k,j} = 1.$$

Note que, para todo $j \in E$, os termos de (1.26) têm a seguinte forma:

$$G(\mathbf{y}) = g(y_1,\ldots,y_n) = \sum_{l=1}^{n} w_l \log y_l,$$

com $h(\mathbf{y}) = \sum_{l=1}^{n} y_l - 1$ e $y_l \geq 0$. Portanto devemos maximizar a função $G(\mathbf{y})$ sujeita a restrição $\sum_{l=1}^{n} y_l = 1$. Assim,

$$\nabla G(\mathbf{y}) + \kappa(\nabla h(\mathbf{y})) = 0 \;\Rightarrow\; \nabla\left(\sum_{l=1}^{n} w_l \log y_l\right) + \kappa\left(\nabla\left(\sum_{l=1}^{n} y_l - 1\right)\right) = 0.$$

Logo,

$$\frac{w_l}{y_l} + \kappa = 0 \qquad \Rightarrow \qquad \kappa = -\frac{w_l}{y_l}, \qquad \forall\ l.$$

Conseqüentemente,

$$\kappa \underbrace{\sum_{l=1}^{n} y_l}_{1} = -\sum_{l=1}^{n} w_l \ \Rightarrow\ \kappa = -\sum_{l=1}^{n} w_l \ \Rightarrow\ y_l = \frac{w_l}{\sum_{l=1}^{n} w_l}.$$

Na equação (1.27) considere:

$$w_i = \frac{P(X_1 = i, \mathbf{Y} \mid \lambda)}{P(\mathbf{Y} \mid \lambda)} \qquad \text{e} \qquad y_i = \widehat{\delta}_i.$$

Assim,

$$y_i = \frac{w_i}{\sum_{i=1}^{m} w_i} \quad \Rightarrow \quad \widehat{\delta}_i = \frac{\frac{P(X_1 = i, \mathbf{Y} \mid \lambda)}{P(\mathbf{Y} \mid \lambda)}}{\sum_{i=1}^{m} \frac{P(X_1 = i, \mathbf{Y} \mid \lambda)}{P(\mathbf{Y} \mid \lambda)}} \quad \Rightarrow \quad \widehat{\delta}_i = \frac{\frac{P(X_1 = i, \mathbf{Y} \mid \lambda)}{P(\mathbf{Y} \mid \lambda)}}{1}$$

$$\Rightarrow \quad \widehat{\delta}_i = P(X_1 = i \mid \mathbf{Y}, \lambda) \quad \Rightarrow \quad \widehat{\delta}_i = e_1(i).$$

Já na equação (1.28) considere:

$$w_j = \sum_{t=1}^{T-1} \frac{P(X_t = i, X_{t+1} = j, \mathbf{Y} \mid \lambda)}{P(\mathbf{Y} \mid \lambda)} \qquad \text{e} \qquad y_j = \widehat{\gamma}_{ij}.$$

Assim,

$$y_j = \frac{w_j}{\sum_{j=1}^{m} w_j} \quad \Rightarrow \quad \widehat{\gamma}_{ij} = \frac{\sum_{t=1}^{T-1} \frac{P(X_t = i, X_{t+1} = j, \mathbf{Y} \mid \lambda)}{P(\mathbf{Y} \mid \lambda)}}{\sum_{j=1}^{m} \sum_{t=1}^{T-1} \frac{P(X_t = i, X_{t+1} = j, \mathbf{Y} \mid \lambda)}{P(v \mid \lambda)}} \quad \Rightarrow$$

$$\widehat{\gamma}_{ij} = \frac{\sum_{t=1}^{T-1} \frac{P(X_t = i, X_{t+1} = j, \mathbf{Y} \mid \lambda)}{P(\mathbf{Y} \mid \lambda)}}{\sum_{t=1}^{T-1} \frac{P(X_t = i, \mathbf{Y} \mid \lambda)}{P(\mathbf{Y} \mid \lambda)}} \quad \Rightarrow \quad \widehat{\gamma}_{ij} = \frac{\sum_{t=1}^{T-1} P(X_t = i, X_{t+1} = j, \mid \mathbf{Y}, \lambda)}{\sum_{t=1}^{T-1} P(X_t = i, \mid \mathbf{Y}, \lambda)}$$

$$\Rightarrow \quad \widehat{\gamma}_{ij} = \frac{\sum_{t=1}^{T-1} a_t(ij)}{\sum_{t=1}^{T-1} e_t(i)}.$$

Por fim, na equação (1.29) considerando,

$$w_k = \sum_{\{t;y_t=k\}}^{T} \frac{P(\mathbf{Y}, X_t = j \,|\, \lambda)}{P(\mathbf{Y} \,|\, \lambda)} \qquad \text{e} \qquad y_k = \widehat{\pi}_{k,j},$$

temos:

$$y_k = \frac{w_k}{\sum_{k=1}^{n} w_k} \quad \Rightarrow \quad \widehat{\pi}_{k,j} = \frac{\sum_{\{t;y_t=k\}}^{T} \frac{P(X_t = i, \mathbf{Y} \,|\, \lambda)}{P(\mathbf{Y} \,|\, \lambda)}}{\sum_{k=1}^{n} \sum_{\{t;y_t=k\}}^{T} \frac{P(X_t = j, \mathbf{Y} \,|\, \lambda)}{P(\mathbf{Y} \,|\, \lambda)}} \quad \Rightarrow$$

$$\widehat{\pi}_{k,j} = \frac{\sum_{\{t;y_t=k\}}^{T} \frac{P(X_t = j, \mathbf{Y} \,|\, \lambda)}{P(\mathbf{Y} \,|\, \lambda)}}{\sum_{t=1}^{T} \frac{P(X_t = j, \mathbf{Y} \,|\, \lambda)}{P(\mathbf{Y} \,|\, \lambda)}} \quad \Rightarrow \quad \widehat{\pi}_{k,j} = \frac{\sum_{\{t;y_t=k\}}^{T} P(X_t = j \,|\, \mathbf{Y}, \lambda)}{\sum_{t=1}^{T} P(X_t = j \,|\, \mathbf{Y}, \lambda)}$$

$$\Rightarrow \quad \widehat{\pi}_{k,j} = \frac{\sum_{\{t;y_t=k\}}^{T} e_t(j)}{\sum_{t=1}^{T} e_t(j)}.$$

Apresentamos um processo de simulação utilizando o pacote `HiddenMarkov` para a obtenção dos estimadores de máxima verossimilhança para o seguinte problema:

Exemplo 1.2 *Seja* $\mathbf{y} = \{y_1, \ldots, y_{600}\}$ *uma seqüência de observações de um modelo de Markov oculto* $\lambda = (\delta, \Gamma, \Pi)$ *com espaço de estados* $\{1, 2, 3\}$,

$$\delta = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad \Gamma = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 1/2 & 1/2 \end{bmatrix}$$

e

$$\pi_{y_{t,1}} = P(Y_t = y_t \,|\, X_t = 1) \sim Poisson(0.1)$$
$$\pi_{y_{t,2}} = P(Y_t = y_t \,|\, X_t = 2) \sim Poisson(2)$$
$$\pi_{y_{t,3}} = P(Y_t = y_t \,|\, X_t = 3) \sim Poisson(4).$$

Como se trata de um processo de simulação, os algoritmos que iremos executar irão fornecer estimativas para os parâmetros do modelo com a amostra específica gerada pela função `dthmm`. Portanto, quando o leitor executar novamente as funções do pacote `HiddenMarkov`, novas estimativas serão apresentadas à amostra gerada pela função `dthmm`, ou seja, cada vez executados os comandos uma nova amostra será gerada e, por conseguinte, novas estimativas serão obtidas.

```
 # Seja lambda=[gama,pi,delta] um Modelo de Markov Oculto
 # com Espaço de Estados: E.
 # Estados da Cadeia: E={1, 2, 3}.
 # Matriz de Transição da Cadeia de Markov:
 gama <- matrix(c(1/2, 1/2, 0, 1/3, 1/3, 1/3, 0, 1/2, 1/2),
 byrow=TRUE, nrow=3)
           [,1]      [,2]      [,3]
[1,]    0.5000000 0.5000000 0.0000000
[2,]    0.3333333 0.3333333 0.3333333
[3,]    0.0000000 0.5000000 0.5000000
 # Distribuição inicial da Cadeia de Markov
 delta <- c(0, 1, 0)
[1] 0 1 0
 # P{obs | estado oculto = 1} ~ Poisson(0.1)
 # P{obs | estado oculto = 2} ~ Poisson(2)
 # P{obs | estado oculto = 3} ~ Poisson(4)
```

A função `dthmm` é utilizada para gerar a cadeia de Markov oculta com os valores

iniciais δ, Γ e os parâmetros $\lambda_1 = 0.1$, $\lambda_2 = 2$ e $\lambda_3 = 4$.

```
x <- dthmm(NULL, gama, delta, "pois", list(lambda=c(0.1, 2, 4)),
discrete = TRUE)
x <- simulate(x, nsim=600)
```

Seqüência Observada

```
$x
  [1]  3 3 0 7 4 3 0 0 0 3 1 0 0 6 1 2 4 1 3 0 0 4 5 2 2 2 0 1 0 0 0 2 2 0
       2 7 3 3 2 2 0 0 1 0 2 5 7 4 3 7
 [51]  3 5 3 1 0 1 0 2 3 4 2 5 5 4 2 3 4 2 2 2 5 0 0 0 1 0 1 3 0 5 3 0 0 0
       2 2 6 2 9 7 1 1 1 0 1 3 0 2 3 2
[101]  0 4 1 3 0 0 0 1 0 1 4 4 9 4 0 2 2 2 0 1 3 0 3 0 0 1 5 2 5 0 3 0 0 1
       2 0 3 4 0 2 1 0 4 7 1 0 0 0 2 1
[151]  0 2 0 0 7 2 6 6 5 3 2 1 5 2 0 1 0 0 0 0 3 0 3 5 2 3 2 4 8 3 3 4 4 0
       1 3 3 2 0 2 2 0 0 1 5 3 4 5 1 3
[201]  2 2 4 1 6 0 1 0 0 3 3 1 0 3 3 3 3 8 0 3 0 0 0 2 2 5 1 1 2 5 4 1 3 2
       7 1 1 0 0 0 0 2 1 0 1 0 0 2 0 0
[251]  4 2 3 1 0 0 0 0 2 5 6 4 7 2 0 0 0 5 1 1 4 2 5 2 0 0 0 0 3 0 1 5 7 1
       3 6 0 3 1 6 0 0 2 0 0 4 6 8 8 8
[301]  2 4 4 3 3 3 3 2 6 3 4 2 1 0 2 3 2 4 1 1 7 1 2 1 7 0 4 4 4 2 6 5 2 4
       4 0 6 1 4 11 0 4 1 0 2 4 0 0 0 2
[351]  7 1 4 2 1 3 4 5 1 3 9 3 3 1 1 1 2 2 5 5 1 0 0 0 1 0 1 5 4 2 2 3 4 2
       3 3 2 3 4 4 1 2 2 0 0 0 0 0 0 1
[401]  0 0 1 1 0 0 0 1 2 1 0 0 0 1 1 0 2 5 3 3 0 3 1 0 0 2 2 2 1 1 2 6 0 0
       0 3 5 0 1 1 2 1 3 3 1 7 7 1 2 4
[451]  4 0 0 0 2 0 2 4 6 2 1 0 2 1 4 4 2 0 0 0 0 0 0 1 2 3 6 2 3 1 3 2 0 1
       1 1 4 5 4 2 3 1 2 0 2 1 2 0 0 2
[501]  4 2 2 0 0 1 0 0 1 4 8 4 3 3 2 4 3 2 5 3 2 7 2 6 0 1 1 2 2 3 1 2 2 3
       2 1 0 3 0 0 0 0 0 2 0 1 1 1 4 0
```

```
[551]  0 0 0 1 0 0 0 2 0 3 9 0 0 4 0 0 0 0 2 3 4 1 1 3 3 2 2 8 2 4 2 4 0 0
       0 1 0 0 0 2 4 1 0 2 0 0 4 5 3 2
```

