EMPREGO DO MODELO DE MARKOV ESCONDIDO NA ANÁLISE DE SÉRIES TEMPORAIS

MEIGAROM DIEGO FERNANDES LOPES

Universidade de São Paulo
Escola de Engenharia de São Carlos
Programa de Graduação em Engenharia Elétrica

São Carlos
2015

MEIGAROM DIEGO FERNANDES LOPES

EMPREGO DO MODELO DE MARKOV ESCONDIDO NA ANÁLISE DE SÉRIES TEMPORAIS

Trabalho de Conclusão de Curso apresentado ao Programa de Graduação da Escola de Engenharia de São Carlos da Universidade de São Paulo como parte dos requisitos para a obtenção do título de Engenheiro Eletricista.

Área de concentração: Sistemas Inteligentes

Orientador: Prof. Dr. Rodrigo Fernandes de Mello

São Carlos
2015

Lopes, Meigarom Diego Fernandes
L864e Emprego do modelo de Markov Escondido na análise de séries temporais / Meigarom Diego Fernandes Lopes; orientadora Rodrigo Fernandes de Mello. São Carlos, .

Monografia (Graduação em Engenharia Elétrica com ênfase em Sistemas de Energia e Automação) -- Escola de Engenharia de São Carlos da Universidade de São Paulo, .

1. Modelo de Markov Oculto. 2. Aprendizado de Maquina. 3. Series Temporais. 4. Cadeia de Markov. 5. Algoritmo de Viterbi. 6. Algoritmo de Maxima Verossimilhanca. 7. Algoritmo de Baum-Welch. 8. Modelo de Misturas. I. Título.

FOLHA DE APROVAÇÃO

Nome: Meigarom Diego Fernandes Lopes

Título: "Emprego do modelo de Markov Escondido na análise de séries temporais"

Trabalho de Conclusão de Curso defendido e aprovado em 01/12/2015,

com NOTA 10,0 (dez), pela Comissão Julgadora:

Prof. Associado Rodrigo Fernandes de Mello - (Orientador - SCC/ICMC/USP)

Prof. Associado Rogério Andrade Flauzino - (SEL/EESC/USP)

Prof. Associado Ivan Nunes da Silva - (SEL/EESC/USP)

Coordenador da CoC-Engenharia Elétrica - EESC/USP:
Prof. Dr. José Carlos de Melo Vieira Júnior

Agradecimentos

Primeiramente à Deus pela saúde e por todos os desafios proporcionados que fortaleceram e deram base pra chegar até aqui. Gostaria de agradecer a minha família pelo apoio incondicional, segurança e tranquilidade que me ajudaram a seguir em frente. Agradecer a todos os amigos que compartilharam boas e ruins experiência e fizeram a caminhada ser mais divertida. Em especial, agradeço aos amigos, Marcus Vinícius dos Reis e Rachid Elias Arab, pela amizade e pelo companheirismo nos diversos momentos durante esses anos de graduação. Finalmente, agradeço à todos os mestres que me ensinaram muito além da teoria, sempre solícitos as requisições de ajuda, principalmente ao professor Rodrigo Fernandes de Mello que me deu a oportunidade de desenvolver esse trabalho sob sua tutoria.

"Nada pode ser obtido sem uma espécie de sacrifício. Para conseguir alguma coisa é preciso oferecer em troca algo de valor equivalente."
(Lei da Troca Equivalente)

Resumo

O Modelo de Markov Escondido é um método estatístico de aprendizado da área de Aprendizado de Máquina. Esse modelo se mostra bastante eficiente na detecção de padrões em séries temporais, pois trata de forma probabilística a variação estrutural dos elementos e ainda possui uma flexibilidade para modelar diferentes naturezas de dados. Este documento apresenta um estudo do Modelo de Markov Escondido com o objetivo de detalhar todo o processo estocástico, através de sua formulação por meio de equações, conceitos e métodos provenientes da Matemática e Estatística. Por fim, as expressões resultantes da formulação teórica serão implementadas utilizando uma linguagem de programação, a fim de se obter um ferramental de funções a serem utilizadas na modelagem experimental de séries.

Palavras-chave: Modelo de Markov Escondido. Aprendizado de Máquina. Método Estatístico. Séries temporais. Processo Estocástico.

Abstract

The Hidden Markov Model is a statistical method from the Machine Learning field. This model turns out to be very efficient to detect patterns embedded in time series, once it deals with the structural variation of data in a probabilistic way and it still has the flexibility to model data differences. This document presents a theoretical study of Hidden Markov Models with the goal of detailing the stochastic process throughout equations, concepts and methods provided by Mathematical and Statistics fields. Therefore, mathematical expressions resultants of theoretical studies will be implemented using a programming language in order to obtain a set of tools to be applied in time series analysis.

Keywords: Hidden Markov Model. Machine Learning. Statistics methods. Time Series. Stochastic Process.

Lista de ilustrações

Lista de tabelas

Sumário

Capítulo 1

Introdução

Um fenômeno físico produz resultados observáveis, os quais podem ser analisados a fim de se obter a regra ou função responsável por gerá-los. Um conjunto dessas observações, ao longo do tempo, formam um sinal [2]. Sinais podem ser de natureza contínua, quando, por exemplo, são obtidos à partir de sensores analógicos, ou discreta, quando o suas amplitudes apresentam subconjunto enumerável de valores. Uma vez manipulados por meio de computadores, esses sinais são, em algum momento, discretizados. Esse é o caso de sinais de áudio, eletrocardiograma, temperaturas de uma região, etc.

Esses dados discretos são tipicamente organizados na forma de séries temporais. Uma série temporal é uma sequência de observações de uma variável ao longo do tempo, coletadas, em geral, em intervalos uniformes [2]. Uma característica importante deste tipo de dado é que ele pode conter dependências entre observações, sendo necessária sua modelagem para melhor compreensão do fenômeno responsável pela geração desses dados.

Os modelos construídos a partir de séries temporais podem ser divididos em: determinísticos e estocásticos. Os modelos determinísticos são capazes de produzir um único conjunto de saídas à partir de uma entrada, enquanto que os estocásticos contam com incertezas relativas à geração de próximas observações [3]. O Mapa Logístico [4] e o Sistema de Lorenz [5] são exemplos de fenômenos que podem ser modelados apenas com componentes determinísticos, enquanto que o movimento de queda do pólen, influenciado pela gravidade, é caracterizado como um movimento Browniano [6] e, portanto, estocástico.

Este trabalho de conclusão de curso emprega o Modelo de Markov Escondido (do inglês *Hidden Markov Model* – **HMM**) para representar fenômenos à partir de séries temporais observáveis. Esse modelo é baseado no ramo estocástico de representação de dados. Mesmo assim, ele produz resultados relevantes para cenários com eventuais dependências temporais.

1.1 Introdução ao Modelo de Markov Escondido

O Modelo de Markov Escondido (**HMM**) considera misturas de distribuições de probabilidade a fim de caracterizar estados que compõem um processo de Markov e suas transições. Tipicamente, busca-se modelar esses estados utilizando distribuições de probabilidade bem conhecidas, tais como a Normal e a Poisson, além disso, o processo de Markov pode ser de primeira ou de maior ordem. Os estados do *HMM* não são conhecidos, ou seja, os parâmetros das distribuições de probabilidade não são definidos, sabe-se que há uma relação de probabilidade entre os estados não-observáveis e sequência de dados observáveis. Esse modelo busca encontrar esses parâmetros utilizando um algoritmo de maximização à partir de um conjunto de dados.

Três problemas de inferência são solucionados pelo **HMM**: i) o primeiro é o problema de avaliação (do inglês *evaluation problem*), o qual diz respeito a uma sequência de observações: dado um modelo de Markov Escondido e uma sequência de observações, calcula-se a probabilidade de que a sequência observada tenha sido produzida pelo modelo. Sua resolução pode ser vista como uma medida sobre quão representativo um dado modelo é para um sequência de observações. Por exemplo, considere o caso no qual um modelo precisa ser escolhido dentre muitos outros candidatos, assim, a solução desse problema permite encontrar o melhor modelo candidato para uma dada sequência de observações; ii) o segundo problema é chamado de decodificação (do inglês *decoding problem*), a qual visa descobrir a melhor sequência de estados capaz de produzir uma sequência de observações. Por exemplo, considere uma sequência de valores que representa a umidade relativa do ar observados diariamente, os estados relacionados a essa sequência poderiam ser: clima ensolarado ou clima chuvoso, por exemplo. A determinação da sequência de estados visitada permite identificar a mudança das fontes que originam os dados, assim, dada uma nova sequência, pode-se prever se dias seguintes serão ensolarados ou chuvosos, através da identificação da mudança de estados; iii) por fim, o terceiro problema é o de aprendizado (do inglês *learning problem*), o qual busca otimizar os parâmetros do modelo, ou seja, tenta encontrar a melhor configuração para as distribuições de probabilidade adotadas e, portanto, que melhor representam a origem da sequência de observações. Nesse cenário, a sequência de dados utilizadas para ajustar os parâmetros do modelo é chamada de sequência de treinamento ou dados de treinamento.

1.2 Objetivo do Trabalho

O objetivo desse trabalho é estudar o funcionamento estocástico do **HMM** representado pelo processo da cadeia de Markov Escondida, a modelagem utilizando variáveis não-conhecidas, os conceitos da Estatística empregados a fim de quantizar esse processo, o entendimento básico de otimização e a aplicação dos resultados em séries temporais

reais com o objetivo de verificar a eficiência e as limitações desse modelo tanto em termos de representatividade dos dados com dependência temporal quanto na realização de previsões.

Primeiramente, foram estudados os conceitos de **HMM**, destacando a teoria Estatística envolvida, o objetivo do modelo e os tipos de problemas que ele tenta solucionar, bem como suas limitações para diferentes cenários. Em conjunto, foi realizada uma revisão da teoria relativa à Cadeias de Markov, a qual é utilizada para compor o Modelo Escondido. Todas essas etapas desse estudo foram transcritas neste documento. Em seguida, os três problemas de inferência solucionados pelo **HMM** foram melhor estudados e aplicados na modelagem de duas séries temporais aqui propostas: uma sem e outra com dependências entre observações. A primeira série temporal foi sinteticamente construída, ou seja, os dados foram originados de distribuições de probabilidades com parâmetros previamente conhecidos a fim de comprovar os resultados teóricos e também a eficiência do modelo. A segunda série temporal utilizada corresponde a medidas de temperaturas máxima diárias entre os dias 01/01/1981 e 31/12/1990 em Melbourne, Austrália [7]. Em particular, essa série ajuda a demonstrar a resolução do segundo problema do HMM, o qual está relacionado à decodificação. Além disso, nas duas séries temporais, o problema de aprendizado também será evidenciado. Cabe ressaltar a relevância de se estudar **HMM**, pois o proponente deste trabalho pretende aplicá-lo em outros problemas, após a conclusão de sua graduação.

1.3 Organização deste Trabalho

As características do **HMM**, bem como suas propriedades e limitações são apresentadas no Capítulo 2, juntamente com uma introdução a Cadeias de Markov. No Capítulo 3, o algoritmo de treinamento do **HMM** é formulado a partir de ferramental estatístico e seguindo os princípios definidos para Cadeias de Markov. Além disso, apresenta-se um método iterativo para a otimização de parâmetros do **HMM**, o qual é conhecido com Algoritmo de Máxima Verossimilhança (do inglês *Maximum Likelihood* – **ML**). O Capítulo 4 apresenta os resultados obtidos para as duas séries estudadas e também as conclusões deste trabalho, bem como direções futuras.

Capítulo 2

Modelo de Markov Escondido

O principal interesse em modelar séries temporais, utilizando o Modelo de Markov Escondido, está em determinar a função de probabilidade geradora dos dados da série. Determinar a função de probabilidade geradora significa encontrar os parâmetros que caracterizam a função de probabilidade. **HMM** permite modelar utilizando qualquer função de distribuição de probabilidades que melhor se adeque ao problema alvo, como por exemplo, a distribuição de Poisson, a Binomial, a Normal, etc., o que permite modelar dados tanto de natureza discreta quanto contínua.

Algumas técnicas são muito utilizadas na modelagem de séries temporais, como por exemplo a técnica de Média Móvel Auto-Regressiva (do inglês *Auto-Regressive Moving Average* – **ARMA**). Resumidamente, **ARMA** ajusta um modelo para descrever um processo estocástico estacionário, em que uma função polinomial é utilizada para a auto-regressão e uma outra função para a média móvel [8]. Contudo, essa técnica tem como limitante o fato de representar apenas dados que seguem a distribuição Normal [1], não sendo adequada para modelar dados discretos, por exemplo.

Nas próximas seções, são apresentadas a formulação do Modelo de Markov Escondido bem como suas premissas e propriedades.

2.1 Modelo de misturas independentes

Os dados de uma série temporal, geralmente, não são originários de apenas uma única distribuição de probabilidade e sim, de múltiplas distribuições diferentes entre si. A formação da série temporal pode ser interpretada do seguinte modo: a cada instante de tempo, um valor da série é gerado por alguma distribuição de probabilidades dentre um conjunto e a escolha do conjunto gerador, para cada instante de tempo, é um processo aleatório.

