Universidade Federal Fluminense

FACULDADE DE ECONOMIA

ADRIELLE BRAGA DE ARAUJO

UMA CONTRIBUIÇÃO À TEORIA DA INFORMAÇÃO E AO CRITÉRIO DE INFORMAÇÃO DE AKAIKE

NITERÓI – RJ

2020

ADRIELLE BRAGA DE ARAUJO

UMA CONTRIBUIÇÃO À TEORIA DA INFORMAÇÃO E AO CRITÉRIO DE INFORMAÇÃO DE AKAIKE

Monografia apresentada ao curso de Bacharelado em Ciências Econômicas da Universidade Federal Fluminense como requisito parcial para conclusão do curso.

Orientador:
Prof. Dr. Jesus Alexei Luizar Obregon

Niterói – RJ

2020

ADRIELLE BRAGA DE ARAUJO

UMA CONTRIBUIÇÃO À TEORIA DA INFORMAÇÃO E AO CRITÉRIO DE INFORMAÇÃO DE AKAIKE

Monografia apresentada ao curso de Bacharelado em Ciências Econômicas da Universidade Federal Fluminense como requisito parcial para conclusão do curso.

Trabalho aprovado em

BANCA EXAMINADORA

Prof. Dr. Jesus Alexei Luizar Obregon
Orientador
Universidade Federal Fluminense

Prof. Dr. Aldo Amilcar Bazan Paracoricona
Universidade Federal Fluminense

Prof. Dr. Hugo Henrique Kegler dos Santos
Universidade Federal Fluminense

Prof. Dr. Luiz Fernando Cerqueira Fonseca
Universidade Federal Fluminense

RESUMO

Sempre que fazemos observações estatísticas ou planejamos e realizamos experimentos estatísticos, buscamos informações. Consequentemente, o conceito de informação desempenha um papel importante em áreas que utilizam sistemas probabilísticos ou estatísticos de observações. Neste trabalho estudamos duas variáveis aleatórias centrais da teoria da informação: a informação de Shannon e a informação de Kullback-Leibler. Mostramos que as respectivas esperanças matemáticas são conhecidas como entropia de Shannon e divergência de Kullback-Leibler. Dedicamos mais espaço para estudar as propriedades matemáticas da divergência de Kullback-Leibler com o objetivo de mostrar sua aplicação na seleção de modelos probabilísticos como uma medida da "distância" entre distribuições de probabilidade. Apresentamos o critério de informação de Akaike como uma metodologia de otimização da esperança da divergência de Kullback-Leibler.

Palavras-chave: Teoria da Informação; Informação de Shannon; Informação de Kullback-Leibler; Entropia de Shannon; Divergência de Kullback-Leibler; Critério de Informação de Akaike.

ABSTRACT

Whenever we make statistical observations or design and conduct statistical experiments, we seek information. Consequently, the concept of information plays an important role in fields that use probabilistic or statistical systems of observations. In this work, we study two central random variables in information theory: Shannon's information and Kullback-Leibler's information. We show that their respective mathematical expectations are known as Shannon's entropy and Kullback-Leibler divergence. We dedicate more space to study the mathematical properties of Kullback-Leibler divergence in order to show its application in the selection of probabilistic models as a measure of the " distance " between probability distributions. We present the Akaike information criterion as a methodology for optimizing the expectation of Kullback-Leibler divergence.

Keywords: Information theory; Shannon's Information; Kullback-Leibler's Information; Shannon's Entropy; Kullback-Leibler Divergence; Akaike Information Criterion.

AGRADECIMENTOS

Agradeço à Universidade Federal Fluminense (UFF) pela oportunidade de cursar minha graduação em uma instituição de qualidade, que abriu espaço para minha formação enquanto indivíduo crítico. O alcance das experiências e do aprendizado que a UFF me proporcionou vai muito além destas páginas. A UFF me mostrou ser também uma generosa fonte de amigos: Ana Leticia, Gustavo, Ivan, Leonardo, Nina e Paula.

Agradeço ao Alexei, cuja orientação exemplar tornou possível minha iniciação no mundo da pesquisa. Sua orientação foi responsável por muito do que aprendi no último ano, seus conselhos serviram não só para a produção desse trabalho de conclusão de curso, como para várias de minhas escolhas e caminhos que resolvi trilhar. Sou francamente grata pelo seu apoio e paciência. Agradeço também a todos os meus professores da UFF, sejam da Economia, Estatística ou Matemática. Particularmente importantes foram os professores Aldo, Felipe, Hugo e Slobodan, que acreditarem em mim desde o começo. O Aldo conseguiu colocar sua presença sobre boa parte da minha formação matemática com a escola de matemática Aldo Bazan. O Felipe pelas aulas de análise que deram início à minha paixão pela matemática. O Hugo pela orientação no mundo da estatística e muitas das nossas conversas. O Slobodan pelas belíssimas aulas de lógica e seu modo encantador de ver a matemática.

Agradeço à minha mãe Isabel, meu pai Leu e meu irmão Daniel por absolutamente tudo, mas em especial pelo exemplo de caráter e por tornar possível a minha formação. A luta e amor de vocês é pressuposto de todas as minhas conquistas.

Dedico esse trabalho de conclusão de curso aos meus pais, tios e avós, operários, trabalhadores do campo, donas de casa. Em especial, dedico àqueles que infelizmente não puderam ver mais um fruto de suas vidas de trabalho.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

A teoria da informação, como estamos interessados, é um ramo da teoria das probabilidades e estatística matemática. Suas formulações abstratas são aplicáveis a qualquer sistema probabilístico ou estatístico de observações. Consequentemente, encontramos a teoria da informação em uma variedade de áreas tal como a probabilidade e a estatística.

A natureza matemática e estatística essencial da teoria da informação foi enfatizada por três grandes pesquisadores, em grande parte responsáveis por seu desenvolvimento e estímulo, Fisher (1956), Shannon (1948) e Wiener (1948).

A informação em um sentido formalmente definido foi introduzida pela primeira vez na estatística por Fisher (1925) em seu trabalho sobre teoria da estimação. A definição de informação de Fisher é bem conhecida entre os estatísticos e suas propriedades são parte fundamental da teoria da estimação. A medida de Fisher fornece a quantidade de informação obtida dos dados de um parâmetro desconhecido.

Shannon (1948) e Wiener (1948), independentemente, desenvolveram trabalhos descrevendo medidas logarítmicas para uso na teoria da comunicação e na biologia, respectivamente. As informações de Shannon e Wiener medem a quantidade de incerteza relacionada a ocorrência de um evento aleatório. A informação de Wiener é essencialmente a mesma que a informação de Shannon embora a motivação seja diferente e aparentemente Shannon (1948) tenha investigado a teoria mais completamente (KULLBACK; LEIBLER, 1951). Shannon (1948) preocupou-se especificamente com as aplicações em engenharia da comunicação enquanto Wiener (1948) concentrou-se em aplicações na biologia (SHANNON; WEAVER, 1949).

A partir dos trabalhos de Shannon (1948) e Shannon e Weaver (1949), Kullback e Leibler (1951) e Good e Osteyee (1974) desenvolveram medidas logarítmicas de informação chamadas informação de Shannon e informação de Kullback-Leibler, respectivamente. Kullback e Leibler (1951) estavam interessados em construir uma medida que pudesse mensurar a "distância" entre populações estatísticas em termos de informação. Good e Osteyee (1974) estudavam medidas para resolver problemas práticos na detecção de sinais e ruídos em sistemas de comunicação.

A esperança da informação de Shannon e a esperança da informação de Kullback-Leibler são chamadas entropia e divergência de Kullback-Leibler, respectivamente. A entropia é a média da quantidade de incerteza associada à ocorrência de um evento aleatório. A divergência de Kullback-Leibler fornece a informação média em virtude da comparação da ocorrência de eventos aleatórios alternativos dado algum evento aleatório condicionante.

Uma extensão da divergência de Kullback-Leibler na estatística se dá com o problema de seleção de modelos estatísticos. Uma dificuldade fundamental na análise de dados estatísticos é escolher um modelo apropriado, estimar e determinar a ordem e dimensão do modelo. A

modelagem de dados estatísticos busca ajustar modelos aos dados sem o conhecimento do verdadeiro processo gerador dos dados. Consequentemente, nas últimas décadas, a literatura estatística reconheceu a necessidade de introduzir o conceito de seleção de modelos ou avaliação de modelos. O problema está posto de forma a escolher o melhor modelo de aproximação dentre uma classe de modelos competitivos com um número diferente de parâmetros por um critério de seleção de modelo adequado dado um conjunto de dados (BOZDOGAN, 1987).

Em muitos problemas estatístico temos um conjunto de observações. Essas observações são os valores de variáveis aleatórias cuja distribuição é usualmente desconhecida. Da informação fornecida pelos dados, fazemos inferência sobre os aspectos desconhecidos da distribuição em questão, como os verdadeiros valores dos parâmetros desconhecidos que governam o processo gerador dos dados observados e futuros. Expressamos um modelo na forma de distribuição de probabilidade e ajustamos o modelo aos dados como uma estimativa da verdadeira distribuição de probabilidade.

Na modelagem de dados, a divergência de Kullback-Leibler funciona como uma medida da "distância" entre o modelo e a verdadeira distribuição de probabilidade, oferecendo um procedimento de inferência para tornar essa "distância" a menor possível. O critério de informação de Akaike (AIC) é um critério de seleção de modelos que lançou as bases do campo moderno da modelagem de dados (BOZDOGAN, 1987). É uma metodologia simples e versátil para estimar a divergência de Kullback-Leibler. O desenvolvimento do AIC tem suas origens na modelagem de séries temporais na qual sua utilidade prática foi amplamente estudada.

No Capítulo 2 fazemos uma breve introdução aos conceitos que são necessários para a construção dos Capítulos 3 e 4. Apresentamos de forma sucinta, sem ser rigorosa sobre todos detalhes, conceitos e resultados relativos à teoria da medida e variáveis aleatórias. Exibimos o Teorema de Bayes e algumas de suas consequências. Consideramos alguns resultados de aproximações assintóticas, o método da máxima verossimilhança e alguns conceitos básicos de séries temporais. No Capítulo 3 discutimos a informação de Shannon e a informação de Kullback-Leibler, suas respectivas esperanças e algumas de suas propriedades. No capítulo 4 fazemos uma demonstração conceitual e matemática do AIC e uma aplicação no problema de seleção da ordem de defasagem de processos autoregressivos para uma classe de processos com diferentes ordens.

2 CONCEITOS PRELIMINARES

Neste capítulo apresentamos algumas definições básicas de teoria das probabilidades e estatística matemática para o desenvolvimento desta monografia. Para se ter uma noção de como é feito o estudo de probabilidade em teoria da informação, na Seção 1 apresentamos alguns conceitos básicos de teoria da medida e variáveis aleatórias. Na Seção 2 exibimos o teorema de Bayes, um resultado fundamental para a construção da informação de Shannon e informação de Kullback-Leibler. Na Seção 3 introduzimos um importante método de estimação pontual conhecido como o método da máxima verossimilhança e suas propriedades assintóticas. Na Seção 4 abordamos algumas propriedades de séries temporais e o processo autoregressivo.

O objetivo deste capítulo é fazer uma breve introdução, logo os resultados não são demonstrados. Para um estudo mais aprofundado do assunto e para ver as demonstrações dos teoremas e proposições apresentados neste capítulo, sugerimos ao leitor ver Brockwell, Davis e Fienberg (1991), Good (1950), Lehmann e Casella (1998) e Taylor (2006).

2.1 Teoria da Medida e Variáveis Aleatórias

Utilizamos a notação $(\Omega, \mathcal{F})$ para o espaço mensurável e $(\Omega, \mathcal{F}, \mu)$ para o espaço de medida, em que Ω é o espaço amostral, $\mathcal{F}$ a $\sigma-$álgebra e μ uma medida definida em $\mathcal{F}$. Se μ satisfaz $\mu(\Omega) = 1$, então μ é chamada de probabilidade e denotamos por P.

Considere $\mathcal{C}$ uma coleção de subconjuntos de um conjunto $\mathcal{S}$. Denotamos por $\sigma(\mathcal{C})$ a menor σ-álgebra contendo $\mathcal{C}$. Se $\mathcal{S}$ é uma espaço topológico e $\mathcal{C}$ são os abertos de $\mathcal{S}$, então $\sigma(\mathcal{C})$ é chamada σ-álgebra de Borel e escrevemos $\mathcal{B}(\mathcal{S})$. Os elementos de $\mathcal{B}(\mathcal{S})$ são chamados borelianos. Por exemplo, se considerarmos a σ-álgebra de Borel $\mathcal{B}(\mathbb{R})$ de $\mathbb{R}$, sendo $\mathbb{R}$ com a topologia usual, os borelianos são os intervalos abertos. A σ-álgebra de Borel $\mathcal{B}(\mathbb{R}^n)$ de $\mathbb{R}^n$ corresponde ao produto de σ-álgebras de n cópias de $\mathcal{B}(\mathbb{R})$. A medida de um intervalo é o comprimento e corresponde a chamada medida de Lebesgue. As extensões para $\mathbb{R}^n$ conduzem a uma forma ingênua de identificar a medida de Lebesgue com áreas, volumes ou hipervolumes. Detalhes mais rigorosos podem ser vistos em Halmos (1974) e Taylor (2006).

Definição 2.1. *Uma medida μ é chamada σ-finita se pudermos encontrar $A_1, A_2, \ldots \in \mathcal{F}$ tal que $\Omega = \cup_{n=1}^{\infty} A_n$ e $\mu(A_n) < \infty$ para cada n.*

Definição 2.2. *Uma medida μ é σ-aditiva se $A_1, A_2, \ldots \in \mathcal{F}$ disjuntos tal que $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$ então $\mu\big(\cup_{n=1}^{\infty} A_n\big) = \cup_{n=1}^{\infty} \mu(A_n)$.*

Dizemos que A ocorre em quase toda parte, se $\mu(\bar{A}) = 0$, onde $\bar{A}$ denota o complementar do conjunto A. No caso em que a medida $\mu = \mathrm{P}$, dizemos que A ocorre quase certamente

($q.c.$). Por outro lado, dizemos que A é μ-negligenciável, se $\mu(A) = 0$. Dada essa nomenclatura, apresentamos as seguintes definições e resultados.

Definição 2.3. *Uma medida μ_1 é chamada absolutamente contínua com respeito a medida μ_2, escrevemos $\mu_1 \ll \mu_2$, se todo conjunto μ_1-negligenciável é também μ_2-negligenciável.*

Definição 2.4. *Dizemos que duas medidas μ_1 e μ_2 são equivalentes se $\mu_1 \ll \mu_2$ e $\mu_2 \ll \mu_1$ e escrevemos $\mu_1 \equiv \mu_2$. Isto é, não existe nenhum conjunto $A \in \mathcal{F}$ tal que $\mu_1(A) = 0$ e $\mu_2(A) \neq 0$ ou $\mu_2(A) = 0$ e $\mu_1(A) \neq 0$.*

Definição 2.5. *Seja $f : (\Omega_1, \mathcal{F}) \to (\Omega_2, \mathcal{A})$ uma função entre dois espaços mensuráveis. Nós dizemos que f é mensurável se*

$$f^{-1}(A) \in \mathcal{F} \quad \forall A \in \mathcal{A} \quad .$$

Se a medida em $(\Omega_1, \mathcal{F})$ é uma probabilidade e $(\Omega_2, \mathcal{A}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, então a função mensurável recebe o nome de vetor aleatório e escrevemos $\boldsymbol{X} = (X_1, \ldots, X_n)$. No caso em que $n = 1$, a função mensurável recebe o nome de variável aleatória e escrevemos X.