Valores iniciais dos parâmetros

```
$Pi
          [,1]      [,2]      [,3]
[1,]   0.5000000 0.5000000 0.0000000
[2,]   0.3333333 0.3333333 0.3333333
[3,]   0.0000000 0.5000000 0.5000000
$delta
[1] 0 1 0
$distn
[1] "pois"
$pm
$pm$lambda
[1] 0.1 2.0 4.0
$pn
NULL
$discrete
[1] TRUE
```

Estados da Cadeia de Markov

```
$y
  [1] 2 2 2 3 3 2 1 1 1 2 2 1 2 3 2 3 3 2 2 1 1 2 3 3 3 2 1 1 1 1 1 2 2 1
      2 2 3 2 2 2 1 1 2 2 3 3 3 3 3 3 2 3 3 2 1 2 1 2 3 3 3 3 3 3 2 2 3 2
      2 2 3 2 1 1 2
 [76] 1 1 2 2 3 3 2 1 1 2 3 2 2 3 3 2 1 1 1 1 2 2 3 3 2 1 2 2 2 1 1 1 2 1
      2 2 3 3 3 2 2 3 2 1 2 2 1 2 1 1 2 3 3 3 2 2 1 1 2 2 1 2 2 1 2 3 2 3
      3 2 1 1 1 2 2
```

```
[151] 1 2 1 2 3 2 3 3 2 3 3 2 3 2 1 2 1 1 1 1 2 1 2 3 2 2 2 2 3 2 2 3 2 3
      3 3 3 2 1 2 2 1 1 2 3 2 2 3 2 2 2 2 3 2 3 2 2 1 1 2 3 2 1 2 3 3 3 3
      3 2 1 1 1 2 3
[226] 3 2 2 3 3 3 2 3 3 3 2 1 1 1 1 1 2 2 1 2 1 1 2 2 1 2 3 2 1 1 1 2 1 2
      2 3 3 3 2 1 1 1 2 2 2 2 3 3 2 1 1 1 1 2 1 2 3 3 2 3 3 2 2 3 2 1 1 2
      2 2 3 3 3 3 2
[301] 2 2 2 2 3 2 3 3 3 3 3 2 1 1 2 2 3 3 2 2 3 2 2 3 3 2 3 3 3 2 3 3 3 2
      3 2 3 2 2 3 2 3 2 1 1 2 1 1 1 2 3 2 2 3 2 3 3 3 3 2 3 3 3 3 2 1 2 2
      3 3 2 1 1 1 1
[376] 1 2 3 3 3 2 3 3 3 3 3 3 3 3 3 2 3 2 1 1 1 1 1 1 1 2 1 2 3 2 2 1 2 3
      2 1 2 1 1 2 1 2 3 3 2 1 2 1 1 1 2 3 2 3 2 2 2 1 1 1 2 2 1 2 2 3 2 3
      2 3 3 3 2 2 3
[451] 2 1 2 1 2 1 2 3 3 2 2 1 2 3 3 3 3 2 1 1 1 1 1 2 2 2 3 3 3 2 2 2 1 2
      3 2 3 3 3 2 3 3 2 1 2 2 2 1 1 2 3 3 2 2 1 2 1 1 2 2 3 2 3 3 2 3 2 2
      3 2 2 3 2 2 1
[526] 2 2 3 3 3 2 3 3 2 3 2 1 2 1 1 1 1 2 2 1 1 2 3 3 2 1 1 1 2 1 1 1 2 1
      2 3 2 1 2 1 1 1 2 3 2 2 2 1 2 3 3 3 3 3 3 3 2 1 1 1 2 1 1 1 2 3 2 1
      2 2 2 3 3 2 3
```

Agora iremos utilizar a função `BaumWelch` para obtermos os estimadores de máxima verossimilhança do modelo.

```
 y <- BaumWelch(x)
$Pi
          [,1]      [,2]      [,3]
[1,] 0.5175848 0.4824152 0.0000000
[2,] 0.2776222 0.3641613 0.3582166
[3,] 0.0000000 0.4872944 0.5127056
$delta
[1] 0 1 0
```

```
$distn
[1] "pois"
$pm
$pm$lambda
[1] 0.1240216 1.8268871 4.0020645
$pn
NULL
$discrete
[1] TRUE
```

Notamos que as estimativas obtidas pelo processo de simulação são próximas dos verdadeiros valores dos parâmetros como mostra a Tabela 1.1.

Tabela 1.1: Estimativas de máxima verossimilhança

Verdadeiro valor do parâmetro	*Estimativas*
$\delta = [0\ 1\ 0]$	$\widehat{\delta} = [0\ 1\ 0]$
$\Gamma = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 1/2 & 1/2 \end{bmatrix}$	$\widehat{\Gamma} = \begin{bmatrix} 0.5175848 & 0.4824152 & 0.0000000 \\ 0.2776222 & 0.3641613 & 0.3582166 \\ 0.0000000 & 0.4872944 & 0.5127056 \end{bmatrix}$
$\lambda_1 = 0.1$	$\widehat{\lambda}_1 = 0.1240216$
$\lambda_2 = 2$	$\widehat{\lambda}_2 = 1.8268871$
$\lambda_3 = 4$	$\widehat{\lambda}_3 = 4.0020645$

Um ponto importante a se comentar é que podemos supor várias outras distribuições para a probabilidade do processo observado, dado o estado oculto da cadeia de Markov (Bernoulli, Binomial, Geométrica, etc), e utilizar as mesmas funções (`dthmm` e `BaumWelch`) para gerar os dados do modelo de Markov oculto e obter as estimativas de máxima verossimilhança.

Capítulo 2

Modelos de Markov Ocultos em Espaço de Estados Geral

No Capítulo 1 trabalhamos com modelos de Markov ocultos $\{X_t, Y_t\}_{t\in\mathbb{N}}$ onde os espaços de estados, tanto de $\{X_t\}_{t\in\mathbb{N}}$ como de $\{Y_t\}_{t\in\mathbb{N}}$, eram finitos. Neste capítulo trabalhamos com modelos de Markov ocultos $\{X_t, Y_t\}_{t\in\mathbb{N}}$ onde os espaços de estados, tanto de $\{X_t\}_{t\in\mathbb{N}}$ como de $\{Y_t\}_{t\in\mathbb{N}}$, são gerais, ou seja, não necessariamente enumeráveis. O objetivo deste capítulo é estimar a densidade comum das variáveis observáveis, para isso vamos dar condições sobre os processos envolvidos de modo a obtermos a convergência do estimador escolhido. Nesta direção relembramos na seção 2.1 os conceitos e resultados das cadeias de Markov em espaço de estados gerais. Também trataremos, na mesma seção, da condição ϕ-*mixing* que mostra o nível de dependência entre as variáveis de um processo e finalizamos falando do estimador que utilizaremos, a saber, o estimador do tipo núcleo. Na seção 2.2 daremos as condições sobre os processos envolvidos para que tenhamos a convergência do estimador para a densidade comum das variáveis observadas.

2.1 Preliminares

2.1.1 Cadeia de Markov em espaço de estados geral

Considere $(\Omega, \mathcal{F}, \mathcal{P})$ um espaço de probabilidade e $\{X_n\}_{n\geq 0}$ uma seqüência qualquer de v.a.s definidas em Ω assumindo valores em $E \subseteq \mathbb{R}^d$, para algum $d \in \mathbb{N}$. Considere as seguintes σ-álgebras induzidas pelas v.a.'s X_j,

$$\mathcal{F}_k = \sigma\Big(\{X_j\},\, j = 0, 1, \ldots, k\Big) \qquad \text{e} \qquad \mathcal{F}_{k+n}^{\infty} = \sigma\Big(\{X_j\},\, j = k+n, k+n+1, \ldots\Big),$$

para todo $k \geq 0$ e $n \geq 0$.

Definição 2.1 *A seqüência $\{X_n\}_{n\geq 0}$, com valores em E, é dita uma cadeia de Markov se, para todo $A \in \mathcal{F}$,*

$$P(X_{n+1} \in A \,|\, \mathcal{F}_n) = P(X_{n+1} \in A \,|\, X_n), \qquad \text{q.c.} \tag{2.1}$$

para qualquer distribuição inicial X_0.

Para o estudo de cadeias de Markov em espaço de estados geral utilizamos a *função de probabilidade de transição* ou *núcleo de transição* no lugar da matriz de transição.

Definição 2.2 *A função $P : E \times \mathcal{F} \to [0, 1]$ é chamada função de probabilidade de transição ou núcleo de transição se,*

i) Para todo $x \in E$, $P(x, \cdot)$ é uma medida de probabilidade em $(E, \mathcal{F})$,

ii) Para todo $A \in \mathcal{F}$, $P(\cdot, A)$ é uma função mensurável.

Ou seja, fixado $x \in E$ temos que P é uma medida de probabilidade em $(E, \mathcal{F})$ e fixado $A \in \mathcal{F}$ temos que P é uma função mensurável.

Seja $P(\cdot,\cdot)$ o núcleo de transição em $(E,\mathcal{F})$. Para cada $n \in \mathbb{N}$ e $A \in \mathcal{F}$, defina a seqüência de funções $\{P^n(\cdot,\cdot)\}_{n\geq 0}$ pelo seguinte processo iterativo,

$$P^{n+1}(x,A) = \int_E P^n(y,A)P(x,dy), \qquad n \geq 0 \tag{2.2}$$

sendo $P^0(x,A) = I_A(x)$.

Definição 2.3 *Dizemos que $P^n(\cdot,\cdot)$, definido por (2.2), é o núcleo de transição de n-passos gerado por $P(\cdot,\cdot)$.*

Consequentemente, se $X_0 = x$ então

$$P(X_n \in A) = P^n(x,A), \qquad \forall\, n \in \mathbb{N}.$$

Uma ferramenta bastante útil neste estudo são as *equações de Chapman-Kolmogorov.*

Proposição 2.1 (Equações de Chapman-Kolmogorov) *Sejam $P(\cdot,\cdot)$ o núcleo de transição em $(E,\mathcal{F})$ e $P^n(\cdot,\cdot)$ definido por (2.2). Então para todo $n,m \geq 0$,*

$$P^{n+m}(x,A) = \int P^n(y,A)P^m(x,dy). \tag{2.3}$$

Na seção 2.2, consideramos $\{X_n\}_{n\geq 0}$ uma cadeia de Markov com valores em E, núcleo de transição $\{P(x,A) : x \in E, A \in \mathcal{F}\}$ e $\pi(\cdot)$ uma distribuição estacionária da cadeia que usaremos como a distribuição inicial do processo. Para isso vamos definir o que seja uma medida de probabilidade estacionária.

Definição 2.4 *Uma medida de probabilidade π em $(E,\mathcal{F})$ é dita estacionária se*

$$\pi(A) = \int_E P(x,A)\pi(dx), \qquad \forall\, A \in \mathcal{F}.$$

Se $\{X_n\}_{n\geq 0}$ é uma cadeia de Markov homogênea e π é uma distribuição estacionária temos que se $X_0 \sim \pi$ então $X_n \sim \pi$ para todo $n \geq 1$, justificando o termo "estacionária".