Para identificar se uma série foi gerada a partir de um conjunto de distribuições escolhidas

aleatoriamente a cada instante de tempo t ou se a série foi gerada a partir de uma única distribuição, podemos utilizar o conceito de **Super Dispersão** (do inglês *Overdispersion*). Em Estatística, Super Dispersão refere-se à presença de maior variabilidade sobre os dados do que aquela esperada pelo modelo teórico [9]. Para entender melhor esse conceito, considere o seguinte conjunto de dados apresenta na Tabela 1.

Tabela 1 – Número anual de ocorrências de terremotos com escalar Richter maior ou igual a 7, no mundo, entre os anos de 1900-2006.

13	14	08	10	16	26	32	27	18	32	36	24	22	23	22	18	25
21	21	14	08	11	14	3	18	17	19	20	22	19	13	26	13	14
22	14	21	22	26	21	23	24	27	41	31	27	35	26	28	36	39
21	17	22	17	19	15	34	10	15	22	18	15	20	15	22	19	16
30	27	29	23	20	16	21	21	25	16	18	15	18	14	10	15	08
15	06	11	08	07	18	16	13	12	13	20	15	16	12	18	15	16
13	15	16	11													

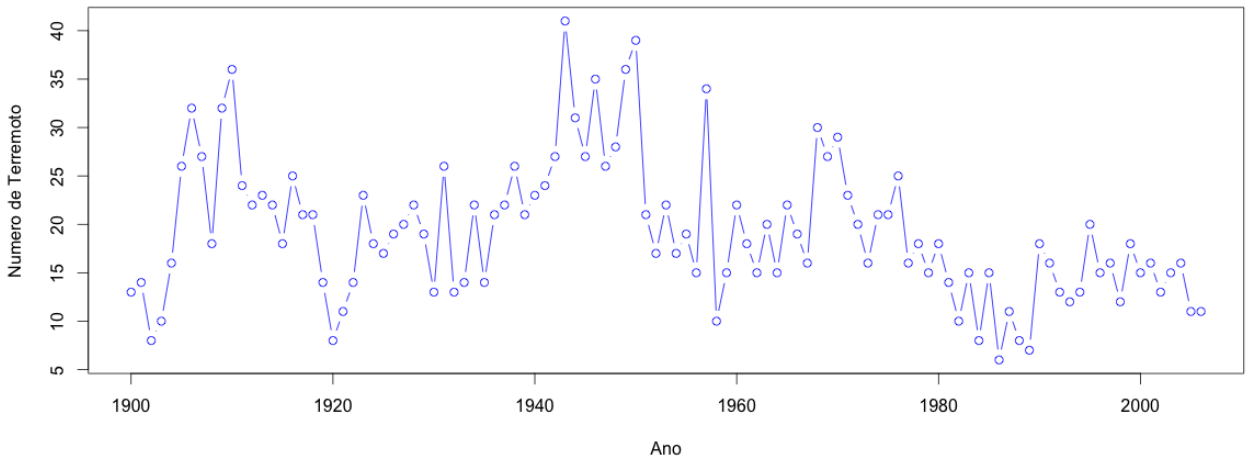

Figura 1 – Número anual de terremotos com escalar Richter maior ou igual a 7, no mundo, entre os anos de 1900-2006.

A Tabela 1 mostra o número de ocorrências de terremotos anuais com escala Richter maior ou igual a 7, entre os anos de 1900 até 2006. Pode-se observar que esses dados são discretos, o que torna distribuições tal como a de Poisson, mais adequadas para modelar essa série.

A Figura 2 representa os valores do conjunto de dados da Tabela 1.

A distribuição de Poisson tem como propriedade o fato de que a média e a variância apresentam valores iguais. Logo, há super dispersão, em um conjunto de dados, que apresenta média diferente do valor da variância. Isso motiva a análise sobre a sequência apresentada na Tabela 1. Calculando-se a média tem-se $\bar{x} \approx 19$ e variância igual a

$s^2 \approx 52$. Portanto, esse dados apresentam Super Dispersão, o que faz com que não seja bem representado por uma única distribuição de probabilidades tal como ilustrado na Figura 2.

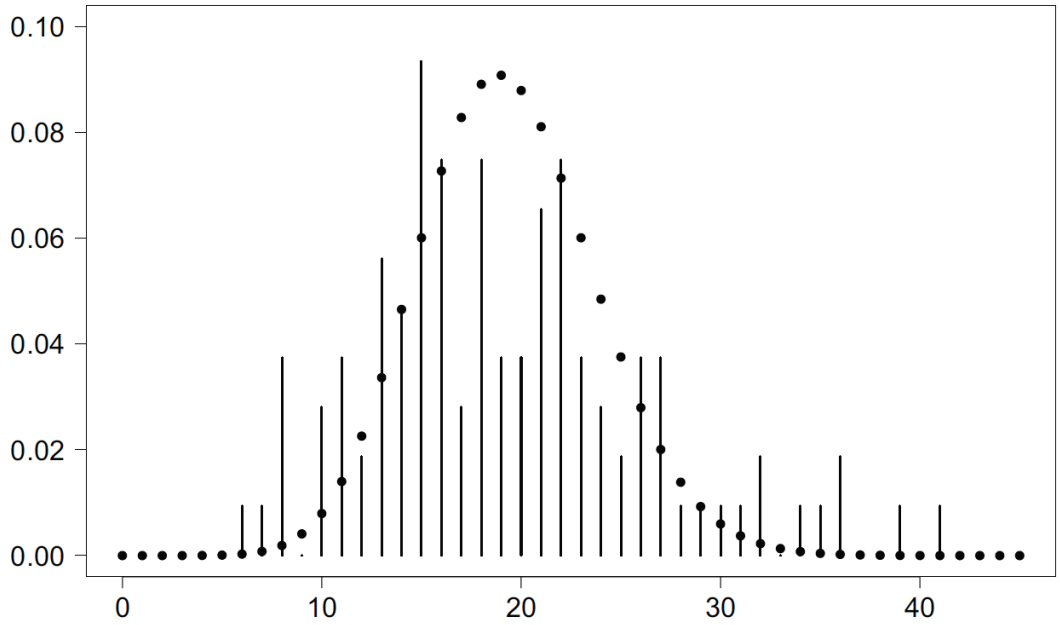

Figura 2 – Modelagem dos dados utilizando apenas 1 distribuição de Poisson (Adaptado de [1])

Um método para lidar com a super dispersão é o **Modelo de Misturas** (do inglês *Mixture Model*), o qual tem a capacidade de representar, por meio de múltiplas distribuições, o comportamento de um conjunto de dados heterogêneo [1]. Por exemplo, considere uma variável aleatória X que representa a compra de cigarros por consumidores em um supermercado, esses consumidores podem ser divididos em 3 grupos distintos: os não-fumantes, os fumantes ocasionais e os fumantes regulares. Neste cenário, mesmo que o número de cigarros comprados pelos consumidores dentro de cada grupo obedeça à uma distribuição de Poisson, a distribuição de X não será modelada por uma única Poisson, pois esse conjunto certamente apresentará super dispersão.

2.2 Parâmetros do Modelo de Misturas

Para exemplificar o Modelo de Misturas, esta seção apresenta um exemplo construído sobre o conjunto de dados apresentado na Tabela 1. Primeiramente, suponha que esses dados foram gerados por duas distribuições de Poisson, cada qual com média igual a λ_1 e λ_2, os quais foram determinados por um processo aleatório denominado **Processo de Parâmetros** (do inglês *Parameter Process*). Seja, também, δ_1 e δ_2 variáveis que indicam a probabilidade de selecionar as distribuições com parâmetro λ_1 e λ_2, respectivamente.

Assim, pode-se definir uma expressão que representa o Modelo de Misturas com essas duas distribuições de Poisson, tal como apresentado na Equação 1, em que a distribuição de Poisson é dada por $p(x) = \frac{e^{-\lambda}\lambda^2}{x!}$, $p_1(x)$ e $p_2(x)$ são as funções distribuição de probabilidades de Poisson com os parâmetros λ_1 e λ_2, respectivamente.

$$p_m(x) = \delta_1 p_1(x) + \delta_2 p_2(x) \tag{1}$$

Cada distribuição de probabilidade é chamada de **Distribuição Componente**(do inglês *Component Distribution*), no contexto de **HMM**. Toda mistura é caracterizada por apresentar um número finito de Distribuições Componentes, no qual m corresponde a esse número na forma $C_{i=1,2,\ldots,m}$.Ao analisar a Equação 1, pode-se dizer que uma observação é produzida pela distribuição componente C_1 com probabilidade δ_1, enquanto que a mesma observação pode ser produzida pela componente C_2 com probabilidade δ_2. Como um exemplo ilustrativo, pode-se supor o lançamento de uma moeda. Neste cenário, δ_1 é a probabilidade da face "Cara"ocorrer, enquanto δ_2 é a probabilidade da face "Coroa". A fim de ilustrar a Equação 1 sobre os dados da Tabela 1, a Figura 3 será utilizada. A coluna denominada "transição de componentes, representa os valores das probabilidades de selecionar umas das distribuições componentes, a segunda coluna ilustra a densidade de probabilidades para cada componente na forma de um gráfico e, a última coluna, representa os valores produzidos por uma das distribuições componentes. A probabilidade componente C_1 é representada por $p_1(x)$ e tem 75% de chance de ser escolhida enquanto que a probabilidade componente C_2, representada por $p_2(x)$, tem 25%. Na primeira linha, pode-se notar que a distribuição componente C_1 foi selecionada, por meio da indicação de um ponto preenchido, e o valor de 14.2 foi produzido (coluna "observações"). A segunda linha indica que, o próximo valor da sequência 29.4 foi gerado pela distribuição C_2, assim sucessivamente.

Essa abordagem pode ser generalizada para qualquer número de componentes. Sendo que $\delta_1, \delta_2, \ldots, \delta_m$ representam as probabilidades de escolher cada uma das distribuições componentes $p_1, p_2, \ldots, p_m$, respectivamente. Sendo também, que a variável aleatória da distribuição de mistura X, pode ser escrita na forma:

$$p(x) = \sum_{i=1}^{m} \delta_i p_i(x). \tag{2}$$

Também pode-se representar a distribuição de misturas para casos discretos, tal como:

$$Pr(X = x) = \sum_{i=1}^{m} Pr(X = x|C = i)Pr(C = i). \tag{3}$$

Após ter encontrado a função que descreve o comportamento da mistura de modelos, precisa-se estimar seus parâmetros δ e λ (neste caso ilustrativo, para a distribuição de

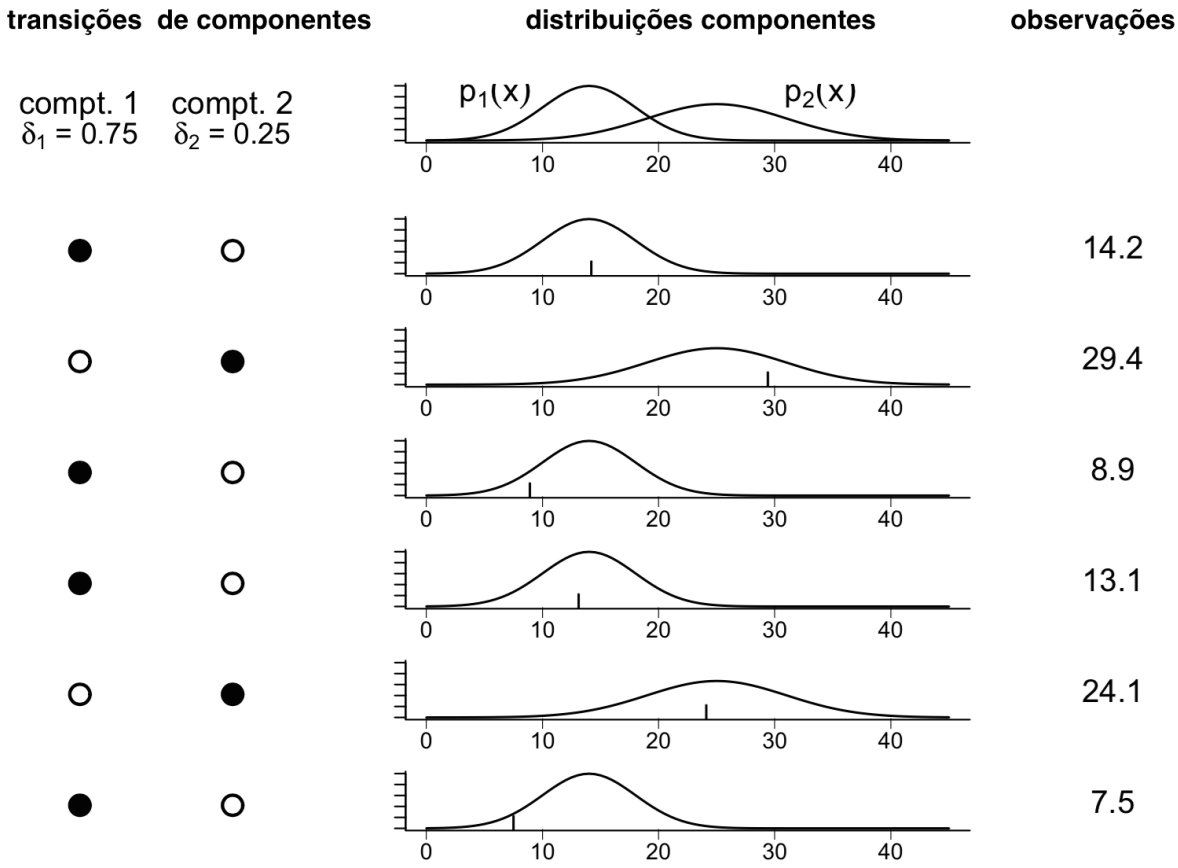

Figura 3 – Estrutura de uma mistura de distribuições componentes (Adaptada de [1])

Poisson). A estimação dos parâmetros da mistura é geralmente realizada pela **Máxima Verossimilhança** (do inglês *Maximum Likelihood* – *Maximum Likelihood* – **ML**).