2.1.1 Esperança Matemática e a Integral de Lebesgue

Um dos conceitos centrais para o desenvolvimento deste trabalho é a esperança matemática. Apresentamos de forma sucinta uma relação entre a esperança matemática e a integral de Lebesgue. A construção rigorosa da integral de Lebesgue necessita de um arcabouço matemático que vai além do espaço disponível neste trabalho. Para uma construção rigorosa precisaríamos demonstrar algumas propriedades de convergência em espaços de Banach, a existência de um funcional linear limitado estendido para o fecho do espaço de funções simples para o corpo dos reais, a existência e unicidade da extensão de Lebesgue e outros. Omitimos estas e outras passagens. Os resultados dessa seção estão baseados em Isnard (2016). Sugerimos ao leitor a verificação dessa construção em Halmos (1974), Isnard (2016) ou Taylor (2006).

Considere $(\Omega, \mathcal{F}, \mu)$ um espaço de medida em que μ é uma medida σ-aditiva e ϕ uma função mensurável em $(\Omega, \mathcal{F}, \mu)$. Dizemos que ϕ é uma função simples se

$$\phi(\omega) = \sum_{i=1}^{n} x_i \mathbb{1}_{A_i}(\omega), \quad \omega \in \Omega \quad ,$$

onde $x_i \in \mathbb{R}$ são distintos, $\{A_i; i = 1, \ldots, n\} \subset \mathcal{F}$ é uma partição de Ω, isto é, $A_i \cap A_j = \emptyset$, se $i \neq j$, e $\cup_i A_{i=1}^n = \Omega$, e $\mathbb{1}_{A_i}$ é a função indicadora de A_i para cada i. O conjunto $\mathcal{H}$ composto pelas funções simples de Ω em $\mathbb{R}$ é um reticulado vetorial e é dito espaço vetorial das funções integráveis à Lebesgue.

Definimos a integral de uma função simples por

$$\mathbb{I}(\phi) = \sum_{i=1}^{n} x_i \mu(A_i) \quad ,$$

sendo $\mathbb{I}$ o funcional linear dado por $\mathbb{I} : \mathcal{H} \to \mathbb{R}$ e utilizamos a notação

$$\mathbb{I}(\phi) = \int \phi d\mu \quad .$$

Considere a função mensurável $f : A \to [0, \infty]$ tal que $A \in \mathcal{F}$. Definimos a integral de Lebesgue de f em A como

$$\mathbb{I}^*(f) = \int_A f d\mu = \sup_{\phi(x) \leq f(x)} \int_A \phi d\mu \quad ,$$

sendo $\mathbb{I}^*$ o funcional linear estendido para o fecho do espaço de funções simples, isto é, $\mathbb{I}^* : \bar{\mathcal{H}} \to \mathbb{R}$. A função f é dita integrável à Lebesgue se sua integral é finita. Quando f é uma função simples, as integrais $\mathbb{I}$ e $\mathbb{I}^*$ coincidem. Lembramos que toda função integrável à Riemann é integrável à Lebesgue e as duas integrais coincidem. Se μ é uma medida de Lebesgue e f uma função integrável à Riemann, então $\int f d\mu = \int f dx$.

No caso em que a função mensurável é uma variável aleatória, a integral de Lebesgue corresponde a esperança matemática da variável aleatória X e escrevemos

$$E[X] = \int X d\mu \quad .$$

Teorema 2.1. ***Radon-Nikodym.*** *Sejam μ_1 e μ_2 duas medidas em $(\Omega, \mathcal{F})$ tal que $\mu_2 \ll \mu_1$ e μ_1 e μ_2 σ-finitas. Então existe uma função real estendida $f : \Omega \to [0, \infty]$ tal que para qualquer $A \in \mathcal{F}$,*

$$\mu_2(A) = \int_A f(s) \mu_1(ds) \quad , \tag{2.1}$$

sendo $\int_A$ a integral de Lebesgue. A função f, chamada derivada de Radon-Nikodym de μ_2 com respeito a μ_1, é única em quase toda parte.

2.2 Teorema de Bayes

Um conceito fundamental na teoria das probabilidades e estatística dada a sua vasta aplicabilidade é o Teorema de Bayes. Para abordar este resultado, consideramos o espaço de probabilidade $(\Omega, \mathcal{F}, \mathrm{P})$. O Teorema de Bayes define a medida de probabilidade condicional admitindo como conhecida a validade do evento $H \in \mathcal{F}$, isto é,

$$\mathrm{P}[E|H] = \frac{\mathrm{P}[E \cap H]}{\mathrm{P}[H]} \quad . \tag{2.2}$$

Segundo Fisher et al. (1950), o termo do lado esquerdo em (2.2) recebe o nome de verossimilhança quando E representa o conjunto de pontos em Ω com características ou propriedades obtidas por processo de experimentação e H é o conjunto de pontos em Ω associados a propriedades entendidas como hipótese.

Uma expressão simétrica a (2.2) condicionada ao evento E fornece a seguinte relação

$$\mathrm{P}[H|E] = \frac{\mathrm{P}[E \cap H]}{\mathrm{P}[E]} \quad ,$$

a qual combinada com (2.2) estabelece que

$$\mathrm{P}[E|H] = \mathrm{P}[E]\frac{\mathrm{P}[H|E]}{\mathrm{P}[H]} \quad . \tag{2.3}$$

Com (2.3) podemos enunciar o princípio da probabilidade inversa a seguir.

Teorema 2.2. *Sejam E e H eventos em $\mathcal{F}$ então*

$$\frac{\mathrm{P}[H|E]}{\mathrm{P}[H]} \propto \mathrm{P}[E|H] \quad , \tag{2.4}$$

onde $\propto$ representa o símbolo de proporcionalidade.

O resultado (2.4) é conhecido como princípio da probabilidade inversa, apresentado por Bayes em 1763. Em (2.4) o denominador do lado esquerdo recebe o nome de probabilidade inicial ou a priori. O numerador é chamado de probabilidade final ou a posteriori. O princípio da probabilidade inversa afirma que a razão das probabilidades finais e iniciais são proporcionais à verossimilhança. A constante de proporcionalidade em (2.4) é $\mathrm{P}[E]$ e escrevemos

$$\mathrm{P}[E|H] \propto \frac{\mathrm{P}[H|E]}{\mathrm{P}[H]} \quad . \tag{2.5}$$

Se considerarmos os eventos H_1, H_2 e E em (2.5) segue que

$$\begin{aligned} \mathrm{P}[E|H_1] &\propto \frac{\mathrm{P}[H_1|E]}{\mathrm{P}[H_1]} \\ \mathrm{P}[E|H_2] &\propto \frac{\mathrm{P}[H_2|E]}{\mathrm{P}[H_2]} \quad , \end{aligned} \tag{2.6}$$

sendo $\mathrm{P}[E]$ a constante de proporcionalidade.

2.3 Teoria Assintótica e Estimadores de Máxima Verossimilhança

As aproximações paramétricas ou modelos que construímos para o vetor aleatório $\boldsymbol{X} = (X_1, \cdots, X_n)$ contém um número de parâmetros $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^t$ tal que $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$.

O princípio da máxima verossimilhança essencialmente assume que a amostra é representativa da população e escolhe como estimador o valor do parâmetro que maximiza a densidade $f(\boldsymbol{x}|\boldsymbol{\theta})$. Em outras palavras, dada uma amostra aleatória, estamos interessados nos valores dos diferentes parâmetros que tornam mais verossímil a ocorrência dessa amostra.

Definição 2.6. *Seja $\boldsymbol{X} = (X_1, \cdots, X_n)$ vetor aleatório com densidade conjunta $f(\boldsymbol{x}|\boldsymbol{\theta})$ tal que $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. Se $\boldsymbol{X}$ é independente e identicamente distribuído com densidade $f(\boldsymbol{x}|\boldsymbol{\theta})$, então as funções*

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}), \quad l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\, f(x_i|\boldsymbol{\theta}) \quad , \tag{2.7}$$

consideradas como funções de $\boldsymbol{\theta}$, são chamadas de função de verossimilhança e função de log-verossimilhança, respectivamente.

Definição 2.7. *O método da máxima verossimilhança consiste em escolher um estimador $\boldsymbol{\theta} \in \Theta$ que maximiza $L(\boldsymbol{\theta})$, isto é,*

$$L(\hat{\boldsymbol{\theta}}) = \sup_{\theta \in \Theta} L(\boldsymbol{\theta}) = \sup_{\theta \in \Theta} l(\boldsymbol{\theta}) \quad . \tag{2.8}$$

Se $\hat{\boldsymbol{\theta}}$ satisfazendo a equação (2.8) existe, é chamado de estimador de máxima verossimilhança. Se a função de log-verossimilhança $l(\boldsymbol{\theta})$ for uma função diferenciável de $\boldsymbol{\theta}$, então se um supremo $\hat{\boldsymbol{\theta}}$ existe, ele deve satisfazer a equação de verossimilhança

$$\frac{\partial\, l(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \mathbf{0} \quad ,$$

sendo $\partial\, l(\boldsymbol{\theta})/\partial\, \boldsymbol{\theta}$ um vetor p-dimensional, com componentes $\partial\, l(\boldsymbol{\theta})/\partial\, \theta_i$ para $i = 1, \cdots, p$.

Exemplo 2.1. *Suponha que $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. A função de densidade de probabilidade é dada por $f(x) = (2\pi\sigma^2)^{-\frac{1}{2}} exp\{-(x-\mu)^2/(2\sigma^2)\}$, $x \in \mathbb{R}$. Então temos que a log-verossimilhança é*

$$l(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \quad .$$

As equações de verosimilhança têm a forma

$$\frac{\partial\, l(\mu, \sigma^2)}{\partial\, \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} x_i - n\mu \right) = 0 \quad ,$$

$$\frac{\partial\, l(\mu, \sigma^2)}{\partial\, \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \quad .$$

Segue que os estimadores de máxima verossimilhança para μ e σ^2 são

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \quad .$$

Agora apresentamos alguns resultados importantes de aproximações assintóticas. Geralmente, expansões de Taylor são utilizadas para derivar momentos de variáveis aleatórias. Uma noção particularmente útil no contexto de qualquer teoria da aproximação é a acurácia da ordem de magnitude das aproximações. Em análise matemática, a ordem da magnitude é uma aproximação controlada pela notação O. Essa notação pode ser estendida para aproximações de probabilidade com pequenas modificações. O objetivo é rever a notação O e considerar a sua extensão para a teoria assintótica e o comportamento assintótico dos estimadores de máxima verossimilhança. Os resultados a seguir podem ser vistos com detalhes em Borovkov (1987) Fuller (2009), Kendall e Stuart (1961) e Lehmann e Casella (1998).

Definição 2.8. *Seja* $\{a_n, b_n, n \in \mathbb{N}\}$ *uma sequência dupla de números reais. A sequência* $\{a_n, n \in \mathbb{N}\}$ *é dita ser no máximo da ordem* b_n*, sendo denotada por*

$$a_n = O(b_n) \quad \textit{quando}\, n \to \infty \quad \textit{se} \lim_{n\to\infty}\left(\frac{|a_n|}{b_n}\right) < K \quad ,$$

para alguma constante $K > 0$.

A notação O pode ser estendida para funções genéricas de valor real $h(\cdot)$ e $g(\cdot)$ com domínio comum $D \neq \emptyset$. Dizemos que $h(x) = O(g(x))$ quando $x \to x_0$ se para uma constante $K > 0$,

$$\lim_{x\to x_0}\left|\frac{h(x)}{g(x)}\right| \leq K, \quad x \in (D - x_0) \quad .$$

Essa notação é particularmente útil no caso de expansões de Taylor, se $h(x)$ for diferenciável de ordem n em $x = x_0$, então

$$\begin{aligned} h(x_0 + \delta) = h(x_0) + h^{(1)}(x_0)\delta + \frac{h^{(2)}(x_0)}{2!}\delta^2 + \cdots \\ + \frac{h^{(n)}(x_0)}{n!}\delta^n + O(\delta^n) \quad \textit{quando} \quad \delta \to 0 \quad . \end{aligned} \tag{2.9}$$

A notação O em (2.9) pode ser estendida para o caso de convergência estocástica, quase certa e em probabilidade.

Definição 2.9. *Sejam* $\{X_n, n \in \mathbb{N}\}$ *uma sequência de variáveis aleatórias e* $\{c_n, n \in \mathbb{N}\}$ *uma sequência de números reais positivos, Dizemos que* X_n *é no máximo da ordem de* c_n *em probabilidade se existe uma sequência não estocástica* $\{a_n, n \in \mathbb{N}\}$ *tal que*

$$a_n = O(1) \quad e \quad \left(\frac{X_n}{c_n} - a_n\right) \to 0 \quad \textit{em probabilidade} \quad ,$$

sendo denotada por $X_n = O_p(c_n)$. *Da mesma forma podemos denotar* $X_n = O_{q.c.}(c_n)$.

Os resultados da ordem de magnitude relacionados a sequências não estocásticas podem ser transformados em ordem de magnitude estocástica usando o teorema que segue.

Teorema 2.3. *Seja* $\{X_n, n \in \mathbb{N}\}$ *uma sequência k-dimensional de vetores aleatórios onde* $\boldsymbol{X}_n \equiv (X_{jn} : j = 1, 2, \ldots, k)$ *tal que*

$$X_{jn} = O_p(c_{jn}), \quad j = 1, 2, \ldots, m, \quad \textit{tal que} \quad m < k.$$

Se $\{g_n(\boldsymbol{X}), n \in \mathbb{N}\}$ *é uma sequência de funções de Borel* $g_n(\cdot) : \mathbb{R}^k \to \mathbb{R}$ *e*

$$g_n(a_n) = O(b_n) \quad ,$$

para alguma sequência não estocástica de vetores k-dimensional $\{\boldsymbol{a}_n, n \in \mathbb{N}\}$ *tal que*

$$a_{jn} = O(c_{jn}), \quad j = 1, 2, \ldots, m \quad \textit{tal que} \quad m < k \quad ,$$

então podemos deduzir a ordem de $g_n(\boldsymbol{X}_n)$ *sendo* $O_p(b_n)$, *isto é,*

$$g_n(\boldsymbol{X}_n) = O_p(b_n) \quad . \tag{2.10}$$

O Teorema (2.10) pode ser usado para transformar resultados não estocásticos relacionados a expansão de Taylor em resultados estocásticos.

Proposição 2.1. *Algumas propriedades relacionadas a ordem O são as seguintes:*

$$\begin{aligned}
&(i)\, Var(X_n) = \sigma_n^2 < \infty \to X_n = O_p(\sigma_n). \\
&(ii)\, X_n = O_p(1/\sqrt{n}) \to X_n = O_p(1). \\
&(iii)\, X_n \xrightarrow{D} X \to X_n = O_p(1).
\end{aligned} \tag{2.11}$$

Com a finalidade de apresentar alguns resultados importantes da teoria assintótica dos estimadores de máxima verossimilhança, denotaremos

$$u(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{\partial \log f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad \text{e} \quad \mathrm{I}(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{\partial^2 \log f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \quad . \tag{2.12}$$

A primeira expressão de (2.12) é um vetor p-dimensional conhecido como função score. A segunda expressão é uma matriz $p \times p$ conhecida como matriz de informação.