2.1.2 Processos ϕ-*mixing*

Existe uma grande classe de processos estocásticos onde as variáveis aleatórias apresentam um certo grau de dependência. A seguir, definiremos uma maneira para estabelecer o grau de dependência entre as variáveis, que diminui com o decorrer do tempo. Tais processos não apresentam apenas importância pelo interesse probabilístico mas, também, pelo potencial em aplicações estatísticas.

Definição 2.5 *A seqüência* $\{X_n\}_{n\geq 0}$ *é dita* ϕ*-mixing com coeficiente* $\phi(n)$ *se,*

$$\sup\left\{\frac{|P(A\cap B)-P(A)P(B)|}{P(A)},\, A\in\mathcal{F}_k,\, B\in\mathcal{F}_{k+n}^{\infty},\, k\geq 1\right\}=\phi(n),$$

desde que $P(A)>0$ *ou, para cada* $A\in\mathcal{F}_k$, $B\in\mathcal{F}_{k+n}^{\infty}$ *e* $k\geq 1$ *tem-se,*

$$|P(A\cap B)-P(A)P(B)|\leq\phi(n)P(A)$$

com $\displaystyle\sum_{n=1}^{\infty}\phi(n)=M<\infty$.

Em outras palavras, as v.a.s X_j tendem a ser assintoticamente independentes, uma vez que a distância entre elas tendem a zero, pois se a $\displaystyle\sum_{n=1}^{\infty}\phi(n)$ converge, então $\phi(n)\to 0$ quando $n\to\infty$.

Uma caracterização alternativa para seqüências ϕ-*mixing* é dada no lema seguinte.

Lema 2.1 (Roussas-Ionnides, 1987) *A fim de que uma seqüência* $\{X_n\}_{n\geq 0}$ *seja* ϕ*-mixing é necessário e suficiente que,*

$$\sup\{\,|P(B\,|\,\mathcal{F}_k)-P(B)|\,,\, B\in\mathcal{F}_{k+n}^{\infty},\, k\geq 1\}=\phi(n).$$

Demonstração. É suficiente mostrar que, para todo $k\geq 1$, $A\in\mathcal{F}_k$ e $B\in\mathcal{F}_{k+n}^{\infty}$, as seguintes inequações são equivalentes, a saber,

$$|P(A\cap B)-P(A)P(B)|\leq\phi(n)P(A) \tag{2.4}$$

e

$$|P(B \,|\, \mathcal{F}_k) - P(B)| \leq \phi(n)P(A). \tag{2.5}$$

Suponhamos, inicialmente, que (2.4) seja verdadeira e que existam $A \in \mathcal{F}_k$ (desde que $P(B \,|\, \mathcal{F}_k)$ seja $\mathcal{F}_k$-mensurável) e $B \in \mathcal{F}_{k+n}^{\infty}$ tais que,

$$|P(B \,|\, \mathcal{F}_k) - P(B)| > \phi(n)P(A) \tag{2.6}$$

com $P(A) > 0$.

Defina os seguintes conjuntos,

$$A^+ = A \cap [P(B \,|\, \mathcal{F}_k) - P(B) > \phi(n)] \qquad \text{e} \qquad A^- = A \cap [P(B \,|\, \mathcal{F}_k) - P(B) < -\phi(n)]$$

Se $P(A^+) > 0$, temos,

$$\int_{A^+} [P(B \,|\, \mathcal{F}_k) - P(B)]\, dP > \int_{A^+} \phi(n) dP$$

$$\Rightarrow \qquad \int_{A^+} P(B \,|\, \mathcal{F}_k) dP - \int_{A^+} P(B) dP > P(A^+)\phi(n)$$

$$\Rightarrow \qquad \int_{A^+} E\left(I_B \,|\, \mathcal{F}_k\right) dP - P(A^+)P(B) > P(A^+)\phi(n)$$

$$\Rightarrow \qquad \int_{A^+} I_B dP - P(A^+)P(B) > P(A^+)\phi(n)$$

$$\Rightarrow \qquad P(B \cap A^+) - P(A^+)P(B) > P(A^+)\phi(n). \tag{2.7}$$

Analogamente, se $P(A^-) > 0$ temos,

$$\int_{A^-} [P(B \,|\, \mathcal{F}_k) - P(B)]\, dP < -\int_{A^-} \phi(n) dP$$

$$\Rightarrow \qquad \int_{A^-} P(B \,|\, \mathcal{F}_k) dP - \int_{A^-} P(B) dP < -P(A^-)\phi(n)$$

$$\Rightarrow \quad \int_{A^-} P(B \,|\, \mathcal{F}_k) dP - \int_{A^-} P(B) dP < -P(A^-)\phi(n)$$

$$\Rightarrow \quad \int_{A^-} E\left(I_B \,|\, \mathcal{F}_k\right) dP - P(A^-)P(B) < -P(A^-)\phi(n)$$

$$\Rightarrow \quad \int_{A^-} I_B dP - P(A^-)P(B) < -P(A^-)\phi(n)$$

$$\Rightarrow \quad P(B \cap A^-) - P(A^-)P(B) < -P(A^-)\phi(n). \tag{2.8}$$

Como a equação (2.4) é válida para todo conjunto $A \in \mathcal{F}_k$, em particular, para A^+ e A^-, (2.7) e (2.8) nos fornecem uma contradição, então (2.4) implica em (2.5).

Reciprocamente, assuma que (2.5) seja válida e provaremos (2.4). De fato,

$$|P(B \,|\, \mathcal{F}_k) - P(B)| \leq \phi(n) \quad \Rightarrow \quad -\phi(n) \leq P(B \,|\, \mathcal{F}_k) - P(B) \leq \phi(n)$$

Integrando em relação $A \in \mathcal{F}_k$ obtemos,

$$-\phi(n)P(A) \leq \int_A \left[P(B \,|\, \mathcal{F}_k) - P(B)\right] dP \leq \phi(n)P(A)$$

$$\Rightarrow \quad -\phi(n)P(A) \leq \int_A P(B \,|\, \mathcal{F}_k) dP - \int_A P(B) dP \leq \phi(n)P(A)$$

$$\Rightarrow \quad -\phi(n)P(A) \leq \int_A E(I_B \,|\, \mathcal{F}_k) dP - P(A)P(B) \leq \phi(n)P(A)$$

$$\Rightarrow \quad -\phi(n)P(A) \leq \int_A I_B dP - P(A)P(B) \leq \phi(n)P(A)$$

$$\Rightarrow \quad -\phi(n)P(A) \leq P(A \cap B) - P(A)P(B) \leq \phi(n)P(A)$$

$$\Rightarrow \quad |P(A \cap B) - P(A)P(B)| \leq \phi(n)P(A).$$

Outro resultado importante sobre processos ϕ-*mixing* é dado no Teorema 2.1 e será fundamental na demonstração de um dos resultados principais do trabalho.

Teorema 2.1 (Roussas-Ionnides, 1987) *Seja η uma v.a. $\mathcal{F}_{k+n}^{\infty}$-mensurável tal que $|\eta| < M$. Então sob a condição ϕ-mixing*

$$|E(\eta \,|\, \mathcal{F}_k) - E(\eta)| \leq 2\phi(n)M \qquad q.c.$$

Demonstração. Suponhamos inicialmente que $\eta \geq 0$ e considere as seguintes partições do intervalo $[0, M)$

$$I_{m_i} = \left[\frac{(i-1)M}{2^m},\, \frac{iM}{2^m}\right), \qquad i = 1, 2, 3, \ldots, 2^m$$

e

$$J_{m_i} = \left[\frac{iM}{2^m},\, M\right), \qquad i = 1, 2, 3, \ldots, 2^m - 1.$$

Sejam $A_{m_i} = \eta^{-1}(I_{m_i})$ e $B_{m_i} = \eta^{-1}(J_{m_i})$, isto é, a imagem inversa de I_{m_i} e J_{m_i} pela v.a. η. Então A_{m_i} e B_{m_i} são conjuntos mensuráveis em $\mathcal{F}_{k+n}^{\infty}$, pois η é $\mathcal{F}_{k+n}^{\infty}$-mensurável.

Defina,

$$\eta_m = \sum_{i=1}^{2^m} \frac{(i-1)M}{2^m} I_{A_{m_i}} = \frac{M}{2^m} \sum_{i=1}^{2^m} (i-1) I_{A_{m_i}}.$$

Como,

$$\begin{aligned}
\sum_{i=1}^{2^m} (i-1) I_{A_{m_i}} &= 0I_{A_{m_1}} + 1I_{A_{m_2}} + 2I_{A_{m_3}} + 3I_{A_{m_4}} + \ldots + (2^m - 1) I_{A_{m_{2^m}}} \\
&= I_{A_{m_2}} + \underbrace{I_{A_{m_3}} + I_{A_{m_3}}}_{2} + \underbrace{I_{A_{m_4}} + I_{A_{m_4}} + I_{A_{m_4}}}_{3} + \ldots + \underbrace{I_{A_{m_{2^m}}} + \ldots + I_{A_{m_{2^m}}}}_{2^m - 1} \\
&= \left(I_{A_{m_2}} + I_{A_{m_3}} + \ldots + I_{A_{m_{2^m}}}\right) + \left(I_{A_{m_3}} + I_{A_{m_4}} + \ldots + I_{A_{m_{2^m}}}\right) + \ldots + I_{A_{m_{2^m}}} \\
&= I_{B_{m_1}} + I_{B_{m_2}} + \ldots + I_{B_{m_{2^m-1}}} = \sum_{i=1}^{2^m - 1} I_{B_{m_i}},
\end{aligned}$$

podemos reescrever,

$$\sum_{i=1}^{2^m}(i-1)I_{A_{m_i}} = \sum_{i=1}^{2^m-1} I_{B_{m_i}}.$$

Logo,

$$\begin{aligned} E(\eta_m \,|\, \mathcal{F}_k) &= E\left[\left(\frac{M}{2^m}\sum_{i=1}^{2^m}(i-1)I_{A_{m_i}}\right) \,|\, \mathcal{F}_k\right] = E\left[\left(\frac{M}{2^m}\sum_{i=1}^{2^m-1} I_{B_{m_i}}\right) \,|\, \mathcal{F}_k\right] \\ &= \frac{M}{2^m}\sum_{i=1}^{2^m-1} E[I_{B_{m_i}} \,|\, \mathcal{F}_k] = \frac{M}{2^m}\sum_{i=1}^{2^m-1} P(B_{m_i} \,|\, \mathcal{F}_k) \end{aligned}$$

e

$$\begin{aligned} E(\eta_m) &= E\left(\frac{M}{2^m}\sum_{i=1}^{2^m}(i-1)I_{A_{m_i}}\right) = E\left(\frac{M}{2^m}\sum_{i=1}^{2^m-1} I_{B_{m_i}}\right) \\ &= \frac{M}{2^m}\sum_{i=1}^{2^m-1} E[I_{B_{m_i}}] = \frac{M}{2^m}\sum_{i=1}^{2^m-1} P(B_{m_i}). \end{aligned}$$

Assim,

$$\begin{aligned} |E(\eta_m \,|\, \mathcal{F}_k) - E(\eta_m)| &= \left|\frac{M}{2^m}\sum_{i=1}^{2^m-1} P(B_{m_i} \,|\, \mathcal{F}_k) - \frac{M}{2^m}\sum_{i=1}^{2^m-1} P(B_{m_i})\right| \\ &= \frac{M}{2^m}\left|\sum_{i=1}^{2^m-1} P(B_{m_i} \,|\, \mathcal{F}_k) - P(B_{m_i})\right| \\ &\leq \frac{M}{2^m}\sum_{i=1}^{2^m-1} |P(B_{m_i} \,|\, \mathcal{F}_k) - P(B_{m_i})| \\ &\leq \frac{M}{2^m}\sum_{i=1}^{2^m-1} \sup\{|P(B_{m_i} \,|\, \mathcal{F}_k) - P(B_{m_i})|\}, \qquad \forall\, B_{m_i} \in \mathcal{F}_{k+n}^{\infty}. \end{aligned}$$