Em geral, a probabilidade de um Modelo de Misturas com m distribuições componentes é dada por:

$$L(\theta_1, ..., \theta_m, ..., \delta_1, ..., \delta_m | x_1, ..., x_n) = \prod_{j=1}^{n} \sum_{i=1}^{m} \delta_i p_i(x_j, \theta_i) \tag{4}$$

Antes de estudar o algoritmo de **ML**, deve-se apresentar o conceito de Cadeia de Markov, bem como suas propriedades, uma vez que será considerada na construção de **HMMs**.

2.3 Cadeia de Markov

A teoria clássica de probabilidades lida com experimentos nos quais os processos são supostos como independentes. Um processo independente significa que seus resultados prévios não influenciam em valores seguintes. Outra vertente da área de probabilidades busca verificar se há essa influência. Esses estudos supõem a dependência entre próximas observações e as anteriores. Em 1907, A. A. Markov começou a estudar esse novo tipo

de processo de dependência, a partir do qual definiu as **Cadeias de Markov** (do inglês *Markov Chain*) [10]. Essas cadeias são utilizadas na formulação do Modelo de Markov escondido.

2.3.1 Definições e Propriedades

Uma sequência de variáveis aleatórias, na Cadeia de Markov são definidas como estados: $C = s_1, s_2, \ldots, s_T$. Os estados são interligados entre si por "arestas direcionadas"ou "caminhos", que partem de cada estado em direção a todos os demais, inclusive o próprio. Esses caminhos têm probabilidades associadas, ou seja, dado um estado qualquer, há a probabilidade de mudar para outro estado ou continuar no mesmo, segundo os valores de probabilidades associados às arestas que partem da origem. O processo começa em um estado qualquer e move-se, sucessivamente, para outro. Cada mudança de estado é chamada de "passo"ou "transição". Por exemplo, seja o estado atual s_i, move-se para o estado s_j no próximo passo com probabilidade definida por p_{ij}. Essa probabilidade é chamada de **Probabilidade de Transição**(do inglês *Transition Probabilities*).

Esses estados caracterizam uma Cadeia de Markov desde que ela satisfaça a propriedade de Markov. Essa propriedade define que: uma sequência de estados, iniciando em algum instante de tempo passado até o instante atual, pode ser representada apenas pelo último estado, ou seja, o estado o mais recente [1].

A partir dessa definição, pode-se formalizar uma Cadeia de Markov tal como:

$$Pr(C_{t+1}|C_t, \ldots, C_1) = Pr(C_{t+1}|C_t) \tag{5}$$

em que $C_t : t \; \varepsilon \; \mathbb{N}$ é uma sequência de variáveis aleatórias discretas.

A Figura 4 ilustra uma Cadeia de Markov. Pode-se observar que todos os estados estão conectados, permitindo a completa mobilidade entre eles. Dado o estado atual, pode-se encontrar a probabilidade de um determinado estado ocorrer no próximo passo.

A probabilidade de transição entre todos os estados pode ser expressada em uma forma matricial, chamada de matriz de **transição de probabilidade t-passo** (do inglês *Transition Probabilities Matrix*):

$$\Gamma = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ \vdots & \ddots & \vdots \\ \gamma_{m1} & \cdots & \gamma_{mm} \end{pmatrix}.$$

No qual, $\gamma_{ij}(t) = Pr(C_{s+t} = j|C_s = i)$.

Além da probabilidade de transição, há a probabilidade de se manter em um determinado estado no instante de tempo t. Essa probabilidade é chamada de **Probabilidade Incon-**

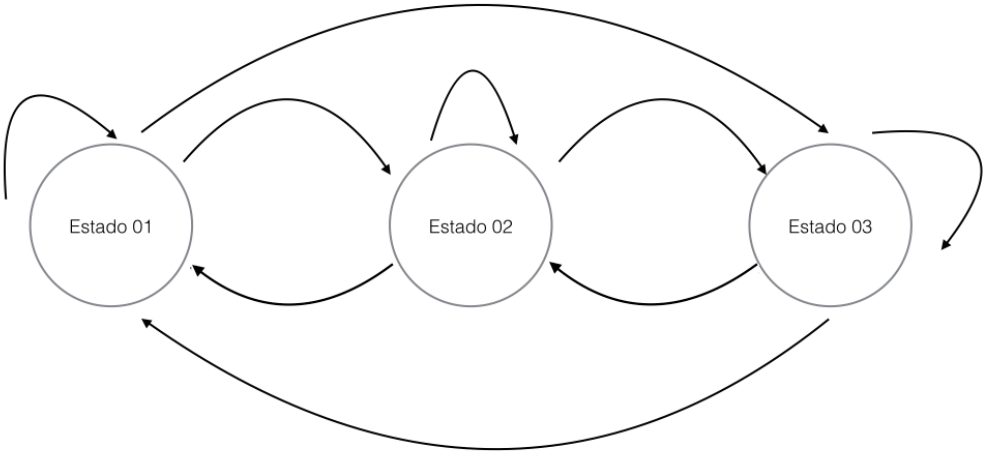

Figura 4 – Ilustração de uma Cadeia de Markov.

dicional (do inglês *Unconditional Probabilities*), $Pr(C_t = j)$, a qual é definida por um vetor de linhas na forma:

$u(t) = Pr(C_t = 1), \ldots, Pr(C_t = m), t\varepsilon\mathbb{N}$.

O conjunto de vetores de linhas para todos os estados da Cadeia de Markov é denominado **Probabilidade Inicial** (do inglês *Initial Probabilities*), a qual será definida como $u(1)$.

2.3.2 Exemplo de uma Cadeia de Markov

Considere a seguinte matriz de probabilidades de transição, apresentada na Tabela 2, a qual ilustra mudanças entre dias ensolarados e chuvosos, em que $C_t = 1$: dia chuvoso e $C_t = 2$: dia ensolarado.

Tabela 2 – Probabilidade de mudança do clima

	dia t+1	
dia t	chuvoso	ensolarado
chuvoso	0.9	0.1
ensolarado	0.6	0.4

Por exemplo, se hoje (dia t) o dia está chuvoso, a probabilidade de amanhã (dia $t + 1$) o dia continuar chuvoso é de 0.9. Ou ainda, se hoje o dia está com o clima chuvoso,

a probabilidade de que amanhã esteja ensolarado é de 0.1 e assim por diante pode-se mapear todas as mudanças entre os dois estados climáticos.

Aumentando o detalhe desse exemplo, suponha que hoje ($t = 1$) o dia está ensolarado. Isso significa que a distribuição do clima de hoje é dada por:

$$u(1) = (Pr(C_1 = 1), Pr(C_1 = 2)) = (0, 1)$$

,

ou seja, a distribuição do clima de amanhã e do dia seguinte pode ser calculada multiplicando-se o estado atual pela matriz de probabilidades de transição, tal como:

$u(2) = (Pr(C_2 = 1), Pr(C_2 = 2)) = u(1)\Gamma = (0,6, 0,4)$

$u(3) = (Pr(C_3 = 1), Pr(C_3 = 2)) = u(1)\Gamma = (0,78, 0,22)$.

2.4 Cadeia de Markov Escondida

Antes de apresentar as características da **Cadeia de Markov Escondida** (do inglês *Hidden Markov Chain*), considere, novamente, o problema e a formulação descritos anteriormente. Considera-se uma série temporal com valores inteiros que representa o número de terremotos com escalar Richter maior ou igual a 7 por ano, contabilizados entre 1900 até 2007. Essas observações discretas podem, portanto, ser representadas por uma distribuição de probabilidades tal como a Poisson. Entretanto, após o cálculo da média e da variância dos dados, observa-se que esses parâmetros possuem valores diferentes, caracterizando a super dispersão dos dados. Logo, eles são formados por mais de uma distribuição de probabilidade, com parâmetros diferentes. Dessa maneira, uma boa escolha para modelar essa série numérica, é por meio de um modelo de misturas de distribuições de Poisson.

Essa mistura supõe que cada valor é gerado por uma das m distribuições que compõem o modelo, as quais são chamadas de distribuições componentes. A escolha dessas distribuições é feita por um mecanismo aleatório secundário, chamado de processo de parâmetros. Entretanto, essa fundamentação assume que os dados gerados pelo modelo de misturas são independentes, logo não se pode utilizar essa técnica sobre os dados de terremotos apresentados na Tabela 1, uma vez que apresentam dependência serial. Essa dependência serial posse ser observada através do resultado da função de autocorrelação aplicada sobre os dados 5.

Uma maneira de permitir essa dependência serial entre as observações é relaxar a premissa de que o processo de parâmetros é serialmente independente. Uma maneira conveniente de fazer isso é assumir que o processo de parâmetros é definido por uma Cadeia de Markov [1].

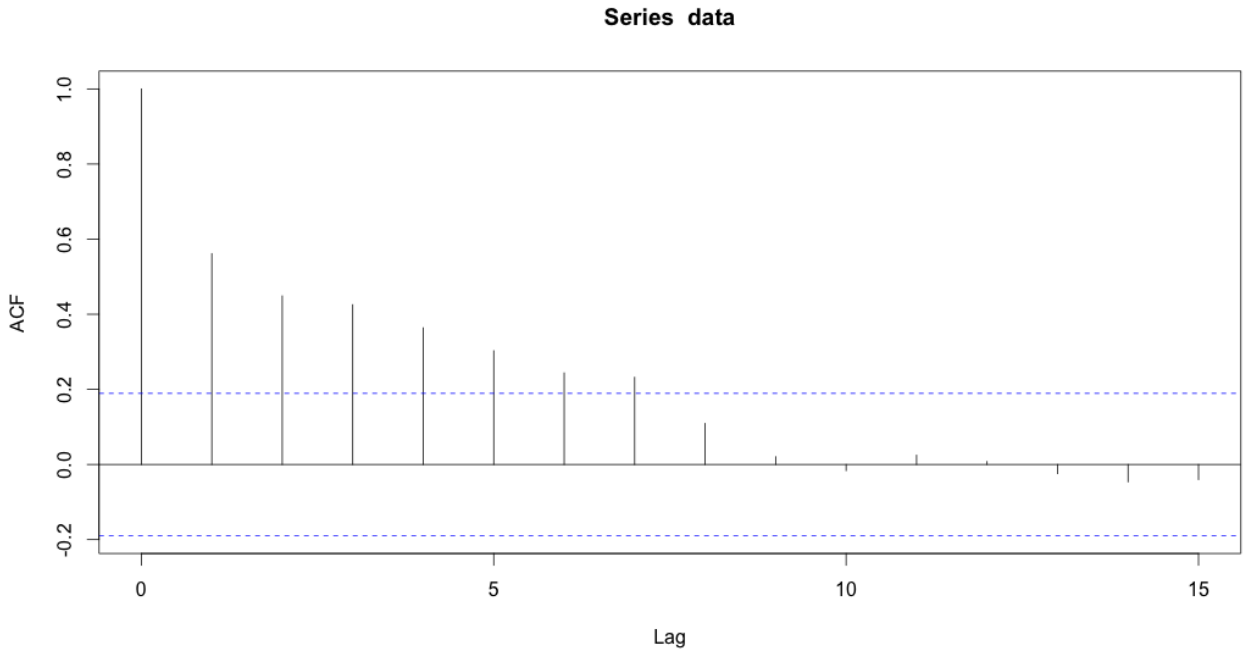

Figura 5 – Função de autocorrelação da série de terremotos.

2.5 Cadeia de Markov Escondia: Definição

A Cadeia de Markov Escondida recebe esse nome devido ao fato de ser capaz de produzir modelos que incorporam variáveis não-observáveis, chamada de *variáveis latentes*, sabe-se que há uma relação probabilística entre esses estados e as observações e entre os próprios estados, essa relação é representada por uma Cadeia de Markov. Matematicamente, os estados escondidos são distribuições de probabilidade cujo os parâmetros e o número de estados são desconhecidos.