Proposição 2.2. *Algumas propriedades assintóticas que os estimadores de máxima verossimilhança satisfazem são as seguintes:*

Suponha que $\hat{\boldsymbol{\theta}}$ *seja o estimador de máxima verossimilhança para o modelo paramétrico contínuo* $\{f(\boldsymbol{x}|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$, *onde* $\boldsymbol{\theta}$ *é o vetor de parâmetros p-dimensional.*

Considere que as seguintes propriedades são válidas para a densidade $f(\boldsymbol{x}|\boldsymbol{\theta})$*:*

(1) A densidade $f(\boldsymbol{x}|\boldsymbol{\theta})$ *é três vezes diferenciável com respeito a* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^t$*.*

(2) Existem funções integráveis em $\mathbb{R}$*,* $F_1(x)$*, e* $F_2(x)$ *e uma função* $H(x)$ *tal que*

$$\int_{-\infty}^{\infty} H(\boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x} < M$$

para um valor real M*, e as seguintes desigualdades valem para qualquer* $\boldsymbol{\theta} \in \Theta$*:*

$$\left|\frac{\partial \log f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_i}\right| < F_1(x), \quad \left|\frac{\partial^2 \log f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_i \theta_j}\right| < F_2(x),$$

$$\left|\frac{\partial^3 \log f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_i \theta_j \theta_k}\right| < H(x), \quad i,j,k = 1, \cdots, p \quad .$$

(3) A seguinte desigualdade vale para qualquer $\boldsymbol{\theta} \in \Theta$*:*

$$0 < \int_{-\infty}^{\infty} f(\boldsymbol{x}|\boldsymbol{\theta}) \frac{\partial \log f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_j}\, d\boldsymbol{x} < \infty, \quad i,j = 1, \cdots, p \quad .$$

Então sob as condições acima, as seguintes propriedades podem ser derivadas:

(a) Assuma que $\boldsymbol{\theta_0}$ *seja uma solução de*

$$\int f(\boldsymbol{x}|\boldsymbol{\theta}) u(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x} = \mathbf{0}$$

e que $X_1, \cdots, X_n$ *são obtidos de acordo com a densidade* $f(\boldsymbol{x}|\boldsymbol{\theta_0})$*. Além disso, suponha que* $\hat{\boldsymbol{\theta}}_n$ *seja o estimador de máxima verossimilhança baseado em* n *observações. Então, as seguintes propriedades valem:*

(i) A equação de verossimilhança

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial \log f(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

tem uma solução que converge para $\boldsymbol{\theta_0}$*.*

(ii) O estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}}_n$ *converge em probabilidade para* $\boldsymbol{\theta_0}$ *quando* $n \to \infty$*.*

(iii) O estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}}_n$ *tem normalidade assintótica, isto é, a distribuição de* $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0})$ *converge em lei para a distribuição normal p-dimensional*

$N_p(\mathbf{0}, \mathcal{J}(\boldsymbol{\theta_0})^{-1})$ com esperança dada pelo vetor $\mathbf{0}$ e a variância pela matriz de covariância $\mathcal{J}(\boldsymbol{\theta_0})^{-1}$. Em outras palavras, quando $n \to \infty$, vale:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0}) = N_p(\mathbf{0}, \mathcal{J}(\boldsymbol{\theta_0})^{-1}) + O_p(n^{-1/2}) \quad , \tag{2.13}$$

O tamanho do termo restante em (2.13) é capturado de forma concisa pela notação O_p, isto é, por (2.11) temos que $\boldsymbol{Z}_n = O_p(n^{-1/2})$ o que significa que $\sqrt{n}\boldsymbol{Z}_n = O_p(1)$ e converge para zero em probabilidade sujeito as mesma condições de regularidade que garantem que os estimadores de máxima verossimilhança são bem comportados. A matriz $\mathcal{J}(\boldsymbol{\theta_0})$ é o valor da matriz $\mathcal{J}(\boldsymbol{\theta})$ em $\boldsymbol{\theta} = \boldsymbol{\theta_0}$, que é dada por

$$\mathcal{J}(\boldsymbol{\theta}) = \int f(\boldsymbol{x}|\boldsymbol{\theta}) u(\boldsymbol{x}|\boldsymbol{\theta}) u(\boldsymbol{x}|\boldsymbol{\theta})^t \, d\boldsymbol{x} \quad .$$

A matriz $\mathcal{J}(\boldsymbol{\theta})$, sob a condição (3), é chamada de matriz de informação de Fisher.

Embora a normalidade assintótica estabelecida acima assuma a existência de $\boldsymbol{\theta}_0 \in \Theta$ que satisfaz a suposição tal que $g(\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta}_0)$, resultados similares, dados abaixo, podem ser obtidos mesmo quando essa suposição não vale:

(b) Assuma que $\boldsymbol{\theta}_0$ é uma solução de

$$\int g(\boldsymbol{x}) u(\boldsymbol{x}|\boldsymbol{\theta}) \, d\boldsymbol{x} = \mathbf{0}$$

e que $X_1, \cdots, X_n$ são obtidos de acordo com a distribuição $g(\boldsymbol{x})$. Neste caso, as seguintes afirmações valem com respeito a $\hat{\boldsymbol{\theta}}_n$:

(i) O estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}}_n$ converge em probabilidade para $\boldsymbol{\theta_0}$ quando $n \to \infty$.

(ii) A distribuição de $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0})$ com respeito ao estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}}_n$ converge em lei para a distribuição normal p-dimensional com esperança dada pelo vetor $\mathbf{0}$ e a variância pela matriz de covariância $J(\boldsymbol{\theta_0})^{-1} K(\boldsymbol{\theta_0}) J(\boldsymbol{\theta_0})^{-1}$ quando $n \to \infty$. Em outras palavras, quando $n \to \infty$, vale:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0}) = N_p(\mathbf{0}, J(\boldsymbol{\theta_0})^{-1} K(\boldsymbol{\theta_0}) J(\boldsymbol{\theta_0})^{-1}) + O_p(n^{-1/2}) \quad , \tag{2.14}$$

O tamanho do termo restante em (2.14) admite o mesmo tamanho de (2.13) e converge para zero em probabilidade. As matrizes $K(\boldsymbol{\theta_0})$ e $J(\boldsymbol{\theta_0})$ são matrizes $p \times p$ avaliadas em $\boldsymbol{\theta} = \boldsymbol{\theta_0}$ e são dadas pelas seguintes equações

$$K(\boldsymbol{\theta}) = \int g(\boldsymbol{x}) u(\boldsymbol{x}|\boldsymbol{\theta}) u(\boldsymbol{x}|\boldsymbol{\theta})^t \, d\boldsymbol{x} \quad , \tag{2.15}$$

$$J(\boldsymbol{\theta}) = -\int g(\boldsymbol{x})\mathrm{I}(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x} \quad . \tag{2.16}$$

2.4 Séries Temporais

De forma intuitiva, uma série temporal é um conjunto de observações x_t, cada uma sendo gravada em um momento específico do tempo t. Uma série temporal discreta é aquela em que o conjunto T_0 dos tempos no qual as observações estão sendo feitas é discreto, como no caso em que as observações são feitas em intervalos fixos do tempo. Séries temporais contínuas são obtidas quando as observações são gravadas continuamente sobre algum intervalo do tempo, por exemplo, $T_0 = [0, 1]$.

Para permitir a possível imprevisibilidade das observações futuras, é natural supor que cada observação x_t é um valor realizado de uma determinada variável aleatória X_t. A série temporal $\{x_t, t \in T_0\}$ é a realização de uma família de variáveis aleatórias $\{X_t, t \in T_0\}$. Estas considerações sugerem modelar os dados como uma realização de um processo estocástico $\{X_t, t \in T\}$ onde $T \supseteq T_0$. Para clarear as ideias, precisamos definir de forma rigorosa o que é um processo estocástico e suas realizações.

Definição 2.10. *Um processo estocástico é uma família de variáveis aleatórias* $\{X_t, t \in T\}$ *definida em um espaço de probabilidade* $(\Omega, \mathcal{F}, \mathrm{P})$.

Recordando a definição de variável aleatória, note que para cada $t \in T$ fixo, X_t é uma função em Ω. Por outro lado, para cada $\omega \in \Omega$, $X_\cdot(\omega)$ é uma função em T. As funções $\{X_\cdot(\omega), \omega \in \Omega\}$ em T são conhecidas como as realizações de um processo estocástico.

Exemplo 2.2. *Sejam* A *e* Θ *variáveis aleatórias independentes com* $A \geq 0$ *e* Θ *uniformemente distribuída no intervalo* $[0, 2\pi)$. *Um processo estocástico* $\{X(t), t \in \mathbb{R}\}$ *pode ser definido em termos de* A *e* Θ *para qualquer* $v \geq 0$ *e* $r > 0$ *por*

$$X_t = r^{-1} A \cos(vt + \Theta) \quad ,$$

ou mais explicitamente,

$$X_t(\omega) = r^{-1} A(\omega) \cos(vt + \Theta(\omega)) \quad , \tag{2.17}$$

onde ω *é um elemento do espaço de probabilidae* Ω *no qual* A *e* Θ *estão definidas. As realizações do processo definidas em* (2.17) *são as funções de* t *obtidas fixando* ω*, isto é, funções da forma*

$$x_t(\omega) = r^{-1} a \cos(vt + \theta) \quad .$$

Para mais exemplos de processos estocásticos, veja Brockwell, Davis e Fienberg (1991) e Fuller (2009).

Definição 2.11. *Se $\{X(t), t \in \mathbb{R}\}$ é um processo tal que $Var(X_t) < \infty$ para cada $t \in T$, então a função de autocovariância γ_X de $\{X_t\}$ é definida como*

$$\gamma_X(r,s) = Cov[X_r, X_s] = E[(X_r - E[X_r])](X_s - E[X_s]), \quad r,s \in T \quad .$$

Definição 2.12. *A série temporal $\{X(t), t \in \mathbb{Z}\}$ é dita ser estacionária se*

$(i)\, E[|X_t|^2] < \infty \quad \forall t \in \mathbb{Z};$

$(ii)\, E[X_t] = m \quad \forall t \in \mathbb{Z};$

$(iii)\, \gamma_X(r,s) = \gamma_X(r+t, s+t) \quad \forall r,s,t \in \mathbb{Z}.$

Se $\{X_t, t \in \mathbb{Z}\}$ é estacionária então $\gamma_X(r,s) = \gamma_X(r-s,0)$ para todo $r,s \in \mathbb{Z}$. É conveniente redefinir a função de autocovariância do processo estacionário como função de uma única variável aleatória,

$$\gamma_X(h) \equiv \gamma_X(h,0) = Cov[X_{t+h}, X_t] \quad \text{para todo } t,h \in \mathbb{Z} \quad .$$

A função $\gamma_X(\cdot)$ é chamada de função de autocovariância de $\{X_t\}$ e $\gamma_X(h)$ como seu valor na defasagem h. A função de autocorrelação de $\{X_t\}$ é definida analogamente como a função cujo valor na defasagem h é

$$\rho_X(h) \equiv \frac{\rho(h)}{\rho_X(0)} = Cov[X_{t+h}, X_t] \text{ para todo } t,h \in \mathbb{Z} \quad .$$

Proposição 2.3. *Se $\gamma(\cdot)$ é a função de autocovariância de um processo estacionário $\{X_t, t \in \mathbb{Z}\}$, então*

$(i)\, \gamma(0) \geq 0,$

$(ii)\, |\gamma(h)| \leq \gamma(0)$ *para todo* $h \in \mathbb{Z}$

$(iii)\, \gamma(h) = \gamma(-h)$ *para todo* $h \in \mathbb{Z} \quad .$

Agora tratamos de uma classe importante de séries temporais $\{X_t, t \in \mathbb{Z}\}$ definidas em termos de equações de diferenças com coeficientes constantes. A imposição dessa estrutura adicional define uma família paramétrica de processos estacionários, o processo autoregressivo de média móvel (ARMA).

Um dos casos mais simples de séries temporais $\{X_t\}$ é o caso em que as variáveis aleatórias $X_t, t \in \mathbb{Z}$, são independentes e identicamente distribuídas com média zero e variância σ^2. Tal processo pode ser identificado com a classe de todos os processos estacionários tendo média zero e função de autocovariância

$$\gamma(h) = \begin{cases} \sigma^2, & \text{se } h = 0 \\ 0, & \text{se } h \neq 0 \end{cases} \quad . \tag{2.18}$$

Definição 2.13. *O processo $\{Z_t\}$ é dito ser um ruído branco com média zero e variância σ^2 e escrevemos*

$$\{Z_t\} \sim WN(0, \sigma^2)$$

se, e só se, $\{Z_t\}$ tem média zero e função de autocovariância como (2.18). Se as variáveis aleatórias Z_t são independentes e identicamente distribuídas com média zero e variância σ^2 então escrevemos

$$\{Z_t\} \sim IID(0, \sigma^2) \quad .$$

Uma classe muito ampla de processos estacionários pode ser gerada usando ruídos brancos. Isto nos leva a definição de um processo autoregressivo de média móvel (ARMA).

Definição 2.14. *O processo $\{X_t, t \in \mathbb{Z}\}$ é dito ser um processo ARMA se $\{X_t\}$ é estacionário e para cada t*

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \quad , \tag{2.19}$$

onde $\{Z_t \sim WN(0, \sigma^2)\}$. Dizemos que $\{X_t\}$ é um processo ARMA(p,q). As equações (2.19) podem ser escritas simbolicamente de uma forma mais compacta como

$$\phi(B)X_t = \theta(B)Z_t, \ t \in \mathbb{Z} \quad , \tag{2.20}$$

onde p^{th} e q^{th} são os graus dos polinômios ϕ e θ e

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$$

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$$

e B é o operador defasagem definido por

$$B^j X_t = X_{t-1} \ j \in \mathbb{Z} \quad .$$

Os polinômios ϕ e θ são chamados de polinômios autoregressivos e de média móvel das equações de diferença (2.20) respectivamente.

Neste trabalho tratamos apenas do caso em que $\theta(z) \equiv 1$, o que nos dá

$$\phi(B)X_t = Z_t \quad ,$$

e o processo é dito ser autoregressivo de ordem p (ou AR(p)).

Exemplo 2.3. *Um caso particular e de grande importância é o caso em que $\phi(z) = 1 - \phi_1(z)$, isto é,*

$$X_t = Z_t + \phi_1 X_{t-1} \quad . \tag{2.21}$$

Iterando (2.21)*, obtemos*

$$\begin{aligned} X_t &= Z_t + \phi_1 Z_{t-1} + \phi_1^2 X_{t-2} \\ &= \dots \\ &= Z_t + \phi_1 Z_{t-1} + \cdots + \phi_1^k Z_{t-k} + \phi_1^{k+1} X_{t-k-1} \quad . \end{aligned}$$

Se $|\phi| < 1$ *e* $\{X_t\}$ *é estacionária então* $||X_t||^2 = E[X_t^2]$ *é constante, logo*

$$\left|\left| X_t - \sum_{j=0}^{k} \phi_1^j Z_{t-j} \right|\right|^2 = \phi_1^{2k+2} ||X_{t-k-1}||^2 \to 0 \quad ,$$

quando $k \to \infty$*. Como* $\sum_{j=0}^{\infty} \phi_1^j Z_{t-j}$ *converge em média quadrática , pelo critério de Cauchy, nós concluímos que*

$$X_t = \sum_{j=0}^{\infty} \phi_1^j Z_{t-j} \quad . \tag{2.22}$$

Assim, $\{X_t\}$ *definida em* (2.22) *é estacionária pois*

$$E[X_t] = \sum_{j=0}^{\infty} \phi_1^j E[Z_{t-j}] = 0 \quad .$$

Além disso, $\{X_t\}$ *como definida em* (2.22) *satisfaz a equação de diferenças em* (2.21) *e, portanto, é a única solução estacionária. O caso em que* $|\phi_1| > 1$*, a série não converge em* L^2*. Se* $|\phi_1| = 1$*, não existe solução estacionária de* (2.21)*. Consequentemente, não existe AR(1) com* $|\phi_1| = 1$ *de acordo com a definição* (2.19)*.*

O caso geral para que um processo autoregressivo AR(p) seja estacionário é equivalente a condição

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0 \text{ para todo } |z| \leq 1 \quad . \tag{2.23}$$

A prova dessa afirmação segue das propriedades elementares das séries de potências e pode ser vista em Brockwell, Davis e Fienberg (1991) e Fuller (2009).

3 TEORIA DA INFORMAÇÃO

Neste capítulo estamos interessados em estudar duas variáveis aleatórias da teoria da informação e suas esperanças. Na Seção 1 discutimos os conceitos de chances a favor e razão de verossimilhança para apresentar as variáveis aleatórias e mostrar que ambas se originam de proporções de verossimilhanças ou chances finais e iniciais. Na Seção 2 apresentamos a variável aleatória informação de Shannon e sua esperança conhecida como entropia. Na Seção 3 apresentamos a variável aleatória informação de Kullback-Leibler e fazemos a construção de sua esperança, a divergência de Kullback-Leibler, a partir do princípio da probabilidade inversa e da derivada de Radon-Nikodym.