Pela Lema 2.1, para todo $B_{m_i} \in \mathcal{F}_{k+n}^{\infty}$ e sabendo que $\phi(n) \geq 0$ temos,

$$\begin{aligned} |E(\eta_m \,|\, \mathcal{F}_k) - E(\eta_m)| &\leq \frac{M}{2^m}\sum_{i=1}^{2^m-1} \sup\{|P(B_{m_i} \,|\, \mathcal{F}_k) - P(B_{m_i})|\} \\ &= \frac{M}{2^m}\sum_{i=1}^{2^m-1} \phi(n) = \frac{M}{2^m}(2^m-1)\phi(n) \\ &= \frac{M}{2^m}2^m\phi(n) - \frac{M}{2^m}\phi(n) \\ &\leq \frac{M}{2^m}2^m\phi(n) = M\phi(n). \end{aligned}$$

Segue do Teorema da Convergência Monótona que,

$$|E(\eta \,|\, \mathcal{F}_k) - E(\eta)| \leq M\phi(n) \qquad q.c.$$

Para o caso em que η é uma variável aleatória qualquer, sejam

$$E(\eta \,|\, \mathcal{F}_k) = E(\eta^+ \,|\, \mathcal{F}_k) - E(\eta^- \,|\, \mathcal{F}_k) \qquad \text{e} \qquad E(\eta) = E(\eta^+) - E(\eta^-),$$

então,

$$\begin{aligned} |E(\eta \,|\, \mathcal{F}_k) - E(\eta)| &= \left|E(\eta^+ \,|\, \mathcal{F}_k) - E(\eta^- \,|\, \mathcal{F}_k) - E(\eta^+) + E(\eta^-)\right| \\ &\leq \left|E(\eta^+ \,|\, \mathcal{F}_k) - E(\eta^+)\right| + \left|E(\eta^- \,|\, \mathcal{F}_k) - E(\eta^-)\right| \\ &\leq M\phi(n) + M\phi(n) = 2M\phi(n). \end{aligned}$$

2.1.3 Estimador do tipo núcleo

Considere $X_1, X_2, \ldots, X_n$ variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) com densidade comum $f(\cdot)$. A teoria de estimação de densidade tem por objetivo construir um estimador para $f(\cdot)$ a partir de uma amostra de uma população. Por exemplo, seja X uma variável aleatória com função densidade $f(\cdot)$, a partir da definição de função densidade, temos quase certamente,

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X \leq x + h).$$

Por outo lado, uma boa aproximação para $P(x - h < X \leq x + h)$ é a proporção das variáveis $X_1, X_2, \ldots, X_n$ que estão no intervalo $(x - h, x + h]$. Portanto, um estimador para $f(x)$ é

$$f_n(x) = \frac{1}{2nh} \#\{X_i; X_i \in (x - h, x + h]\}, \tag{2.9}$$

sendo # a cardinalidade. Esse estimador é considerado estimador ingênuo.

Motivado por esse estimador, Parzen (1962) define uma nova classe de estimadores de densidades em $\mathbb{R}$ denominados *estimadores do tipo núcleo*,

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \qquad h_n \overset{n}{\to} 0 \quad \text{quando} \quad n \longrightarrow \infty, \tag{2.10}$$

sendo K uma função real, denominada função núcleo e $h_n = h$ uma seqüência de constantes positivas, ambas, apropriadamente escolhidas.

Tomando o núcleo

$$K(x) = \begin{cases} \dfrac{1}{2}, & x \in [-1, 1) \\ 0, & x \notin [-1, 1) \end{cases}$$

podemos rescrever (2.9) na forma (2.10).

Os resultados assintóticos de Parzen são estendidos para o caso $\mathbb{R}^d$ por Cacoullos (1964) substituindo nh por nh^d. Em Rao (1983) é feito um estudo sobre essa classe de estimadores tanto no caso univariado como no caso multivariado.

Esses estimadores também são usados na estimação de densidades associadas a cadeias de Markov. Os primeiros resultados foram obtidos por Billingsley (1961) para o caso em que o espaço de estados é finito. Para estimar a densidade associada a cadeias de Markov estritamente estacionárias e com espaço de estados real, Roussas (1969) considera o estimador dado em (2.10).

Para a obtenção da consistência fraca e normalidade assintótica, além das condições estabelecidas sobre a função núcleo é assumido que a cadeia satisfaz a condição D_0 de Doob (1953), isto é, existe uma medida μ em $(\mathbb{R}, \mathcal{B})$, um inteiro $n_0 \geq 1$ e $\epsilon > 0$ tais que

$$P^{n_0}(x, A) \leq 1 - \epsilon \qquad \text{se} \qquad \mu(A) \leq \epsilon.$$

Rosenblatt (1970) estuda esses estimadores quando a condição D_0 é enfraquecida e substituída pela condição ϕ-*mixing*. Resultados mais recentes referentes aos estimadores do tipo núcleo são obtidos por Roussas (1991) e em Athreya-Atuncar (1998).

Esses estimadores são generalizados por Campos-Dorea (2001) que estabelecem as condições para que os estimadores fossem assintoticamente não viciados, fracamente consistentes, consistentes em média quadrática, fortemente consistente e assintoticamente normal distribuído para a densidade de uma seqüência de variáveis aleatórias $X_1, X_2, \ldots$, independentes e identicamente distribuídas. Os mesmos resultados são obtidos por Campos-Dorea (2005) para a densidade estacionária e a densidade limite de uma cadeia de Markov.

2.2 Estimação da densidade das variáveis observadas

Nesta seção consideramos um modelo de Markov oculto em espaço de estados geral, ou seja, $\{X_t\}_{t\geq 0}$ um processo de Markov homogêneo assumindo valores em $\mathcal{X} \subseteq \mathbb{R}^d$ que descreve a evolução de um processo $\{Y_t\}_{t\geq 0}$ não observado, sendo $\mathcal{X} \subseteq \mathbb{R}^d$ é um conjunto mensurável e $\mathcal{B}(\mathcal{X})$ a σ-álgebra de Borel gerada pelo conjunto $\mathcal{X}$. Na prática não conhecemos a densidade de Y_k, apenas um conjunto de observações dessa variável. Portanto, buscamos estabelecer resultados acerca da densidade dessa variável aleatória.

Definição 2.6 *Um modelo de Markov oculto (HMM) em espaço de estados geral é um processo estocástico duplo $\{(X_t, Y_t)\}_{t\in\mathbb{N}}$ tal que:*

1. *$\{X_t\}_{t\in\mathbb{N}}$ é um processo de Markov homogêneo não observável, de espaço de estados $\mathcal{X} \subseteq \mathbb{R}^d$ e núcleo de transição $\{P(x, A); x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})\}$.*

2. *$\{Y_t\}_{t\in\mathbb{N}}$ é uma seqüência de variáveis aleatórias condicionalmente independentes com a distribuição de Y_j dependendo somente de X_j, ou seja, para $A_j \in \mathcal{B}(\mathbb{R}^d)$,*

$$P(Y_0 \in A_0, \ldots, Y_n \in A_n \,|\, X_0, \ldots, X_n) = \prod_{j=0}^{n} P(Y_j \in A_j \,|\, X_j). \tag{2.11}$$

Consideramos que a cadeia tem distribuição estacionária π,

$$\pi(A) = \int_{\mathbb{R}^d} P^n(x, A)\pi(dx), \qquad \forall\, n, \forall\, A \in \mathcal{B}(\mathcal{X}),$$

sendo $P^n(\cdot\,,\cdot)$ a n-ésima iteração do núcleo de transição $P(\cdot\,,\cdot)$, isto é,

$$P^n(x,A) = P(X_n \in A \,|\, X_0 = x) = \int_{\mathbb{R}^d} P^k(y,A)P^{n-k}(x,dy), \quad 1 \leq k \leq n-1,$$

e que o processo observável $\{Y_n\}_{n\in\mathbb{N}}$ é uma seqüência de variáveis aleatórias que assume valores em $\mathbb{R}^d$ e para todo $A \in \mathcal{B}(\mathbb{R}^d)$,

$$P(Y_n \in A \,|\, Y_0, \ldots, Y_{n-1}, X_0, \ldots, X_n) = \int_A f(y \,|\, X_n)dy. \tag{2.12}$$

Como as variáveis $X_i's$ são não observadas, a inferência é baseada na amostra observada $Y_0, \ldots, Y_n$. Seja

$$f(y) = \int_{\mathbb{R}^d} f(y \,|\, x)\pi(dx) \tag{2.13}$$

a densidade de Y_k.

Para se estimar $f(\cdot)$ usaremos os estimadores do tipo núcleo apresentados na seção 2.1.3,

$$f_n(y) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{Y_i - y}{h}\right),$$

sendo K uma função densidade em $\mathbb{R}^d$ e $h = h_n$ satisfaz,

$$h_n \to 0 \qquad \text{e} \qquad nh_n \to \infty \qquad \text{quando} \qquad n \to \infty.$$

Inicialmente, verificamos que,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} g(X_k) = \int g(z)\pi(dz) \qquad \text{q.c.} \tag{2.14}$$

para uma função limitada e não negativa $g(\cdot)$.

E quando assumimos que o processo $\{X_t, Y_t\}_{t\geq 0}$ é um modelo de Markov oculto mostramos que,

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} |f_n(y) - f(y)|\, dy = 0 \qquad \text{q.c.} \tag{2.15}$$

A notação P_π será usada para indicar a probabilidade da cadeia quando a distribuição inicial é π. De modo análogo, E_π será usada para denotar a esperança da cadeia quando a distribuição inicial é π.

Na demonstração de (2.14) fazemos uso dos seguintes resultados:

Lema 2.2 (Devroye, 1991) *Sejam $\mathcal{F}_0 = \{\emptyset, \mathcal{X}\} \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_n$ uma seqüência de σ-álgebras encaixadas e U uma variável aleatória integrável e $\mathcal{F}_n$-mensurável. Defina a martingale de Doob $U_l = E(U \,|\, \mathcal{F}_l)$ e assuma que para $l = 1, \ldots, n$ existam variáveis aleatórias V_l, $\mathcal{F}_{l-1}$-mensuráveis e constantes a_l tais que $V_l \leq U_l \leq V_l + a_l$. Então, dado $\epsilon > 0$,*

$$P(|U - EU| \geq \epsilon) \leq 4 \exp\left\{-2\epsilon^2 / \sum_{l=1}^{n} a_l^2\right\}.$$

Condição 2.1 *Assuma que a cadeia $\{X_t\}_{t\geq 0}$ admite uma distribuição estacionária π e satisfaz a condição ϕ-mixing, ou seja, para todo $A \in \mathcal{F}_k$, $B \in \mathcal{F}_{k+n}^{\infty}$ e $k \geq 1$,*

$$|P_\pi(A \cap B) - P_\pi(A)P_\pi(B)| \leq \phi(n)P_\pi(A),$$

com $\sum_{n\geq 1} \phi(n) = M_1 < \infty$.