A Figura 6 ilustra a estrutura do **HMM** e representa as dependências entre as variáveis do modelo.Neste cenário, pode-se definir, respectivamente, a relação entre os estados, definidas anteriormente como probabilidade de transição e também a relação entre os estados e as observações definidas como *probabilidades de emissão* (do inglês *Emission Probabilities*):

$$Pr(C_t|C_{(t-1)}) = Pr(C_t|C_{(t-1)}), t = 2, 3, \ldots$$

$$Pr(X_t|X_{(t-1)}, C_{(t)}) = Pr(X_t|C_t), t \, \epsilon \, \mathbb{N}.$$

Este modelo apresenta duas variáveis aleatórias. A primeira é estado $C = c$ e a segunda refere-se aos dados $X = x$ (observações). A fim formular a relação entre elas, deve-se utilizar a teoria de probabilidade conjunta.

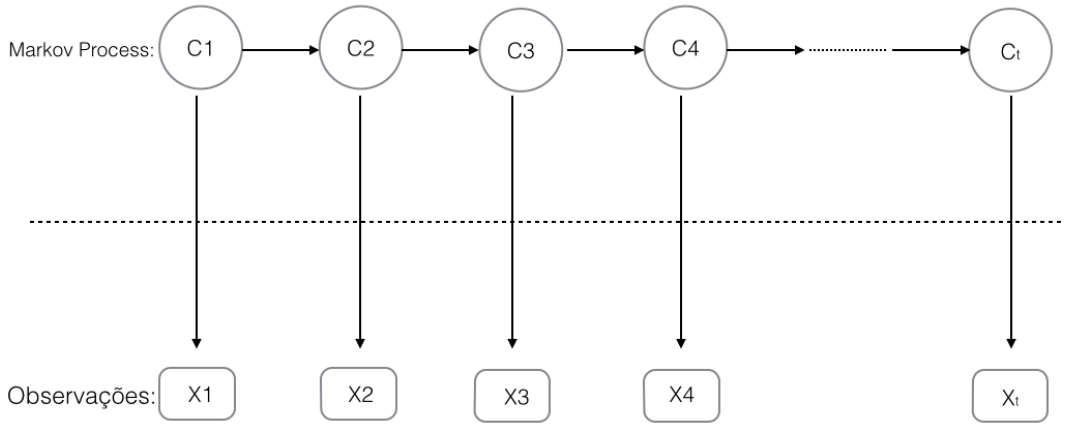

Figura 6 – Ilustração de uma Cadeia de Markov Escondida.

Até este ponto, definiu-se todas as variáveis envolvidas no Modelo de Markov Escondido, agora serão utilizadas algumas definições da Estatística e também a técnica de maximização da verossimilhança para encontrar o melhor conjunto de parâmetros para o modelo de misturas, de acordo com conjunto de dados fornecido.

Capítulo 3

Algoritmo de Treinamento

Este capítulo apresenta a formulação do algoritmo de treinamento para o **HMM**. Esse algoritmo busca encontrar o melhor conjunto de parâmetros para as probabilidades de estados iniciais, de transição e para os parâmetros das distribuições componentes de mistura. A série temporal 1 será utilizada no seu desenvolvimento.

3.1 Algoritmo de Máxima Verossimilhança

O algoritmo de otimização baseado em Máxima Verossimilhança (**ML**), tenta encontrar o maior valor de probabilidade de um conjunto de parâmetros sobre uma dada sequência de dados, considerando um cenário em que esses dados são incompletos ou não-observáveis (variáveis latentes.) Existem duas principais aplicações desse algoritmo. A primeira ocorre quando há, de fato, valores desconhecidos, devido a problemas ou limitações relacionadas ao processo de observação. A segunda aplicação se dá no cenário em que a otimização analítica de uma função de distribuição de probabilidades é muito difícil [11].

ML possui basicamente dois passos chamados de **E step** (do inglês *Expectation step*) e *M step* (do ingês *Maximization step*). O primeiro passo é responsável por computar a esperança com respeito às variáveis latentes, dados valores iniciais e um conjunto de observações, enquanto o segundo passo estima o novo conjunto de parâmetros de acordo com o critério de Máxima Verossimilhança.

Determinando a função de verossimilhança (do inglês *likelihood function*), obtém-se:

$$P(X|\theta) = \prod_{i=1}^{N} P(X_i|\theta) = L(\theta|X) \tag{6}$$

em que θ é o conjunto de parâmetros para as distribuições de probabilidade, X é a sequência de observações e N, sua cardinalidade.

Assim, o algoritmo de **ML** objetiva encontrar os parâmetros que maximizam o valor da função, tal que:

$$\theta^* = \arg\max_{\theta} L(\theta|X) \tag{7}$$

em que θ^* é o conjunto de parâmetros ótimos encontrado.

Suponha que na Equação 6, um conjunto de dados $X = (x_1, x_2, \ldots, x_T)$ observável foi gerado por alguma distribuição com parâmetros desconhecidos. Assim, a Equação 6 pode ser escrita como uma probabilidade conjunta das variáveis X e C, em que C define os estados para a Cadeia de Markov:

$$\begin{aligned}
P(X|\theta) &= \sum_{i=1}^{m} P(X = x_t, C = c_i|\hat{\theta}) = && (8)\\
L(\theta|X) &= \prod_{t=1}^{n}\sum_{i=1}^{m} P(X = x_t, C = c_i|\hat{\theta}) = && (9)\\
\log(L(\theta|X)) &= \log(\sum_{t=1}^{n}\sum_{i=1}^{m} P(X = x_t, C = c_i|\hat{\theta})) = && (10)\\
&= \log(\sum_{t=1}^{n}\sum_{i=1}^{m} P(C = c_i|X = x_t, \theta)\frac{P(X = x_t, C = c_i|\hat{\theta}))}{P(C = c_i|X = x_t, \theta)} = && (11)\\
&\leq \sum_{t=1}^{n}\sum_{i=1}^{m} P(C = c_i|X = x_t, \theta)\log\frac{P(X = x_t, C = c_i|\hat{\theta}))}{P(C = c_i|X = x_t, \theta)} = && (12)\\
&\leq \sum_{t=1}^{n}\sum_{i=1}^{m} P(C = c_i|X = x_t, \theta)\log P(X = x_t, C = c_i|\hat{\theta}) - && (13)\\
&- \sum_{t=1}^{n}\sum_{i=1}^{m} P(C = c_i|X = x_t, \theta)\log P(C = c_i|X = x_t, \theta) && (14)\\
&= E_{p(C|X,\theta)}[logP(X = x_t, C = c_i|\hat{\theta}) - H(P(C|X, \theta)) && (15)
\end{aligned}$$

A equação 3.6 introduz a dependência da variável latente ao modelo. A formulação do **HMM** torna-se mais fácil ao se considerar essa variável. No passo seguinte, a igualdade dos termos é transformada em uma desigualdade, a fim de usar o resultado da **Desigualdade de Jensen** [12] para garantir que a função logaritmo passe a operar dentro do somatório. Assim, a equação 3.10 é composta por dois termos: o primeiro termo representa uma função auxiliar à ser maximizada e o segundo termo é entropia do mútua entre o modelo e a distribuição real dos dados, se há independência entre os estados escondidos e as observações, esse termo se torna zero [13].

A função auxiliar $Q(\theta, \hat{\theta})$ é definida tal como:

$$Q(\theta, \hat{\theta}) = \sum_{t=1}^{n}\sum_{i=1}^{m} P(C = c_i|X = x_t, \theta) \log P(X = x_t, C = c_i|\theta) = \quad (16)$$

$$= \sum_{t=1}^{n}\sum_{i=1}^{m} \frac{P(C = c_i, X = x_t|\theta)}{P(X = x_t|\theta)} \log P(X = x_t, C = c_i|\hat{\theta}). \quad (17)$$

$$(18)$$

em que $\hat{\theta}$ é uma estimação inicial para os parâmetros θ.

Antes de prosseguir com a maximização da função auxiliar $Q(\theta, \hat{\theta})$, uma função que expressa $P(X, C|\theta)$ precisa ser definida a fim de aplicar a maximização. Utilizando a Cadeia de Markov de primeira ordem, ilustrada na Figura 6, a distribuição de um conjunto de variáveis aleatórias V_i é definida utilizando a regra da cadeia:

$$Pr(V_1, V_2, \ldots, V_n) = \prod_{i=1}^{n} Pr(V_i|pa(V_i))$$

.

Utilizando a nomenclatura de Teoria de Grafos, $pa(V_i)$ é denominado 'nó pai' de V_i. Assim, examinando o grafo direto da Figura 6, para quatro variáveis aleatórias, observa-se: $X_t, X_{t+k}, C_t, C_{t+k}$. Ainda tem-se que, $pa(C_t)$ não possui um 'nó pai', $pa(X_t) = C_t$, $pa(C_{t+k}) = C_t$ e $pa(X_{t+k}) = C_{t+k}$. Logo:

$$Pr(X_t, X_{t+k}, C_t, C_{t+k}) = Pr(C_t)Pr(X_t|C_t)Pr(C_{t+k}|C_t)Pr(X_{t+k}|C_{t+k}).$$

Portanto, essa definição pode ser generalizada para mais estados:

$P(X_t, X_{t+k}) =$
$= \sum_{i=1}^{m}\sum_{j=1}^{m} Pr(X_t, X_{t+k}, C_t = i, C_{t+k} = j)$
$= \sum_{i=1}^{m}\sum_{j=1}^{m} Pr(C_t = i)Pr(X_t|C_t = i)Pr(C_{t+k} = j|C_t = j)Pr(X_{t+k}|C_{t+k})$
$= \sum_{i=1}^{m}\sum_{j=1}^{m} u_i(t)p_i(x_i)\gamma_{ij}(k)p_j(w).$

Agrupando-se os termos, tem-se:

$$P(X = x_t, C = c_i|\hat{\theta}) = \hat{\delta}_{c_1} \prod_{t=1}^{T-1} \hat{\delta}_{c_1,c_{t+1}} \prod_{t=1}^{T} p_{c_t}(x_t) \quad (19)$$

em que $\hat{\delta} = Pr(C_t = i)$ e $p_{c_t}(x_t) = Pr(X_t|C_t = i)$.

Aplicando log em ambos os lados, tem-se:

$$\log P(X = x_t, C = c_i|\hat{\theta}) = \log \hat{\delta}_{c_1} + \sum_{t=1}^{T-1} \hat{\delta}_{c_1,c_{t+1}} + \sum_{t=1}^{T} p_{c_t}(x_t). \quad (20)$$

Utilizando a mesma formulação para a função auxiliar, tem-se a probabilidade definida em termos das probabilidades da Cadeia de Markov Escondida. Assim, a função auxiliar pode ser maximizada em relação a todos os parâmetros da função. Substituindo a Equação 20 na Equação 16, tem-se:

$$Q(\theta,\hat{\theta}) = \sum_{t=1}^{n}\sum_{i=1}^{m}\frac{P(X=x_t,C=c_i|\theta)}{P(X|\theta)}(\log\hat{\delta}_{c_1}+\sum_{t=1}^{n-1}\hat{\gamma}_{c_1,c_{t+1}}+\sum_{t=1}^{n}p_{c_t}(x_t)) \quad (21)$$

$$= Q_\delta(\hat{\delta},\theta)+Q_\gamma(\hat{\gamma},\theta)+Q_p(\hat{p},\theta). \quad (22)$$

3.2 Maximização dos termos separadamente

A função auxiliar definida na Equação 16 pode ser decomposta como a soma de três funções independentes:

$$Q_\delta(\hat{\delta},\theta) = \sum_{i=1}^{m}\frac{P(X,C_1=i|\theta)}{P(X=x_t|\theta)}\log\hat{\delta}_{c_1} \quad (23)$$

$$Q_\gamma(\hat{\gamma},\theta) = \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{t=1}^{n}\frac{P(X,C_t=i,C_{t+1}=j|\theta)}{P(X|\theta)}\log\hat{\gamma}_{ij} \quad (24)$$

$$Q_p(\hat{p},\theta) = \sum_{j=1}^{m}\sum_{t=1}^{T}\frac{P(X,C=j|\theta)}{P(X|\theta)}\log\hat{p_{c_j}}(x_t). \quad (25)$$

Assim, maximiza-se, separamente, os três termos individuais anteriores utilizando **Multiplicadores de Lagrange** e sujeitos às seguintes restrições:

$$\sum_{i=1}^{N}\hat{\delta}_i = 1, \quad (26)$$

$$\sum_{j=1}^{N}\hat{\gamma_i j} = 1, \forall i \quad (27)$$

$$\sum_{k=1}^{M}\hat{p_j}(x_k) = 1, \forall j \quad (28)$$

Como resultado tem-se:

1. Para a parâmetro δ:

$$\hat{\delta}_i = \frac{P(X, C_1 = i|\theta)}{P(X|\theta)} \tag{29}$$

2. Para a parâmetro γ:

$$\hat{\gamma_i j} = \frac{\sum_{t=1}^{T-1} P(X, C_t = i, C_{t+1} = j|\theta)}{P(X|\theta)} \log \hat{\gamma_i j} \tag{30}$$

3. Para o último termo, a maximização dependerá da função de distribuição de probabilidades escolhida para a criação do modelo. Considerando a distribuição de probabilidades de Poisson e Normal, tem-se:

   a) Para a distribuição de probabilidades de Poisson, tem-se:

$$\hat{\lambda_j} = \frac{\sum_{t=1}^{T} P(C_t = i, X|\theta) X_t}{P(C_t = i, X|\theta)} \tag{31}$$

   b) Para a distribuição de probabilidades Normal, tem-se:

$$\hat{\mu_j} = \frac{\sum_{t=1}^{T} P(C_t = i, X|\theta) X_t}{P(C_t = i, X|\theta)} \tag{32}$$

$$\hat{\sigma_j} = \frac{\sum_{t=1}^{T} P(C_t = i, X|\theta)(X_t - \hat{\mu_j})^2}{P(C_t = i, X|\theta)} \tag{33}$$

3.3 Propagação e Retropropagação das probabilidades

Da seção anterior, as expressões para encontrar os valores máximos das variáveis do **HMM** foram determinadas à partir do resultado da otimização utilizando os multiplicadores de Lagrange. Contudo, as expressões ainda estão em termos de probabilidades. Deve-se, então, reformular esse resultado a fim de expressá-lo em termos das probabilidades que definem uma Cadeia de Markov Escondida. A partir da análise da Cadeia de Markov Escondida, observa-se dois processos que sintetizam o funcionamento das mudanças de estados, os quais dão origem às probabilidades: de propagação (do inglês *forward probability*) e de retropropagação (do inglês *backward probability*). Essas probabilidades são introduzidas a seguir.