3.1 Chances a Favor e Razão de Verossimilhança

O método natural para decidir qual evento é o mais provável de ocorrer à luz de resultados de um experimento é o princípio da probabilidade inversa (GOOD, 1950). No contexto dos testes de hipóteses, Jeffreys (1948) explica o uso do princípio considerando que H_1 e H_2 são duas hipóteses sobre um experimento. Se não há motivo para acreditar em uma em vez da outra, então as probabilidades a priori são iguais. A hipótese mais provável, quando houver evidência disponível, será a que mais provavelmente levará a essa evidência. Por outro lado, se os dados forem igualmente prováveis de ocorrer em qualquer hipótese, eles dizem nada novo e deve se retornar a opinião anterior. A situação em que os eventos são igualmente prováveis torna não recomendável fazer prognósticos. O princípio da probabilidade inversa lida com situações mais complexas. O ponto levantado por Jeffreys (1948) é que o princípio fornece uma regra geral de acordo com o senso comum que nos guia no uso da experiência para decidir entre hipóteses.

Good (1950) propõe discutir os testes de hipóteses de acordo com o conceito de chances a favor ou chances de ocorrência. Em termos gerais, a chance a favor de um evento H com probabilidade $\mathrm{P}[H]$ é definida como

$$\mathcal{O}[H] = \frac{\mathrm{P}[H]}{1 - \mathrm{P}[H]} \quad ,$$

onde a nomenclatura $\mathcal{O}$ é oriunda da palavra inglesa *odds*. A chance $\mathcal{O}$ do evento H condicionado pelo evento E é

$$\mathcal{O}[H|E] = \frac{\mathrm{P}[H|E]}{1 - \mathrm{P}[H|E]} \quad .$$

Seguindo a nomenclatura da Seção 2.2, $\mathcal{O}[H]$ recebe o nome de chance inicial e $\mathcal{O}[H|E]$ chance final. Se pensarmos os eventos H_1 e H_2 em (2.6) como complementares, isto é, $H_1 = H$

e $H_2 = \bar{H}$, temos a razão das verossimilhanças

$$\begin{aligned}\frac{\mathrm{P}[E|H]}{\mathrm{P}[E|\bar{H}]} &= \frac{\mathrm{P}[E \cap H]}{\mathrm{P}[H]} \frac{\mathrm{P}[\bar{H}]}{\mathrm{P}[E \cap \bar{H}]} \\ &= \frac{\mathrm{P}[E \cap H]}{\mathrm{P}[E \cap \bar{H}]} \frac{\mathrm{P}[\bar{H}]}{\mathrm{P}[H]} = \frac{1}{\mathcal{O}[H]} \frac{\mathrm{P}[H|E]}{\mathrm{P}[E \cap \bar{H}]} \\ &= \frac{1}{\mathcal{O}[H]} \frac{\mathrm{P}[H|E]\mathrm{P}[\mathrm{E}]}{\mathrm{P}[\bar{H}|E]\mathrm{P}[\mathrm{E}]} = \frac{1}{\mathcal{O}[H]} \frac{\mathrm{P}[H|E]}{1 - \mathrm{P}[H|E]} \\ &= \frac{\mathcal{O}[H|E]}{\mathcal{O}[H]} \quad .\end{aligned}$$

Dizemos que a razão das chances finais e iniciais é igual a razão das verossimilhanças, o que nos dá,

$$\frac{\mathcal{O}[H|E]}{\mathcal{O}[H]} = \frac{\mathrm{P}[E|H]}{\mathrm{P}[E|\bar{H}]} \quad . \tag{3.1}$$

A igualdade em (3.1) da origem ao fator

$$\frac{\mathcal{O}[H|E]}{\mathcal{O}[H]} \equiv f \equiv \frac{\mathrm{P}[E|H]}{\mathrm{P}[E|\bar{H}]} \quad , \tag{3.2}$$

que pode ser interpretado como o fator necessário para que as chances iniciais adquiram o valor das chances finais. Uma outra interpretação para (3.2) é como o fator a favor do evento H em virtude dos condicionantes E.

Uma proposta de mensuração do fator f foi dada por Turing (GOOD, 1950) que sugeriu como a melhor forma de capturar a diferença entre o estágio final e o inicial como sendo a aplicação da função logarítmica em (3.2)

$$\log f = \log \frac{\mathcal{O}[H|E]}{\mathcal{O}[H]} = \log \frac{\mathrm{P}[E|H]}{\mathrm{P}[E|\bar{H}]} \quad . \tag{3.3}$$

Com o desenvolvimento da teoria da informação e comunicação, Good e Osteyee (1974) consolidam o trabalho de Good (1950). Interessados em medidas da teoria da informação e especialmente em problemas relacionadas com a detecção de sinais e ruídos na teoria da comunicação, Good e Osteyee (1974) estudam o conceito de informação. Com o intuito de comparar a razão das chances finais e iniciais, no caso em que $H = E$, Good e Osteyee (1974) propõe

$$\log \frac{\mathrm{P}[H|E]}{\mathrm{P}[E]} = \log \frac{\mathrm{P}[H|H]}{\mathrm{P}[H]} \quad ,$$

o que à luz de (3.3) fornece

$$\mathrm{I}[H] = -\log \mathrm{P}[H] \quad . \tag{3.4}$$

Kullback e Leibler (1951) generalizaram o conceito de informação proposto por Shannon (1948) e Shannon e Weaver (1949). Em um trabalho clássico sobre informação e suficiência

estatística, Kullback e Leibler (1951) propuseram uma medida da "distância" ou divergência entre populações estatísticas em termos de medidas da teoria da informação. Posteriormente, Kullback (1958) aprofunda seus estudos sobre teoria da informação e estatística e aborda o trabalho de Good (1950) por meio da equação (3.3). Na comparação de dois eventos alternativos H_1 e H_2, a equação (3.3) se torna

$$\mathcal{KL}[H_1/H_2 : E] = \log \frac{\mathcal{O}[H_1/H_2|E]}{\mathcal{O}[H_1/H_2]} \quad . \tag{3.5}$$

As expressões (3.4) e (3.5) são variáveis aleatórias chamadas informação de Shannon e informação de Kullback-Leibler, respectivamente. Elas são objeto de desenvolvimento nas seções que seguem deste capítulo.

3.2 A Variável Aleatória Informação de Shannon

Definição 3.1. *Sejam* $(\Omega, \mathcal{F}, \mathrm{P})$ *um espaço de probabilidade e* $E \in \mathcal{F}$ *um evento. A fim de que* $\mathrm{I} : (\Omega, \mathcal{F}, \mathrm{P}) \to \mathbb{R}$ *seja o ganho de informação* $\mathrm{I}[E]$ *sobre o evento* E *quando* E *ocorre é necessário que*

(i) seja uma função decrescente de probabilidade,

(ii) e o ganho de informação da ocorrência conjunta de eventos independentes deve ser igual a soma dos ganhos de informação individuais de cada evento.

Teorema 3.1. *Uma função que satisfaz as duas propriedades acima é*

$$\mathrm{I}[E] = -\log \mathrm{P}[E] \quad , \tag{3.6}$$

sendo chamada de informação de Shannon ou auto-informação.

Demonstração. (i) Suponha que $\mathrm{P}[E] < \mathrm{P}[F]$. Então

$$\mathrm{I}[E] = -\log \mathrm{P}[E] > -\log \mathrm{P}[F] = \mathrm{I}[F] \quad .$$

(ii) Suponha que $E_1, \ldots, E_n$ são eventos independentes. Então

$$\begin{aligned} \mathrm{I}\Big(\bigcap_{i=1}^{n} E_i\Big) &= -\log \mathrm{P}\Big(\bigcap_{i=1}^{n} E_i\Big) = -\log \prod_{i=1}^{n} \mathrm{P}[E_i] \\ &= -\sum_{i=1}^{n} \log \mathrm{P}[E_i] = \sum_{i=1}^{n} \mathrm{I}[E_i] \quad . \end{aligned}$$

□

Proposição 3.1. *As únicas funções que satisfazem as condições acima são as funções logarítmicas.*

A prova dessa afirmação consiste em demonstrar a unicidade das funções logarítmicas e pode ser vista em Shannon (1948). Os logaritmos são caracterizados por duas propriedades simples e naturais, de modo que a escolha do processo de apresentá-los é uma questão de preferência. Uma vez que valham as duas propriedades, só existe uma maneira de alterar uma função logarítmica: multiplicá-la por uma constante. Isto decorre do fato de que toda função logarítmica é uma bijeção entre $\mathbb{R}^+$ e $\mathbb{R}$. Segue das propriedades da função logarítmica que dada uma função logarítmica $L : \mathbb{R}^+ \to \mathbb{R}$, existe um único elemento $a > 0$ tal que $L(a) = 1$. Este número é chamado base da função L. Nesse sentido, qualquer função que satisfaça as duas propriedades de (3.6) deve ser proporcional a $-\log \mathrm{P}[E]$. Neste trabalho utilizamos o logaritmo na base 10. Porém, se a base fosse 2, então a unidade seria um bit. Por exemplo, se uma moeda honesta é jogada e o resultado é cara, então um bit de informação é fornecido (GOOD, 1950).

Proposição 3.2. *Algumas das propriedades que* $\mathrm{I}[E]$ *satisfaz são as seguintes:*

$(i)\, \mathrm{I}[\Omega] = 0.$

$(ii)\, \mathrm{I}[E] \geqslant 0\,,$ *para todo* $E \in \mathcal{F}$.

$(iii)\, \mathrm{P}[E] = 0 \to \mathrm{I}[E] = \infty.$

3.2.1 Informação Mútua

Definição 3.2. *O ganho de informação condicional* $\mathrm{I}[E|F]$ *é definido como o ganho de informação quando* E *ocorre dado que* F *ocorreu, sendo* $\mathrm{P}[F] > 0$*, como mostrada a seguir*

$$\begin{aligned} \mathrm{I}[E|F] &= -\log \mathrm{P}[E|F] = -\log \frac{\mathrm{P}[E \cap F]}{\mathrm{P}[F]} \\ &= -\log \mathrm{P}[E \cap F] + \log \mathrm{P}[F] = \mathrm{I}[E \cap F] - \mathrm{I}[F] \quad . \end{aligned} \tag{3.7}$$

De forma análoga, temos o ganho de informação quando F *ocorre dado que* E *ocorreu, como*

$$\mathrm{I}[F|E] = \mathrm{I}[E \cap F] - \mathrm{I}[E] \quad . \tag{3.8}$$

Proposição 3.3. *Algumas propriedades que* $\mathrm{I}[E|F]$ *satisfaz são as seguintes:*

$(i)\, F \subseteq E \to \mathrm{I}[E|F] = 0.$

$(ii)\, E \subseteq F \to \mathrm{I}[E|F] = \mathrm{I}[E] - \mathrm{I}(F).$

$(iii)\, E \perp F \to \mathrm{I}[E|F] = \mathrm{I}[E].$

$(iv)\, \mathrm{P}[E \cap F] = 0 \to \mathrm{I}[E|F] = \infty.$

Observamos que (3.7) e (3.8) fornecem duas formas de calcular a informação da ocorrência conjunta de dois eventos, isto é,

$$\begin{aligned} \mathrm{I}[E \cap F] &= \mathrm{I}[E|F] + \mathrm{I}[F] \\ \mathrm{I}[E \cap F] &= \mathrm{I}[F|E] + \mathrm{I}[E] \quad . \end{aligned}$$

A combinação destas últimas resulta na definição de informação mútua como apresentada a seguir.

Definição 3.3. *A informação mútua entre eventos E e F é definida como*

$$\mathrm{I}[E:F] = \mathrm{I}[E] - \mathrm{I}[E|F] = \mathrm{I}[F] - \mathrm{I}[F|E] \quad , \tag{3.9}$$

sendo $\mathrm{P}[E] > 0 \wedge \mathrm{P}[F] > 0$ *por* (3.7)*. Então,* $\mathrm{I}[E:F]$ *é simétrica em* E *e* F*.*

A informação mútua entre dois eventos é uma medida de decrescimento (se positiva) ou crescimento (se negativa) da incerteza sobre a ocorrência de E (ou F) causada pela ocorrência de F (ou E). Também pode ser pensada como uma medida do decrescimento (se positiva) ou crescimento (se negativa) do ganho de informação quando E (ou F) ocorre, causada pela ocorrência de F (ou E).

Nesse sentido, $\mathrm{I}[E:F]$ representa uma medida da variação da informação do evento E dado a ocorrência do evento F. Aplicando (3.6) a (3.9), temos que $\mathrm{I}[E:F]$ também pode ser expressa como

$$\begin{aligned} \mathrm{I}[E:F] &= -\log \mathrm{P}[E] + \log \mathrm{P}[E|F] = \log \frac{\mathrm{P}[E|F]}{\mathrm{P}[E]} \\ &= -\log \mathrm{P}[F] + \log \mathrm{P}[F|E] = \log \frac{\mathrm{P}[F|E]}{\mathrm{P}[F]} = \mathrm{I}[F:E] \quad . \end{aligned} \tag{3.10}$$

Segue que $\mathrm{I}[E:F] > 0 \leftrightarrow \mathrm{P}[E|F] > \mathrm{P}[E]$ ou equivalentemente $\mathrm{I}[F:E] > 0 \leftrightarrow \mathrm{P}[F|E] > \mathrm{P}[F]$.

Proposição 3.4. *As propriedades que* $\mathrm{I}[E:F]$ *satisfaz são as seguintes:*

$(i)\ E \perp F \rightarrow \mathrm{I}[E:F] = 0.$
$(ii)\ F \subseteq E \rightarrow \mathrm{I}[E:F] = \mathrm{I}[E].$
$(iii)\ \mathrm{P}[E] > 0 \wedge \mathrm{P}[F] > 0 \wedge \mathrm{P}[E \cap F] = 0 \rightarrow \mathrm{I}[E:F] = -\infty.$

Seguindo a construção feita em (2.5) pelo princípio da probabilidade inversa, temos que a informação mútua entre dois eventos é o $\log$ da verossimilhança ou o $\log$ da razão entre as probabilidades finais e iniciais. Good (1956) também apresenta a informação mútua entre dois eventos como o logaritmo do fator de associação entre os eventos.

A definição de informação mútua está ligada com o princípio da suficiência estatística. Se E é um resultado de um experimento e θ uma estatística, isto é, uma função de E, então o parâmetro populacional $\hat{\theta}$ é dito ser suficiente para θ se $\mathrm{P}[E|\hat{\theta}] = \mathrm{P}[E|\theta, \hat{\theta}]$. Pode ser deduzido que $\mathrm{I}[\theta:E] = \mathrm{I}[\theta:\hat{\theta}]$ e isso dá um significado rigoroso ao fato que $\hat{\theta}$ fornece toda a informação a respeito de θ que é dado pelo experimento ou evidência (GOOD, 1956).

3.2.2 Entropia: Esperança da Informação de Shannon

Consideramos X a variável aleatória discreta informação de Shannon associada ao espaço de probabilidade $(\Omega, \mathcal{F}, \mathrm{P})$. Na abordagem proposta por Shannon (1948), temos um espaço particionado em uma quantidade finita de eventos E_k cujas probabilidades são assumidas como conhecidas. A probabilidade de um evento E_k é tal que $\mathrm{P}[E_k] = p_k$. A função $\mathrm{I} \equiv X : (\Omega, \mathcal{F}, \mathrm{P}) \to \mathbb{R}$ é dada por $\mathrm{I}[\omega] \equiv X(\omega) = -\log p_k$, $\omega \in E_k \in \mathcal{F}$. A entropia é a esperança da informação de Shannon e representa a média da quantidade de informação por evento. Simbolicamente, denotamos a entropia por

$$\mathrm{H}[X] = \mathrm{H}[\mathrm{I}] = E[\mathrm{I}] = E[X] = -\sum_{k=1}^{n} p_k \log p_k \quad .$$

Proposição 3.5. *Algumas das propriedades que* $\mathrm{H}[X]$ *satisfaz são as seguintes:*

(i) $\mathrm{H} = 0$ *se, e só se, todos os* p_k *são zero, exceto um, este possuindo a unidade de valor. Assim, apenas quando tivermos certeza do resultado, H é zero. Caso contrário,* H *é positivo.*

(ii) Para um dado n*,* H *é máximo e é igual a* $\log n$ *quanto todos os* p_k *são iguais (isto é,* $\frac{1}{n}$*). Esta é intuitivamente a situação mais incerta.*

A generalização do caso unidimensional para o caso n-dimensional pode ser considerada como uma regra de indução para a derivação de H para qualquer espaço de probabilidade de dimensão finita. No caso bidimensional para variáveis aleatórias discretas X e Y temos que

$$\mathrm{H}[X, Y] = -\sum_{ij} p(i, j) \log p(i, j) \quad .$$

Enquanto

$$\mathrm{H}[X] = -\sum_{i,j} p(i, j) \log \sum_{j} p(i, j)$$

$$\mathrm{H}[Y] = -\sum_{i,j} p(i, j) \log \sum_{i} p(i, j) \quad .$$

É fácil ver que

$$\mathrm{H}[X, Y] \leq \mathrm{H}[X] + \mathrm{H}[Y] \quad ,$$

com igualdade se, e só se, as variáveis aleatórias são independentes (isto é, $p(i, j) = p(i)p(j)$). A incerteza conjunta é menor ou igual a soma das incertezas individuais.

Também podemos definir a entropia condicional $\mathrm{H}[X|Y]$, que é a entropia de uma variável aleatória condicionada ao conhecimento de outra variável aleatória,

$$\mathrm{H}[X|Y] = -\sum_{i,j} p(i, j) \log p(i|j) \quad .$$

Quando consideramos variáveis aleatórias contínuas, a entropia é uma extensão do caso discreto das definições de entropia para o caso contínuo, de forma similar ao que é feito quando se estende probabilidades discretas para contínuas. Assim, a entropia para a variável aleatória contínua X é dada por

$$\mathrm{H}[X] = -\int_{\mathbb{R}} f(x)\log f(x)dx \quad .$$

No caso bidimensional, temos

$$\mathrm{H}[X,Y] = -\int_{\mathbb{R}}\int_{\mathbb{R}} f(x,y)\log f(x,y)dxdy$$

$$\mathrm{H}[X|Y] = -\int_{\mathbb{R}}\int_{\mathbb{R}} f(x,y)\frac{\log f(x,y)}{f_Y(y)}dxdy \quad .$$

Para o caso n-dimensional, isto é, $X_1,\ldots,X_n$ sequência de variáveis aleatórias, com densidade $f(x_1,\ldots,x_n)$, a entropia é definida como

$$\mathrm{H}[X_1,\ldots,X_n] = -\int_{\mathbb{R}}\cdots\int_{\mathbb{R}} f(x_1,\ldots,x_n)\log f(x_1,\ldots,x_n)dx_1\ldots dx_n \quad .$$

Todas as definições acima estão condicionadas a existência das correspondentes integrais. Para mais informações sobre entropia, veja Cover e Thomas (2012) Shannon (1948) e Reza (1994).

Exemplo 3.1. *Seja X variável aleatória com distribuição exponencial com parâmetro λ. Então a função densidade é dada por $f(x) = \lambda e^{-\lambda x}, x \geq 0$. Temos que a entropia de X é*

$$\begin{aligned}
\mathrm{H}[X] &= -\int_0^\infty \lambda e^{-\lambda x}\log(\lambda e^{-\lambda x})dx \\
&= -\left(\int_0^\infty (\log\lambda)\lambda e^{-\lambda x}dx + \int_0^\infty (-\lambda x)\lambda e^{-\lambda x}dx\right) \\
&= -\log\lambda\int_0^\infty f(x)dx + \lambda E[X] \\
&= -\log\lambda + 1 \quad .
\end{aligned}$$

3.3 A Variável Aleatória Informação de Kullback-Leibler

O principal objetivo dessa seção é apresentar a variável aleatória informação de Kullback-Leibler e a sua esperança, a divergência de Kullback-Leibler. Definida a variável aleatória informação de Kullback-Leibler, mostramos que a sua esperança é construída a partir do princípio da probabilidade inversa e da derivada de Radon-Nikodym.

Definição 3.4. *Sejam H_1 e H_2 duas hipóteses relacionadas a alguma evidência E. H_1 e H_2 devem ser pensadas como eventos em espaços de probabilidade a priori $(\Omega, \mathcal{F}, \mathrm{P})$ e $(\Omega, \mathcal{F}, \mathrm{P}_E)$, sendo P_E é a probabilidade condicional dado E. Então a variável aleatória informação de Kullback-Leibler (ou simplesmente informação de Kullback-Leibler) mede o peso de evidência em favor de H_1 em oposição a H_2 dado E, definida como*

$$\mathcal{KL}[H_1/H_2 : E] = \log \frac{\mathcal{O}[H_1/H_2|E]}{\mathcal{O}[H_1/H_2]} \quad , \tag{3.11}$$

sendo $\mathcal{O}[H_1/H_2|E]$ a chance de ocorrência em favor de H_1 em oposição a H_2 dado E, e $\mathcal{O}[H_1/H_2]$ é a chance de ocorrência em favor de H_1 em oposição a H_2. A informação de Kullback-Leilber pode assumir valores positivos ou negativos.

Outra equação para $\mathcal{KL}[H_1/H_2 : E]$ é, de (3.11) e (3.9),

$$\begin{aligned} \mathcal{KL}[H_1/H_2 : E] &= \log \frac{\mathrm{P}[H_1|E]}{\mathrm{P}[H_2|E]} - \log \frac{\mathrm{P}[H_1]}{\mathrm{P}[H_2]} \\ &= \log \frac{\mathrm{P}[H_1|E]}{\mathrm{P}[H_1]} - \log \frac{\mathrm{P}[H_2|E]}{\mathrm{P}[H_2]} \\ &= \mathrm{I}[H_1 : E] - \mathrm{I}[H_2 : E] \quad . \end{aligned} \tag{3.12}$$

A informação de Kullback-Leibler pode ser interpretada como a diferença em informação sobre H_1 comparado com H_2 dado E. Se considerarmos E um evento nos espaços de probabilidade $(\Omega, \mathcal{A}, \mathrm{P}_{H_1})$ e $(\Omega, \mathcal{A}, \mathrm{P}_{H_2})$, outra equação para $\mathcal{KL}[H_1/H_2 : E]$ é

$$\begin{aligned} \mathcal{KL}[H_1/H_2 : E] &= \mathrm{I}[E : H_1] - \mathrm{I}[E : H_2] \\ &= \log \frac{\mathrm{P}[E|H_1]}{\mathrm{P}[E]} - \log \frac{\mathrm{P}[E|H_2]}{\mathrm{P}[E]} \\ &= \log \frac{\mathrm{P}[E|H_1]}{\mathrm{P}[E|H_2]} \quad . \end{aligned} \tag{3.13}$$

Como podemos ver em (3.13), a informação de Kullback-Leibler também representa a medida da diferença de informação entre duas verossimilhanças a posteriori, uma condicionada por H_1 e outra por H_2.

3.3.1 Divergência: Esperança da Informação de Kullback-Leibler

A construção da divergência de Kullback-Leibler tem como base o sistema (2.6) junto com a aplicação da derivada de Radon-Nikodym (2.1) como mostramos a seguir. Primeiro reescrevemos (2.6) considerando que as probabilidades condicionadas por H_1 e H_2 induzem

medidas finitas μ_1 e μ_2,

$$\mu_i(E) \propto \frac{\mathrm{P}[H_i|E]}{\mathrm{P}[H_i]}, \quad , i = 1, 2 \quad . \tag{3.14}$$

Em seguida, introduzimos uma terceira medida λ tal que $\lambda \equiv \mu_1$ e $\lambda \equiv \mu_2$. Por exemplo, λ pode ser μ_1, ou μ_2, ou $(\mu_1 + \mu_2)/2$. Então por (2.1) existem funções $f_i, i = 1, 2$, mensuráveis finitas em relação a λ, únicas para $\lambda(E) = 0$, tal que

$$\mu_i(E) = \int_E f_i \, d\lambda, \quad i = 1, 2 \quad ,$$

para todo conjunto mensurável $E \in \mathcal{F}$. O símbolo $[\lambda]$, pronunciado como módulo λ, segue de uma afirmação em relação aos elementos de Ω, que significa que a afirmação é verdadeira exceto para um conjunto E tal que $E \in \mathcal{F}$ e $\lambda(E) = 0$. As funções f_i recebem o nome de funções de densidade generalizadas (ou ainda, derivadas de Radon-Nikodym) e, assim como as funções de densidade usuais, representam uma derivada generalizada da função de distribuição associada à medida de probabilidade μ_i. Escrevemos

$$d\mu_i = f_i d\lambda \quad ou \quad \frac{d\mu_i}{d\lambda} = f_i, \quad i = 1, 2 \quad . \tag{3.15}$$

Teorema 3.2. *Seja $\mathcal{KL}[H_1/H_2 : E]$ a informação de Kullback-Leibler e μ_1 a medida condicionada pela hipótese H_1, então*

$$\begin{aligned} \mathcal{D}[f_1 : f_2] &= \int \mathcal{KL}[H_1/H_2 : E] \, d\mu_1 \\ &= \int \log \frac{f_1(x)}{f_2(x)} \, d\mu_1 \end{aligned} \tag{3.16}$$

é chamada de divergência de Kullback-Leibler.

Demonstração. Considerando o módulo λ em (3.14), escrevemos

$$\mu_i(E) \propto \frac{\mathrm{P}[H_i|E]}{\mathrm{P}[H_i]}[\lambda], \quad i = 1, 2 \quad .$$

Por (3.15), temos que $f_i = \frac{d\mu_i}{d\lambda}, i = 1, 2$. Considerando que a constante de proporcionalidade é a mesma para $i = 1, 2$, escrevemos

$$\frac{f_1(E)}{f_2(E)} = \frac{\mathrm{P}[H_1|E]}{\mathrm{P}[H_1]} \frac{\mathrm{P}[H_2]}{\mathrm{P}[H_2|E]}[\lambda] \quad . \tag{3.17}$$

Agora aplicando o logaritmo em (3.17), vem que

$$\begin{aligned} \log \frac{f_1(E)}{f_2(E)} &= \log \frac{\mathrm{P}[H_1|E]}{\mathrm{P}[H_1]} + \log \frac{\mathrm{P}[H_2|E]}{\mathrm{P}[H_2]}[\lambda] \\ &= \log \frac{\mathrm{P}[H_1|E]}{\mathrm{P}[H_2|E]} - \log \frac{\mathrm{P}[H_1]}{\mathrm{P}[H_2]}[\lambda] \quad , \end{aligned} \tag{3.18}$$

que corresponde a (3.12), o peso de evidência a favor de H_1 em oposição à H_2 dado E. Assim (3.18) pode ser calculada pelo logaritmo da proporção das densidades generalizadas de E dado H_1 e H_2. Sabemos que $\mathcal{KL}[H_1/H_2 : E]$ representa a diferença das informações mútuas que, por definição, é uma função em μ_1, logo seu valor esperado é

$$\begin{aligned}\mathcal{D}[f_1 : f_2] &= \int \mathcal{KL}[H_1/H_2 : E]\, d\mu_1 \\ &= \int \log \frac{f_1(x)}{f_2(x)}\, d\mu_1 = \int f_1(x)\, \log \frac{f_1(x)}{f_2(x)}\, d\lambda(x) \quad .\end{aligned} \tag{3.19}$$

A equação (3.19) é a divergência de Kullback-Leibler em relação a μ_1 ou a divergência de Kullback-Leibler de μ_1 com respeito a μ_2. De forma análoga, podemos definir

$$\begin{aligned}\mathcal{D}[f_2 : f_1] &= \int \mathcal{KL}[H_1/H_2 : E]\, d\mu_2 \\ &= \int \log \frac{f_2(x)}{f_1(x)}\, d\mu_2 = \int f_2(x)\, \log \frac{f_2(x)}{f_1(x)}\, d\lambda(x) \quad ,\end{aligned} \tag{3.20}$$

como a esperança da informação de Kullback-Leibler em relação a μ_2 ou a divergência de Kullback-Leibler de μ_2 com respeito a μ_1. □

Podemos definir $\bar{\mathcal{D}}[f_1 : f_2]$ por

$$\begin{aligned}\bar{\mathcal{D}}[f_1 : f_2] =& \mathcal{D}[f_1 : f_2] + \mathcal{D}[f_2 : f_1] \\ &= \int (f_1(x) - f_2(x)) \log \frac{f_1(x)}{f_2(x)} d\lambda(x) \\ &= \int \log \frac{P[H_1|E]}{P[H_2|E]} d\mu_1 - \int \log \frac{P[H_1|E]}{P[H_2|E]} d\mu_2 \quad .\end{aligned} \tag{3.21}$$

$\bar{\mathcal{D}}[f_1 : f_2]$ é uma medida da diferença de informação entre duas hipóteses H_1 e H_2 ou entre μ_1 e μ_2. Além disso, $\bar{\mathcal{D}}[f_1 : f_2]$ é simétrica com respeito à μ_1 e μ_2, e as probabilidades a priori $P[H_i], i = 1, 2$, não aparecem. A informação de (3.21) tem todas as propriedades de distância como definido na topologia exceto a desigualdade triangular e portanto não é uma distância. Para mais detalhes, veja Kullback (1958).

Geralmente, na literatura estatística, (3.19) é representada por

$$\begin{aligned}\mathcal{D}[f_1 : f_2] &= E_{f_1}\left[\log \frac{f_1(X)}{f_2(X)}\right] \\ &= E_{f_1}[\log f_1(X)] - E_{f_1}[\log f_2(X)] \quad ,\end{aligned} \tag{3.22}$$

onde a esperança está sendo tomada sob f_1.

Exemplo 3.2. *Suponha que estamos interessados em calcular a divergência de Kullback-Leibler entre duas funções de probabilidade e que as mesmas são normalmente distribuídas. Sejam* $f_1 \sim N(\xi, \tau^2)$ *e* $f_2 \sim N(\mu, \sigma^2)$*. Se* E_{f_1} *é uma esperança com respeito a* f_1*, então a variável aleatória* X *segue distribuição* $N(\xi, \tau^2)$ *e a equação vale*

$$\begin{aligned} E_{f_1}[(X-\mu)^2] &= E_{f_1}[(X-\xi)^2 + 2(X-\xi)(\xi-\mu) + (\xi-\mu)^2] \\ &= \tau^2 + (\xi-\mu)^2 \quad . \end{aligned}$$

Para a distribuição normal, sabemos que $f_1(x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\{-(x-\mu)^2/(2\sigma^2)\}$, $x \in \mathbb{R}$*, o que nos dá*

$$\begin{aligned} E_{f_1}[\log\ f_2(X)] &= E_X\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(X-\mu)^2}{2\sigma^2}\right] \\ &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{\tau^2 + (\xi-\mu)^2}{2\sigma^2} \quad . \end{aligned}$$

Em particular, se tomarmos $\mu = \xi$ *e* $\sigma^2 = \tau^2$ *nesta equação, segue que*

$$E_{f_1}[\log\ f_1(X)] = -\frac{1}{2}\log(2\pi\tau^2) - \frac{1}{2} \quad .$$

A divergência de Kullback-Leibler de f_1 *com respeito a* f_2 *é dada por*

$$\begin{aligned} \mathcal{D}[f_1 : f_2] &= E_{f_1}[\log\ f_1(X)] - E_{f_1}[\log\ f_2(X)] \\ &= \frac{1}{2}\left[\log\frac{\sigma^2}{\tau^2} + \frac{\tau^2 + (\xi-\mu)^2}{\sigma^2} - 1\right] \quad . \end{aligned}$$

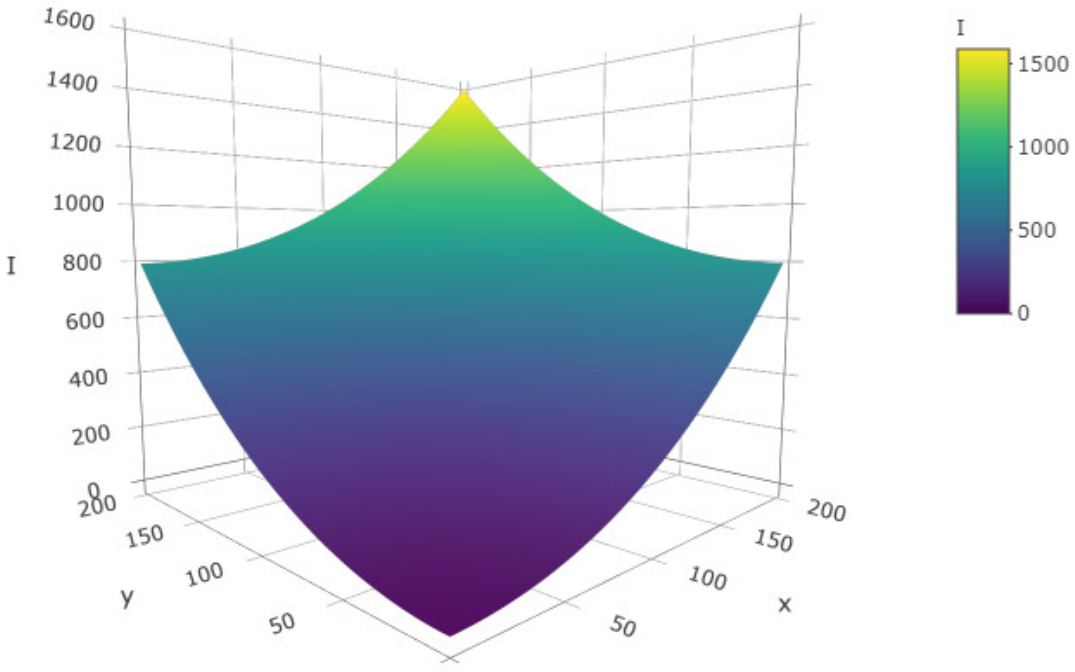
A Figura-1 representa a variação da divergência de Kullback-Leibler supondo que $f_1 \sim N(0,1)$ e os parâmetros de f_2 variam. Denotamos $x = \xi$ e $y = \tau^2$ e

$$\mathcal{D}[f_1; f_2] = \frac{1}{2}\left[-2\log\tau + \tau^2 + \xi^2 - 1\right] \quad .$$

Conforme ξ e τ^2 se aproximam dos parâmetros de f_1, a divergência de Kullback-Leibler tende a zero.

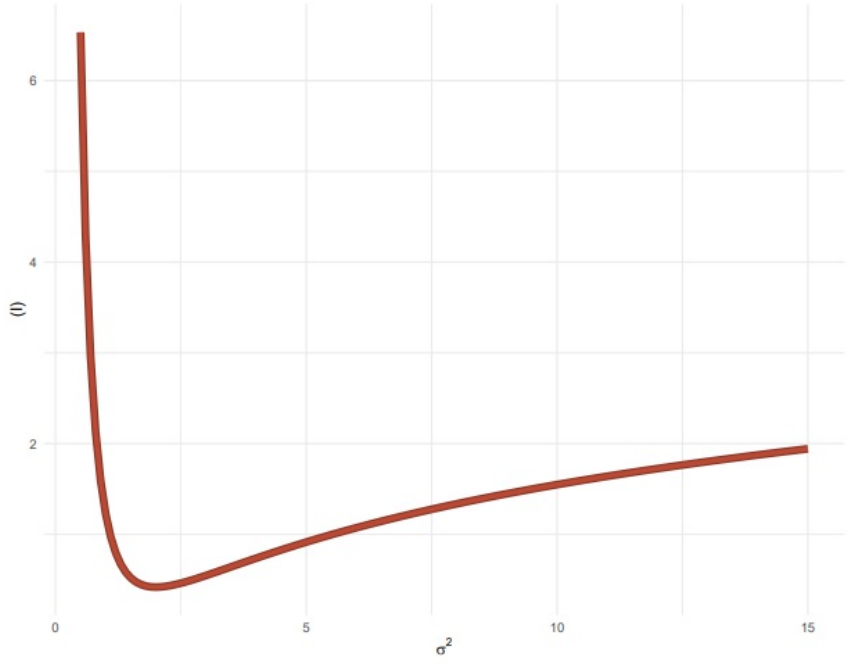
Exemplo 3.3. *Agora suponha que estamos interessados em calcular a divergência de Kullback-Leibler entre duas funções de probabilidade mas que* $f_1 \sim Laplace(\mu, b) = Laplace(0,1)$ *e que* $f_2 \sim N(\mu, \sigma^2)$*.*

Figura 1 – Comparação de $\mathcal{D}[f_1; f_2]$ assumindo que $f_1 \sim N(0,1)$ e $f_2 \sim N(\xi, \tau^2)$



Fonte: Vital (2019)

Figura 2 – Comparação de $\mathcal{D}[f_1; f_2]$ assumindo que $f_1 \sim Laplace(0,1)$ e $f_2 \sim N(0, \sigma^2)$



Fonte: Vital (2019)

Para a distribuição de Laplace com parâmetros $\mu = 0$ *e* $b = 1$*, sabemos que* $f_1(x) = \frac{1}{2}\exp\{-|x|\}, x \in \mathbb{R}$*. Então*

$$\begin{aligned} E_{f_1}[\log f_1(X)] &= -\log 2 - \frac{1}{2}\int_{-\infty}^{\infty} |x| e^{-|x|} dx \\ &= -\log 2 - \int_0^{\infty} x e^{-x} dx \\ &= -\log 2 - 1 \end{aligned}$$

$$\begin{aligned} E_{f_1}[\log f_2(X)] &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{4\sigma^2}\int_{-\infty}^{\infty} (x-\mu)^2 e^{-|x|} dx \\ &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{4\sigma^2}(4 + 2\mu^2) \quad . \end{aligned}$$

Então a divergência de Kullback-Leibler de f_1 *com respeito a* f_2 *é*

$$\mathcal{D}[f_1; f_2] = \frac{1}{2}\log(2\pi\sigma^2) + \frac{2+\mu^2}{2\sigma^2} - \log 2 - 1 \quad .$$

A Figura-2 representa a variação da divergência Kulback-Leibler supondo que $f_1 \sim Laplace(0, 1)$ e $f_2 \sim N(0, \sigma^2)$. Então temos que

$$\mathcal{D}[f_1; f_2] = \frac{1}{2}\log(\pi\sigma^2) + \frac{1}{\sigma^2} - 1 \quad .$$

Vemos que supondo distribuições de probabilidade diferentes, a divergência Kullback-Leibler não assume o valor zero. Ainda que $\sigma^2 = 1$, temos que $\mathcal{D}[f_1; f_2] = \frac{1}{2}\log(\pi) > 0$.

3.3.2 Propriedades da Divergência de Kullback-Leibler

Nesta subseção estudamos algumas propriedades da divergência de Kullback-Leibler e examinamos as suas implicações. Para mais detalhes sobre as propriedades da divergência de Kullback-Leibler, o leitor pode consultar Kullback (1958).

Teorema 3.3. $\mathcal{D}[f_1 : f_2]$ *é aditiva para variáveis aleatórias independentes, isto é, para* X *e* Y *independentes sob as hipóteses* H_i*,* $i = 1, 2$*,*

$$\mathcal{D}[f_1 : f_2; X, Y] = \mathcal{D}[f_1 : f_2; X] + \mathcal{D}[f_1 : f_2; Y] \quad .$$

Demonstração.

$$\begin{aligned}\mathcal{D}[f_1 : f_2; X, Y] &= \int f_1(x,y)\ \log \frac{f_1(x,y)}{f_2(x,y)} d\lambda(x,y) \\ &= \int g_1(x)h_1(y)\ \log \frac{g_1(x)h_1(y)}{g_2(x)h_2(y)} d\mu(x)d\nu(y) \\ &= \int g_1(x)\ \log \frac{g_1(x)}{g_2(x)} d\mu(x) + \int h_1(y)\ \log \frac{h_1(y)}{h_2(y)} d\nu(y) \\ &= \mathcal{D}[f_1 : f_2; X] + \mathcal{D}[f_1 : f_2; Y] \quad ,\end{aligned} \tag{3.23}$$

sendo, por causa da independência, $f_i(x,y) = g_i(x)h_i(y), i = 1,2$, $d\lambda(x,y) = d\mu(x)d\nu(y)$, $\int g_i\mu_i(x) = 1, \int g_i(y)d\nu(y) = 1, i = 1,2$. □

Aditividade da divergência para eventos independentes é intuitivamente uma propriedade fundamental e de fato é postulado como uma propriedade necessária em muitos desenvolvimentos axiomáticos da teoria da informação. Veja, por exemplo, Barnard (1949), Fisher et al. (1950), Good (1950) e Shannon (1948). Da mesma forma como apresentamos em (3.6), a divergência de Kullback-Leibler satisfaz essa condição. Uma amostra de n observações independentes da mesma população fornece n vezes a informação média em uma única observação.

Se X e Y não são independentes, uma propriedade aditiva ainda existe, porém em termos de informação condicional como definimos abaixo. Para simplificar o argumento e evitar problemas da teoria da medida para probabilidades condicionais, devemos lidar com funções de densidade de probabilidade e a medida de Lebesgue.

Teorema 3.4.

$$\begin{aligned}\mathcal{D}[f_1 : f_2; X, Y] &= \mathcal{D}[f_1 : f_2; X] + \mathcal{D}[f_1 : f_2; Y|X] \\ &= \mathcal{D}[f_1 : f_2; Y] + \mathcal{D}[f_1 : f_2; X|Y] \quad .\end{aligned}$$

Demonstração.

$$\begin{aligned}\mathcal{D}[f_1 : f_2; X, Y] &= \int f_1(x,y)\ \log \frac{f_1(x,y)}{f_2(x,y)} dxdy \\ &= \int g_1(x) \log \frac{g_1(x)}{g_2(x)}\, dx + \int g_1(x) \left[\int h_1(y|x) \log \frac{h_1(y|x)}{h_2(y|x)} dy \right] dx \quad ,\end{aligned}$$

sendo $g_i(x) = \int f_i(x,y)dy$, $h_i(y|x) = f_i(x,y)/g_i(x)$, $i = 1,2$.

Agora definimos

$$\mathcal{D}[f_1 : f_2; Y|X = x] = \int h_1(y|x) \log \frac{h_1(y|x)}{h_2(y|x)} dy \quad ,$$

$$\mathcal{D}[f_1 : f_2; Y|X] = E_{g_1}[\mathcal{D}[f_1 : f_2; Y|X = x]] = \int g_1(x)\mathcal{D}[f_1 : f_2; Y|X = x]dx \quad ,$$

sendo $\mathcal{D}[f_1 : f_2; Y|X = x]$ definido como a divergência condicional em Y em favor de H_1 em oposição a H_2 quando $X = x$, sob H_1, e $\mathcal{D}[f_1 : f_2; Y|X]$ é o valor médio da informação condicional de Kullback-Leibler sob H_1. Nós obtemos o mesmo resultado trocando o procedimento com respeito a X e Y. □

Teorema 3.5. *$\mathcal{D}[f_1 : f_2]$ é quase positiva definida; isto é, $\mathcal{D}[f_1 : f_2] \geq 0$ com igualdade se e somente se $f_1(x) = f_2(x)[\lambda]$.*

Demonstração. Seja $g(x) = f_1(x)/f_2(x)$. Então

$$\begin{aligned}\mathcal{D}[f_1 : f_2] &= \int f_2(x)g(x)\log g(x)d\lambda(x) \\ &= \int g(x)\log g(x)d\mu_2(x) \quad .\end{aligned}$$

com $d\mu_2(x) = f_2(x)d\lambda(x)$.

Tomando $\phi(t) = t\log t$, como $0 < g(x) < \infty[\lambda]$, podemos escrever

$$\phi(g(x)) = \phi(1) + [g(x) - 1]\phi'(1) + \frac{1}{2}[g(x) - 1]^2\phi''(h(x))[\lambda] \quad ,$$

sendo que $h(x)$ está entre $g(x)$ e 1, então $0 < h(x) < \infty[\lambda]$. Já que $\phi(1) = 0$, $\phi'(1) = 1$, e

$$\int g(x)d\mu_2(x) = \int f_1(x)d\lambda(x) = 1 \quad , \tag{3.24}$$

encontramos

$$\int \phi(g(x))d\mu_2(x) = \frac{1}{2}\int [g(x) - 1]^2\phi''(h(x))d\mu_2(x) \quad ,$$

onde $\phi''(t) = 1/t > 0$ para $t > 0$. Nós vemos por (3.24) que

$$\int g(x)\log g(x)d\mu_2(x) = \int f_1(x)\log\frac{f_1(x)}{f_2(x)}d\lambda(x) \geq 0 \quad , \tag{3.25}$$

nós temos igualdade se e somente se $g(x) = f_1(x)/f_2(x) = 1[\lambda]$. □

A desigualdade em (3.25) nos diz que na média a informação obtida de observações estatísticas é positiva. Não há diferença de informação se a distribuição das observações são as mesmas sob as duas hipóteses módulo λ.

3.3.2.1 Divergência de Kullback-Leibler e Estimadores de Máxima Verossimilhança

Existem várias formas de se mensurar a proximidade entre uma aproximação paramétrica $f_2(\boldsymbol{x}|\boldsymbol{\theta})$ e f_1, porém a divergência de Kullback-Leibler está relacionada com o método da máxima verossimilhança (CLAESKENS; HJORT et al., 2008). Decompondo a divergência de Kullback-Leibler em termos de f_1 e de $f_2(\boldsymbol{x}|\boldsymbol{\theta})$ temos

$$\mathcal{D}[f_1 : f_2(\boldsymbol{x}|\boldsymbol{\theta})] = E_{f_1}[\log\ f_1(\boldsymbol{X})] - E_{f_1}[\log\ f_2(\boldsymbol{X}|\boldsymbol{\theta})] \quad . \tag{3.26}$$

Aplicando a lei dos grandes números em

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\ f_2(x_i|\boldsymbol{\theta}) \quad ,$$

para cada valor do vetor de parâmetros $\boldsymbol{\theta}$, temos que

$$n^{-1} l(\boldsymbol{\theta}) \to E_{f_1}[\log\ f_2(\boldsymbol{X}|\boldsymbol{\theta})] \quad ,$$

a convergência é quase certa, isto é, com probabilidade 1. O estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}}$ que maximiza $l(\boldsymbol{\theta})$ irá convergir quase certamente, sob algumas condições, para o minimizador $\boldsymbol{\theta_0}$ de 3.26 de f_1 para f_2,

$$\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta_0} = \inf_{\boldsymbol{\theta} \in \Theta} \mathcal{D}[f_1 : f_2(\boldsymbol{x}|\boldsymbol{\theta})] \quad . \tag{3.27}$$

4 O CRITÉRIO DE INFORMAÇÃO DE AKAIKE (AIC)

Neste capítulo tratamos do problema de estimar a dimensão de um modelo probabilístico baseado na informação de Kullback-Leibler. Akaike (1973) construiu um estimador da esperança da informação de Kullback-Leibler, isto é, da divergência de Kullback-Leibler, baseado na função de log-verossimilhança em seu ponto de máximo. Essa medida é conhecida como critério de informação de Akaike (AIC).

Na Seção 1 fazemos uma derivação do AIC. Primeiro fazemos uma discussão conceitual da metodologia do AIC. Em seguida a demonstração matemática. Na Seção 2 apresentamos uma aplicação do AIC para a seleção da ordem de defasagem de processos autoregressivos.

4.1 Uma Derivação Geral do AIC

No capítulo 3 calculamos a divergência de Kullback-Leibler em (3.2) e (3.3) mas em ambos os casos tínhamos conhecimento prévio da estrutura dos dados envolvidos. Em geral, lidamos como problemas mais complexos em que existe incerteza tanto na escolha da distribuição probabilística quanto em seus parâmetros. Nessa seção estamos interessados em calcular um estimador da divergência de Kullback-Leibler.

A demonstração dessa seção é baseada em Anderson e Burnham (2004). Outras demonstrações podem ser encontradas em Akaike (1973), Bozdogan (1987), Chow et al. (1979), Sawa (1978), Shibata (1989), Sugiura (1978) e Stone (1982).

4.1.1 Derivação Conceitual do AIC

Para uma derivação conceitual geral do AIC a partir da divergência de Kullback-Leibler para o melhor modelo da classe de modelos $f(\boldsymbol{x}|\boldsymbol{\theta})$, começamos com

$$\mathcal{D}[g : f(\cdot|\boldsymbol{\theta_0})] = \int g(\boldsymbol{x}) \, \log \frac{g(\boldsymbol{x})}{f(\boldsymbol{x}|\boldsymbol{\theta})} \, d\boldsymbol{x} \quad . \tag{4.1}$$

Note que embora não conheçamos $\boldsymbol{\theta}$ para o modelo, o valor da divergência de Kullback-Leibler para a classe de modelos é tomado como $\mathcal{D}[g : f]$ avaliado em $\boldsymbol{\theta_0}$, pois o valor do parâmetro estará estimando $\boldsymbol{\theta_0}$. Também, note a notação expandida em (4.1), então podemos representar $\mathcal{D}[g : f]$ como dependente, em geral, do valor do parâmetro desconhecido, dado a estrutura do modelo.

Dado que temos $Y_1, \ldots, Y_n$ como amostra de $g(\cdot)$, o passo lógico seria encontrar o estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{y})$ e calcular uma estimativa de $\mathcal{D}[g : f(\cdot|\boldsymbol{\theta_0})]$ como

$$\mathcal{D}[g : f(\cdot|\hat{\boldsymbol{\theta}}(\boldsymbol{y}))] = \int g(\boldsymbol{x}) \, \log \frac{g(\boldsymbol{x})}{f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}(\boldsymbol{y}))} \, d\boldsymbol{x} \quad .$$

Se pudéssemos encontrar $\boldsymbol{\theta_0}$ que minimiza a divergência de Kullback-Leibler, saberíamos que o nosso objetivo seria $\mathcal{D}[g : f] = 0$, isto é, não haveria perda de informação quando assumimos a distribuição f para aproximar a distribuição g. Então poderíamos avaliar a qualidade do modelo com base no valor absoluto zero. Porém, na análise de dados, os parâmetros do modelo devem ser estimados, e geralmente, há incerteza substancial nesta estimativa.

Modelos baseados em parâmetros estimados, consequentemente $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ não $\boldsymbol{\theta}$, são diferentes de modelos com parâmetros conhecidos. Qualquer valor de $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ que não seja $\boldsymbol{\theta_0}$ resulta em $\mathcal{D}[g : f(\cdot|\hat{\boldsymbol{\theta}}(\boldsymbol{y}))] > \mathcal{D}[g : f(\cdot|\boldsymbol{\theta_0})]$. Assim, mesmo que tivéssemos a estrutura correta do modelo, porque precisamos estimar $\boldsymbol{\theta}$, devemos pensar em termos da divergência de Kullback-Leibler como tendo, em média, um valor > 0. Nesse sentido, buscamos o modelo não para minimizar $\mathcal{D}[g : f(\cdot|\boldsymbol{\theta_0})]$, mas o valor ligeiramente maior, em média, dado por

$$E_{\boldsymbol{Y}}[\mathcal{D}[g : f(\cdot, \hat{\boldsymbol{\theta}}(\boldsymbol{Y}))]] > \mathcal{D}[g : f(\cdot|\boldsymbol{\theta_0})] \quad . \tag{4.2}$$

(Todas as esperanças são tomadas com respeito a g, independente da notação para variável aleatória envolvida.) Então dado que devemos estimar $\boldsymbol{\theta}$, adotaremos o critério de selecionar o modelo f que minimiza $E_{\boldsymbol{Y}}[\mathcal{D}[g : f(\cdot, \hat{\boldsymbol{\theta}}(\boldsymbol{y})]]$. Portanto, nosso objetivo deve ser minimizar o valor esperado da divergência de Kullback-Leibler estimada.

Reescrevendo o valor esperado da divergência de Kullback-Leibler estimada, temos

$$\begin{aligned} E_{\boldsymbol{Y}}[\mathcal{D}[g : f(\cdot, \hat{\boldsymbol{\theta}}(\boldsymbol{Y})]] &= \int g(\boldsymbol{x}) \log g(\boldsymbol{x})\, d\boldsymbol{x} - E_{\boldsymbol{Y}}\left[\int g(\boldsymbol{x}) \log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}(\boldsymbol{y}))\, d\boldsymbol{x}\right] \\ E_{\boldsymbol{Y}}[\mathcal{D}[g : f(\cdot, \hat{\boldsymbol{\theta}}(\boldsymbol{Y}))]] &= \int g(\boldsymbol{x}) \log g(\boldsymbol{x})\, d\boldsymbol{x} - E_{\boldsymbol{Y}} E_{\boldsymbol{X}}[\log f(\boldsymbol{X}|\hat{\boldsymbol{\theta}}(\boldsymbol{Y}))] \quad . \end{aligned} \tag{4.3}$$

Nós podemos estimar $E_{\boldsymbol{Y}} E_{\boldsymbol{X}}[\log f(\boldsymbol{X}|, \hat{\boldsymbol{\theta}}(\boldsymbol{Y}))]$, e portanto, selecionar o modelo que minimiza (4.3). De forma simples, podemos dizer que estamos selecionando o melhor modelo Kullback-Leibler estimado usando AIC.

4.1.2 Derivação Matemática do AIC

Agora apresentamos uma derivação matemática do AIC a partir da divergência de Kullback-Leibler. A abordagem mais geral para derivar o AIC usa a expansão da série de Taylor para segunda ordem. Para detalhes sobre expansão da série de Taylor, o leitor pode consultar Tao (2009) e Rudin et al. (1964).

$$\mathcal{D}[g : f(\cdot|\boldsymbol{\theta_0})] = E_{\boldsymbol{X}}[\log g(\boldsymbol{X})] - E_{\boldsymbol{X}}[\log f(\boldsymbol{X}|\boldsymbol{\theta_0})] \quad .$$

Nosso objetivo não deve ser selecionar um modelo com base no ínfimo da divergência de Kullback-Leibler com $\boldsymbol{\theta_0}$ conhecido dado f, mas selecionar o modelo com $\boldsymbol{\theta}$ baseado em minimizar uma esperança da divergência de Kullback-Leibler. Denotaremos por Q como sendo

o nosso critério para seleção de modelos. Então devemos estimar, sem viés, para cada modelo, o valor

$$Q = \int g(\boldsymbol{y}) \left[\int g(\boldsymbol{x}) \log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}(\boldsymbol{y}))d\boldsymbol{x} \right] d\boldsymbol{y} \quad .$$

O problema de seleção de modelos baseado na divergência de Kullback-Leibler é encontrar uma expressão útil e um estimador de

$$Q = E_{\hat{\boldsymbol{\theta}}} E_{\boldsymbol{X}}[\log f(\boldsymbol{X}|\hat{\boldsymbol{\theta}})] \quad , \tag{4.4}$$

sendo entendido que o estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}}$ é baseado na amostra $Y_1, \ldots, Y_n$ e as duas esperanças são com respeito a distribuição g. Note que Q é uma dupla esperança baseada, conceitualmente, em duas amostras independentes.

Proposição 4.1. *Sejam $\hat{\boldsymbol{\theta}}$ estimador de máxima verosimilhança de $\boldsymbol{\theta}$ e $\Sigma = E_{\hat{\boldsymbol{\theta}}}\left[[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}][\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}]^t\right]$ matriz de covariância assintótica de $\boldsymbol{\theta}$, então*

$$Q = E_{\boldsymbol{X}}[\log f(\boldsymbol{X}|\boldsymbol{\theta_0})] - \frac{1}{2} tr\left[J(\boldsymbol{\theta_0})\,\Sigma\right] + O_p(n^{-1}) \quad .$$

Demonstração. Para encontrarmos uma expressão para Q, o primeiro passo é aplicar a expansão da série de Taylor em $\log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}})$ em torno de $\boldsymbol{\theta_0}$ para qualquer $\boldsymbol{x}$, então

$$\begin{aligned} \log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) = \log f(\boldsymbol{x}|\boldsymbol{\theta_0}) + u(\boldsymbol{x}|\boldsymbol{\theta})^t[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}] \\ + \frac{1}{2}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}]^t \mathrm{I}(\boldsymbol{x}|\boldsymbol{\theta_0})[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}] + O_p(n^{-1}) \quad . \end{aligned} \tag{4.5}$$

Com o objetivo de relacionar (4.4) com (4.5), tomamos o valor esperado com respeito a $X_1, \ldots, X_n$.

$$\begin{aligned} E_{\boldsymbol{X}}[\log f(\boldsymbol{X}|\hat{\boldsymbol{\theta}})] = E_{\boldsymbol{X}}[u(\boldsymbol{X}|\boldsymbol{\theta_0})]^t[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}] \\ + \frac{1}{2}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}]^t [E_{\boldsymbol{X}} \mathrm{I}(\boldsymbol{X}|\boldsymbol{\theta_0})][\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}] + O_p(n^{-1}) \quad . \end{aligned} \tag{4.6}$$

O termo linear em (4.6) é zero. Também, usando (2.16) no termo quadrático de (4.6), temos

$$E_{\boldsymbol{X}}[\log f(\boldsymbol{X}|\hat{\boldsymbol{\theta}})] = E_{\boldsymbol{X}}[\log f(\boldsymbol{X}|\boldsymbol{\theta_0})] - \frac{1}{2}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}]^t J(\boldsymbol{\theta_0})[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}] + O_p(n^{-1}) \quad . \tag{4.7}$$

Tomando a esperança em (4.7) com respeito a $\hat{\boldsymbol{\theta}}$ e usando a função traço, temos

$$E_{\hat{\boldsymbol{\theta}}} E_{\boldsymbol{X}}[\log f(\boldsymbol{X}|\hat{\boldsymbol{\theta}})] = E_{\boldsymbol{X}}[\log f(\boldsymbol{X}|\boldsymbol{\theta_0})] - \frac{1}{2} tr\left[J(\boldsymbol{\theta_0})\left[E_{\hat{\boldsymbol{\theta}}}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}][\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}]^t\right]\right] + O_p(n^{-1}) \quad . \tag{4.8}$$

O lado esquerdo de (4.8) é (4.4) e $E_{\hat{\boldsymbol{\theta}}}\left[[\hat{\boldsymbol{\theta}}-\boldsymbol{\theta_0}][\hat{\boldsymbol{\theta}}-\boldsymbol{\theta_0}]^t\right]=\Sigma$ é a matriz de covariância assintótica de $\boldsymbol{\theta}$, porque a esperança está sendo tomada com respeito a distribuição g. Então nós temos

$$Q = E_{\boldsymbol{X}}[\log\ f(\boldsymbol{X}|\boldsymbol{\theta_0})] - \frac{1}{2}tr\left[J(\boldsymbol{\theta_0})\,\Sigma\right] + O_p(n^{-1}) \quad . \tag{4.9}$$

□

O próximo passo requer um resultado que ainda não derivamos: uma relação entre Q e $E_{\boldsymbol{X}}[\log\ f(\boldsymbol{X}|\hat{\boldsymbol{\theta}}(\boldsymbol{x})]$, que é a esperança da função de log-verossimilhança avaliada em $\hat{\boldsymbol{\theta}}$.

Proposição 4.2. *Sejam $\hat{\boldsymbol{\theta}}$ estimador de máxima verossimilhança de $\boldsymbol{\theta}$ e $\Sigma = E_{\hat{\boldsymbol{\theta}}}\left[[\hat{\boldsymbol{\theta}}-\boldsymbol{\theta_0}][\hat{\boldsymbol{\theta}}-\boldsymbol{\theta_0}]^t\right]$ matriz de covariância assintótica de $\boldsymbol{\theta}$, então*

$$E_{\boldsymbol{X}}[\log\ f(\boldsymbol{X}|\boldsymbol{\theta_0})] = E_{\boldsymbol{X}}[\log\ f(\boldsymbol{X}|\hat{\boldsymbol{\theta}}(\boldsymbol{x}))] - \frac{1}{2}tr\left[[J(\boldsymbol{\theta_0})]\Sigma]\right] + O_p(n^{-1}) \quad .$$

Demonstração. Fazemos uma segunda expansão, dessa vez em $\log\ f(\boldsymbol{x}|\boldsymbol{\theta_0})$ em torno de $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$, tratando $X_1,\ldots,X_n$ como amostra, consequentemente temos que o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ para $X_1,\ldots,X_n$. Aplicando a aproximação da série de Taylor em torno de $\hat{\boldsymbol{\theta}}$, temos

$$\begin{aligned}\log\ f(\boldsymbol{x}|\boldsymbol{\theta_0}) = \log\ f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) + u(\boldsymbol{x}|\hat{\boldsymbol{\theta}})^t[\boldsymbol{\theta_0}-\hat{\boldsymbol{\theta}}] \\ + \frac{1}{2}[\boldsymbol{\theta_0}-\hat{\boldsymbol{\theta}}]^t\mathrm{I}(\boldsymbol{x}|\hat{\boldsymbol{\theta}})[\boldsymbol{\theta_0}-\hat{\boldsymbol{\theta}}] + O_p(n^{-1}) \quad .\end{aligned} \tag{4.10}$$

Como o estimador de máxima verossimilhança é solução da equação de verossimilhança, tomando a esperança em (4.10) e usando as propriedades da função traço, temos

$$E_{\boldsymbol{X}}[\log\ f(\boldsymbol{X}|\boldsymbol{\theta_0})] = E_{\boldsymbol{X}}[\log\ f(\boldsymbol{X}|\hat{\boldsymbol{\theta}})] - \frac{1}{2}tr\left[E_{\boldsymbol{X}}\left[\hat{J}(\hat{\boldsymbol{\theta}})\right][\boldsymbol{\theta_0}-\hat{\boldsymbol{\theta}}][\boldsymbol{\theta_0}-\hat{\boldsymbol{\theta}}]^t\right] + O_p(n^{-1}) \quad , \tag{4.11}$$

sendo que $\hat{J}(\hat{\boldsymbol{\theta}})$ denota a matriz Hessiana da função de log verosimilhança avaliada no estimador de máxima verossimilhança.

Para progredirmos com (4.11), usaremos a aproximação $\hat{J}(\hat{\boldsymbol{\theta}}) = J(\boldsymbol{\theta_0}) + O_p(n^{-1})$. É óbvio que $E[\hat{J}(\boldsymbol{\theta_0})] = J(\boldsymbol{\theta_0})$ sob a distribuição g. Quando $X_1,\ldots,X_n$ é amostra de $g(\cdot)$, $\hat{J}(\boldsymbol{\theta_0})$ converge para $J(\boldsymbol{\theta_0})$ quando $n\to\infty$. Assim, um estimador para $J(\boldsymbol{\theta_0})$ é $\hat{J}(\hat{\boldsymbol{\theta}})$. Pelo fato de $\hat{\boldsymbol{\theta}}$ ser o estimador de máxima verossimilhança sob o modelo $f(\boldsymbol{x}|\boldsymbol{\theta})$, $\hat{\boldsymbol{\theta}}$ converge para $\boldsymbol{\theta_0}$ quando $n\to\infty$, e consequentemente $\hat{J}(\hat{\boldsymbol{\theta}})$ converge para $J(\boldsymbol{\theta_0})$.

Usando $\hat{J}(\hat{\boldsymbol{\theta}}) = J(\boldsymbol{\theta_0}) + O_p(n^{-1})$, obtemos

$$\begin{aligned}E_{\boldsymbol{X}}\left[\hat{J}(\hat{\boldsymbol{\theta}})\right][\boldsymbol{\theta_0}-\hat{\boldsymbol{\theta}}][\boldsymbol{\theta_0}-\hat{\boldsymbol{\theta}}]^t &= [J(\boldsymbol{\theta_0})]\left[E_{\boldsymbol{X}}[\boldsymbol{\theta_0}-\hat{\boldsymbol{\theta}}][\boldsymbol{\theta_0}-\hat{\boldsymbol{\theta}}]^t\right] + O_p(n^{-1}) \\ &= [J(\boldsymbol{\theta_0})]\left[E_{\boldsymbol{X}}[\hat{\boldsymbol{\theta}}-\boldsymbol{\theta_0}][\hat{\boldsymbol{\theta}}-\boldsymbol{\theta_0}]^t\right] + O_p(n^{-1}) \\ &= [J(\boldsymbol{\theta_0})]\Sigma + O_p(n^{-1}) \quad .\end{aligned} \tag{4.12}$$

Usando (4.12) em (4.11), obtemos

$$E_{\boldsymbol{X}}[\log\ f(\boldsymbol{X}|\boldsymbol{\theta_0})] = E_{\boldsymbol{X}}[\log\ f(\boldsymbol{X}|\hat{\boldsymbol{\theta}}(\boldsymbol{x}))] - \frac{1}{2}tr\,[[J(\boldsymbol{\theta_0})]\Sigma]] + O_p(n^{-1}) \quad .$$

□

Proposição 4.3. *Seja $\hat{\boldsymbol{\theta}}$ estimador de máxima verossimilhança de $\boldsymbol{\theta}$, então*

$$Q = E_{\boldsymbol{X}}[\log\ f(\boldsymbol{X}|\hat{\boldsymbol{\theta}}(\boldsymbol{x}))] - tr[[J(\boldsymbol{\theta_0})]\Sigma] + O_p(n^{-1}) \quad .$$

Demonstração. Usando (4.12) em (4.11), obtemos

$$E_{\boldsymbol{X}}[\log\ f(\boldsymbol{X}|\boldsymbol{\theta_0})] = E_{\boldsymbol{X}}[\log\ f(\boldsymbol{X}|\hat{\boldsymbol{\theta}}(\boldsymbol{x}))] - \frac{1}{2}tr\,[[J(\boldsymbol{\theta_0})]\Sigma]] + O_p(n^{-1}) \quad . \tag{4.13}$$

Considerando (4.9)

$$Q = E_{\boldsymbol{X}}[\log\ f(\boldsymbol{X}|\boldsymbol{\theta_0})] - \frac{1}{2}tr\,[J(\boldsymbol{\theta_0})\,\Sigma] + O_p(n^{-1}) \quad .$$

Substituindo (4.13) em (4.9), temos um resultado chave conhecido na literatura

$$Q = E_{\boldsymbol{X}}[\log\ f(\boldsymbol{X}|\hat{\boldsymbol{\theta}}(\boldsymbol{x}))] - tr[[J(\boldsymbol{\theta_0})]\Sigma] + O_p(n^{-1}) \quad . \tag{4.14}$$

□

Proposição 4.4. *O critério de informaçao de Akaike (AIC) é definido como*

$$AIC = -2\ \log\ f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) + 2p \quad .$$

Demonstração. A literatura geralmente não apresenta (4.14), mas uma forma alternativa equivalente

$$Q = E_{\boldsymbol{X}}[\log\ f(\boldsymbol{X}|\hat{\boldsymbol{\theta}}(\boldsymbol{x}))] - tr[K(\boldsymbol{\theta_0})[J(\boldsymbol{\theta_0})^{-1}]] + O_p(n^{-1}) \quad . \tag{4.15}$$

A notação $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ ao invés de simplesmente $\boldsymbol{\theta}$ é usada em (4.15) para enfatizar que no lado direito de (4.15) somente o vetor aleatório $X_1, \ldots, X_n$ está envolvido, e pode ser tomada para referir aos dados disponíveis. De (4.14) ou (4.15), podemos inferir que um critério para selecionar modelos é estruturalmente da forma

$$\hat{Q} = \log\ f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) - \hat{tr}[[J(\boldsymbol{\theta_0})]\Sigma] + O_p(n^{-1}) \quad ,$$

ou

$$\hat{Q} = \log\ f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) - \hat{tr}[K(\boldsymbol{\theta_0})[J(\boldsymbol{\theta_0})^{-1}]] + O_p(n^{-1}) \quad .$$

A função de log-verossimilhança $\log\ f(\boldsymbol{x}|\hat{\boldsymbol{\theta}})$ em (4.15) é um estimador não viesado de $E_{\boldsymbol{X}}[\log\ f(\boldsymbol{X}|\hat{\boldsymbol{\theta}}(\boldsymbol{x}))]$, porém é viesado como estimador de Q. Consequentemene, o próximo passo é termos um estimador não viesado (ou com pouco viés) do termo envolvendo o traço, ou

pelo menos um estimador com pequeno erro quadrático médio. Então o melhor modelo será aquele com o maior valor de $\hat{Q}$, porque irá produzir um modelo com a menor divergência de Kullback-Leibler estimada esperada. Por uma questão de convenção, o critério é apresentado como

$$-2 \log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) + 2\hat{tr}\left[K(\boldsymbol{\theta_0})[J(\boldsymbol{\theta_0})^{-1}]\right] \quad .$$

Se g é um subconjunto de f, isto é, $f = g$ ou se g está contido em f, então $\mathcal{I}(\boldsymbol{\theta_0}) \equiv J(\boldsymbol{\theta_0}) = K(\boldsymbol{\theta_0}) = \Sigma^{-1}$, e consequentemente $tr[[J(\boldsymbol{\theta_0})]\Sigma] = p$. Mesmo que f não seja uma boa aproximação para g, a literatura indica que o melhor estimador é usar $\hat{tr}[[J(\boldsymbol{\theta_0})]\Sigma] = p$, para mais informações, veja Shibata (1989). Assim, temos que o AIC pode ser escrito como

$$AIC = -2 \log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) + 2p \quad . \tag{4.16}$$

□

4.2 Seleção da Ordem de Defasagem de Processos Autoregressivos

Sabemos que se temos uma série temporal estacionária, isto é, uma série temporal para a qual as propriedades estatísticas como a média, autocorrelação e variância não dependem do tempo, um processo autoregressivo AR é geralmente utilizado para modelar os dados. Assim, escrevemos

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t \quad ,$$

onde Z_t é um ruído branco. Assumiremos que o processo AR(p) é estacionário, isto é, satisfaz a condição (2.23). O valor da ordem do processo autoregressivo indica o quanto as observações do passado influenciam no valor atual de X_t. Se p é pequeno, apenas observações do passado recente influenciam no presente. Por outro lado, se p é grande, efeitos de longo prazo no passado ainda influenciam na observação presente (CLAESKENS; HJORT et al., 2008). Conhecer a ordem do processo autoregressivo é especialmente importante para fazer previsões, isto é, prever os valores de $X_{t+1}, X_{t+2}, \ldots$ quando observamos a série até o tempo t. O AIC pode ser usado para selecionar uma ordem p apropriada. Para um número $p = 1, 2, \ldots$ de candidatos a ordem, construímos para cada p o AIC(p) tomando duas vezes o valor da log-verossimilhança no ponto de máximo para aquele modelo p, penalizando em duas vezes o número de parâmetros estimados, o qual é $p + 1$ (adicionando um para o desvio padrão σ estimado). Assim, para cada p, o AIC computa o valor

$$AIC(p) = -2 \log \hat{\sigma}_p - 2(p + 1) \quad ,$$

onde $\hat{\sigma}$ é o estimador de máxima verossimilhança do desvio padrão no modelo com ordem p. O maior valor do AIC(p) calculado corresponde ao melhor modelo. O melhor, de acordo com o método do AIC, corresponde a aquele que tem a menor esperança da divergência de Kullback-Leibler estimada em relação ao verdadeiro processo gerador dos dados.

5 CONSIDERAÇÕES FINAIS

Neste trabalho apresentamos uma proposta de sistematização e organização de conceitos da teoria da informação visando a sua extensão para o problema de seleção de modelos probabilísticos com a construção e aplicação do critério de informação de Akaike (AIC). Começamos com conceitos que acreditamos serem necessário para o desenvolvimento do trabalho. Abordamos alguns resultados da teoria da medida e variáveis aleatórias, o Teorema de Bayes, o método da verossimilhança e suas propriedades assintóticas e noções de séries temporais.

Com a base teórica previamente colocada, estabelecemos uma relação entre os conceitos de chances a favor e razão de verossimilhança e mostramos que conduzem a formação das variáveis aleatórias informação de Shannon e informação de Kullback-Leibler. A informação de Shannon se origina da comparação entre a razão das chances finais e iniciais quando consideramos apenas um evento. Já a informação de Kullback-Leibler da comparação da razão de chances finais e iniciais de dois eventos alternativos em virtude de um evento condicionante. Mostramos que a esperança da informação de Shannon é conhecida como a entropia de Shannon enquanto a esperança da informação de Kullback-Leibler é a divergência de Kullback-Leibler sendo construída a partir do princípio da probabilidade inversa e da aplicação da derivada de Radon-Nikodym.

Discutimos o uso da divergência Kullback-Leibler como uma medida da "distância" entre distribuições de probabilidade de modo a oferecer um procedimento de inferência para a seleção de modelos probabilísticos. Demonstramos que o AIC é uma metodologia que calcula uma estimativa da esperança da divergência Kullback-Leibler para uma classe de modelos e seleciona aquele que produz a menor perda de informação de Kullback-Leibler. Por fim, fizemos uma aplicação do AIC em séries temporais de modo a possibilitar uma ferramenta de seleção da ordem de defasagem de processos autoregrresivos.

6 REFERÊNCIAS

AKAIKE, H. Information theory and an extension of the maximum likelihood principle,[w:] proceedings of the 2nd international symposium on information, bn petrow, f. *Czaki, Akademiai Kiado, Budapest*, p. 267—-281, 1973. Citado na página 39.

ANDERSON, D.; BURNHAM, K. *Model selection and multi-model inference*. [S.l.]: pringer, 2004. Citado na página 39.

BARNARD, G. A. Statistical inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, v. 11, n. 2, p. 115–149, 1949. Citado na página 36.

BOROVKOV, A. *Mathematical Statistics*. [S.l.]: Gordon and Breach Science publish, 1987. Citado na página 15.

BOZDOGAN, H. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, Springer, v. 52, n. 3, p. 345–370, 1987. Citado 2 vezes nas páginas 9 e 39.

BROCKWELL, P. J.; DAVIS, R. A.; FIENBERG, S. E. *Time series: theory and methods*. [S.l.]: Springer Science & Business Media, 1991. Citado 3 vezes nas páginas 10, 19 e 22.

CHOW, G. C. et al. A comparison of the information and posterior probability criteria for model selection. Princeton University, Econometric Research Program, 1979. Citado na página 39.

CLAESKENS, G.; HJORT, N. L. et al. *Model selection and model averaging*. [S.l.]: Cambridge University Press, 2008. Citado 2 vezes nas páginas 38 e 44.

COVER, T. M.; THOMAS, J. A. *Elements of information theory*. [S.l.]: John Wiley & Sons, 2012. Citado na página 29.

FISHER, R. A. Theory of statistical estimation. Cambridge University Press, v. 22, n. 5, p. 700–725, 1925. Citado na página 8.

FISHER, R. A. *Statistical methods and scientific inference.* [S.l.]: Hafner Publishing Co., 1956. Citado na página 8.

FISHER, R. A. et al. *Statistical methods for research workers.* [S.l.]: Oliver and Boyd, Edinburgh, 1950. Citado 2 vezes nas páginas 13 e 36.

FULLER, W. A. *Introduction to statistical time series*. [S.l.]: John Wiley & Sons, 2009. v. 428. Citado 3 vezes nas páginas 15, 19 e 22.

GOOD, I. *Probability and the weighing of evidence. Charles Griffin*. [S.l.]: London/Hafner Press, New York, 1950. Citado 6 vezes nas páginas 10, 23, 24, 25, 26 e 36.

GOOD, I. Some terminology and notation in information theory. *Proceedings of the IEE-Part C: Monographs*, IET, v. 103, n. 3, p. 200–204, 1956. Citado na página 27.

GOOD, I. J.; OSTEYEE, D. *Information, weight of evidence. The singularity between probability measures and signal detection*. [S.l.]: Springer, 1974. v. 376. Citado 2 vezes nas páginas 8 e 24.

HALMOS, P. R. *Measure theory*. [S.l.]: Springer-Verlag NewYork· Heidelberg· Berlin, 1974. Citado 2 vezes nas páginas 10 e 11.

ISNARD, C. *Introdução à Medida e Integração*. [S.l.]: Impa, 2016. Citado na página 11.

JEFFREYS, H. *The theory of probability*. [S.l.]: OUP Oxford, 1948. Citado na página 23.

KENDALL, M.; STUART, A. *The advanced theory of statistics: Inference and relationship, vol. 2*. [S.l.]: Griffin, London, 1961. Citado na página 15.

KULLBACK, S. *Information theory and statistics*. [S.l.]: Dover Publications, 1958. Citado 3 vezes nas páginas 25, 32 e 35.

KULLBACK, S.; LEIBLER, R. A. On information and sufficiency. *The annals of mathematical statistics*, JSTOR, v. 22, n. 1, p. 79–86, 1951. Citado 3 vezes nas páginas 8, 24 e 25.

LEHMANN, E. L.; CASELLA, G. *Theory of point estimation*. [S.l.]: Springer Science & Business Media, 1998. Citado 2 vezes nas páginas 10 e 15.

REZA, F. M. *An introduction to information theory*. [S.l.]: Courier Corporation, 1994. Citado na página 29.

RUDIN, W. et al. *Principles of mathematical analysis*. [S.l.]: McGraw-hill New York, 1964. v. 3. Citado na página 40.

SAWA, T. Information criteria for discriminating among alternative regression models. *Econometrica: Journal of the Econometric Society*, JSTOR, p. 1273–1291, 1978. Citado na página 39.

SHANNON, C. E. A mathematical theory of communication. *Bell system technical journal*, Wiley Online Library, v. 27, n. 3, p. 379–423, 1948. Citado 6 vezes nas páginas 8, 24, 26, 28, 29 e 36.

SHANNON, C. E.; WEAVER, W. *The Mathematical Theory of Communication*. [S.l.]: The University of Illinois Press, 1949. Citado 2 vezes nas páginas 8 e 24.

SHIBATA, R. *Statistical aspects of model selection*. [S.l.]: Springer, 1989. Citado 2 vezes nas páginas 39 e 44.

STONE, C. J. Local asymptotic admissibility of a generalization of akaike's model selection rule. *Annals of the Institute of Statistical Mathematics*, Springer, v. 34, n. 1, p. 123–133, 1982. Citado na página 39.

SUGIURA, N. Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. *Communications in Statistics-Theory and Methods*, Taylor & Francis, v. 7, n. 1, p. 13–26, 1978. Citado na página 39.

TAO, T. Analysis ii, texts and readings in mathematics, vol. 38. *Hindustan Book Agency, New Delhi*, 2009. Citado na página 40.

TAYLOR, M. E. *Measure theory and integration*. [S.l.]: American Mathematical Soc., 2006. Citado 2 vezes nas páginas 10 e 11.

VITAL, G. Trabalho de Conclusão de Curso, *Aplicações de Técnicas de Análise de Sentimentos às Atas do Comitê de Política Monetária*. 2019. Citado na página 34.

WIENER, N. *Cybernetics or Control and Communication in the Animal and the Machine*. [S.l.]: The M.I.T. press, 1948. Citado na página 8.