Teorema 2.2 (Dorea-Zhao, 2002) *Sob a Condição 2.1 e se $0 \leq g(.) \leq M < \infty$ então para cada $\epsilon > 0$ existem constantes $c_1 = c_1(\epsilon, M) > 0$ e $c_2 = c_2(\epsilon, M) > 0$ tais que*

$$P_\pi\left(\left|\frac{1}{n}\sum_{k=1}^{n} g(X_k) - \int_{\mathbb{R}^d} g(z)\pi(dz)\right| \geq \epsilon\right) \leq c_1 e^{-c_2 n}. \tag{2.16}$$

Um dos pontos chave na demonstração do Teorema 2.2 é assegurar que a equação

$$\hat{g}(x) - \int_{\mathbb{R}^d} \hat{g}(y)P(x, dy) = g(x) - \int_{\mathbb{R}^d} g(y)\pi(dy), \qquad \forall x \in \mathcal{X}, \tag{2.17}$$

tem uma solução limitada $\hat{g}$ sendo g uma função em $\mathcal{X}$ tal que $\int_{\mathbb{R}^d} g(y)\pi(dy) < \infty$. A equação (2.17) é conhecida na literatura como equação de Poisson. Ver mais detalhes em Meyn-Tweedie (1993).

Demonstração. Supondo que a cadeia satisfaça a condição ϕ-*mixing* do Teorema 2.1, se η é $\mathcal{F}_{k+n}^{\infty}$-mensurável e $|\eta| \leq M$ então

$$|E_\pi(\eta \,|\, \mathcal{F}_k) - E_\pi(\eta)| \leq 2\phi(n)M \qquad q.c.$$

Fazendo $k = 0$ a variável aleatória $g(X_n)$ é $\mathcal{F}_n^{\infty}$-mensurável e como π é uma distribuição estacionária para a cadeia temos,

$$E_\pi\Big(g(X_n) \,|\, \mathcal{F}_0\Big) = E_\pi\Big(g(X_n) \,|\, \sigma(X_0)\Big) = E_\pi\Big(g(X_n) \,|\, X_0\Big) = \int g(y)P^n(X_0, dy)$$

$$\Rightarrow E_\pi\Big(g(X_n) \,|\, X_0 = x\Big) = \int g(y)P^n(x, dy)$$

e

$$E_\pi\Big(g(X_n)\Big) = \int g(y)P(X_n, dy) = \int g(y)P(X_0, dy) = \int g(y)\pi(dy).$$

Como $|g| \leq M$,

$$\Big|E_\pi\Big(g(X_n) \,|\, X_0\Big) - E_\pi\Big(g(X_n)\Big)\Big| \leq 2\phi(n)M$$

$$\Rightarrow \quad \left|\int g(y)P^n(X_0, dy) - \int g(y)\pi(dy)\right| \leq 2\phi(n)M$$

$$\Rightarrow \quad \left|\int g(y)\Big[P^n(X_0, dy) - \pi(dy)\Big]\right| \leq 2\phi(n)M \qquad q.c.$$

No passo seguinte iremos mostrar que a equação (2.17) admite uma solução limitada. Seja

$$\hat{g}(x) = g(x) - g_0 + \sum_{n\geq 1}[P^n g(x) - g_0], \tag{2.18}$$

com

$$g_0 = E_\pi\Big(g(X_n)\Big) = \int g(y)\pi(dy) \quad \text{e} \quad P^n g(x) = E\left(g(X_n) \,|\, X_0 = x\right) = \int g(y)P^n(x, dy).$$

Inicialmente, vamos mostrar que $\hat{g}(x)$ é limitada. De fato,

$$\begin{aligned}
|\hat{g}(x)| &= \left| g(x) - g_0 + \sum_{n\geq 1}[P^n g(x) - g_0] \right| \leq |g(x) - g_0| + \sum_{n\geq 1} |P^n g(x) - g_0| \\
&\leq |g(x)| + |g_0| + \sum_{n\geq 1} |P^n g(x) - g_0| \leq |g(x)| + \int |g(x)|\, \pi(dx) + \sum_{n\geq 1} |P^n g(x) - g_0| \\
&\leq M + M \int \pi(dx) + \sum_{n\geq 1} |P^n g(x) - g_0| = 2M + \sum_{n\geq 1} |P^n g(x) - g_0| \,.
\end{aligned}$$

Como,

$$|P^n g(x) - g_0| = \Big| E_\pi\Big(g(X_n) \,|\, X_0 = x \Big) - E_\pi\Big(g(X_n) \Big) \Big| \leq 2\phi(n) M, \tag{2.19}$$

e a série $\sum_{n\geq 1} \phi(n)$ é convergente,

$$\begin{aligned}
|\hat{g}(x)| &\leq 2M + \sum_{n\geq 1} 2\phi(n) M = 2M + 2M \sum_{n\geq 1} \phi(n) \\
&\leq 2M + 2MM_1 = 2M(1 + M_1) = L.
\end{aligned}$$

Ou seja, $\hat{g}$ é limitada. Agora, iremos mostrar que (2.18) é uma solução da equação (2.17). Com efeito,

$$\begin{aligned}
P\hat{g}(x) &= \int \hat{g}(y) P(x, dy) \\
&= \int \Big(g(y) - g_0 + \sum_{n\geq 1} [P^n g(y) - g_0] \Big) P(x, dy) \\
&= \underbrace{\int g(y) P(x, dy)}_{Pg(x)} - \int g_0 P(x, dy) + \int \Big(\sum_{n\geq 1} [P^n g(y) - g_0] \Big) P(x, dy),
\end{aligned}$$

segue de (2.19) que,

$$\begin{aligned}
P\hat{g}(x) &= Pg(x) - g_0 \int P(x, dy) + \sum_{n\geq 1} \int \Big(P^n g(y) - g_0 \Big) P(x, dy) \\
&= Pg(x) - g_0 + \sum_{n\geq 1} \Big[\int P^n g(y) P(x, dy) - \int g_0 P(x, dy) \Big] \\
&= Pg(x) - g_0 + \sum_{n\geq 1} \Big[\int \int g(z) P^n(y, dz) P(x, dy) - \int g_0 P(x, dy) \Big] \\
&= Pg(x) - g_0 + \sum_{n\geq 1} \Big[\int \int g(z) P(x, dy) P^n(y, dz) - g_0 \Big]
\end{aligned}$$

$$
\begin{aligned}
&= Pg(x) - g_0 + \sum_{n\geq 1}\Big[\int g(z)\int P^n(y,dz)P(x,dy) - g_0\Big] \\
&= Pg(x) - g_0 + \sum_{n\geq 1}\Big[\int g(z)P^{n+1}(x,dz) - g_0\Big] \\
&= Pg(x) - g_0 + \sum_{n\geq 1}\Big[P^{n+1}g(x) - g_0\Big] \\
&= \sum_{n\geq 1}\Big[P^n g(x) - g_0\Big],
\end{aligned}
$$

logo,

$$\hat{g}(x) - P\hat{g}(x) = g(x) - g_0 + \sum_{n\geq 1}\Big[P^n g(x) - g_0\Big] - \sum_{n\geq 1}\Big[P^n g(x) - g_0\Big] = g(x) - g_0.$$

Então, (2.18) é solução da equação de Poisson. Conseqüentemente, podemos escrever:

$$
\begin{aligned}
\sum_{k=1}^{n}\Big[g(X_k) - g_0\Big] &= \sum_{k=1}^{n}\Big[\hat{g}(X_k) - P\hat{g}(X_k)\Big] \\
&= \hat{g}(X_1) - P\hat{g}(X_1) + \hat{g}(X_2) - P\hat{g}(X_2) + \ldots + \hat{g}(X_n) - P\hat{g}(X_n) \\
&= \hat{g}(X_1) - P\hat{g}(X_n) + \Big[\hat{g}(X_2) - P\hat{g}(X_1) + \cdots + \hat{g}(X_n) - P\hat{g}(X_{n-1})\Big] \\
&= \hat{g}(X_1) - P\hat{g}(X_n) + \sum_{k=2}^{n}\Big[\hat{g}(X_k) - P\hat{g}(X_{k-1})\Big].
\end{aligned}
$$

Notemos que,

$$
\begin{aligned}
\frac{1}{n}\sum_{k=1}^{n} g(X_k) - \int g(z)\pi(dz) &= \frac{1}{n}\sum_{k=1}^{n} g(X_k) - \frac{n}{n}\int g(z)\pi(dz) \\
&= \frac{1}{n}\Big[\sum_{k=1}^{n} g(X_k) - n\int g(z)\pi(dz)\Big] \\
&= \frac{1}{n}\sum_{k=1}^{n}\Big[g(X_k) - \int g(z)\pi(dz)\Big] \\
&= \frac{1}{n}\sum_{k=1}^{n}\Big[g(X_k) - g_0\Big]
\end{aligned}
$$

Logo,

$$\begin{aligned} & P_\pi\left(\frac{1}{n}\left|\sum_{k=1}^{n} g(X_k) - \int g(z)\pi(dz)\right| \geq \epsilon\right) = P_\pi\left(\frac{1}{n}\left|\sum_{k=1}^{n}\Big[g(X_k) - g_0\Big]\right| \geq \epsilon\right) \\ \leq \;& P_\pi\left(\frac{1}{n}\left|\hat{g}(X_1) - P\hat{g}(X_n)\right| \geq \frac{\epsilon}{2}\right) + P_\pi\left(\frac{1}{n}\left|\sum_{k=2}^{n}\Big[\hat{g}(X_k) - P\hat{g}(X_{k-1})\Big]\right| \geq \frac{\epsilon}{2}\right) \\ = \;& P_\pi\left(\left|\hat{g}(X_1) - P\hat{g}(X_n)\right| \geq \frac{n\epsilon}{2}\right) + P_\pi\left(\left|\sum_{k=2}^{n}\Big[\hat{g}(X_k) - P\hat{g}(X_{k-1})\Big]\right| \geq \frac{n\epsilon}{2}\right). \end{aligned}$$

Como $g(x)$ é limitada, segue que,

$$\lim_{n\to\infty} P_\pi\left(\left|\hat{g}(X_1) - P\hat{g}(X_n)\right| \geq \frac{n\epsilon}{2}\right) = 0.$$

Assim, para provarmos (2.16) é suficiente mostramos que,

$$P_\pi\left(\left|\sum_{k=2}^{n}\Big[\hat{g}(X_k) - P\hat{g}(X_{k-1})\Big]\right| \geq \frac{n\epsilon}{2}\right) \leq c_1 e^{-nc_2}.$$

O objetivo é aplicarmos o Lema 2.2. Para isso, definamos,

$$U = \sum_{k=2}^{n}\Big[\hat{g}(X_k) - P\hat{g}(X_{k-1})\Big] \qquad \text{e} \qquad U_l = E_\pi\Big(U \,|\, \mathcal{F}_l\Big),$$

sendo U $\mathcal{F}_n$-mensurável e integrável. Vamos mostrar que existem v.a.s V_l, $\mathcal{F}_{l-1}$ mensuráveis e constantes a_l tais que,

$$V_l \leq U_l \leq V_l + a_l. \tag{2.20}$$

De fato,

$$\begin{aligned} U_l = E_\pi\Big(U \,|\, \mathcal{F}_l\Big) &= E_\pi\left(\sum_{k=2}^{n}\Big[\hat{g}(X_k) - P\hat{g}(X_{k-1})\Big] \,|\, \mathcal{F}_l\right) \\ &= \sum_{k=2}^{n}\Big[E\Big(\hat{g}(X_k) \,|\, \mathcal{F}_l\Big) - E_\pi\Big(P\hat{g}(X_{k-1}) \,|\, \mathcal{F}_l\Big)\Big] \\ &= \sum_{k=2}^{l}\Big[E_\pi\Big(\hat{g}(X_k) \,|\, \mathcal{F}_l\Big) - E_\pi\Big(P\hat{g}(X_{k-1}) \,|\, \mathcal{F}_l\Big)\Big] + \\ &+ \sum_{k=l+1}^{n}\Big[E_\pi\Big(\hat{g}(X_k) \,|\, \mathcal{F}_l\Big) - E_\pi\Big(P\hat{g}(X_{k-1}) \,|\, \mathcal{F}_l\Big)\Big]. \end{aligned} \tag{2.21}$$