3.3.1 Probabilidade de propagação

O objetivo da propagação é o de calcular a probabilidade conjunta $Pr(X = x_t, C = \mathbf{c_t})$, em que $\mathbf{C_t} = (\mathbf{c_1}, \mathbf{c_2}, \ldots, \mathbf{c_t})$ é um sequência de estados. Ou seja, dado uma sequência de observações, encontra-se a probabilidade desse sequencia ter sido gerada pro todos

as sequências de estados possíveis. A probabilidade conjunta é calculada utilizando a distribuição marginal para todas as possíveis sequências de estados. Considerando um conjunto de três estados C_1, C_2 e C_3 e uma única observação para a variável aleatória X_1, por exemplo, a probabilidade da observação x_1 ser originada do estado C_1 por todos os possíveis caminhos é dada por:

$$\begin{aligned} P(X_1, C_1, C_2, C_3) =& P(X_1, C_1) + P(X_1, C_2) + P(X_1, C_3) = \\ &= \delta_1 p_1(x_1)\gamma_{11} + \delta_2 p_2(x_1)\gamma_{21} + \delta_3 p_3(x_1)\gamma_{31}, \end{aligned} \tag{34}$$

em que δ um vetor que contém as probabilidade iniciais dos estados $\delta_1, \delta_2, \delta_3$.

Define-se, então:

$$\alpha_i(t) = p(X_1 = x_1, \ldots, X_t = x_t, C_t = i|\theta), \tag{35}$$

Logo, repete-se esse processo para todos os estados do modelo e utilizando uma notação matricial, pode-ser generalizar a probabilidade de propagação :

Para $t = 1, 2, \ldots, T$ e $j = 1, 2, \ldots, m$

$$\begin{aligned} \alpha_t(j) &= Pr(\mathbf{X}^{(t)} = \mathbf{x}^{(t)}, C_t = j) = \\ &= \delta\mathbf{P}(x_1)\mathbf{\Gamma P}(x_2)\ldots\mathbf{\Gamma P}(x_t) = \delta\mathbf{P}(x_1)\prod_{s=2}^{t}\mathbf{\Gamma P}(x_s). \end{aligned} \tag{36}$$

Finalmente, pode-se resumir a função de propagação como a probabilidade de uma sequência terminar em um estado i no instante de tempo t.

3.3.2 Probabilidade de retropropagação

A retropropagação tem um procedimento bem similar à anterior, porém, de maneira reversa, ela calcula a probabilidade de que uma sequência de observações dado que o processo foi iniciado em um estado i no instante de tempo t. Define-se a probabilidade de retropropagação na forma:

$$\beta_i(t) = p(X_{t+1} = x_{t+1}, \ldots, X_T = x_T|C_t = i, \theta), \tag{37}$$

e em notação matricial

Para $t = 1, 2, \ldots, T$ e $j = 1, 2, \ldots, m$

$$\beta_t(j) = Pr(\mathbf{X}_{t+1}^{(t)} = \mathbf{x}_{t+1}^{(t)}|C_t = j) \tag{38}$$

$$= \mathbf{\Gamma P}(x_{t+1}), \mathbf{\Gamma P}(x_{t+2}), \ldots, \mathbf{\Gamma P}(x_T) = (\prod_{s=t+1}^{T}\Gamma P(x_s))1'. \tag{39}$$

Agora, as probabilidades dos resultados da maximização podem ser escritos em termos das probabilidades de propagação e retropropagação.

A fim de simplificar a notação, define-se as variáveis $u_t(i)$ e $v_{jk}(t)$:

$$u_t(i) = \frac{P(C_t = i, X|\theta)}{P(X|\theta)} = \frac{P(C_t = i, X|\theta)}{\sum_{j=1}^{N} P(X|\theta)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \tag{40}$$

$$v_{jk}(t) = \frac{P(C_t = i, C_{t+1} = j, X|\theta)}{P(X|\theta)} = \frac{P(C_t = i, C_{t+1} = j, X|\theta)}{\sum_{i=1}^{N}\sum_{j=1}^{N} P(C_t = i, C_{t+1} = j, X|\theta)} = \tag{41}$$

$$= \frac{\alpha_t(i)\alpha_{ij}(i)b_j(X_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)\alpha_{ij}(i)b_j(X_{t+1})\beta_{t+1}(j)}. \tag{42}$$

Finalmente, as expressões maximizadas para cada variável do Modelo de Markov Escondido são dadas por:

1. Para o parâmetro δ :
$$\hat{\delta}_i = \frac{P(X, C_1 = i|\theta)}{P(X|\theta)} = u_t(i) \tag{43}$$

2. Para a parâmetro γ :
$$\hat{\gamma_i j} = \frac{\sum_{t=1}^{T-1} P(X, C_t = i, C_{t+1} = j|\theta)}{P(X|\theta)} = \frac{\sum_{t=1}^{T-1} v_{jk}(i,j)}{\sum_{t=1}^{T-1} u_t(i)} \tag{44}$$

3. Para o parâmetro μ, λ e σ :
$$\hat{\mu_j} = \hat{\lambda_j} = \frac{\sum_{t=1}^{T} P(C_t = i, X|\theta)X_t}{P(C_t = i, X|\theta)} = \frac{\sum_{t=1}^{T} u_t(i)x_t}{\sum_{t=1}^{T} u_t(i)} \tag{45}$$

$$\hat{\sigma_j} = \frac{\sum_{t=1}^{T} P(C_t = i, X|\theta)(X_t - \hat{\mu_j})^2}{P(C_t = i, X|\theta)} = \frac{\sum u_t(i)(x_t - u_t(i))^2}{\sum u_t(i)} \tag{46}$$

,

as quais foram utilizadas para a implementação do **HMM** considerado neste trabalho.

3.3.3 Global Decoding

Os resultados da formulação do **HMM** obtidos até agora, serão usados para mostrar o algoritmo de decodificação global (do inglês **global decoding**), mais conhecido como *Algoritmo de Viterbi*. Este método está interessado em determinar a sequência de mais provável de estados da cadeia de Markov que deram origem a sequência de observações. O resultado do Viterbi possui grande aplicações principalmente quando os estados possuem significados importantes para os modelos, como por exemplo na modelagem de sistemas

de comunicação, rastreamento de alvos, etc. Outros exemplos da utilização do algoritmo de Viterbi podem ser vistos nesse artigo [14].

O desenvolvimento do algoritmo é similar a probabilidade de propagação, exceto que ele possui um passo retroativo que será esclarecido adiante. Assim, inicialmente defini-se:

$$\xi_{1i} = Pr(C_1 = i, X_1 = x_1) = \delta_i p_i(x_1) \tag{47}$$

para $t = 2, 3, \ldots, T$, tem-se:

$$\xi_{ti} = \max_{c_1, c_2, \ldots, c_{t-1}} Pr(\mathbf{C}^{(t-1)} = c^{(t-1)}, C_t = i, \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) \tag{48}$$

As probabilidades de cada caminho visitado, é dado por:

$$\xi_{tj} = (\max_i(\xi_{t-1}\gamma_{ij})p_j(x_t)) \tag{49}$$

Por fim, a maximização da sequência de estados $i_1, i_2, \ldots, i_T$ pode ser determinada recursivamente:

$$i_T = \arg \max_{i=1,\ldots,m} \xi_{T_i} \tag{50}$$

para $t = T-1, T-2, \ldots, 1$ tem-se:

$$i_t = \arg \max_{i=1,\ldots,m} (\xi_{ti}\gamma_{i,i_{t+1}}) \tag{51}$$

3.3.4 Análise de sequência

A fim de avaliar uma sequência, a partir da formulação anterior, deriva-se a fórmula para a distribuição de X_t condicionada à todas as outras observações. Assim, define-se:

$$\mathbf{X}^{(-t)} = (X_1, \ldots, X_{t-1}, X_{t+1}, \ldots, X_T). \tag{52}$$

Usa-se a probabilidade de propagação e de retropropagação, logo tem-se:

$$Pr(X_t = x|\mathbf{X}^{-t} = \mathbf{x}^{-t}) = \frac{\delta\mathbf{P}(x_1)\mathbf{B}_2 \ldots \mathbf{B}_{t-1}\Gamma\mathbf{P}(x)\mathbf{B}_{t+1} \ldots \mathbf{B}_T\mathbf{1}'}{\delta\mathbf{P}(x_1)\mathbf{B}_2 \ldots \mathbf{B}_{t-1}\Gamma\mathbf{B}_{t+1} \ldots \mathbf{B}_T\mathbf{1'}} \approx \alpha_{t-1}\Gamma\mathbf{P}(x)\beta_t^{`} \tag{53}$$

Sendo: $\alpha_t = \delta P(x_1)B_2 \ldots B_t$; $\beta_t = B_{t+1} \ldots B_T 1'$; $B_t = \Gamma P(x_t)$

A distribuições condicionais acima é a razão entre duas probabilidades de **HMM**: o numerador é a probabilidade de um observação e o denominador é a probabilidade de

uma observação considerando x não-observável. Assim, a expressão para a avaliação de uma sequência por um **HMM** é dada por:

$$Pr(X_t = x|\mathbf{X}^{(-t)} = x^{(-t)}) = \frac{\sum_i^m \alpha_t \Gamma \beta_t \mathbf{p}(x_t)}{\sum_j^m \alpha_t \Gamma \beta_t} \tag{54}$$

A equação 54 mostra a probabilidade de uma determinada sequência pertencer a um dado **HMM**, no qual resolver o problema de avaliação do modelo.

Capítulo 4

Resultados Experimentais

Neste capítulo, as equações resultantes da formulação do Modelo de Markov Escondido (**HMM**), são implementadas utilizando uma linguagem de programação a fim de se construir um ferramental de funções que permita modelar séries temporais. A implementação foi realizada utilizando a linguagem de programação **R** [15] devido à sua facilidade e praticidade em manipular e operar matrizes, além de possuir alguns pacotes com funções para **HMM** já implementadas e que podem ser utilizadas sob a licença **GNU/GPL**. Após a implementação, dois experimentos serão realizados a fim de demonstrar os resultados da modelagem.

4.1 Pacotes utilizados

Inicialmente, dois pacotes do **R** foram instalados e testados. O primeiro é chamado de **HMM package** [16] o qual contém funções para modelar **HMM** utilizando apenas séries discretas. Além disso, não há qualquer conhecimento sobre o tipo de distribuição de probabilidade que modela os estados nem a opção de escolha dessas distribuições como parâmetro estipulado pelo usuário. Essas limitações não permitem explorar a flexibilidade do **HMM** em modelar tanto séries discretas como contínuas. Outra limitação desse pacote está na ausência de funções disponíveis para avaliar os modelos, como os critérios de seleção **AIC** e **BIC**, por exemplo. Por fim, os parâmetros ótimos, conseguidos após a convergência, não são acessíveis. Devido a essas limitações, um segundo pacote, chamado **depmixS4** [17] foi instalado e testado. Esse pacote modela **HMM** com variáveis latentes, utilizando tanto variáveis categóricas quanto dados contínuos, além de possuir flexibilidade na escolha dos parâmetros do modelo como por exemplo, a família de distribuição de probabilidade dos estados, número de iterações máxima para o algoritmo de convergência e aplicação dos critério de seleção **AIC** e **BIC**. Todos os resultados relativos à convergência são acessíveis. Além desses dois pacotes, a implementação das equações foi realizada tendo como base a formulação apresentada no Capitulo 3 e a implementação disponível

no apêndice do livro [1].