Como,

$$\begin{aligned} E_\pi\Big(\hat{g}(X_k)\,|\,\mathcal{F}_{k-1}\Big) &= E_\pi\Big(\hat{g}(X_k)\,|\,\sigma(X_0,\ldots,X_{k-1})\Big) = E_\pi\Big(\hat{g}(X_k)\,|\,\sigma(X_{k-1})\Big) \\ &= E_\pi\Big(\hat{g}(X_k)\,|\,X_{k-1}\Big) = \int_{\mathcal{X}} \hat{g}(y)P(X_{k-1},dy) = P\hat{g}(X_{k-1}). \end{aligned}$$

Para $k \geq l+1$ temos,

$$E_\pi\Big(P\hat{g}(X_{k-1})\,|\,\mathcal{F}_l\Big) = E_\pi\Big(E_\pi\Big(\hat{g}(X_k)\,|\,\mathcal{F}_{k-1}\Big)\,|\,\mathcal{F}_l\Big) = E_\pi\Big(\hat{g}(X_k)\,|\,\mathcal{F}_l\Big)$$

Logo, a segunda parcela de (2.21) é zero e

$$U_l = E_\pi\Big(U\,|\,\mathcal{F}_l\Big) = \sum_{k=2}^{l}\Big[E_\pi\Big(\hat{g}(X_k)\,|\,\mathcal{F}_l\Big) - E_\pi\Big(P\hat{g}(X_{k-1})\,|\,\mathcal{F}_l\Big)\Big].$$

Para $k < l+1$, $\mathcal{F}_k \subseteq \mathcal{F}_l$ e como X_k e X_{k-1} são $\mathcal{F}_l$-mensurável,

$$U_l = E_\pi\Big(U\,|\,\mathcal{F}_l\Big) = \sum_{k=2}^{l}\Big[\hat{g}(X_k) - P\hat{g}(X_{k-1})\Big]. \tag{2.22}$$

De maneira análoga,

$$U_{l-1} = E_\pi\Big(U\,|\,\mathcal{F}_{l-1}\Big) = \sum_{k=2}^{l-1}\Big[\hat{g}(X_k) - P\hat{g}(X_{k-1})\Big]. \tag{2.23}$$

Por outro lado,

$$\begin{aligned} |P\hat{g}(x)| &= \left|\int \hat{g}(y)P(x,dy)\right| \leq \int |\hat{g}(y)|\,P(x,dy) \\ &\leq \int LP(x,dy) = L\int P(x,dy) = L, \end{aligned} \tag{2.24}$$

logo, por (2.22), (2.23) e (2.24),

$$|U_l - U_{l-1}| = |\hat{g}(X_l) - P\hat{g}(X_{l-1})| \leq |\hat{g}(X_{l-1})| + |P\hat{g}(X_{l-2})| \leq 2L.$$

Fazendo $V_{l-1} = U_{l-1} - 2L$ e $a_{l-1} = 4L$ obtemos (2.20). E mais,

$$E_\pi\Big(P\hat{g}(X_{k-1})\Big) = \int\int \hat{g}(y)P(z,dy)\pi(dz) = \int \hat{g}(y)\int P(z,dy)\pi(dz) = \int \hat{g}(y)\pi(dy)$$

$$\Rightarrow E_\pi\Big(\hat{g}(X_k)\Big) = E_\pi\Big(P\hat{g}(X_{k-1})\Big).$$

Assim,

$$E_\pi(U) = \sum_{k=2}^{n}\Big[E_\pi\Big(\hat{g}(X_k)\Big) - E_\pi\Big(P\hat{g}(X_{k-1})\Big)\Big] = \sum_{k=2}^{n}\Big[E_\pi\Big(\hat{g}(X_k)\Big) - E_\pi\Big(\hat{g}(X_k)\Big)\Big] = 0$$

Então, como conseqüência do Lema 2.2,

$$P_\pi\left(\left|\sum_{k=2}^{n}\Big[\hat{g}(X_k) - P\hat{g}(X_{k-1})\Big]\right| \geq \frac{n\epsilon}{2}\right) \leq 4\exp\left\{\frac{-n\epsilon^2}{32L}\right\}.$$

Tomando $c_1 = 4$ e $c_2 = \dfrac{\epsilon^2}{32L}$, obtemos o resultado desejado. ∎

Como conseqüência do Teorema 2.2 e da proposição seguinte, obtemos (2.14). Vejamos:

Proposição 2.2 *Seja $\{X_n\}_{n\geq 0}$ uma seqüência de v.a.s em $(\Omega, \mathcal{F}, P)$. Dado $\epsilon > 0$ se*

$$\sum_{n=1}^{\infty} P(|X_n| > \epsilon) < \infty,$$

então

$$P\left(\lim_{n\to\infty} X_n = 0\right) = 1.$$

A prova será omitida. Para maiores detalhes ver Athreya-Lahiri (2006).

Corolário 2.1 *Sob as hipóteses do Teorema 2.2 temos*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} g(X_k) = \int g(z)\pi(dz) \qquad \text{q.c.}$$

desde que a série $\sum_{n\geq 1} c_1 e^{-nc_2}$ seja convergente.

Demonstração. Segue do Teorema 2.2 que,

$$P_\pi\left(\left|\frac{1}{n}\sum_{k=1}^{n} g(X_k) - \int g(z)\pi(dz)\right| \geq \epsilon\right) \leq c_1 e^{-nc_2}.$$

Sob a hipótese de que $\sum_{n\geq 1} c_1 e^{-nc_2} < \infty$ temos,

$$\sum_{n\geq 1} P_\pi\left(\left|\frac{1}{n}\sum_{k=1}^{n} g(X_k) - \int g(z)\pi(dz)\right| \geq \epsilon\right) \leq \sum_{n\geq 1} c_1 e^{-nc_2} < \infty.$$

Usando a Proposição 2.2 obtemos,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} g(X_k) = \int g(z)\pi(dz) \qquad \text{q.c.}$$

Na demonstração do Teorema 2.2 supomos que a distribuição inicial da cadeia é a distribuição estacionária e assumimos que a cadeia satisfaz a condição ϕ-*mixing*. Nos exemplos a seguir discutimos a importância da hipótese da distribuição inicial ser a distribuição estacionária.

Nos exemplos seguintes considere m como a medida de Lebesgue.

Exemplo 2.1 *Sejam* $\mathcal{X}$= $[0,2]$ *e* $\{X_n\}_{n\geq 0}$ *uma cadeia de Markov com núcleo de transição dado por*

$$P(x,dy) = \left[I_{[0,1]^2}(x,y) + I_{(1,2]^2}(x,y)\right] dy.$$

a) Considere $\pi(x) = I_{[0,1]}(x)$. Logo,

$$\pi(A) = \int_A \pi(x)dx = \int_A I_{[0,1]}(x)dx = \int_{A\cap[0,1]} dx = m(A\cap[0,1])$$

e

$$\begin{aligned}\pi P(A) &= \int_{\mathcal{X}} P(x,A)\pi(x)dx = \int_{\mathcal{X}\cap[0,1]} P(x,A)dx = \int_{[0,1]}\int_A P(x,dy)dx \\ &= \int_{[0,1]}\int_{A\cap[0,1]} dydx = m(A\cap[0,1]) = \pi(A),\end{aligned}$$

ou seja, $\pi(x)$ é uma densidade estacionária. Então

$$P_\pi(X_0 \in A) = P_\pi(X_n \in A) = m(A\cap[0,1]), \qquad \forall\, n \geq 1 \tag{2.25}$$

Por outro lado,

$$\begin{aligned} P_\pi(X_0 \in A, X_1 \in B) &= \int_B \int_A P(x, dy)\pi(x)dx = \int_B \int_{A\cap[0,1]} P(x, dy)dx \\ &= \int_{B\cap[0,1]} \int_{A\cap[0,1]} dydx = m(A \cap [0, 1])m(B \cap [0, 1]) \end{aligned}$$

e

$$\begin{aligned} & P_\pi(X_0 \in A, X_2 \in B) = \int_B \int_{\mathcal{X}} \int_A P(y, dz)P(x, dy)\pi(x)dx \\ = & \int_B \int_{\mathcal{X}} \int_{A\cap[0,1]} P(y, dz)P(x, dy)dx = \int_B \int_{[0,1]} \int_{A\cap[0,1]} P(y, dz)dydx \\ = & \int_{B\cap[0,1]} \int_{[0,1]} \int_{A\cap[0,1]} dzdydx = m(A \cap [0, 1])m(B \cap [0, 1]). \end{aligned}$$

De maneira análoga obtemos,

$$P_\pi(X_0 \in A, X_n \in B) = m(A \cap [0, 1])m(B \cap [0, 1]), \qquad \forall n > 2.$$

Pelo fato da cadeia ser homogênea e por (2.25), para todo $i \geq 0$ e $k \geq 1$, temos

$$\begin{aligned} & P_\pi(X_i \in A, X_{i+k} \in B) = P_\pi(X_{i+k} \in B \,|\, X_i \in A)P(X_i \in A) \\ = & \; P_\pi(X_k \in B \,|\, X_0 \in A)P(X_i \in A) = P_\pi(X_k \in B \,|\, X_0 \in A)P(X_0 \in A) \\ = & \; P_\pi(X_k \in B, X_0 \in A) = m(A \cap [0, 1])m(B \cap [0, 1]). \end{aligned} \tag{2.26}$$

Segue de (2.25) e (2.26) que,

$$P_{\pi_1}(X_i \in A, X_{i+k} \in B) - P_{\pi_1}(X_i \in A)P_{\pi_1}(X_{i+k} \in B) = 0, \tag{2.27}$$

então a cadeia $\{X_n\}_{n\geq 0}$ é ϕ-*mixing* com coeficiente $\phi(n) = 0$.

b) Agora, considere $\pi(x) = 2I_{[0,\frac{1}{2}]}(x)$ como a densidade inicial do processo. Sendo assim,

$$\pi(A) = \int_A \pi(x)dx = 2\int_A I_{[0,\frac{1}{2}]}(x)dx = 2\int_{A\cap[0,\frac{1}{2}]} dx = 2m(A \cap [0, \frac{1}{2}]),$$

porém,

$$\begin{aligned}\pi P(A) &= \int_{\mathcal{X}} P(x,A)\pi(x)dx = 2\int_{[0,\frac{1}{2}]} P(x,A)dx = 2\int_{[0,\frac{1}{2}]}\int_A P(x,dy)dx \\ &= 2\int_{[0,\frac{1}{2}]}\int_{A\cap[0,1]} dydx = m(A\cap[0,1]).\end{aligned}$$

Portanto π não é uma densidade estacionária. Por outro lado,

$$P_\pi(X_0 \in A) = \int_A \pi(x)dx = 2m(A\cap[0,\frac{1}{2}]) \tag{2.28}$$

e

$$\begin{aligned}P_\pi(X_1 \in A) &= \int_{\mathcal{X}} P(x,A)\pi(x)dx = 2\int_{\mathcal{X}} P(x,A)I_{[0,1/2]}dx = 2\int_{[0,1/2]} P(x,A)dx \\ &= 2\int_{[0,1/2]}\int_A P(x,dy)dx = 2\int_{[0,1/2]}\int_{A\cap[0,1]} dydx = m(A\cap[0,1]).\end{aligned}$$

Para $x \in [0,1]$

$$P^2(x,A) = \int_{\mathcal{X}} P(y,A)P(x,dy) = \int_{[0,1]} P(y,A)dy = \int_{[0,1]}\int_A P(y,dz)dy = m(A\cap[0,1])$$

Conseqüentemente,

$$P_\pi(X_2 \in A) = \int_{\mathcal{X}} P^2(x,A)\pi(x)dx = 2\int_{[0,1/2]} P^2(x,A)dx = m(A\cap[0,1]).$$

De maneira análoga,

$$P_\pi(X_n \in A) = m(A\cap[0,1]), \qquad \forall\, n \geq 1, \tag{2.29}$$

isto é, $X_1, X_2, \ldots$ são uniformemente distribuidas em $[0,1]$.