4.2 Experimentação

Como mencionado no Capítulo 1, são modelado duas séries temporais: uma sem e outra com dependência temporal. Essas séries ajudam a verificar a eficiência do modelo, utilizando as resoluções dos três problemas na utilização de **HMM** apresentados anteriormente. Primeiramente, para cada série será gerado um conjunto de modelos candidatos, diferenciados entre si pelo número de estados escondidos. Em seguida, um critério de seleção será aplicado, a fim de escolher o melhor modelo dentre os candidatos, o qual é utilizado na resolução do problema de análise, decodificação e aprendizado. Os dois critérios de seleção considerados são explicados à seguir:

4.2.1 Métodos de Seleção de Modelos

O número de estados está diretamente relacionado ao número de parâmetros do modelo. Considere uma distribuição de probabilidade Normal, a cada novo estado acrescentado, o número de parâmetros é duplicado, pois essa distribuição possui dois parâmetros. Por um lado, com aumento do número de estados tem-se uma melhoria no modelo, pelo outro, aumenta-se o número de parâmetros, os quais podem comprometer a convergência do algoritmo de **ML**. Logo, há um balanço entre o número de estados e a eficiência do modelo. O equilíbrio é encontrado utilizando um critério de seleção.

Os critérios de seleção, em geral, usados para o **HMM** são: *Akaik information criterion (**AIC**)* e *Bayesian information criterion (**BIC**)*. **AIC** é uma medida relativa de qualidade entre modelos estatísticos. Ele estima a perda de informação relativa quando um modelo é usado para representar o processo gerador dos dados e é definido tal como:

$$AIC = -2\log L + 2p \tag{55}$$

$\log L$ é o logaritmo da verossimilhança de um modelo ajustado e p é o número de parâmetros do modelo. Dado um conjunto de modelos candidatos, o modelo escolhido será aquele que apresentar o menor valor para **AIC**, pois o primeiro termo que mostra o quão bom o modelo se ajusta aos dados, diminui com o aumento dos estados e, o segundo termo, chamado de **termo de penalidade**, aumenta com número de parâmetros. **BIC** é um critério de seleção bem semelhante ao **AIC**, diferenciando-se apenas pelo termo de penalidade. **BIC** é definido como:

$$BIC = -2\log L + p\log n \tag{56}$$

em que n é a cardinalidade da série. Comparado com **AIC**, o termo de penalidade de **BIC** é mais rigoroso, ou seja, ele favorece o modelo com menor número de parâmetros.

4.2.1.1 Experimento 1

O primeiro conjunto de dados contém 1000 valores, nos quais 300 foram amostrados a partir de uma distribuição de probabilidade Normal, θ_1, com média e variância iguais a, respectivamente: $\mu_1 = 10$ e $\sigma_1 = 5$ e os outros 700 restantes foram amostrados de outra distribuição de probabilidade Normal θ_2, com média e variância iguais a: $\mu_2 = 130$ e $\sigma_2 = 10$. Assim, quatro modelos foram criados, sendo que o primeiro contém 2 estados, o segundo 3, sucessivamente até o último, que contém 5.

Os valores do logaritmo da verossimilhança, **AIC** e **BIC** foram calculados e podem ser vistos na Tabela 3. Pode-se observar que os valores para esses critérios aumentam à medida que o número de estados cresce. O modelo com 2 estados escondidos foi escolhido, pois apresenta os menores valores, como pode ser observado na Figura 7. Este resultado é totalmente esperado, pois a série é conhecida previamente, ou seja, sabe-se que ela é composta por duas distribuições de probabilidades Normal, logo o melhor modelo contém 2 estados escondidos, cada um modelando uma distribuição de probabilidade.

Tabela 3 – Valores do logaritmo da probabilidade, AIC e BIC para os modelos candidatos.

	Número de estados			
	2	3	4	5
$\log L$	3551.32	3550.27	3540.81	3542.92
AIC	7112.65	7122.54	7119.62	7143.85
BIC	7137.19	7176.53	7212.87	7286.17

À partir da escolha do **HMM** com dois estados escondidos, pode-se resolver o três problemas de inferência discutidos no Capítulo 1.

4.2.1.2 Problema de aprendizagem

O problema de aprendizado é um resultado direto do treinamento do modelo, utilizando um conjunto de dados. Os parâmetros inicias para as variáveis e aqueles obtidos após a convergência do algoritmo **ML** são mostrados na Tabelas 4 e 5.

Tabela 4 – Convergência dos parâmetros.

μ_1	μ_2	σ_1	σ_2	δ_1	δ_2	μ_1^*	μ_2^*	σ_1^*	σ_2^*	δ_1^*	σ_2^*
67	84	50	44	0.333	0.335	10.16	99.85	4.81	10.62	1.00	0.00

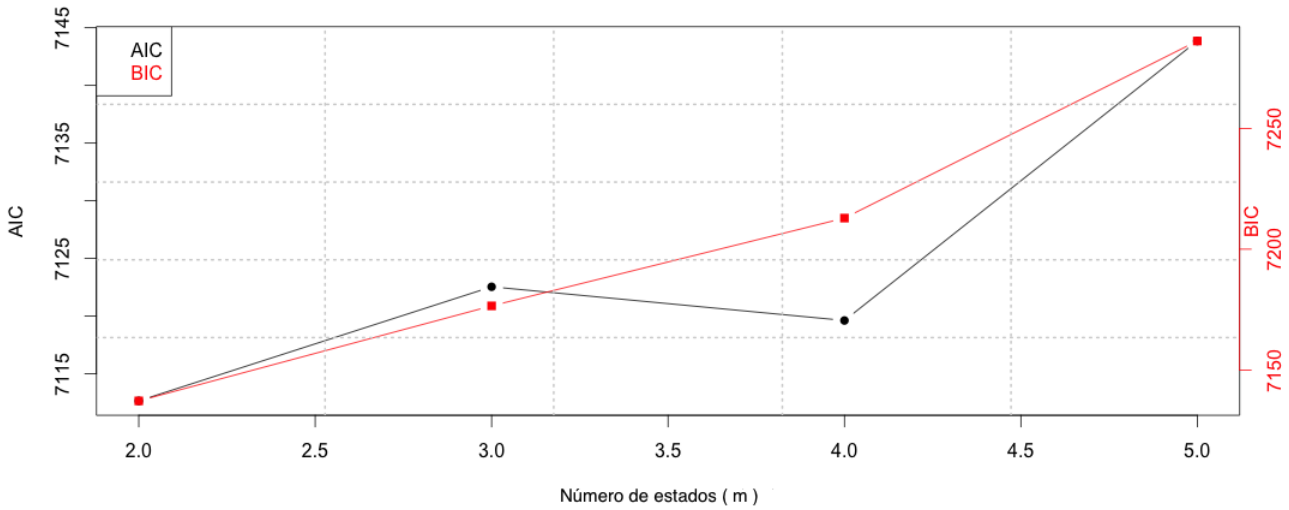

Figura 7 – Valores de AIC e BIC para cada modelo em função do número de estados.

Tabela 5 – Convergência das probabilidades de transição.

	γ		γ^*	
	1	2	1	2
1	0.0006	0.0017	0.997	0.003
2	0.0020	0.0068	0.000	1.000

Esses resultados evidenciam uma grande mudança entre os valores iniciais e os ótimos, isso mostra a efetividade do algoritmo de **ML** para a otimização dos parâmetros. Além disso, observa-se que os valores da média e da variância ótimos, convergiram para valores muito próximos da média e da variância real dos dados. Outro resultado interessante pode ser observado na Tabela 5, que contém os valores ótimos da matriz de probabilidades de transição: a probabilidade de permanecer no estado 1 ou no estado 2 é praticamente 1, isso significa que não há transição entre os estados, o que está totalmente de acordo com o conjunto de dados, pois ele é formado por duas distribuições de probabilidades Normal com médias muito distantes entre si e desvio padrão muito pequeno.

A Figura 8, ilustra o ajuste do modelo em relação aos dados, por meio do histograma da série e do traço do modelo.

4.2.2 Problema da decodificação

A resolução do problema da decodificação provém da própria formulação do **HMM**. A decodificação global (do inglês, *Global Decoding*) propõe encontrar a melhor sequência de estados que deram origem a série. Esse método também é conhecido como o algoritmo de Viterbi (do inglês *Viterbi algorithm*) e foi apresentado no Capítulo 3. Assim, a aplicação do

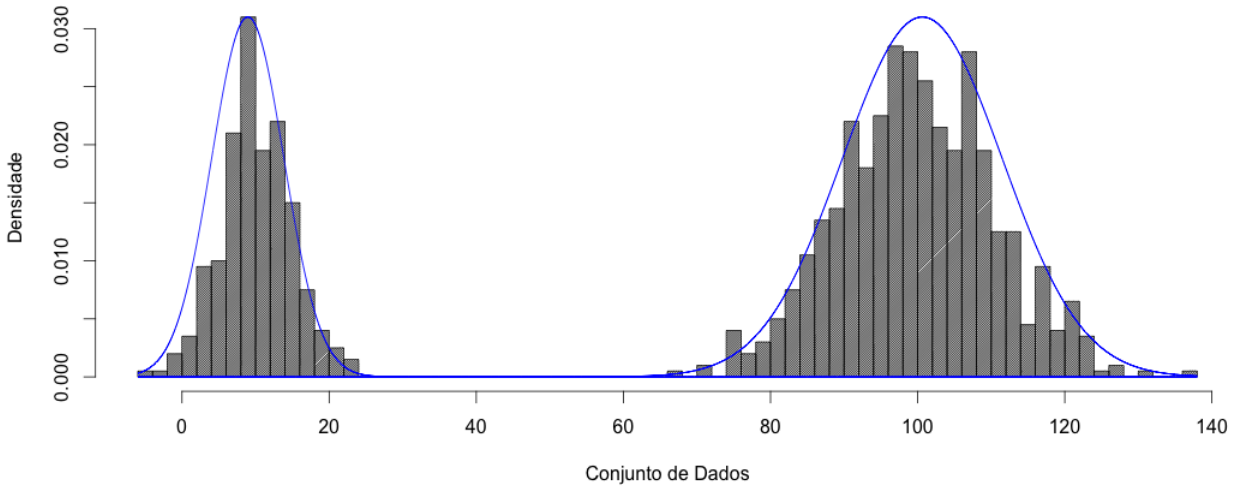

Figura 8 – Modelo ajustado sobre o conjunto de dados.

algoritmo de Viterti sobre a série considerada apresenta uma sequência ótima de estados visitados, a qual pode ser visualizada na Figura 9.

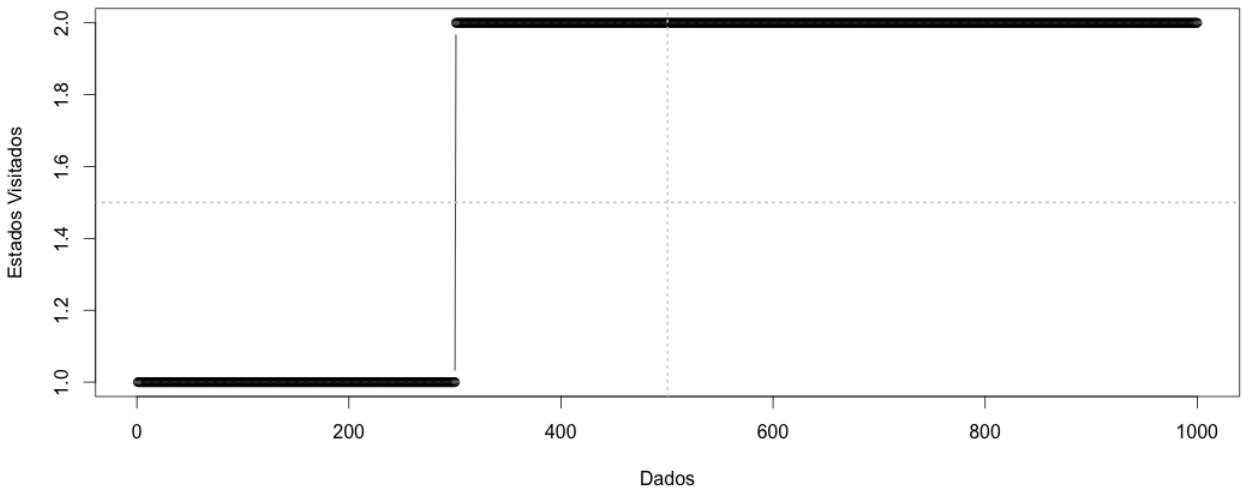

Figura 9 – Sequência ótima de estados visitados.

A Figura 9 mostra a posição dos valores na série no eixo horizontal, enquanto que no eixo vertical mostra qual estado os dados foram originados. Pode-se observar que todos os valores da sequência da posição 1 até a posição 300, foram originadas do estado 1 e as observações posteriores foram originadas do estado 2. Este resultado está totalmente de acordo com o esperado, pois as primeiras 300 observações foram amostradas a partir de uma distribuição de probabilidades Normal θ_1, e as outras de θ_2. A fim de ilustrar

ainda melhor a mudança dos estados visitados, criou-se um conjunto de dados da seguinte maneira: 300 dados gerados de uma distribuição de probabilidade Normal d_1, com média e variância igual à $\mu = 10$ e $\sigma = 5$; 700 dados gerados de uma distribuição de probabilidade Normal d_2, com média e variância igual à $\mu = 50$ e $\sigma = 2$; 700 dados gerados de uma distribuição de probabilidade Normal d_3, com média e variância igual à $\mu = 130$ e $\sigma = 10$; 300 dados gerados de uma distribuição de probabilidade Normal d_4, com média e variância igual à $\mu = 90$ e $\sigma = 2$.

O conjunto de dados final é composto pela concatenação dos dados acima, na seguinte ordem d_1,d_2, d_3 e d_4. A sequência de dados após a aplicação do algoritmo de Viterbi é mostrada na Figura 10:

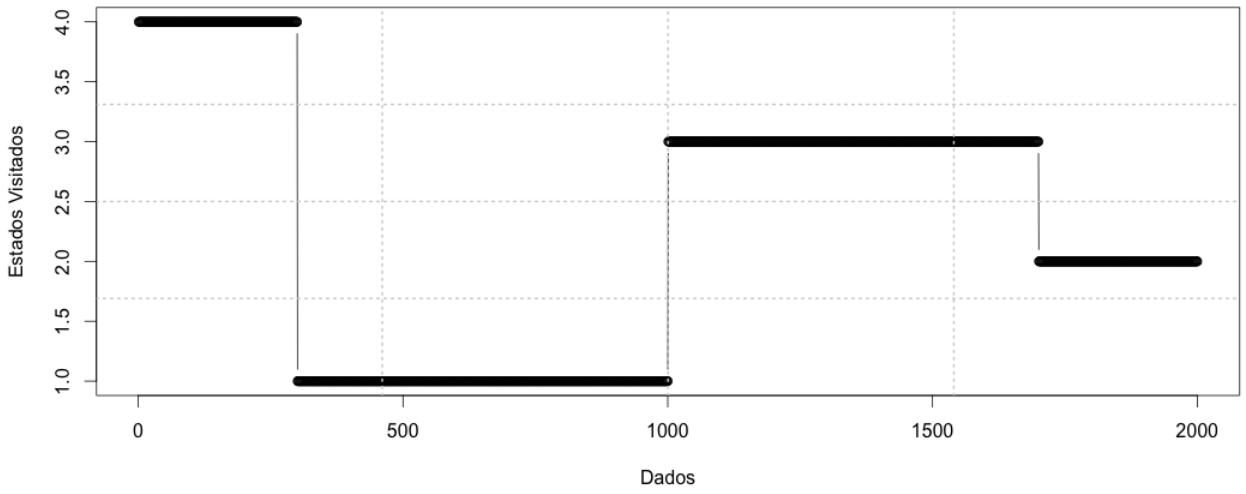

Figura 10 – Sequência de estados visitados.

O resultado mostra que há uma troca de estados na posição de concatenação das distribuições de probabilidades, então conclui-se que houve uma detecção da troca dos estados. É importante ressaltar que os estados do **HMM** não seguem uma ordem, ou seja, a primeiro estado não precisa representar, necessariamente, a primeira distribuição de probabilidades que forma o conjunto.

4.2.3 Problema de análise

A formulação do **HMM** também origina a resolução do problema da análise. Utiliza-se apenas a probabilidade de propagação para determinar a probabilidade de uma nova sequência ter sido gerada a partir de um dado modelo. A formulação desse problema está proposta no Capítulo 3 desse documento. Para mostrar esse resultado utiliza-se duas sequências, X_1 e X_2, produzidas sinteticamente, as quais podem ser vistas nas Tabelas 4.2.3 e 4.2.3.

Tabela 6 – Valores da sequência 1: X_1

18.0	19.0	10.0	11.0	12.0	15.0	8.0	2.0	5.0	7.0
12.0	15.0	13.0	13.20	1.20	1.5	7.0	10.0	11.0	13.0

Tabela 7 – Valores da sequência 2: X_2

96.55	83.82	89.81	7.20	101.11	93.69	5.80	8.65	110.73	97.27
97.97	1.20	99.92	97.08	90.92	2.04	12.86	19.85	100.85	101.25

Essas sequências foram geradas sinteticamente de modo que a maior parte seus valores pertençam a uma das duas distribuições de probabilidades componentes que formam o conjunto de treinamento.

A sequência X_1 4.2.3, contém valores originados por uma distribuição de probabilidades, cujos parâmetros são os mesmos da distribuição componente θ_1. Assim, espera-se que o resultado da avaliação indique um valor de probabilidade maior para θ_1 em comparação à θ_2. Da mesma maneira, a sequência X_2 4.2.3, contém a maior parte de seus valores próximos à distribuição componente θ_2 do que θ_1. Portanto, espera-se que o resultado da avaliação indique que a probabilidade da sequência ter sido gerada por θ_2 será muito maior do que θ_1. A fim de confirmar essas hipóteses, o algoritmo foi aplicado sobre as duas sequências:

Para a sequência X_1:

1. $P(X_1 = 18, 19, \ldots, 13 \mid \theta_1^*) = 1.00$
2. $P(X_1 = 18, 19, \ldots, 13 \mid \theta_2^*) = 0.00$

Para a sequência X_2:

1. $P(X_2 = 126.55, \ldots, 128.25 \mid \theta_1^*) = 25.58$
2. $P(X_2 = 126.55, \ldots, 128.25 \mid \theta_2^*) = 74.42$

A porcentagem da contribuição de cada distribuição componente é mostrada anteriormente, a sequência X_1 é composta por 100% do estado θ_1 e a sequência X_2 é composta por aproximadamente 25% do estado θ_1 e 75% do estado θ_2.

4.2.4 Experimento 2

O segundo conjunto de dados contém dependência temporal e representa o comportamento da temperatura em Melboure, Austrália, no qual medidas da temperatura máxima diária

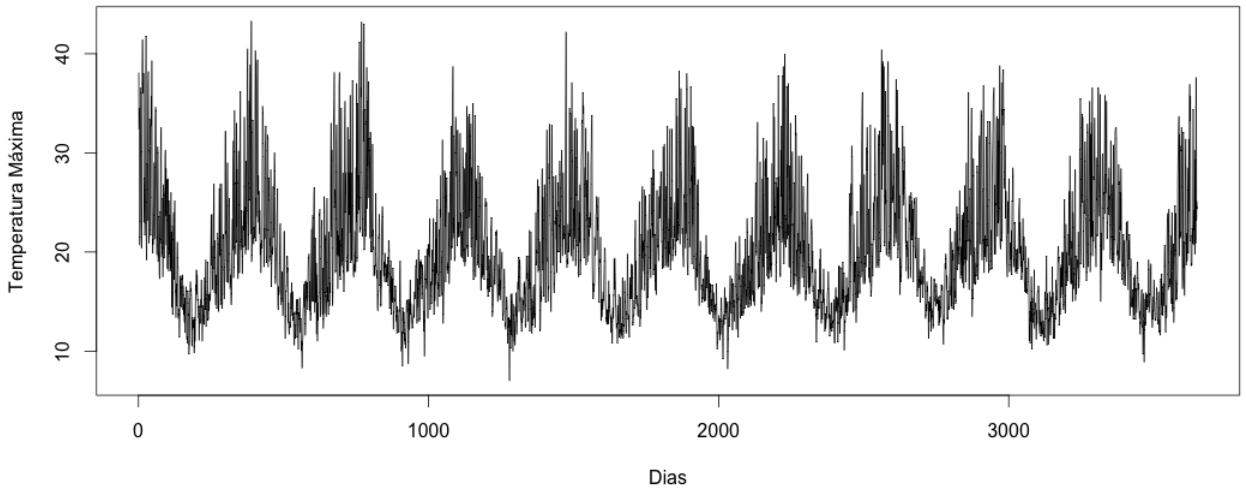

Figura 11 – Sequência de estados visitados.

entre os anos de 1981 e 1990 foram coletadas [7], contabilizando 3650 amostras. Os dados podem ser vistos na Figura 11.

Primeiramente, um conjunto de 10 modelos candidatos, no qual o primeiro modelo contém 2 estados escondidos e o último 11 estados foram criados. Foram utilizados os critérios de seleção **AIC** e **BIC**. A Tabela 8 mostra os valores para cada estado do conjunto.

Tabela 8 – Valores do logaritmo da verossimilhança, **AIC** e **BIC**.

	Número dos estados									
	2	3	4	5	6	7	8	9	10	11
$\log L$	8610	8226	8030	7940	7884	7844	7803	7762	7742	7717
AIC	1723	1647	1609	1593	1585	1579	1574	1570	1570	1569
BIC	1726	1654	1621	1611	1609	1613	1617	1623	1635	1648

A Figura 12 mostra a relação entre o número de estados com os valores **AIC** e **BIC**. Observa-se que esses valores decrescem com o aumento do número de estados até o modelo com 6 estados, a partir do qual, os valores começam a aumentar. Assim, esse deve ser escolhido entre os candidatos.

4.2.5 Problema de aprendizado para a série de temperatura

O problema de aprendizado é resultado direto da convergência dos parâmetros. As Tabelas 9, 10, 11 e 12 mostram os seus valores iniciais e ótimos.

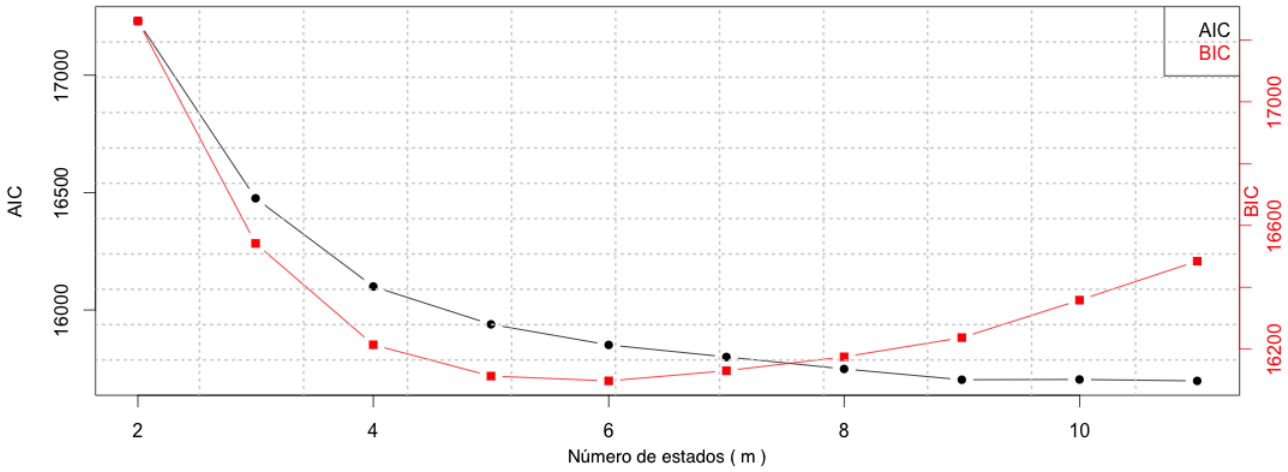

Figura 12 – Valores de AIC e BIC em função do número de estados.

Tabela 9 – Convergência dos valores da média do modelo.

	Valor da média inicial						Valor da média ótima					
	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_1^*	μ_2^*	μ_3^*	μ_4^*	μ_5^*	μ_6^*
	24	21	17	16	26	29	18.16	20.91	15.50	13.12	26.02	34.06

Tabela 10 – Convergência dos valores da variância do modelo.

	Valor do variância inicial						Valor da variância ótima					
	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_1^*	σ_2^*	σ_3^*	σ_4^*	σ_5^*	σ_6^*
	24	21	17	16	26	29	1.23	1.69	1.18	1.48	3.06	3.54

Tabela 11 – Convergência dos valores da probabilidade inicial do modelo

	Valor do variância inicial						Valor da variância ótima					
	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_1^*	δ_2^*	δ_3^*	δ_4^*	δ_5^*	δ_6^*
	0.14	0.24	0.06	0.16	0.34	0.45	0.0	0.0	0.0	0.0	0.0	1.0

Tabela 12 – Convergência dos valores da probabilidade de transição do modelo

	γ						γ^*					
	1	2	3	4	5	6	1	2	3	4	5	6
1	0.4	0.007	0.005	0.006	0.002	0.1	0.54	0.24	0.12	0.046	0.03	0.00
2	0.006	0.004	0.1	0.80	0.002	0.80	0.04	0.5	0.07	0.00	0.03	0.006
3	0.9	0.005	0.02	0.20	0.007	0.006	0.16	0.08	0.66	0.8	0.00	0.00
4	0.004	0.2	0.06	0.14	0.004	0.005	0.00	0.00	0.13	0.86	0.00	0.00
5	0.004	0.002	0.009	0.09	0.001	0.003	0.13	0.18	0.018	0.00	0.45	0.21
6	0.1	0.005	0.5	0.6	0.007	0.4	0.02	0.30	0.00	0.00	0.2	0.45

As Tabelas 9 e 10 mostram os valores da média e da variância dos estados enquanto que a Tabela 11 mostra a probabilidade do modelo iniciar em cada um dos 6 estados escondidos. Nota-se que o modelo começa pelo estado de número 6 pois sua probabilidade é de 100%, evidenciando que o primeiro valor da série foi produzido por esse estado. Por fim, a Tabela 12 mostra a probabilidade de transição dos estados que contém resultados interessantes como por exemplo: não há transição do estado 1, para o 6, com médias $\mu_1^* = 18.16$ e $\mu_6^* = 34.06$, respectivamente, pois a probabilidade de ocorrer essa transição é 0%. Analisando essa informação do ponto de vista do significado dos dados, conclui-se que não há aumento ou diminuição brusca da temperatura.

4.2.6 Problema de decodificação para a série de temperatura

Encontra-se a melhor sequência de estados visitados utilizando o algoritmo de Viterbi [18]. Após sua aplicação, a melhor sequência de estados é mostrada na Figura 13 junto com o gráfico dos dados. O primeiro gráfico apresenta os valores máximos de temperatura em função da sequência de amostras e o segundo gráfico mostra a transição dos estados também em função das amostras. Comparando os dois gráficos, observa-se que as mudanças de estados seguem o comportamento dos dados, ou seja, o modelo contém estados escondidos que mapeam os valores da série. A fim de ficar mais visível essa observação, a Figura 14 exibe as 150 primeiras amostras.

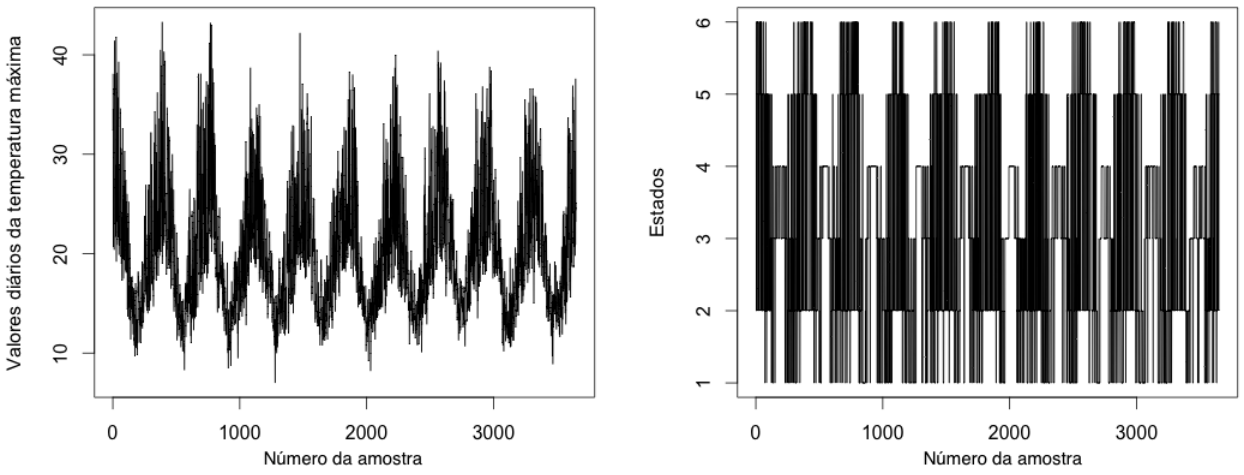

Figura 13 – Sequência de estados visitados.

As distribuições de probabilidades Normal de todos os estados foram grafadas sobre a série na Figura 15, indicando o alcance do mapeamento de cada estado, sendo que a linha contínua representa o valor da média e a linha tracejada são os limites superiores e inferiores do desvio padrão.

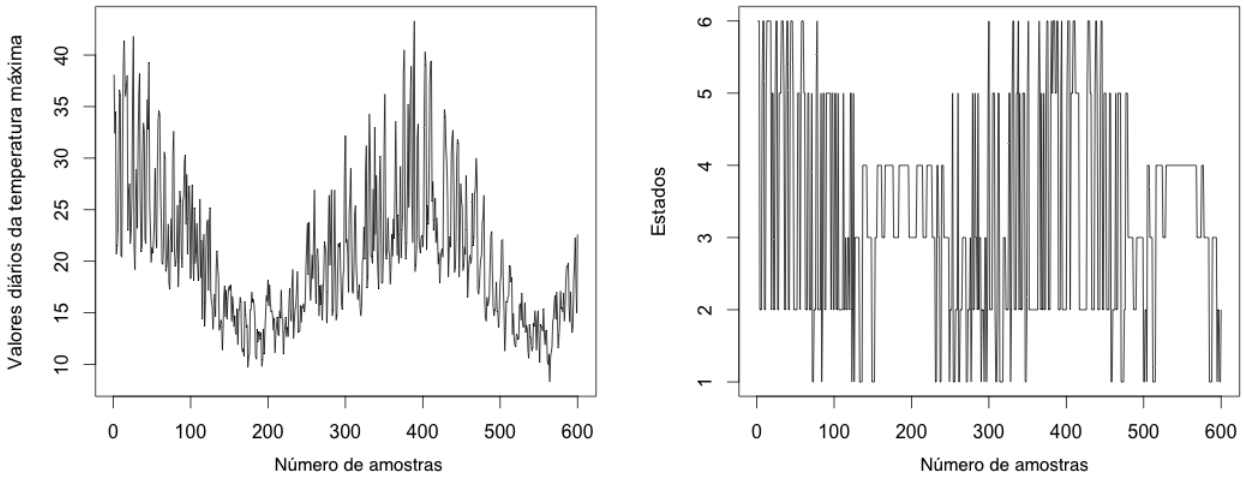

Figura 14 – Sequência de estados visitados para as primeiras 150 amostras da série.

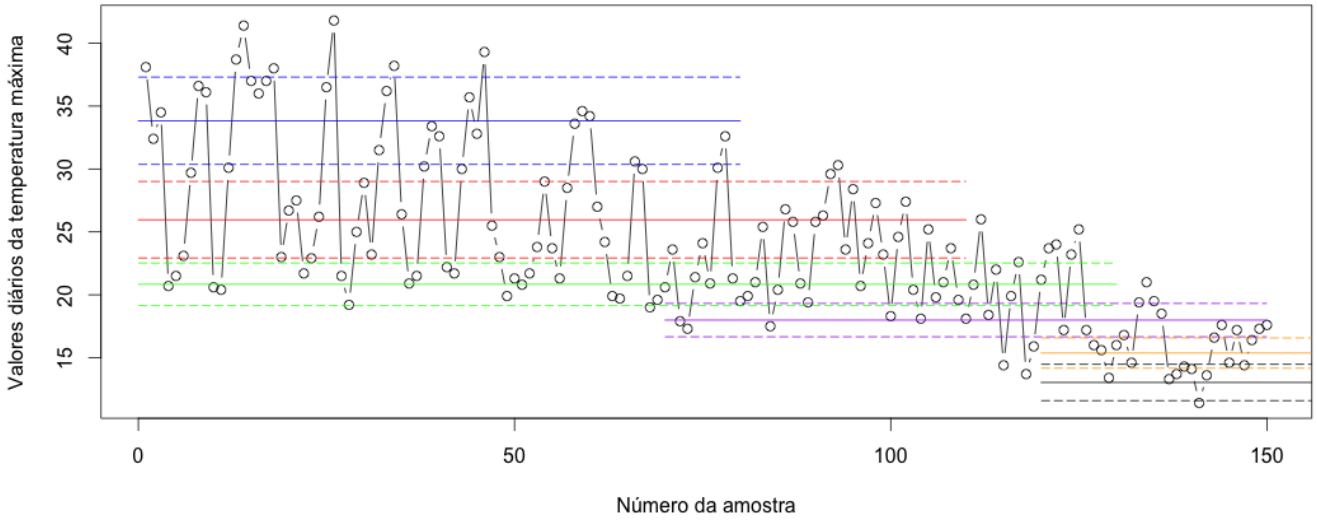

Figura 15 – Intervalos de representação dos estados

Esse resultado ilustra a principal característica do **HMM** de encontrar um número de distribuições de probabilidades que seja representativo sobre toda a série.

Capítulo 5

Conclusão

O trabalho de conclusão de curso apresentou um estudo sobre o modelo de Markov Escondido (**HMM**) iniciando pelo problema de motivação que é o de modelar séries que contém super dispersão. A solução proposta utilizou um modelo de misturas, no qual a cadeia de Markov é incorporada para justificar a transição entre as distribuições componentes do modelo, e as probabilidade de geração dos valores da série pora cada um dos estados. Logo, definidos esses parâmetros, o algoritmo de Máxima Verossimilhança é aplicado com o objetivo de encontrar o conjunto de parâmetros ótimos em relação a uma série de dados. O modelo de Markov Escondido não define um número de estados escondidos ótimos, então, utiliza-se um critério de seleção a fim de encontrar o melhor modelo dentro um conjunto de candidatos. Uma vez definido o número de estados e os parâmetros ótimos do modelo, pode-se resolver os três problemas inerentes da modelagem utilizando **HMM**, conhecidos como: problema de avaliação, decodificação e aprendizado.

Dois experimentos foram realizados: no primeiro, uma série temporal sintética, composta por duas distribuições de probabilidade Normal com médias e variâncias bem distantes entre si e conhecidas previamente, foi utilizada para verificar claramente os resultados dos problemas citados, sendo que o problema de aprendizado é uma consequência do algoritmo de Máxima Verossimilhança (**ML**). O segundo experimento, utilizou a série temporal de dados reais, para criar um **HMM** que evidenciasse o problema de aprendizado e de decodificação através do mapeando da série utilizando os valores ótimos dos parâmetros e, também, utilizando a melhor sequência de estados visitados como resultado do algoritmo de Viterbi. O problema da análise não foi resolvido devido a grande amplitude do estudo, pois o proponente acredita que um análise sobre os dados, principalmente o estudo do espaço dos dados e os possíveis espaços no qual esses mesmos dados seriam melhor representados, deveria ser realizada antes da aplicação do **HMM**.

Os principais aspectos positivos desse trabalho foram: os estudos sobre as teorias da Estatística, o entendimento básico do funcionamento do algoritmo de otimização, programação utilizando a linguagem **R** e principalmente o contato com artigos e livros relacionados ao

tema. Os aspectos negativos foram: a falta de referências com explicações mais simples tanto na fase da formulação do **HMM**, quanto na realização de testes com séries temporais.

5.0.7 Trabalhos futuros

Como possíveis trabalhos futuros pode-se apontar:

1. O Estudo do espaço dos dados das séries a fim de verificar se há a necessidade de mudança do espaço utilizando transformação linear e não-linear, baseadas nos conceitos de Aprendizado Estatístico;
2. Modelar séries temporais que contém dados em duas ou mais dimensões;
3. Implementar uma cadeia de Markov de ordem maior e estudar se houve melhoria no modelo;
4. Implementar outras distribuições de probabilidade além da Poisson e Normal a fim de verificar se há melhoria do **HMM** para diferentes séries temporais
5. Utilizar o **HMM** tanto para classificação quanto para previsão de dados de séries temporais, utilizando o algoritmo de avaliação de sequência e Viterbi;
6. Estudar o **HMM** para a extração de características de um conjunto de dados utilizados para classificação.

Referências

[1] Walter Zucchini Iain L. MacDonald. *Hidden Markov Models for Time Series, An Introduction Using R*. Chapman & Hall/RC, 2009.

[2] Peter J.; Richard A. Davis Brockwell. *Introduction to Time Series and Forecasting*. New York Springer 1996, 1996.

[3] Samuel Karlin Howard M.Taylor. *An Introduction to Stochastic Modeling*. Academy Press, 3rd edition.

[4] Michel Dirickx Marcel Ausloos. The logistic map and the route to chaos, 2006.

[5] Edward Lorenz. *Essence of Chaos*. University of Washington Press.

[6] Edward Nelson. *Dynamical Theories of Brownian Motion*. Princeton University Press., 2001.

[7] Daily maximum temperatures in melbourne, austrália, 1981-1990. `https://datamarket.com/data/set/2323/daily-maximum-temperatures-in-melbourne-australia-1981-1990#!ds=2323&display=line`.

[8] Gregory C. Reinsel George E. P. Box, Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. 4th edition, 2008.

[9] Joseph M. Hilbe. *Modeling Count Data*. 2014.

[10] Charles M. Grinstead J. Laurie Snell. *Introduction to Probability*. American Mathematical Society, 1997.

[11] Jeff A. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 1998.

[12] Gábor Lugosi Luc Devroye, László Gyorfi. *A Probabilitic Theory of Pattern Recognition*. Springer, 1997.

[13] Miguel Barão. Entropia, entropia relativa e informação mútua. 2003.

[14] M. S. Ryan and G. R. Nudd. The viterbi algorithm. 1993.

[15] R Core Team. R: A language and environment for statistical computing. `http://www.R-project.org/`.

[16] Dr. Lin Himmelmann. *Package HMM*, 2015.

[17] Ngmar Visser and Maarten Speekenbrink. *Dependent Mixture Models - Hidden Markov Models of GLMs and Other Distributions in S4*, 2014.

[18] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. 1989.