Para a distribuição conjunta de X_0 e X_2 temos,

$$\begin{aligned}&P_\pi(X_0 \in A, X_2 \in B) = \int_B\int_{\mathcal{X}}\int_A P(y,dz)P(x,dy)\pi(x)dx \\ &= 2\int_B\int_{\mathcal{X}}\int_{A\cap[0,1/2]} P(y,dz)P(x,dy)dx = 2\int_B\int_{[0,1]}\int_{A\cap[0,1/2]} P(y,dz)dydx \\ &= 2\int_{B\cap[0,1]}\int_{[0,1]}\int_{A\cap[0,1/2]} dzdydx = 2m(A\cap[0,\frac{1}{2}])m(B\cap[0,1]),\end{aligned}$$

então,

$$P_\pi(X_0 \in A, X_2 \in B) = P_\pi(X_0 \in A)P_\pi(X_2 \in B).$$

Da mesma forma podemos verificar que

$$P_\pi(X_0 \in A, X_n \in B) = 2m(A \cap [0, \frac{1}{2}])m(B \cap [0,1]) \qquad \forall n \geq 1,$$

e assim, por (2.28) e (2.29) temos,

$$P_\pi(X_0 \in A, X_n \in B) = P_\pi(X_0 \in A)P_\pi(X_n \in B) \qquad \forall\, n \geq 1.$$

Para a distribuição conjunta de X_1 e X_3 temos

$$\begin{aligned} P_\pi(X_1 \in A, X_3 \in B) &= \int_B \int_{\mathcal{X}} \int_{A\cap[0,1]} P(y,dz)P(x,dy)dx = \int_B \int_{[0,1]} \int_{A\cap[0,1]} P(y,dz)dydx \\ &= \int_{B\cap[0,1]} \int_{[0,1]} \int_{A\cap[0,1]} dzdydx = m(A \cap [0,1])m(B \cap [0,1]). \end{aligned}$$

De modo análogo,

$$P_\pi(X_1 \in A, X_n \in B) = m(A \cap [0,1])m(B \cap [0,1]), \qquad \forall n \geq 2. \tag{2.30}$$

Como a cadeia é homogênea e de (2.30) e (2.29), para todo $i \geq 1$ e $k \geq 1$,

$$\begin{aligned} & P_\pi(X_i \in A, X_{i+k} \in B) = P_\pi(X_{i+k} \in B \,|\, X_i \in A)P_\pi(X_i \in A) \\ = \;& P_\pi(X_{k+1} \in B \,|\, X_1 \in A)P_\pi(X_i \in A) = P_\pi(X_{k+1} \in B \,|\, X_1 \in A)P_\pi(X_1 \in A) \\ = \;& P_\pi(X_{k+1} \in B, X_1 \in A) = m(A \cap [0,1])m(B \cap [0,1]). \end{aligned}$$

Por (2.29),

$$P_\pi(X_i \in A)P_\pi(X_{i+k} \in B) = P_\pi(X_i \in A, X_{i+k} \in B), \qquad i \geq 1 \quad e \quad k \geq 1. \tag{2.31}$$

O que nos garante que a cadeia é ϕ-*mixing* com $\phi(n) = 0$.

Portanto π não é uma densidade estacionária, mas a cadeia satisfaz a condição ϕ-*mixing*.

Em ambas situações, seja $g(x) = x$ definida em $[0,1]$ então g é limitada e,

$$\frac{1}{n}\sum_{k=1}^{n} g(X_k) = \frac{1}{n}\sum_{k=1}^{n} X_k,$$

por (2.27) e (2.31), $X_1, X_2, \ldots$ são variáveis aleatórias independentes e identicamente distribuídas com distribuição uniforme em $[0,1]$. Pela Lei Forte dos Grandes Números (Kintchine)

$$\frac{1}{n}\sum_{k=1}^{n} g(X_k) \to \frac{1}{2} \qquad \text{quando} \qquad n \to \infty.$$

Para $\pi(x) = I_{[0,1]}(x)$

$$\int g(x)\pi(x)dx = \int g(x)I_{[0,1]}(x)dx = \int_0^1 x dx = \frac{1}{2}.$$

No entanto, para $\pi(x) = 2I_{[0,1/2]}(x)$,

$$\int g(x)\pi(x)dx = 2\int g(x)I_{[0,1/2]}(x)dx = 2\int_0^{1/2} x dx = \frac{1}{4}.$$

Portanto a condição *ϕ-mixing* não é suficiente para assegurar os resultados de convergência se a densidade inicial da cadeia não for estacionária.

Na demonstração do próximo teorema utilizamos os seguintes resultados:

Lema 2.3 (Hoeffding, 1963) *Sejam $\xi_1, \ldots, \xi_n$ variáveis aleatórias independentes com $a_i \leq \xi_i \leq b_i$. Então dado $\epsilon > 0$,*

$$P\left(\left|\sum_{i=1}^{n}\Big(\xi_i - E(\xi_i)\Big)\right| \geq \epsilon\right) \leq 2\exp\left\{-2\epsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2\right\}.$$

Lema 2.4 (Dorea-Zhao, 2002) *Sejam K e g densidades em $\mathbb{R}^d$ e $\{\eta_m\}$ uma seqüência arbitrária de variáveis aleatórias. Suponhamos que, dado $\epsilon > 0$, existam constantes $d_1 = d_1(\epsilon)$ e $d_2 = d_2(\epsilon)$ tais que para cada $A \in B(\mathcal{X})$,*

$$P\left(\frac{1}{n}\left|\sum_{k=1}^{n}\left[I_A(\eta_k) - \int_A g(y)dy\right]\right| \geq \epsilon\right) \leq d_1 e^{-d_2 n}.$$

Então se $h = h_n$ é uma seqüência de números positivos satisfazendo

$$h_n \to 0, \quad nh_n^d \to \infty \qquad quando \qquad n \to \infty$$

e

$$g_n(y) = \frac{1}{nh^d} \sum_{k=1}^{n} K\left(\frac{\eta_k - y}{h}\right)$$

temos,

$$P\left(\int |g_n(y) - g(y)|\, dy \geq \epsilon\right) \leq d_1 e^{-d_2 n}.$$

Agora temos as ferramentas que precisamos para estabelecer as condições necessárias para a convergência em L_1 do estimador do tipo núcleo para a densidade das variáveis do processo observado, que é o objetivo deste capítulo.

Teorema 2.3 (Dorea-Zhao, 2002) *Sob a Condição 2.1, se $\{Y_n\}_{n\geq 0}$ satisfaz (2.11) e (2.12) e $h = h(n)$ é uma seqüência de constantes positivas tal que,*

$$h_n \to 0 \qquad \text{e} \qquad nh_n \to \infty \qquad \text{quando} \qquad n \to \infty,$$

então para f e f_n definidas por:

$$f(y) = \int_{\mathbb{R}^d} f(y \,|\, x)\pi(dx) \qquad \text{e} \qquad f_n(y) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{Y_i - y}{h}\right),$$

existem constantes $c_1^{'} = c_1^{'}(\epsilon) > 0$ e $c_2^{'} = c_2^{'}(\epsilon) > 0$ tais que,

$$P_\pi\left(\int_{\mathbb{R}^d} |f_n(y) - f(y)|\, dy \geq \epsilon\right) \leq c_1' e^{-c_2' n}. \tag{2.32}$$

Demonstração. Pelo Lema 2.4, para provarmos (2.32) é suficiente mostrarmos que

$$P_\pi\left(\left|\frac{1}{n}\sum_{k=1}^{n} I_A(Y_k) - \int_A f(y)dy\right| \geq \epsilon\right) \leq c_1^{'} e^{-c_2^{'} n}, \qquad \forall\, A \in \mathcal{B}(\mathcal{X}). \tag{2.33}$$

Notemos que,

$$\begin{aligned} P_\pi\left(\left|\frac{1}{n}\sum_{k=1}^{n} I_A(Y_k) - \int_A f(y)dy\right| \geq \epsilon\right) \;&\leq\; P_\pi\left(\left|\frac{1}{n}\sum_{k=1}^{n} [I_A(Y_k) - g(X_k)]\right| \geq \frac{\epsilon}{2}\right) \\ &+\; P_\pi\left(\left|\frac{1}{n}\sum_{k=1}^{n} g(X_k) - \int_A f(y)dy\right| \geq \frac{\epsilon}{2}\right). \end{aligned}$$

Seja $g(x) = \int_A f(y \mid x)dy$, então

$$\begin{aligned}\int_{\mathbb{R}^d} g(x)\pi(dx) &= \int_{\mathbb{R}^d}\left[\int_A f(y \mid x)dy\right]\pi(dx)\\ &= \int_A\left[\int_{\mathbb{R}^d} f(y \mid x)\pi(dx)\right]dy\end{aligned}$$

e por (2.13) temos

$$\int_{\mathbb{R}^d} g(x)\pi(dx) = \int_A f(y)dy. \tag{2.34}$$

Logo,

$$\begin{aligned}P_\pi\left(\left|\frac{1}{n}\sum_{k=1}^n I_A(Y_k) - \int_A f(y)dy\right| \geq \epsilon\right) &\leq P_\pi\left(\left|\frac{1}{n}\sum_{k=1}^n [I_A(Y_k) - g(X_k)]\right| \geq \frac{\epsilon}{2}\right)\\ &+ P_\pi\left(\left|\frac{1}{n}\sum_{k=1}^n g(X_k) - \int_{\mathbb{R}^d} g(x)\pi(dx)\right| \geq \frac{\epsilon}{2}\right).\end{aligned}$$

Por outro lado, $g(x) = \int_A f(y \mid x)dy$ é limitada pois, para todo $A \in \mathcal{B}(\mathbb{R}^d)$,

$$0 \leq g(x) = \int_A f(y \mid x)dy \leq \int_{\mathbb{R}^d} f(y \mid x)dy = 1. \tag{2.35}$$

Segue do Teorema 2.2 que

$$P_\pi\left(\left|\frac{1}{n}\sum_{k=1}^n g(X_k) - \int_{\mathbb{R}^d} g(x)\pi(dx)\right| \geq \frac{\epsilon}{2}\right) \leq d_1 e^{-d_2 n}. \tag{2.36}$$

Defina $Z_k = I_A(Y_k) - g(X_k)$, $k = 1, \ldots, n$. De (2.34) temos,

$$E\left(g(X_k)\right) = \int_{\mathbb{R}^d} g(x)\pi(dx) = \int_A f(y)dy = P\left(Y_k \in A\right) = E\left(I_A(Y_k)\right) \quad \Rightarrow \quad E\left(Z_k\right) = 0$$

Como a distribuição de Y_k só depende do correspondente X_k e Z_k é a diferença de funções mensuráveis a Borel então $Z_1, \ldots, Z_n$ são variáveis aleatórias independentes. E mais, se $Y_k \in A$ então $Z_k = 1 - g(X_k)$ e como $0 \leq g(.) \leq 1$ segue que $Z_k \leq 1$. Caso $Y_k \notin A$ teremos $Z_k = -g(X_k)$ e como $0 \leq g(.) \leq 1$ então $Z_k \geq -1$. Ou seja, $-1 \leq Z_k \leq 1$.

Tomando $a_k = -1$ e $b_k = 1$ para todo $k = 1, \ldots, n$ e aplicando o Lema 2.3 temos,

$$\begin{aligned} P_\pi\left(\left|\frac{1}{n}\sum_{k=1}^{n}[I_A(Y_k) - g(X_k)]\right| \geq \frac{\epsilon}{2}\right) &= P_\pi\left(\left|\sum_{k=1}^{n}[I_A(Y_k) - g(X_k)]\right| \geq \frac{n\epsilon}{2}\right) \\ &\leq 2\exp\left\{\frac{\frac{-n^2\epsilon^2}{2}}{\sum_{k=1}^{n}(1-(-1))^2}\right\} = 2\exp\left\{\frac{\frac{-n^2\epsilon^2}{2}}{4n}\right\} \\ &= 2\exp\left\{\frac{-n^2\epsilon^2}{2}\frac{1}{4n}\right\} = 2\exp\left\{\frac{-n\epsilon^2}{8}\right\}. \end{aligned}$$

Fazendo $d_1^{'} = 2$ e $d_2^{'} = \frac{\epsilon^2}{8}$ concluimos que,

$$P_\pi\left(\left|\frac{1}{n}\sum_{k=1}^{n}[I_A(Y_k) - g(X_k)]\right| \geq \frac{\epsilon}{2}\right) \leq d_1^{'}e^{-d_2^{'}n}. \tag{2.37}$$

De (2.36) e (2.37),

$$P_\pi\left(\left|\frac{1}{n}\sum_{k=1}^{n}I_A(Y_k) - \int_A f(y)dy\right| \geq \epsilon\right) \leq d_1e^{-d_2n} + d_1^{'}e^{-d_2^{'}n}. \tag{2.38}$$

Sejam $c_1^{'} = d_1 + d_1^{'}$ e $c_2^{'} = \min\{d_2, d_2^{'}\}$. Então

$$\begin{aligned} P_\pi\left(\left|\frac{1}{n}\sum_{k=1}^{n}I_A(Y_k) - \int_A f(y)dy\right| \geq \epsilon\right) &\leq d_1e^{-d_2n} + d_1^{'}e^{-d_2^{'}n} \\ &\leq d_1e^{-c_2^{'}n} + d_1^{'}e^{-c_2^{'}n} \\ &= (d_1 + d_1^{'})e^{-c_2^{'}n} = c_1^{'}e^{-c_2^{'}n}. \end{aligned}$$

■

Como conseqüência do Teorema 2.3 e da Proposição 2.2 segue o seguinte corolário.

Corolário 2.2 *Sob as hipóteses do Teorema 2.3 temos*

$$\lim_{n\to\infty}\int_{\mathbb{R}^d}|f_n(y) - f(y)|\,dy = 0 \qquad \text{q.c.}$$

desde que a série $\sum_{n\geq 1}c_1^{'}e^{-nc_2^{'}}$ *seja convergente.*

Demonstração. Pelo Teorema 2.3;

$$P_\pi\left(\int_{\mathbb{R}^d}|f_n(y) - f(y)|\,dy \geq \epsilon\right) \leq c_1^{'}e^{-nc_2^{'}}.$$

Sob a hipótese de que $\sum_{n\geq 1} c_1^{'} e^{-nc_2^{'}} < \infty$ temos,

$$\sum_{n\geq 1} P_\pi \left(\int_{\mathbb{R}^d} |f_n(y) - f(y)|\, dy \geq \epsilon \right) \leq \sum_{n\geq 1} c_1^{'} e^{-nc_2^{'}} < \infty.$$

Usando a Proposição 2.2 obtemos,

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} |f_n(y) - f(y)|\, dy = 0 \qquad \text{q.c.}$$

Considerações Finais

Neste trabalho estudamos os modelos de Markov ocultos sob duas perspectivas diferentes: o paramétrico e o não-paramétrico.

No primeiro caso supomos que o espaço de estados tanto do processo observado $\{Y_t\}_{t\geq 0}$ quanto do processo de Markov $\{X_t\}_{t\geq 0}$ são finitos. Neste caso, calculamos a probabilidade da sequência de observações pelos procedimentos *forward* e *backward* e encontramos os estimadores de máxima verossimilhança via algoritmo EM. Além disso, avaliamos esses estimadores via simulação. No segundo caso, supomos que o processo de Markov $\{X_t\}_{t\in\mathbb{N}}$ e o processo observado assumem valores em um conjunto não enumerável, ou seja, em espaços de estados geral. Neste caso, estabelecemos as condições a respeito da cadeia e do processo $\{Y_t\}_{t\geq 0}$ para que

$$\lim_{n\to\infty}\int_{\mathbb{R}^d}|f_n(y)-f(y)|\,dy=0 \qquad \text{q.c.}$$

sendo $f(\cdot)$ é a densidade de Y_k e $f_n(\cdot)$ o estimador do tipo núcleo, ou seja, para uma densidade K em $\mathbb{R}^d$,

$$f_n(y)=\frac{1}{nh^d}\sum_{i=1}^{n}K\left(\frac{Y_i-y}{h}\right),$$

com $h=h_n$ uma sequência de constantes positivas satisfazendo

$$h_n \xrightarrow{n} 0 \qquad \text{e} \qquad nh_n \xrightarrow{n} \infty \qquad \text{quando} \quad n\longrightarrow\infty.$$

Ou seja, o estimador do tipo núcleo em $\mathbb{R}^d$ converge na norma L_1 para a densidade das variáveis Y_k do processo observado quase certamente.

Em Dorea-Zhao (2002) também é obtida a convergência na norma L_1 do estimador do tipo núcleo para a densidade do processo observado quando é retirada a suposição de que a distribuição inicial do processo de Markov é a estacionária. Em outras palavras, eles mostram os mesmos resultados dos Teoremas 2.2 e 2.3, supondo uma distribuição inicial qualquer para o processo de Markov.

Referências Bibliográficas

ATHREYA, K. B.; ATUNCAR, G. S. Kernel estimations for real-valued Markov chains. *Sankhya: The Indian Journal of Statistics*. Vol. 60, p. 1-17, 1998.

ATHREYA, K. B.; LAHIRI, S. N. *Measure theory and probability theory*. 1. ed. Springer. 2006. 618 p.

BAUM, L. E.; PETRIE, T. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*. Vol. 37, p. 1554-1563, 1966.

BICKEL, P. J.; RITOV, Y. Inference in hidden Markov models I: Local asymptotic Normality in the stationary case. *Bernoulli*. Vol. 2, p. 199-228, 1996.

BICKEL, P. J.; RITOV, Y.; RYDEN, T. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics*. Vol. 26, p. 1614-1635, 1998.

BILLINGSLEY, P. Statistical methods in Markov chains. *The Annals of Mathematical Statistics*. Vol. 37, p. 12-40, 1961.

CACOULLOS, T. Estimation of a multivariate density. *Ann. Inst. Statist. Math.*. Vol. 18, p. 178-179, 1964.

CAMPOS, V. S. M.; DOREA, C. C. Y. Kernel density estimation: the general case. *Statistics and Probability Letters*. Vol. 55, p. 173-180, 2001.

CAMPOS, V. S. M.; DOREA, C. C. Y. Kernel estimation for stationary density of Markov chains with general state space. *Ann. Inst. Statist. Math.*. Vol. 57, p. 443-453, 2005.

CAPPE, O.; MOULINES, E.; RYDEN, T. *Inference in hidden Markov models*. 1. ed. New York: Springer, 2005. 652 p.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*. Vol. 39, p. 1-38, 1977.

DEVROYE, L. Exponential inequalities in nonparametric estimation, *Nonparametrie Functional Estimations and Related Topics*. (ed. G. G. Roussas). *Kluwer Academic Publishers*. 31-44. 1991.

DEVROYE, L. The equivalence of weak, strong and complete convergence in L_1 for kernel density estimates. *Annals of Statistics*. Vol. 11, p. 896-904, 1983.

DEVROYE, L. GYORFI, L. *Nonparametric Density Estimation: The L_1 View*. John Wiley & Sons. 1985.

DOOB, J. L. *Stochastic Processes*. John Wiley & Sons. New York. 1953.

DOREA, C. C. Y.; ZHAO, L. C. Nonparametric density estimation in hidden Markov models. *Statistical Inference for Stochastic Processes*. Vol. 5, p. 55-64, 2002.

FREDKIN, D. R.; RICE, J. A. Maximum likelihood estimation and identification directly from single-channel recordings. *Proc. Royal Soc. London*. Vol. 249, p. 125-132. 1992.

HOEFFDING, W. Inequalities for sums of bounded random variables. *Journal of the American Statistical Association*. Vol. 58, p. 13-30, 1963.

HOLLAND, J. H. *Adaptation in Natural and Artificial Systems*. Michigan: University of Michigan Press, 1975.

HUANG, S.; AHMADI, M.; SID-AHMED, M. A. A hidden Markov model-based character extraction method. *The Journal of the Pattern Recognition Society*. Vol. 9, p. 2890-2900, 2008.

ISAACSON, D. L.; MADSEN, R. W. *Markov chains theory and applications*. Arrangement. Florida. 1985. 256 p.

LEROUX, B. G. Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*. Vol. 40, p. 127-143, 1992.

LEROUX, B. G.; PUTERMAN, M. L. Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*. Vol. 48, p. 545-558, 1992.

LINDGREN, G. Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics*. Vol. 5, p. 81-91, 1978.

MACDONALD, I. L.; ZUCCHINI, W. *Hidden Markov and other models for discrete-valued time series*. 1. ed. Chapman & Hall/crc. 1997. 236 p.

MEYN, S. L.; TWEEDIE, R. L. *Markov chains and stochastic stability*. 1. ed. Springer. 1993. 550 p.

PARZEN, E. On estimation of a Probability Function and its mode. *Annals of Mathematical Statistics*. Vol. 33, p. 1065-1076, 1962.

PETRIE, T. Probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*. Vol. 40, p. 97-115, 1969.

RABINER, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*. Vol.77, p. 257-284, 1989.

RAO, P. B. L. S. *Nonparametric functional estimation*. New York: Academic Press, 1983.

ROSENBLATT, M. Density estimates and Markov sequences. *Annals Math. Statistic*. Vol. 27, p. 832-837, 1970.

ROUSSAS, G. G. Estimation of transition distribution function and its quantiles in Markov processes: strong consistency and Asymptotic Normality. *Nonparametric Functional Estimation and Related Topics*. (ed. G. G. Roussas). Kluwer Academic Publishers. Dordrecht. p. 443-462, 1991.

ROUSSAS, G. G.; IONNIDES, D. Moment inequalities for mixing sequences of random variables. *Stochastic Analysis and Applications*. Vol. 5, p. 61-120, 1987.

ROUSSAS, G. G. Nonparametric estimation in Markov processes. *Annals of the Institute of Statistical Mathematics*. Vol. 21, p. 73-87, 1969.

RYDEN, T. Consistent and asymptotically normal parameter estimates for hidden Markov models. *Annals of Statistics*. Vol. 22, p. 1884-1895, 1994.

TIERNEY, L. Markov chains for exploring posterior distributions. *Annals of Statistics*. Vol 22, p. 1701-1728, 1994.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo