



**UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA DE TELEINFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TELEINFORMÁTICA**

DAVID CARNEIRO DE SOUZA

**VARIEDADES ESTATÍSTICAS GENERALIZADAS E UMA
MODIFICAÇÃO DO ESTIMADOR DE MÁXIMA VEROSSIMILHANÇA**

FORTALEZA

2016

DAVID CARNEIRO DE SOUZA

VARIÉDADES ESTATÍSTICAS GENERALIZADAS E UMA MODIFICAÇÃO DO
ESTIMADOR DE MÁXIMA VEROSSIMILHANÇA

Tese apresentada ao Programa de Pós-graduação em Engenharia de Teleinformática do Departamento de Engenharia de Teleinformática da Universidade Federal do Ceará, como parte dos requisitos necessários para a obtenção do título de Doutor em Engenharia de Teleinformática. Área de concentração: Sinais e Sistemas.

Orientador: Prof. Dr. Charles Casimiro Cavalcante

Coorientador: Prof. Dr. Rui Facundo Vigellis.

FORTALEZA

2016

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

- S237v Souza, David Carneiro de.
Variedades Estatísticas Generalizadas e uma Modificação do Estimador de Máxima Verossimilhança /
David Carneiro de Souza. – 2016.
121 f.
- Tese (doutorado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação
em Engenharia de Teleinformática, Fortaleza, 2016.
Orientação: Prof. Dr. Charles Casimiro Cavalcante.
Coorientação: Prof. Dr. Rui Facundo Vigelis.
1. Variedades estatísticas, phi-famílias, phi-divergência e estimadores.. I. Título.
- CDD 621.38
-

DAVID CARNEIRO DE SOUZA

VARIEDADES ESTATÍSTICAS GENERALIZADAS E UMA MODIFICAÇÃO DO
ESTIMADOR DE MÁXIMA VEROSSIMILHANÇA

Tese apresentada ao Programa de Pós-graduação em Engenharia de Teleinformática do Departamento de Engenharia de Teleinformática da Universidade Federal do Ceará, como parte dos requisitos necessários para a obtenção do título de doutor em Engenharia de Teleinformática. Área de concentração: Sinais e sistemas.

Aprovada em: 15/12/2016.

BANCA EXAMINADORA

Prof. Dr. Charles Casimiro Cavalcante (Orientador)
Universidade Federal do Ceará (UFC)

Prof^a Dra. Ireni Sueli Rodrigues Costa
Universidade Estadual de Campinas (UNICAMP)

Prof. Dr. Renato da Rocha Lopes
Universidade Estadual de Campinas (UNICAMP)

Prof. Dr. Rui Facundo Vigelis
Universidade Federal do Ceará (UFC)

Prof. Dr. Guilherme Alencar Barreto
Universidade Federal do Ceará (UFC)

Prof. Dr. Juvêncio Santos Nobre
Universidade Federal do Ceará (UFC)

Prof. Dr. Jorge Herbert Soares de Lira
Universidade Federal do Ceará (UFC)

Dedico esse trabalho ao meu Deus, pela convicção de sua permissão para esse momento. À minha esposa Natália e meu filho Natanael, por todo amor, força e compreensão nesse período intenso de trabalho. Aos meus pais por sempre terem investido em meus estudos e confiarem em meu potencial.

AGRADECIMENTOS

A Deus, meu Senhor e Salvador, o qual tenho convicção de sua ação misericordiosa e de infinita graça sobre minha vida.

À minha família, nas pessoas de minha esposa Natália e meu filho Natanael, por serem a minha maiores fontes de inspiração, amor, paciência e alegria em minha vida. Amo vocês para sempre...

Aos professores Dr. Charles Casimiro Cavalcante e Dr. Rui Facundo Vigelis pelas orientações como orientador e coorientador, respectivamente. Com certeza, o fato de terem feito jus às funções me permitiu terminar esse trabalho.

Aos professores Dr. Charles Casimiro Cavalcante, Dr. Guilherme Alencar Barreto, Dr. Juvêncio Santos Nobre, Dr. Jorge Herbert Soares de Lira, Dr. Rui Facundo Vigelis, Dra. Ireni Sueli Rodrigues Costa e Dr. Renato da Rocha Lopes por comporem a banca examinadora, contribuindo com suas valiosas críticas, colaborações e sugestões.

Aos amigos de doutorado que, ao longo do tempo, compartilhamos conversas contribuições, ajudas, risadas e sofrimentos.

"Quero saber como Deus criou este mundo. Não estou interessado neste ou naquele fenômeno, no espectro deste ou daquele elemento. Quero saber os Seus pensamentos. O resto são detalhes." *Albert Einstein.*

RESUMO

Esta tese é dedicada ao estudo das variedades estatísticas generalizadas, dentro do contexto de geometria, principalmente, e de inferência estatística. Nossa contribuição está na generalização e adaptação de alguns resultados dentro dos contextos citados. Com o uso de uma nova função, denominada por φ -função, que é usada inicialmente para famílias não paramétricas e que tem similaridades com a função exponencial, parametrizamos uma variedade estatística \mathcal{P} , a qual será o objeto de nossos estudos nessa tese. A partir da φ -função, definimos uma função de $\mathcal{D}_\varphi(\cdot \parallel \cdot)$ entre duas distribuições de probabilidade, chamada φ -divergência, com a qual podemos extrair uma métrica g_{ij} e recuperar um par de conexões duais $D^{(-1)}$ e $D^{(1)}$ em \mathcal{P} . Com essas conexões, definimos uma família de conexões $D^{(\alpha)}$, as quais também podem ser recuperadas a partir de uma classe de funções de divergência $\mathcal{D}_\varphi^{(\alpha)}(\cdot \parallel \cdot)$. Além disso, generalizaremos a divergência de Rényi e a de Kullback–Leibler. Além do aspecto geométrico e considerando que \mathcal{P} é uma família paramétrica, estudaremos questões sobre inferência estatística. Mais precisamente, propomos uma interpretação adaptada às φ -funções de método de estimação de máxima verossimilhança (MLE), o qual definimos por quase-verossimilhança generalizada, em menção ao MLE. Este é desenvolvido para qualquer φ -função dada, na qual associamos uma função de verossimilhança \mathcal{L}_φ . Usaremos um algoritmo desenvolvido para solucionar o problema de busca de raízes de uma equação que maximiza a função de verossimilhança, diante da restrição de nem sempre encontrar uma solução analítica que depende da complexidade da φ -função. Ademais, alguns resultados experimentais serão apresentados. E por fim, estudaremos questões sobre convergência assintótica dos estimadores. Dados esses fatos, as contribuições dessa tese em aspectos geométricos está na generalização das famílias exponenciais, o que nos possibilitará a ter mais ferramentas para estudar problemas como os de otimização, processamento de sinais e outros. Em aspectos de inferência estatística, propomos uma versão mais geral do método de estimação, onde podemos atacar problemas mais diversos.

Palavras-chave: Divergência de Rényi . φ -função. φ -divergência. φ -família. Variedades estatísticas. Geometria da informação. Estimador de máxima verossimilhança. φ -gaussiana.

ABSTRACT

This thesis is devoted to study generalized statistical manifolds, within the context of geometry mainly and statistical manifold. Our principal contribution is the generalization of some results within these contexts cited before. Using the φ -function, which is used primarily in nonparametric families and has similarity with the exponential function, we parameterize the statistical manifold \mathcal{P} , which is the focus of study here. From the φ -function, we define a function $\mathcal{D}_\varphi(\cdot \parallel \cdot)$ between two probability distributions, called φ -divergence, with which we can define a metric g_{ij} and recover a pair of dual connections $D^{(-1)}$ e $D^{(1)}$ in \mathcal{P} . With these connections, we define a family of connections $D^{(\alpha)}$, which may also be recovered from a class of divergence $\mathcal{D}_\varphi^{(\alpha)}(\cdot \parallel \cdot)$. Furthermore, we generalize the Rényi divergence and Kullback–Leibler divergence. In addition to the geometric aspects and considering that \mathcal{P} is a parametric family, we will study questions about statistics inference. More precisely, we propose a more general interpretation of the method maximum likelihood estimator (MLE), which we define by $\varphi^{(-1)}$ -likelihood, in reference to MLE. This is developed for any φ -function given which we associate a likelihood function \mathcal{L}_φ . We will use an algorithm developed to solve the search problem of an equation roots that maximizes the likelihood function, in the presence of restriction not always find an analytical solution that depends of complexity of φ -function. Furthermore, some experimental results are showed. Finally, we study assumptions about asymptotic convergence estimators. Given these facts, the contribution of the thesis in geometric aspects is the generalization of exponential families, which will enable us to have more tools to study problems such as optimization, signal processing and others. In aspects of statistical inference, we propose a more general version of estimation method where we can attack various problems.

Keywords: Rényi divergence. φ -function. φ -divergence. φ -family. Statistical manifolds. Information geometry. Maximum likelihood estimator. φ -Gaussian.

LISTA DE FIGURAS

Figura 1 – Modelo paramétrico.	17
Figura 2 – Transporte paralelo sem conexão dual.	64
Figura 3 – Transporte paralelo com conexão dual.	65
Figura 4 – Teorema de Pitágoras.	74
Figura 5 – Teorema da Projeção.	75
Figura 6 – Parametrização diferenciável.	113
Figura 7 – Função diferenciável f	114

LISTA DE GRÁFICOS

Gráfico 1 – Estimador de máxima verossimilhança generalizado para $\varphi(u) = \log(\exp(u) + 1)$	95
Gráfico 2 – Estimador de máxima verossimilhança generalizado para $\varphi(u) = \exp_{\kappa}(u)$, com $\kappa = 0.2$	95

SUMÁRIO

1	INTRODUÇÃO	13
2	GEOMETRIA DA INFORMAÇÃO	16
2.1	Modelos Paramétricos	17
2.2	Variedades Estatísticas	18
2.2.1	<i>α-Conexão de Fisher</i>	21
2.2.2	<i>Família Exponencial</i>	23
2.2.3	<i>Família Mistura</i>	26
2.2.4	<i>Conexão Métrica</i>	28
2.2.5	<i>Entropia em Variedades Estatísticas</i>	29
3	GEOMETRIA DAS VARIEDADES ESTATÍSTICAS GENERALIZADAS	35
3.1	φ -Função	35
3.2	Variedades Estatísticas Generalizadas	39
3.3	φ -Família	43
3.4	Generalização da Divergência de Rényi	45
3.5	Geometria Induzida de $\mathcal{D}_\varphi(\cdot \parallel \cdot)$	54
3.6	Dualidade	61
3.6.1	<i>Conexões Duais</i>	62
3.6.2	<i>Coordenadas Duais</i>	68
3.7	Síntese	75
4	ESTIMADOR DE MÁXIMA VEROSSIMILHANÇA MODIFICADO	77
4.1	Estimador de Máxima Verossimilhança Generalizado	77
4.1.1	<i>Preliminares de Teoria da Estimção e Análise de Convergência</i>	77
4.2	Desigualdade de Cramér–Rao	96
4.3	Síntese	99
5	CONCLUSÕES, PERSPECTIVAS e APLICAÇÕES	100
5.1	Conclusões	100
5.2	Perspectivas	100
5.3	Aplicações	102
	REFERÊNCIAS	104
6	APÊNDICE-INTRODUÇÃO À GEOMETRIA RIEMANNIANA	112
6.1	Variedades Diferenciáveis	112
6.2	Fibrado Tangente e Campos Vetoriais	117
6.3	Variedades Riemannianas	118

6.4	Conexões	119
6.5	Geodésicas	124
6.6	Curvaturas	125

1 INTRODUÇÃO

Geometria da informação é uma área de pesquisa que estuda modelos estatísticos munidos de uma estrutura geométrica. Em 1945, RAO (1992) deu início a esses estudos, ao introduzir uma métrica riemanniana, induzida pela matriz de Informação de Fisher, aplicando-a em famílias de distribuições de probabilidade FISHER (1922). Alguns problemas são costumeiramente abordados em contextos euclidianos, como tratamento de imagem, filtragem adaptativa, processamento de sinais, otimização dentre outros AMARI (2013); AMARI and KUMON (1988); ?); VAN TREES (2004). Porém, caso a geometria euclidiana não suporte a modelagem, é natural estender para um ambiente mais robusto denominado variedade estatística.

Modelos estatísticos são equipados com uma estrutura geométrica induzida de uma função de divergência. Um caso clássico é quando a divergência de Kullback–Leibler induz a métrica de Fisher e um par de conexões, denominadas conexão exponencial e conexão mistura. AMARI and NAGAOKA (2000) também considerou uma família de α -divergências que induz uma família de α -conexões. Também outros pesquisadores AMARI, OHARA, and MATSUZOE (2012); MATSUZOE and HENMI (2013) têm investigado essas estruturas geométricas em modelos estatísticos diferentes, baseados, por exemplo, em funções exponenciais generalizadas (também chamado de ϕ -exponencial). Alguns exemplos de exponencial deformada tiveram suas propriedades evidenciadas, como o caso da κ -exponencial analisada por PISTONE (2009) e a q -exponencial investigada em AMARI and OHARA (2011). Em NAUDTS (2004) as exponenciais deformadas no contexto da física estatística. Uma variedade não paramétrica, chamada de φ -família, foi proposta nos trabalhos de VIGELIS and CAVALCANTE (2013b,a), nos quais os autores generalizam as famílias exponenciais em um conjunto não paramétrico CENA and PISTONE (2007); GRASSELLI (2010). Esse último exemplo será um dos focos de nossas discussões.

Baseados nas similaridades entre a família exponencial e a φ -família, definimos uma divergência, chamada de φ -divergência, com a qual é possível mostrar que a divergência de Kullback–Leibler é um caso particular, ao mesmo tempo que induzimos a geometria da φ -família. Esses resultados são obtidos a partir de uma φ -função, função esta que possui similaridades com a função exponencial. Com a φ -função, definimos $f_\theta = \varphi^{-1}(p_\theta)$ que irá substituir $l_\theta = \log(p_\theta)$ e nos permitirá definir uma nova métrica para as variedades estatísticas generalizadas. Além do mais, mostraremos que essa métrica também é induzida a partir de uma φ -divergência.

Propomos uma generalização da divergência de Rényi VAN ERVEN and HARREMOËS (2014); HARREMOËS (2006). Com isso, podemos encontrar uma família de conexões induzida pela generalização da divergência de Rényi. Em um caso limitante,

a divergência de Rényi se reduz à φ -divergência. Em VIGELIS, SOUZA, and CAVALCANTE (2015) a φ -divergência dá origem a um par de conexões duais $D^{(1)}$ e $D^{(-1)}$, em que mostramos que a α -conexão $D^{(\alpha)}$, com $\alpha \in [-1, 1]$, é proveniente da combinação linear convexa de $D^{(-1)}$ e $D^{(1)}$, ou seja

$$D^{(\alpha)} = \frac{1 - \alpha}{2} D^{(-1)} + \frac{1 + \alpha}{2} D^{(1)},$$

e coincide com a α -conexão proveniente da generalização da divergência de Rényi.

Um outro problema que se segue de variedades estatísticas é o de estimação de parâmetros. O método de estimação de máxima verossimilhança foi desenvolvido por Fisher ALDRICH (1997) e é utilizado amplamente em várias aplicações CAMERON and TRIVEDI (1986); KAY (2013, 1993). Esse método tem sido uma importante ferramenta utilizada em problemas de processamento de sinais, comunicações, dentre outros KAY (2013, 1993). Nesse método, os parâmetros de um modelo estatístico são estimados pela observação de amostras.

A partir de uma φ -família, desenvolvemos o método de máxima verossimilhança generalizado, que para $\varphi(\cdot) = \exp(\cdot)$ coincide com o método de verossimilhança clássico. Esse método se mostrará eficiente para uma φ -família VIGELIS, SOUZA, and CAVALCANTE (2015). Estabeleceremos uma condição necessária para estimar θ de tal modo que ele maximize função de verossimilhança generalizada. Em contra partida, veremos que alguns entraves, como não exibir uma relação analítica entre o estimador e a amostra, induz-nos a desenvolver métodos numéricos. Especificamente, para uma φ -gaussiana, a qual tem similaridades com a função de densidade gaussiana padrão, desenvolvemos um algoritmo que depende das propriedades geométricas intrínsecas da φ -gaussiana. Através de simulações, vimos que o método converge, o que nos motivou a fazer uma análise de convergência.

A organização bem como a estrutura dessa tese são apresentadas da seguinte forma. No Capítulo 2, fazemos uma breve revisão dos principais conceitos e resultados da geometria de informação clássica. Além de estudarmos a métrica de Fisher, caracterizamos as famílias exponencial e mistura, definimos uma família de conexões, discutimos um pouco sobre dualidade e versamos sobre divergência. Tal capítulo se faz necessário também para efeitos de comparação, já que algumas de nossas contribuições generalizam alguns resultados de Amari. No Capítulo 3, apresentamos as nossas principais contribuições. Definiremos φ -função e uma φ -família. Definiremos uma chamada φ -divergência com a qual recuperamos um par de conexões duais. Essas conexões, quando combinadas de forma convexa, permite-nos definir uma família de conexões (α -conexão). Generalizamos a conhecida divergência de Rényi e veremos que a partir dela induz-se a mesma α -conexão. Alguns resultados que dependem de um sistema de coordenadas duais também serão abordados, a citar o Teorema de Pitágoras e o Teorema da Projeção. No

Capítulo 4, fazemos uma breve revisão de alguns resultados de inferência, desenvolvemos o nosso método de máxima verossimilhança generalizado e fazemos a respectiva análise de convergência. Na segunda seção, falamos sobre uma versão estendida da Desigualdade de Cramér–Rao para as φ -família, mostrando em seguida que o caso clássico torna-se um caso particular do que propomos. No capítulo 5, indicamos, principalmente, as ações que esse trabalho pode se desenvolver. Por fim, apresentamos um apêndice que revisa os principais conceitos e resultados de Geometria Riemanniana que são utilizados ao longo dessa tese.

2 GEOMETRIA DA INFORMAÇÃO

Geometria da Informação é um ramo da Matemática que se utiliza de ferramentas geométricas para estudar modelos estatísticos e suas devidas aplicações SUN *et al.* (2011). Muitas dessas ferramentas são utilizadas em problemas aplicados, a citar: otimização, processamento de sinais, inferência estatística, filtragem, classificação, dentre outros AMARI (2013); VAN TREES (2004). Um típico exemplo é a técnica baseada no critério de máxima verossimilhança para detecção de sinais não ortogonais com superposição espectral de banda limitada MOTA, CAVALCANTE, and LUCENA (2005). Em sua grande maioria, essas aplicações são desenvolvidas em espaços sem curvatura, como os euclidianos, o que motiva o estudo de modelos estatísticos em ambientes mais complexos, como é caso das variedades estatísticas AMARI and NAGAOKA (2000). Com essa motivação é que se desenvolve a Geometria da Informação.

Geometria da Informação foi introduzida inicialmente com os trabalhos de Rao RAO (1992) e, atualmente, tem sido desenvolvida por vários pesquisadores, como podemos testificar nos trabalhos de AMARI (2013); AMARI and NAGAOKA (2000); MURRAY and RICE (1993); SHIMA (2007). Este capítulo tem por intuito apresentar um resumo das principais ferramentas de geometria da informação que serão utilizadas nesta tese. Além do mais, no capítulo seguinte, estenderemos alguns resultados para uma classe de família mais geral do que a exponencial o que torna persistente esse capítulo preliminar.

De um modo geral, caracteriza-se uma família de distribuições de probabilidade identificando-as por com um vetor paramétrico $\theta = (\theta^1, \dots, \theta^n)^T$, $n \geq 1$, em que tais parâmetros são vistos como um sistema de coordenadas O'NEILL (1983). Tal identificação é feita através de uma função de parametrização. Podemos ilustrar essa identificação com o exemplo abaixo, em que consideramos a família de distribuições normais que variam segundo o vetor de parâmetros $\theta = (\mu, \sigma)$ da forma

$$p_{\mu, \sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right). \quad (1)$$

Para efeito de simplificação, usaremos o termo parâmetro, fazendo-se entender vetor paramétrico. O parâmetro $\theta = (\mu, \sigma)$ identifica-se biunivocamente com a distribuição $p_{\mu, \sigma}$ em (1), devido às propriedades intrínsecas da função a exponencial. Especificamente nesse exemplo, a função exponencial funciona como uma parametrização de um sistema de coordenadas em uma variedade bi-dimensional. É nesse contexto que vamos discorrer ao longo desse capítulo.

2.1 Modelos Paramétricos

Seja (T, Σ, μ) um espaço de medida, onde T é um conjunto, Σ é uma σ -álgebra de subconjuntos de T e $\mu: \Sigma \rightarrow [0, \infty]$ é a função de medida satisfazendo as seguintes condições abaixo JAMES (1996):

- (a) $\mu(\emptyset) = 0$;
- (b) $\mu(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i)$ para qualquer união disjunta enumerável de conjuntos $\{E_i\} \subseteq \Sigma$.

Um exemplo clássico do espaço de medida é o espaço de probabilidade.

Portanto, consideremos o modelo paramétrico n -dimensional \mathcal{S} de uma família de distribuições de probabilidade dado por

$$\mathcal{S} = \{p_{\theta}(x) := p(x; \theta) : \theta = (\theta^1, \dots, \theta^n) \in \Theta\}, \quad (2)$$

na qual cada parâmetro $\theta \in \Theta \subset \mathbb{R}^n$ se associa com uma função densidade de probabilidade p_{θ} e possui dimensão n . Podemos visualizar como opera essa definição na figura abaixo.

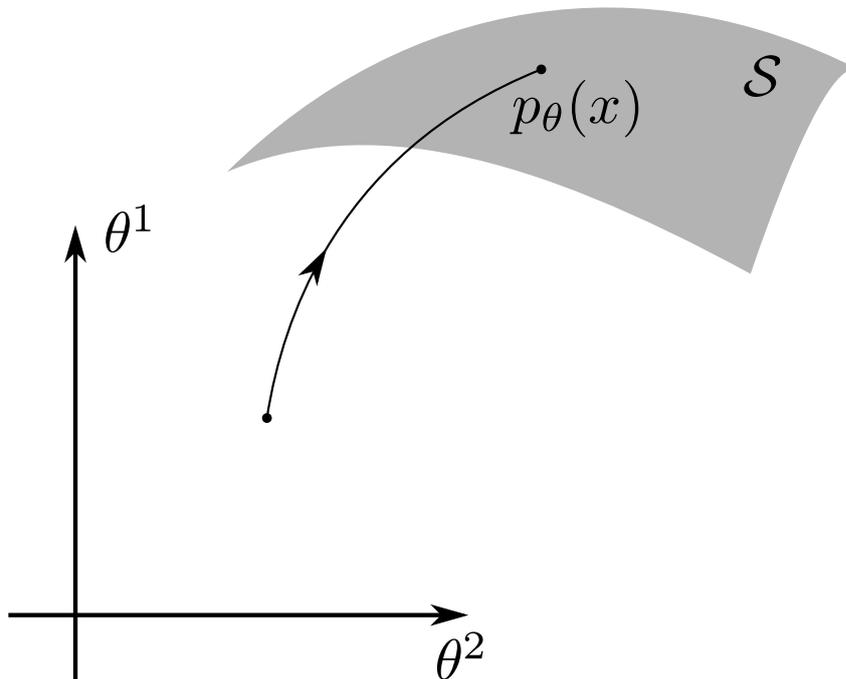


Figura 1: Modelo paramétrico.

Dado um modelo paramétrico estatístico $\mathcal{S} = \{p_{\theta} : \theta \in \Theta\}$, a aplicação $\varrho: \mathcal{S} \rightarrow \mathbb{R}^n$ definida por $\varrho(p_{\theta}) = \theta$ nos permite considerar $\varrho = (\theta^i)$ como um sistema de coordenadas. Além do mais, assume-se algumas condições de regularidade sobre \mathcal{S} , a saber, e as expressões $\frac{\partial p_{\theta}}{\partial \theta^i}$ e $\frac{\partial^2 p_{\theta}}{\partial \theta^i \partial \theta^j}$ existem e que as ordens de integração ou diferenciação

possam ser permutadas. Conclui-se que

$$\int_T \frac{\partial p_\theta}{\partial \theta^i} d\mu = \frac{\partial}{\partial \theta^i} \int_T p_\theta d\mu = \frac{\partial}{\partial \theta^i} 1 = 0. \quad (3)$$

Além disso, considerando \mathcal{S} como a imagem de uma função injetiva $\iota: \Theta \rightarrow \mathcal{S}$, $\iota(\theta) = p_\theta$, faz sentido definir a função ι como a inversa de ϱ . Ademais, podemos inserir outra parametrização para \mathcal{S} . Seja $v: \Theta \rightarrow v(\Theta) \subseteq \mathbb{R}^n$ um difeomorfismo LIMA (1976). Então, $v \circ \varrho: \mathcal{S} \rightarrow v(\Theta) \subset \mathbb{R}^n$ é um outro sistema de coordenadas. Os dois exemplos de famílias paramétricas mais usuais na literatura são LINDSAY *et al.* (1983); LINDSAY (1986). Estas serão citadas na seção seguinte.

2.2 Variedades Estatísticas

Sejam (T, Σ, μ) um espaço de medida e L^0 o conjunto de todas as funções reais, mensuráveis em T . Uma variedade estatística LAURITZEN (1984) é uma coleção de distribuições de probabilidades paramétrica, equipada com objetos geométricos, como métrica e conexões, por exemplo, a qual está contida em

$$\mathcal{P}_\mu = \left\{ p \in L^0 : p > 0 \text{ e } \int_T p d\mu = 1 \right\}.$$

Isto é, \mathcal{P}_μ representa o conjunto de todas as medidas de probabilidade em T as quais são equivalentes com a medida μ . Como caso usual, se $T = \mathbb{R}$ (ou T é um conjunto discreto), a medida μ é identificada como a medida de Lebesgue. Para a caracterização da variedade diferenciável das famílias exponenciais, consideremos o conjunto das famílias de distribuições de probabilidade da forma

$$\mathcal{P} = \{p_\theta : \theta \in \Theta\}, \quad (4)$$

em que a função de probabilidade p_θ é dada em termos dos parâmetros $\theta = (\theta^1, \dots, \theta^n) \in \Theta$ por uma aplicação injetiva como já foi citado na introdução deste capítulo. A família paramétrica \mathcal{P} é chamada de *variedade estatística* se satisfaz as seguintes condições abaixo:

- (A1) Θ é aberto e conexo em \mathbb{R}^n ;
- (A2) p_θ é diferenciável com respeito ao parâmetro θ ;
- (A3) As operações de derivação com respeito ao parâmetro θ^i e integração com respeito a medida μ permutam entre si;
- (A4) A *matriz de informação de Fisher*, ou *métrica de Fisher*, definida por AMARI (2013)

$$g_{ij}(\theta) = -E_\theta \left[\frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} \right], \quad (5)$$

é positiva definida em cada ponto $\theta \in \Theta$, em que

$$l_\theta = \log p_\theta \quad (6)$$

e $E_\theta[\cdot]$ denota o operador esperança com respeito à distribuição p_θ mensurado por μ , isto é, $E_\theta[\cdot] = \int_T (\cdot) p_\theta d\mu$.

Sob a condição (A4), pode-se mostrar que a expressão (5) define uma métrica riemanniana em \mathcal{P} AMARI and NAGAOKA (2000). Além disso, como veremos a seguir, a métrica de Fisher (5) pode ser identificada com outras expressões, onde em uma delas, identificaremos o espaço tangente em $p_\theta T_{p_\theta} \mathcal{P}$. Com efeito, observe que a condição (A3) nos permite mostrar que

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta^i} \int_T p_\theta d\mu \\ &= \int_T \frac{\partial p_\theta}{\partial \theta^i} d\mu \\ &= \int_T \frac{\partial l_\theta}{\partial \theta^i} p_\theta d\mu. \end{aligned} \quad (7)$$

Isto é, a *função score* $\frac{\partial l_\theta}{\partial \theta^i}$ SEN, SINGER, and DE LIMA (2010) tem esperança zero:

$$E_\theta \left[\frac{\partial l_\theta}{\partial \theta^i} \right] = 0. \quad (8)$$

A informação de Fisher é definida como

$$I_F(\theta) = E \left[\frac{\partial}{\partial \theta} \ln(\theta) \frac{\partial}{\partial \theta} \ln(\theta) \right] = -Cov \left[\frac{\partial}{\partial \theta} \ln(\theta) \right].$$

Além disso, ao derivarmos a expressão (7) com respeito ao parâmetro θ^j , obtemos

$$0 = \int_T \frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} p_\theta d\mu + \int_T \frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} p_\theta d\mu, \quad (9)$$

de onde podemos concluir que a matriz de Fisher (5) é também identificada por

$$g_{ij}(\theta) = E_\theta \left[\frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} \right]. \quad (10)$$

O produto interno entre os vetores

$$X = \sum_i a^i \frac{\partial}{\partial \theta^i} \quad \text{e} \quad Y = \sum_j b^j \frac{\partial}{\partial \theta^j}$$

pode ser expresso como

$$g(X, Y) = \sum_{i,j} g_{ij}(\theta) a^i b^j = \sum_{i,j} E_\theta \left[\frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} \right] a^i b^j = E_\theta[\tilde{X}\tilde{Y}],$$

em que

$$\tilde{X} = \sum_i a^i \frac{\partial l_\theta}{\partial \theta^i} \quad \text{e} \quad \tilde{Y} = \sum_j b^j \frac{\partial l_\theta}{\partial \theta^j}.$$

Portanto, observa-se que o espaço tangente $T_{p_\theta} \mathcal{P}$ pode ser identificado com um outro $\tilde{T}_{p_\theta} \mathcal{P}$, o qual tem por vetores geradores as funções $\frac{\partial l_\theta}{\partial \theta^i}$, equipado com o produto interno definido por $\langle \tilde{X}, \tilde{Y} \rangle_\theta = E_\theta[\tilde{X}\tilde{Y}]$. Observe que ao escrevermos o vetor \tilde{X} como combinação linear dos vetores $\frac{\partial l_\theta}{\partial \theta^i}$, a expressão (8) implica que $E_\theta[\tilde{X}] = 0$. Além do mais, a expressão (10) implica que (g_{ij}) é sempre semi-positiva definida, desde que

$$\sum_{i,j} g_{ij} a^i a^j = E_\theta \left[\left(\sum_i a^i \frac{\partial l_\theta}{\partial \theta^i} \right)^2 \right] \geq 0. \quad (11)$$

Vejamos que a expressão da métrica de Fisher dada em (10) ainda assume outra forma. Dessa outra versão, também podemos obter a desigualdade (11). Com efeito

$$\begin{aligned} g_{ij}(\theta) &= \int_T \frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} p_\theta d\mu \\ &= \int_T \frac{\frac{\partial p_\theta}{\partial \theta^i} \frac{\partial p_\theta}{\partial \theta^j}}{p_\theta p_\theta} p_\theta d\mu \\ &= 4 \int_T \frac{\frac{\partial p_\theta}{\partial \theta^i} \frac{\partial p_\theta}{\partial \theta^j}}{2\sqrt{p_\theta} 2\sqrt{p_\theta}} d\mu \\ &= 4 \int_T \frac{\partial \sqrt{p_\theta}}{\partial \theta^i} \frac{\partial \sqrt{p_\theta}}{\partial \theta^j} d\mu. \end{aligned} \quad (12)$$

Sejam um ponto $\theta \in \Theta$ e $v \in T_{p_\theta} \mathcal{P}$ um vetor no espaço tangente de \mathcal{P} em p_θ . Com relação à simetria, a obtemos tendo em vista que $g_{ij}(\theta) = g_{ji}(\theta)$ é consequência direta da definição (10). Através da expressão (12), podemos escrever que

$$\begin{aligned} g(v, v) &= \sum_{i,j} g_{ij} v^i v^j \\ &= 4 \sum_{i,j} \left(\int_T v^i \frac{\partial}{\partial \theta^i} \sqrt{p_\theta} v^j \frac{\partial}{\partial \theta^j} \sqrt{p_\theta} \right) d\mu \\ &= 4 \int_T \left(\sum_i v^i \frac{\partial \sqrt{p}}{\partial \theta^i} \right)^2 dx \geq 0, \end{aligned} \quad (13)$$

e concluimos que a matriz $[g_{ij}(\theta)]$ é não negativa. Além do mais,

$$\begin{aligned}
g(v, v) = 0 &\Leftrightarrow \int_T \left(\sum_i v^i \frac{\partial \sqrt{p}}{\partial \theta^i} \right)^2 d\mu = 0 \\
&\Leftrightarrow \sum_i v^i \frac{\partial \sqrt{p}}{\partial \theta^i} = 0 \\
&\Leftrightarrow \sum_i v^i \frac{\partial p}{\partial \theta^i} = 0 \\
&\Leftrightarrow v^i = 0, \quad \forall i = 1, \dots, n
\end{aligned} \tag{14}$$

desde que os vetores $\left\{ \frac{\partial p_\theta}{\partial \theta^i} \right\}$ sejam linearmente independentes. Isto é, g é não degenerativa. Logo, assumindo que g não satisfaz o caso degenerativo citado em (14), a expressão (13) se limita à desigualdade estrita. A representação $\frac{\partial p_\theta}{\partial \theta^i}$ usada para vetor tangente difere um pouco do que é abordado tradicionalmente na geometria riemanniana. Como será informado no Capítulo 6, dada uma variedade riemanniana \mathcal{M} CARMO (1976a) e um sistema de coordenadas (θ^i) de um ponto $p \in \mathcal{M}$, o operador campo vetorial está definido no espaço das funções diferenciáveis \mathfrak{F} em \mathcal{M} e é relacionado com $f \in \mathfrak{F}$ da seguinte forma

$$f \mapsto \sum_{i=1}^n a^i(p) \left(\frac{\partial}{\partial \theta^i} \right)_p f,$$

em que a^i são funções de valores reais CARMO (1976a). Convencionalmente, os vetores $\frac{\partial}{\partial \theta^i}$ são denominados os vetores da base do espaço tangente.

Um outro elemento de caráter geométrico de uma variedade é a conexão com seus respectivos símbolos de Christoffel CARMO (1976a). Nas variedades estatísticas, podemos definir uma família de conexões associada a uma métrica SHIMA (2007). É o que veremos a seguir.

2.2.1 α -Conexão de Fisher

De uma forma geral, uma conexão é um operador que permite mensurar o nível de variação de um campo vetorial em relação a um outro campo, isto é, a mesma ideia das derivadas direcionais. Além disso, através da conexão podemos estender a noção de linha reta em quaisquer variedades. Intuitivamente, se a derivada direcional deste campo vetorial se anular na direção da curva, então esta curva é chamada de *geodésica* O'NEILL (1983). Dados dois campos vetoriais v e w definidos em um espaço tangente de uma variedade, a derivada de w na direção de v é dada por CARMO (1976a)

$$D_v w = \sum_k \left(\sum_i v^i \frac{\partial w^k}{\partial \theta^i} + \sum_{i,j} v^i w^j \Gamma_{ij}^k \right) \frac{\partial}{\partial \theta^k}, \tag{15}$$

o que mostra que seus componentes dependem dos vetores v e w e de seus símbolos de Christoffel Γ_{ij}^k .

Sobre as variedades estatísticas, AMARI and NAGAOKA (2000) define uma família de conexões (α -conexão) $D^{(\alpha)}$, $\alpha \in [-1, 1]$, cujos símbolos de Christoffel são definidos por

$$\Gamma_{ij}^{(\alpha)k} = \sum_m g^{mk} \Gamma_{ijm}^{(\alpha)},$$

em que

$$\Gamma_{ijk}^{(\alpha)} = E_\theta \left[\left(\frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} + \frac{1 - \alpha}{2} \frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} \right) \frac{\partial l_\theta}{\partial \theta^k} \right]. \quad (16)$$

Note que esta expressão define uma conexão de torção livre, isto é, vale a permutação dos índices dos símbolos das conexões, isto é, $\Gamma_{ij}^{(\alpha)k} = \Gamma_{ji}^{(\alpha)k}$. Esses símbolos nos garantem uma família de conexões afins $D^{(\alpha)}$ em \mathcal{P} definidas por

$$\Gamma_{ijk}^{(\alpha)} = g \left(D_{\frac{\partial}{\partial \theta^i}}^{(\alpha)} \frac{\partial}{\partial \theta^j}, \frac{\partial}{\partial \theta^k} \right), \quad (17)$$

em que g é a métrica de Fisher. Chamamos $D^{(\alpha)}$ de α -conexão. Dados duas conexões $D^{(\alpha)}$ e $D^{(\beta)}$, a relação entre elas é expressa por AMARI and NAGAOKA (2000)

$$\Gamma_{ijk}^{(\beta)} = \Gamma_{ijk}^{(\alpha)} + \frac{\alpha - \beta}{2} T_{ijk}, \quad (18)$$

em que o tensor covariante de grau 3 O'NEILL (1983) é simétrico e definido em qualquer ponto θ por

$$T_{ijk} = E_\theta \left[\frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} \frac{\partial l_\theta}{\partial \theta^k} \right]. \quad (19)$$

A expressão (16) tem três casos especiais de nosso interesse que são obtidos para $\alpha = -1$, $\alpha = 0$ e $\alpha = 1$, em que as conexões são chamadas de mistura, métrica e exponencial, respectivamente AMARI and NAGAOKA (2000). É comum simbolizá-las por $D^{(m)} := D^{(-1)}$, $D^{(metric)} := D^{(0)}$ e $D^{(e)} := D^{(1)}$. Por fim, é possível mostrar que a α -conexão $D^{(\alpha)}$ pode ser combinada com $D^{(m)}$ e $D^{(e)}$ por

$$D^{(\alpha)} = \frac{1 + \alpha}{2} D^{(e)} + \frac{1 - \alpha}{2} D^{(m)}.$$

Um conceito que será utilizado nessa tese será o de conexão *flat*.

Definição 2.1 *Uma conexão D é dita ser flat se seu tensor torção T e o tensor curvatura R são identicamente nulos.*

Ainda não definimos o conceito de curvatura que será feito no Capítulo 6. Porém, antecipadamente, afirmamos que se uma conexão tem curvatura identicamente nula, a curvatura R associada a essa conexão também será. A seguir, veremos algumas propriedades dos casos particulares das conexões citadas. Mostraremos alguns exemplos que nos ajudarão a perceber algumas características importantes, como verificar que as

conexões $D^{(m)}$ e $D^{(e)}$ são *flat* nas famílias exponencial e mistura que serão citadas a seguir.

2.2.2 Família Exponencial

Uma família de distribuição de probabilidades $\mathcal{E}_p = \{p_\theta : \theta \in \Theta\} \subseteq \mathcal{P}_\mu$ é dita ser uma *família exponencial centrada em $p \in \mathcal{P}_\mu$* se seus elementos são descritos da forma

$$p_\theta = \exp\left(\sum_{i=1}^n \theta^i u_i - \psi(\theta)\right) p, \quad (20)$$

em que $u_1, \dots, u_n : T \rightarrow \mathbb{R}$ são funções mensuráveis satisfazendo algumas propriedades, que mostraremos a seguir, e $\psi(\theta)$ é uma função introduzida na função (20) de tal forma que $\int_T p_\theta d\mu = 1$. A função ψ é chamada de *função normalizadora*. O conjunto $\Theta \subset \mathbb{R}^n$ é considerado ser maximal ROGERS (1967). Em outras palavras, definimos Θ de tal modo que contenha todos os vetores $\theta = (\theta^i)$, tal que

$$E_p \left[\exp\left(\lambda \sum_{i=1}^n \theta^i u_i\right) \right] < \infty, \quad \text{para algum } \lambda > 1,$$

em que $E_p[\cdot]$ denota a integral com respeito à medida $p \cdot \mu$. Além disso, cada função u_k satisfaz as seguintes condições abaixo:

- (U1) $E_p[u_k] = 0$ e
- (U2) existe um $\varepsilon > 0$ tal que

$$E_p[\exp(\lambda u_k)] < \infty, \quad \text{para todo } \lambda \in (-\varepsilon, \varepsilon).$$

Essas condições serão de interesse particular no capítulo que se segue. Considerando as condições (U1)–(U2), a família exponencial é uma subvariedade LEE (2003) da família exponencial não paramétrica CENA and PISTONE (2007); PISTONE and ROGANTIN (1999); PISTONE and SEMPI (1995). A condição (U1) garante que a função ψ é não negativa, o que é de nosso interesse. A condição (U2) indica que o conjunto Θ é aberto. Deste fato, como vemos em VIGELIS and CAVALCANTE (2013b), temos que a função $\psi(\theta)$ é convexa.

Em \mathcal{E}_p , a matriz de informação de Fisher (g_{ij}) é a matriz hessiana de ψ , o que engloba \mathcal{E}_p como uma variedade hessiana SHIMA (2007), e a conexão exponencial é identicamente nula. Com efeito, as expressões abaixo

$$\frac{\partial l_\theta}{\partial \theta^i} = u_i(x) - \frac{\partial \psi}{\partial \theta^i} \quad \text{e} \quad \frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} = -\frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}$$

implicam que

$$g_{ij}(\theta) = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}.$$

Um fato interessante é que a conexão exponencial $D^{(e)}$ é identicamente nula, desde que

$$\Gamma_{ijk}^{(e)} = E_{\theta} \left[\frac{\partial^2 l_{\theta}}{\partial \theta^i \partial \theta^j} \frac{\partial l_{\theta}}{\partial \theta^k} \right] = - \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j} E_{\theta} \left[\frac{\partial l_{\theta}}{\partial \theta^k} \right] = 0. \quad (21)$$

Vale também salientar que a função geradora de momentos tem relações com a divergência de Kullback–Leibler. Desde que para p_{θ} seja descrita como em (20) e do fato de ser uma função densidade de probabilidade, temos que

$$\psi(\theta) = \sum_{i=1}^n \theta^i u_i + \log \frac{p}{p_{\theta}}. \quad (22)$$

Ao aplicarmos o operador esperança em ambos os lados de (22) e da propriedade (U1), obtemos

$$\psi(\theta) = \int_T \log \left(\frac{p}{p_{\theta}} \right) p d\mu. \quad (23)$$

Vejam os dois exemplos abaixo que mostram os resultados citados de forma aplicada em distribuições de probabilidade usuais.

Exemplo 1 Consideremos o conjunto discreto $T = \{0, 1, \dots, n\}$ e o domínio

$$\Lambda = \left\{ \lambda = (\lambda_i) \in \mathbb{R}^{n+1} : \sum_{i=0}^n \lambda_i \text{ e } \lambda_i > 0 \right\}.$$

A família $\mathcal{P} = \{p(i; \lambda) : \lambda \in \Lambda\}$, cujos membros são dados por $p(i; \lambda) = \lambda_i$ é exemplo de uma família exponencial. Fixe qualquer distribuição de probabilidade $p(i)$ em \mathcal{P} . Seja $u_j(i) = \delta_{ij} - \delta_{i0} p(j)/p(0)$, onde δ_{ij} é a função delta de Kronecker. Verifica-se que $E_p[u_j] = 0$. Então,

$$p(i; \lambda) = \exp \left\{ \sum_{j=1}^n \left(\log \frac{\lambda_j}{p(j)} - \log \frac{\lambda_0}{p(0)} \right) u_j(i) + \log \frac{\lambda_0}{p(0)} \right\} p(i). \quad (24)$$

Denotando

$$\theta^j = \log \frac{\lambda_j}{p(j)} - \log \frac{\lambda_0}{p(0)},$$

a expressão (24) resulta em

$$p(i; \theta) = \exp \left(\sum_{j=1}^n \theta^j u_j(i) - \psi(\theta) \right) p(i),$$

onde

$$\psi(\theta) = \log \left(p(0) + \sum_{j=1}^n p(j) \exp \theta^j \right).$$

Então \mathcal{P} é uma família exponencial.

Exemplo 2 A distribuição de Poisson com parâmetro $\lambda > 0$ é dada por

$$p(i; \lambda) = \frac{\lambda^i}{i!} e^{-\lambda}, \quad i \in \{0, 1, 2, \dots\}.$$

Afirmamos que a família de distribuições de Poisson também formam uma família exponencial. Com efeito, percebe-se que

$$\frac{\lambda^i}{i!} e^{-\lambda} = \exp[(i-1)\log(\lambda) - (\lambda - \log(\lambda) - 1)] \frac{e^{-1}}{i!}.$$

Denotando

$$u(i) = i - 1, \quad p(i) = \frac{e^{-1}}{i!},$$

e

$$\psi(\theta) = e^\theta - \theta - 1, \quad \theta = \log \lambda,$$

temos

$$p(i; \theta) = \exp(\theta u(i) - \psi(\theta)) p(i), \quad \theta \in \mathbb{R}.$$

Mais uma vez, com algumas manipulações mostra-se que $E_p[u] = 0$ e $E_p[\exp(\lambda u_k)] < \infty$, para todo $\lambda > 0$.

Em AMARI and NAGAOKA (2000), cita-se a família exponencial de uma outra forma com a diferença de não necessariamente ser centrada em p . Vamos rapidamente descrevê-la. Consideremos $n+1$ funções de valores reais $\{1, F_1, \dots, F_n\}$, onde 1 denota uma função constante. A família exponencial é o conjunto de todas as distribuições de probabilidade dadas por

$$p_\theta = \exp\left(C + \sum_{i=1}^n \theta^i F_i - \psi(\theta)\right). \quad (25)$$

De igual modo, a função $\psi: \Theta \rightarrow \mathbb{R}$ tem o papel de normalizadora da função, isto é, $\int_T p_\theta d\mu = 1$. Com algumas manipulações, verifica-se que a função $\psi(\theta)$ é dada por

$$\psi(\theta) = \log \int_T \exp\left(C + \sum_{i=1}^n \theta^i F_i\right) du. \quad (26)$$

Um exemplo clássico da família exponencial é a família das funções gaussianas.

Exemplo 3 A família de gaussianas dadas como em (1) são exemplos de família exponencial, quando assumimos que

$$C(x) = 0, \quad F_1(x) = x, \quad F_2(x) = x^2, \quad \theta^1 = \frac{\mu}{\sigma^2},$$

$$\theta^2 = \frac{1}{2\sigma^2}, \quad \psi(\theta) = \frac{\mu^2}{2\sigma^2} + \log(\sqrt{2\pi}\sigma) = -\frac{(\theta^1)^2}{4\theta^2} + \frac{1}{2} \log\left(-\frac{\pi}{\theta^2}\right).$$

Uma observação importante em (25) é que se valendo de condições de regu-

laridade e supondo que $E_\theta[F_i] < \infty$, em que o operador esperança é dado por $E_\theta[\cdot] := \int_T (\cdot) p_\theta d\mu$, podemos caracterizar a média das funções F_j da seguinte forma que se segue

$$\begin{aligned} \int_T \frac{\partial}{\partial \theta^i} p_\theta d\mu &= 0 \Leftrightarrow \\ \int_T p_\theta \left(F_j - \frac{\partial \psi}{\partial \theta^j}(\theta) \right) d\mu &= 0 \Leftrightarrow \\ \int_T p_\theta F_j d\mu &= \frac{\partial}{\partial \theta^j} \psi(\theta) \int_T p_\theta d\mu \Leftrightarrow \\ E_\theta[F_j] &= \frac{\partial}{\partial \theta^j} \psi(\theta). \end{aligned}$$

Desta feita, observe que na família exponencial,

$$l_\theta = C + \sum_{i=1}^n \theta^i F_i - \psi(\theta),$$

a qual com algumas manipulações algébricas, mais uma vez temos

$$\frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} = - \frac{\partial^2 \psi(\theta)}{\partial \theta^i \partial \theta^j}. \quad (27)$$

Também podemos observar que a conexão $\Gamma_{ijk}^{(e)}$ é identicamente nula sobre a família exponencial (25). De fato,

$$\begin{aligned} \Gamma_{ijk}^{(1)} &= E_\theta \left[\frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} \frac{\partial l_\theta}{\partial \theta^k} \right] \\ &= -E_\theta \left[\frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j} \frac{\partial l_\theta}{\partial \theta^k} \right] \\ &= - \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j} E_\theta \left[\frac{\partial l_\theta}{\partial \theta^k} \right] \\ &= - \frac{\partial^2 \psi(\theta)}{\partial \theta^i \partial \theta^j} \frac{\partial}{\partial \theta^k} E_\theta[p_\theta] \\ &= 0. \end{aligned} \quad (28)$$

2.2.3 Família Mistura

Uma família de distribuições $\mathcal{M} = \{p(t; \eta) : \eta \in \Theta\} \subseteq \mathcal{P}_\mu$ é chamada de *família mistura* se

$$p(t; \eta) = \sum_{i=1}^n \eta_i p_i(t) + \left(1 - \sum_{i=1}^n \eta_i \right) p_0(t), \quad (29)$$

em que p_0 e p_i são distribuições de probabilidade e $\eta = (\eta_i)$ é tomado tal que o lado direito de (29) define uma distribuição de probabilidade em \mathcal{P}_μ . Claramente, o conjunto $\{\eta = (\eta_i) : \eta_i \in (0, 1) \text{ e } \sum_{i=1}^n \eta_i < 1\}$ está contido em Θ . Diferenciando $l(t; \eta) = \log p(t; \eta)$

com respeito a η_i , tomamos

$$\frac{\partial l(t; \eta)}{\partial \eta_i} = \frac{p_i(t) - p_0(t)}{p(t; \eta)},$$

e então

$$g_{ij} = E_\eta \left[\frac{p_i - p_0}{p_\eta} \frac{p_j - p_0}{p_\eta} \right].$$

O Exemplo 1 também pode ser reescrito como família mistura. Vejamos no exemplo abaixo:

Exemplo 4 Consideremos as hipóteses do Exemplo 1. Sejam p_0, p_1, \dots, p_n distribuições de probabilidade independentes e identicamente distribuídas em \mathcal{P} . Então

$$p(j; \eta) = \sum_{i=1}^n \eta_i p_i(j) + \left(1 - \sum_{i=1}^n \eta_i\right) p_0(j)$$

é uma parametrização para \mathcal{P} . Para estas distribuições, temos

$$p(j; \eta) = \eta_j,$$

em que $\eta_0 = 1 - \sum_{i=1}^n \eta_i$. Assim a matriz de informação de Fisher é dada por

$$\begin{aligned} g_{ij} &= E_\eta \left[\frac{p_i - p_0}{p_\eta} \frac{p_j - p_0}{p_\eta} \right] \\ &= \sum_{k=0}^n \frac{p_i(k) - p_0(k)}{p(k; \eta)} \frac{p_j(k) - p_0(k)}{p(k; \eta)} p(k; \eta) \\ &= \sum_{k=0}^n \frac{(\delta_{ik} - \delta_{0k})(\delta_{jk} - \delta_{0k})}{\eta_k} \\ &= \frac{\delta_{ij}}{\eta_i} + \frac{1}{\eta_0}, \end{aligned}$$

cuja inversa é dada por

$$g^{ij} = \eta_i \delta_{ij} - \eta_i \eta_j.$$

A conexão mistura $D^{(m)}$ é identicamente nula em \mathcal{M} , desde que

$$\frac{\partial^2 l_\eta}{\partial \eta_i \partial \eta_j} = -\frac{\partial l_\eta}{\partial \eta_i} \frac{\partial l_\eta}{\partial \eta_j},$$

de forma que

$$\begin{aligned} \Gamma_{ijk}^{(m)} &= \Gamma_{ijk} + T_{ijk} \\ &= E_\theta \left[\frac{\partial^2 l_\eta}{\partial \eta_i \partial \eta_j} \frac{\partial l_\eta}{\partial \eta_k} \right] + E_\theta \left[\frac{\partial l_\eta}{\partial \eta_i} \frac{\partial l_\eta}{\partial \eta_j} \frac{\partial l_\eta}{\partial \eta_k} \right] \\ &= 0. \end{aligned} \tag{30}$$

2.2.4 Conexão Métrica

A conexão métrica é aquela cujo tensor métrico $g_{ij}^{(metric)}$ é dado por

$$g_{ij}^{(metric)} = \frac{1}{2} \left(\frac{\partial g_{ik}}{\partial \theta^j} + \frac{\partial g_{jk}}{\partial \theta^i} - \frac{\partial g_{ij}}{\partial \theta^k} \right), \quad (31)$$

em que g_{ij} é a métrica de Fisher dada em (10), a expressão (31) é satisfeita. De fato, calculando cada uma das parcelas em 31, temos

$$\begin{aligned} \frac{\partial g_{ik}}{\partial \theta^j} &= \frac{\partial}{\partial \theta^j} \int_T \frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^k} p_\theta d\mu \\ &= \int_T \left(\frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^k} \frac{\partial p_\theta}{\partial \theta^j} + \frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} \frac{\partial l_\theta}{\partial \theta^k} + \frac{\partial l_\theta}{\partial \theta^i} \frac{\partial^2 l_\theta}{\partial \theta^j \partial \theta^k} \right) p_\theta d\mu \\ &= \int_T \left(\frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} \frac{\partial l_\theta}{\partial \theta^k} + \frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} \frac{\partial l_\theta}{\partial \theta^k} + \frac{\partial l_\theta}{\partial \theta^i} \frac{\partial^2 l_\theta}{\partial \theta^j \partial \theta^k} \right) p_\theta d\mu \end{aligned} \quad (32)$$

De forma análoga,

$$\frac{\partial g_{jk}}{\partial \theta^i} = \int_T \left(\frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} \frac{\partial l_\theta}{\partial \theta^k} + \frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} \frac{\partial l_\theta}{\partial \theta^k} + \frac{\partial l_\theta}{\partial \theta^j} \frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^k} \right) p_\theta d\mu \quad (33)$$

] e

$$\frac{\partial g_{ij}}{\partial \theta^k} = \int_T \left(\frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} \frac{\partial l_\theta}{\partial \theta^k} + \frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^k} \frac{\partial l_\theta}{\partial \theta^j} + \frac{\partial l_\theta}{\partial \theta^i} \frac{\partial^2 l_\theta}{\partial \theta^j \partial \theta^k} \right) p_\theta d\mu \quad (34)$$

Portanto, substituindo as expressões (32), (33) e (34) em (31), obtemos

$$\begin{aligned} \Gamma_{ijk}^{(0)} &= \frac{1}{2} \int_T \left(\frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} \frac{\partial l_\theta}{\partial \theta^k} + 2 \frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} \frac{\partial l_\theta}{\partial \theta^k} \right) p_\theta d\mu \\ &= E_\theta \left[\left(\frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^k} + \frac{1}{2} \frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} \right) \frac{\partial l_\theta}{\partial \theta^k} \right]. \end{aligned} \quad (35)$$

Observe que a expressão (35) acima coincide com (16) quando consideramos $\alpha = 0$. Nesse caso, a variedade estatística \mathcal{P} adicionada desta conexão métrica se comporta como uma variedade riemanniana. Dadas as expressões para as conexões métrica, mistura e exponencial, podemos ver que ambas se relacionam pela equação abaixo

$$\begin{aligned} \Gamma_{ijk}^{(-1)} + \Gamma_{ijk}^{(1)} &= E_\theta \left[\left(\frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} + \frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} \right) \frac{\partial l_\theta}{\partial \theta^k} \right] + E_\theta \left[\frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} \frac{\partial l_\theta}{\partial \theta^k} \right] \\ &= E_\theta \left[\left(2 \frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} + \frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} \right) \frac{\partial l_\theta}{\partial \theta^k} \right] \\ &= 2E_\theta \left[\frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} + \frac{1}{2} \frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} \right] \\ &= 2\Gamma_{ijk}^{(0)}. \end{aligned} \quad (36)$$

2.2.5 Entropia em Variedades Estatísticas

Os estudos de entropia, medida de incerteza de uma variável aleatória, tiveram seus inícios com o trabalho de SHANNON (2001). No estudo de Teoria da Informação, a teoria de entropia é uma ferramenta de grande utilidade para os mais diversos problemas de aplicação COVER and THOMAS (2012, 2006). Iremos focar alguns conceitos e propriedades de entropia para modelos probabilísticos.

Mais uma vez, tomemos a variedade estatística (2). A função de entropia $H: \Theta \rightarrow \mathbb{R}$ é definida por

$$H(\theta) = \begin{cases} - \int_T p_\theta \log p_\theta dx, & \text{se } T \text{ é absolutamente contínuo;} \\ - \sum_{x \in T} p_\theta(x) \log p_\theta(x), & \text{se } T \text{ é discreto.} \end{cases} \quad (37)$$

Especialmente para essa tese, focaremos no caso absolutamente contínuo. Para efeito de entendimento e aplicação, vejamos o exemplo abaixo.

Exemplo 5 *Considere a função de distribuição normal dada por*

$$p_{\mu, \sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right). \quad (38)$$

Usando a definição de entropia para variáveis contínuas (37) na expressão (38), obtemos que

$$H(\mu, \sigma) = \int_{\mathbb{R}} \left(\frac{1}{2} \log(2\pi) + \log(\sigma) + \frac{(x - \mu)^2}{2\sigma^2} \right) p_{\mu, \sigma}(x) dx.$$

Com rápidas manipulações, obtemos que $H(\mu, \sigma) = \log(\sigma\sqrt{2\pi e})$. É interessante observar que a função de entropia não depende da média μ , mas apenas do desvio-padrão σ .

Um interesse natural quando estudamos entropia é de maximizá-la, pois significa que recuperamos o máximo de informação. Mais precisamente, maximizar uma entropia significa encontrar uma função densidade de probabilidade que dê o maior grau de incerteza. Vejamos as seguintes caracterizações abaixo.

Seja f uma função definida sobre a variedade estatística \mathcal{P} . Consideremos a notação $\partial_i f := \frac{\partial}{\partial \theta^i} (f \circ p_\theta)$ que mostra a variação que se tem na distribuição de probabilidade em relação a um parâmetro θ^i . Aplicaremos esse conceito para a entropia.

Definição 2.2 *Um ponto $q \in \mathcal{P}$ é um ponto crítico da função entropia H se*

$$\partial_i H = 0, \quad \forall \partial_i \in T_q \mathcal{P}. \quad (39)$$

Obtemos uma expressão para a entropia, quando desenvolvemos a definição

dada na expressão (39). Com efeito,

$$\begin{aligned}
\partial_i H &= -\partial_i \int_T p_\theta \log(p_\theta) dx \\
&= -\int_T [\log(p_\theta) + 1] \partial_i p_\theta dx \\
&= -\int_T \log(p_\theta) \partial_i p_\theta dx \\
&= -E_\theta[l_\theta \partial_i l_\theta].
\end{aligned} \tag{40}$$

Para classificarmos uma distribuição de tal forma que ela maximize ou minimize a função de entropia, precisa-se estudar a informação proveniente da matriz Hessiana HIRIART-URRUTY, Strodiot, and Nguyen (1984). Dessa forma, diferenciando a expressão (40) em relação ao parâmetro θ^j , temos

$$\begin{aligned}
\partial_{ij} H &= -\partial_j \int l_\theta \partial_i l_\theta p_\theta dx \\
&= -\int_T (l_\theta \partial_i l_\theta \partial_j p_\theta + \partial_j l_\theta \partial_i l_\theta p_\theta + l_\theta \partial_i \partial_j l_\theta p_\theta) dx \\
&= -E_\theta[\partial_i \partial_j l_\theta] - E_\theta[(\partial_j l_\theta \partial_i l_\theta + \partial_i \partial_j l_\theta) l_\theta] \\
&= -g_{ij}(\theta) - E_\theta[(\partial_j l_\theta \partial_i l_\theta + \partial_i \partial_j l_\theta) l_\theta].
\end{aligned} \tag{41}$$

Isto é, (41) representa a matriz hessiana da função entropia H .

Uma outra questão de extrema relevância para geometria da informação se concentra nas *funções de divergência* AMARI and NAGAOKA (2000). Este conceito tem por princípio estabelecer uma relação entre duas funções densidade de probabilidade. As funções de divergência tem como um dos exemplos clássicos a *entropia relativa de Kullback–Leibler* ou divergência de Kullback–Leibler COVER and THOMAS (2006). Esta medida é utilizada nos estudos que envolvem entropia. A seguir, vamos definir alguns conceitos e propriedades preliminares de entropia, para avançarmos nas questões de divergência de Kullback–Leibler.

Um outro ponto importante é corrigir o fato de que, para algumas funções, a medida de entropia possa ser negativa. Com esse propósito, define-se a *entropia relativa* VAN ERVEN and HARREMOÉS (2014). A mais conhecida função de entropia relativa é a *entropia relativa de Kullback–Leibler*. Está é uma medida não comutativa da diferença entre duas densidades de probabilidade p e q em uma mesma variedade estatística e é definida por

$$\mathcal{D}_{\text{KL}}(p \parallel q) = \begin{cases} \sum_{x_i \in T} p(x_i) \log \frac{p(x_i)}{q(x_i)}, & \text{se } T \text{ é discreto;} \\ \int_T p \log\left(\frac{p}{q}\right) dx, & \text{se } T \text{ é contínuo.} \end{cases} \tag{42}$$

Um fato de extrema relevância para as nossas contribuições é perceber que a função

geradora de cumulantes dada em (23) coincide com a expressão para a divergência. Um outro exemplo é dado abaixo.

Exemplo 6 *Considere duas funções de densidade exponencial*

$$p_{\theta^1}(x) = \theta^1 e^{-\theta^1 x} \quad e \quad p_{\theta^2}(x) = \theta^2 e^{-\theta^2 x}.$$

Observe que como

$$\log\left(\frac{p_{\theta^1}(x)}{p_{\theta^2}(x)}\right) = \log\left(\frac{\theta^1}{\theta^2}\right) + (\theta^2 - \theta^1)x,$$

temos

$$\begin{aligned} \mathcal{D}_{\text{KL}}(p_{\theta^1} \parallel p_{\theta^2}) &= \int_0^\infty p_{\theta^1} \log\left(\frac{p_{\theta^1}}{p_{\theta^2}}\right) dx \\ &= \int_0^\infty \log\left(\frac{\theta^1}{\theta^2}\right) p_{\theta^1} dx + (\theta^2 - \theta^1) \int_0^\infty x p_{\theta^1}(x) dx \\ &= \frac{\theta^2}{\theta^1} - \log\left(\frac{\theta^2}{\theta^1}\right) - 1. \end{aligned}$$

Logo, observando o gráfico da função $f(x) = x - \log x - 1$, concluímos que $\mathcal{D}_{\text{KL}}(p_{\theta^1} \parallel p_{\theta^2}) \geq 0$. Obviamente, $\mathcal{D}_{\text{KL}}(p_{\theta^1} \parallel p_{\theta^2}) \neq \mathcal{D}_{\text{KL}}(p_{\theta^2} \parallel p_{\theta^1})$, com igualdade caso $p_{\theta^1} = p_{\theta^2}$. Além disso, um importante ponto é que nesse exemplo, não vale a desigualdade triangular. Caso tomemos uma terceira função densidade de probabilidade p_{θ^3} , associada ao parâmetro θ^3 , observamos que a desigualdade abaixo

$$\mathcal{D}_{\text{KL}}(p_{\theta^1} \parallel p_{\theta^2}) + \mathcal{D}_{\text{KL}}(p_{\theta^2} \parallel p_{\theta^3}) \geq \mathcal{D}_{\text{KL}}(p_{\theta^1} \parallel p_{\theta^3}). \quad (43)$$

nem sempre é satisfeita. De fato, para tal, deveríamos ter que

$$\begin{aligned} \mathcal{D}_{\text{KL}}(p_{\theta^1} \parallel p_{\theta^2}) + \mathcal{D}_{\text{KL}}(p_{\theta^2} \parallel p_{\theta^3}) \geq \mathcal{D}_{\text{KL}}(p_{\theta^1} \parallel p_{\theta^3}) &\Leftrightarrow \\ f\left(\frac{\theta^2}{\theta^1}\right) + f\left(\frac{\theta^3}{\theta^2}\right) \geq f\left(\frac{\theta^3}{\theta^1}\right) &\Leftrightarrow \\ \log\left(\frac{\theta^3}{\theta^1}\right) - \log\left(\frac{\theta^2}{\theta^1}\right) - \log\left(\frac{\theta^3}{\theta^2}\right) \geq 1 + \frac{\theta^3}{\theta^1} - \frac{\theta^3}{\theta^2} - \frac{\theta^2}{\theta^1} &\Leftrightarrow \\ 0 \geq 1 + \frac{\theta^3}{\theta^1} - \frac{\theta^3}{\theta^2} - \frac{\theta^2}{\theta^1} & \end{aligned}$$

o que nem sempre é obtido. Portanto, essas interpretações observadas na entropia relativa de Kullback–Leibler permitem definir um conceito mais geral que é essencial para construir a geometria de uma variedade estatística.

Definição 2.3 *Seja \mathcal{S} uma variedade estatística e considere a função suave $\mathcal{D}: \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ SHANKS and PURSELL (1954) satisfazendo*

$$\mathcal{D}(p \parallel q) \geq 0, \quad e \quad \mathcal{D}(p \parallel q) = 0 \Leftrightarrow p = q$$

para quaisquer $p, q \in \mathcal{S}$. Nessas condições, a função \mathcal{D} é chamada de divergência AMARI and NAGAOKA (2000).

Como vimos, a entropia relativa descrita na equação (42) não satisfaz a desigualdade triangular. Por esse motivo, a caracterização da divergência dada na Definição 2.3 não cita a desigualdade triangular. Além disso, também não vale a simetria. Esse comentário é pertinente para evidenciar que a função divergência não é uma função distância definida entre dois pontos de \mathcal{S} . Além disso, o fato de maior relevância da divergência é que a partir dela, com algumas manipulações, podemos caracterizar os elementos da matriz da métrica da variedade e os símbolos de Christoffel. Para os passos iniciais desse entendimento, seja, mais uma vez função de divergência $\mathcal{D}(p_{\theta^1} \parallel p_{\theta^2}) = \frac{\theta^2}{\theta^1} - \log \frac{\theta^2}{\theta^1} - 1$ obtida no Exemplo 6. As duas primeiras derivadas com respeito ao parâmetro θ^2 são

$$\frac{\partial}{\partial \theta^2} \mathcal{D}(p_{\theta^1} \parallel p_{\theta^2}) = \frac{1}{\theta^1} - \frac{1}{\theta^2} \quad \text{e} \quad \frac{\partial^2}{\partial \theta^1 \partial \theta^2} \mathcal{D}(p_{\theta^1} \parallel p_{\theta^2}) = \frac{1}{(\theta^2)^2}.$$

Avaliando as derivadas parciais em $\theta^2 = \theta^1$, obtemos

$$\left[\frac{\partial}{\partial \theta^2} \mathcal{D}(p_{\theta^1} \parallel p_{\theta^2}) \right]_{\theta^2=\theta^1} = 0 \quad \text{e} \quad \left[\frac{\partial^2}{\partial \theta^1 \partial \theta^2} \mathcal{D}(p_{\theta^1} \parallel p_{\theta^2}) \right]_{\theta^2=\theta^1} = \frac{1}{(\theta^1)^2}, \quad (44)$$

em que $\frac{1}{(\theta^1)^2}$ coincide com o valor da métrica de Fisher $g_{11}(\theta^1)$. Mais geralmente,

$$\begin{aligned} \frac{\partial^2}{\partial \theta^1 \partial \theta^2} \mathcal{D}_{\text{KL}}(p_{\theta^1} \parallel p_{\theta^2}) &= \frac{\partial^2}{\partial \theta^1 \partial \theta^2} \int_T p_{\theta^2} \log(p_{\theta^2}) dx - \frac{\partial^2}{\partial \theta^1 \partial \theta^2} \int_T p_{\theta^1} \log(p_{\theta^2}) dx \\ &= \int_T p_{\theta^1} \frac{\partial^2 l_{\theta^2}}{\partial \theta^1 \partial \theta^2} dx. \end{aligned}$$

Fazendo $\theta^2 = \theta^1$, vemos que a derivada segunda coincide com a métrica, como comprovamos a seguir

$$\begin{aligned} \left[\frac{\partial^2}{\partial \theta^1 \partial \theta^2} \mathcal{D}_{\text{KL}}(p_{\theta^1} \parallel p_{\theta^2}) \right]_{\theta^2=\theta^1} &= - \int_T p_{\theta^1} \frac{\partial l_{\theta^1}}{\partial \theta^1 \partial \theta^2} dx \\ &= - E_{\theta^1} \left[\frac{\partial l_{\theta^1}}{\partial \theta^1 \partial \theta^2} \right] \\ &= g_{ij}(\theta^1), \end{aligned} \quad (45)$$

em que esta última igualdade decorre da notação utilizada em (5). Dessa forma, a expressão (45) nos sugere obter a matriz de informação de Fisher, ao considerar quaisquer parâmetros θ^i e θ^j , como vemos a seguir.

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^i\partial\theta^j}\mathcal{D}_{\text{KL}}(p_{\theta^1} \parallel p_{\theta^2}) &= \frac{\partial^2}{\partial\theta^i\partial\theta^j} \int_T p_{\theta^1} \log(p_{\theta^1}) dx - \frac{\partial^2}{\partial\theta^i\partial\theta^j} \int_T p_{\theta^1} \log(p_{\theta^2}) dx \\
&= - \int_T p_{\theta^1} \frac{\partial^2 l_{\theta^2}}{\partial\theta^i\partial\theta^j} dx \Leftrightarrow \\
\left[\frac{\partial^2}{\partial\theta^i\partial\theta^j} \mathcal{D}_{\text{KL}}(p_{\theta^1} \parallel p_{\theta^2}) \right]_{\theta^2=\theta^1} &= - \int_T \frac{\partial^2 l_{\theta^1}}{\partial\theta^i\partial\theta^j} p_{\theta^1} dx \\
&= -E_{\theta^1} \left[\frac{\partial^2 l_{\theta^1}}{\partial\theta^i\partial\theta^j} \right] \\
&= g_{ij}(\theta^1), \tag{46}
\end{aligned}$$

em que a última igualdade é concluída da expressão (5).

Além disso, também se pode relacionar o símbolo de Christoffel da conexão $D^{(1)}$ com a $\mathcal{D}_{\text{KL}}(\cdot \parallel \cdot)$. De fato, derivando (46) em relação a outro parâmetro θ^k e avaliando em $\theta^2 = \theta^1$, temos

$$\begin{aligned}
\left[\frac{\partial^3}{\partial\theta^i\partial\theta^j\partial\theta^k} \mathcal{D}_{\text{KL}}(p_{\theta^1} \parallel p_{\theta^2}) \right]_{\theta^2=\theta^1} &= \frac{\partial}{\partial\theta^k} \int_T \frac{\partial l_{\theta^1}}{\partial\theta^i\partial\theta^j} dx \\
&= E_{\theta^1} \left[\frac{\partial^2 l_{\theta^1}}{\partial\theta^i\partial\theta^j} \frac{\partial l_{\theta^1}}{\partial\theta^k} \right] \\
&= \Gamma_{ijk}^{(1)}(\theta^1),
\end{aligned}$$

em que essa última igualdade é válida desde que

$$g\left(D_{\frac{\partial}{\partial\theta^i}}^{(1)} \frac{\partial}{\partial\theta^j}, \frac{\partial}{\partial\theta^k}\right) = \Gamma_{ijk}^{(1)} = E_{\theta} \left[\left(\frac{\partial l_{\theta}}{\partial\theta^i\partial\theta^j} \right) \left(\frac{\partial l_{\theta}}{\partial\theta^k} \right) \right]$$

é fato pela definição e pela fórmula da α -conexão citada em (16).

Por fim nessa seção, introduzimos a noção de distância de *distância de Fisher* $\mathbf{dist}(p, q)$, em que p e q são duas distribuições de probabilidade na variedade estatística \mathcal{P} . Essa medida representa a distância da informação entre p e q . Esta medida é definida como a menor curva, geodésica, em \mathcal{P} que conecta os pontos p e q . É fato que a divergência de Kullback–Leibler $\mathcal{D}_{\text{KL}}(p \parallel q)$ e a distância de Fisher $\mathbf{dist}(p, q)$. Para melhor entendimento, vejamos o seguinte exemplo:

Exemplo 7 Consideremos duas distribuições exponenciais, isto é, $p_{\theta_1}(x) = \theta_1 e^{-\theta_1 x}$ e $p_{\theta_2}(x) = \theta_2 e^{-\theta_2 x}$ e $\gamma(s)$ a geodésica ligando. Portanto, $\mathbf{dist}(p_{\theta_1}, p_{\theta_2}) = \ln\left(\frac{\theta_2}{\theta_1}\right)$ e $\mathcal{D}_{\text{KL}}(p_{\theta_1} \parallel p_{\theta_2}) = \frac{\theta_1}{\theta_2} - \ln\left(\frac{\theta_1}{\theta_2}\right) - 1$, em que $0 < \theta_2 < \theta_1$. Algumas manipulações mostram que

$$\lim_{\theta_2 \downarrow \theta_1} \frac{\mathcal{D}_{\text{KL}}(p_{\theta_1} \parallel p_{\theta_2})}{\mathbf{dist}(p_{\theta_1}, p_{\theta_2})} = 0.$$

O exemplo acima nos permite estabelecer a seguinte relação entre distância de Fisher e divergência de Kullback–Leibler através do teorema que se encontra, com sua

respectiva demonstração, em CALIN and UDRISTE (2014) e se segue.

Teorema 2.1 *Seja $d = \mathbf{dist}(p, q)$ a distância de Fisher entre duas funções de densidade p e q e $\mathcal{D}_{KL}(p \parallel q)$ é a divergência de Kullback–Leibler. Então*

$$\mathcal{D}_{KL}(p \parallel q) = \frac{1}{2}d^2(p, q) + o(d^2(p, q)).$$

Com essas ferramentas, nosso próximo passo será apresentar uma classe de funções de densidade de probabilidade mais geral que as famílias exponenciais. Em seguida, vamos caracterizar geometricamente essa nova variedade. Apresentaremos também o conceito de dualidade que estende a relação entre conexões mistura e exponencial aplicadas nessa abordagem mais geral.

3 GEOMETRIA DAS VARIEDADES ESTATÍSTICAS GENERALIZADAS

Neste capítulo apresentaremos as principais contribuições dessa tese relativas à Geometria da Informação. Utilizando-se de uma função que chamaremos de φ -função, a qual possui características similares à função exponencial, propomos uma estrutura de variedade para uma família parametrizada \mathcal{P} , denominada de φ -família, a qual generaliza a família exponencial citada no Capítulo 2. Além disso, propomos uma função de divergência entre distribuições de probabilidade. Com esta será possível definir uma nova métrica e uma família de conexões para variedades estatísticas. Mostraremos que essas entidades geométricas estão intrinsecamente ligadas à φ -função. Ademais, exibiremos uma família de divergências $\mathcal{D}_\varphi^{(\alpha)}(\cdot \parallel \cdot)$, indexadas por um parâmetro α , com a qual obtemos uma outra maneira de recuperar uma família de conexões, em especial, um par de conexões duais. Por fim, alguns resultados clássicos também serão adaptados para \mathcal{P} , tais como os Teoremas de Pitágoras e da Projeção.

3.1 φ -Função

No capítulo anterior, percebemos que uma métrica (5) e uma família de conexões, α -conexão, (16) são induzidas a partir da função $l_\theta = \log(p_\theta)$, com as quais podemos caracterizar o conjunto \mathcal{P} como uma variedade. De igual modo, iremos definir essas entidades geométricas a partir da mudança do logaritmo pela inversa de uma φ -função. Portanto, a generalização citada na introdução desse capítulo provém dessa mudança. Portanto, definamos o que seja uma φ -função.

Definição 3.1 *Uma função $\varphi: \mathbb{R} \rightarrow (0, \infty)$ é denominada φ -função se satisfaz as seguintes condições VIGELIS, SOUZA, and CAVALCANTE (2015):*

- (a1) $\varphi(\cdot)$ é convexa;
- (a2) $\lim_{u \rightarrow -\infty} \varphi(u) = 0$ e $\lim_{u \rightarrow \infty} \varphi(u) = \infty$;
- (a3) existe uma função mensurável $u_0: T \rightarrow (0, \infty)$ tal que

$$\int_T \varphi(c + \lambda u_0) d\mu < \infty, \quad \text{para todo } \lambda > 0, \quad (47)$$

para toda função mensurável $c: T \rightarrow \mathbb{R}$ satisfazendo $\int_T \varphi(c) d\mu = 1$.

A maneira que $\varphi(\cdot)$ é definida nos garante que a sua imagem esteja em um espaço de probabilidades. Além disso, em (a3), o fato de u_0 ser mensurável é uma condição necessária e suficiente para que $\varphi(\cdot)$ também seja mensurável. Observa-se que a função exponencial com $u_0 = 1$ satisfaz as condições (a1)–(a3). Um outro exemplo de φ -função é dado a seguir.

Exemplo 8 (KANIADAKIS and SCARFONE (2002)) A função exponencial de Kaniadakis $\exp_\kappa: \mathbb{R} \rightarrow (0, \infty)$ para $\kappa \in [-1, 1]$ é definida como

$$\exp_\kappa(u) = \begin{cases} (\kappa u + \sqrt{1 + \kappa^2 u^2})^{1/\kappa}, & \text{se } \kappa \neq 0, \\ \exp(u), & \text{se } \kappa = 0. \end{cases}$$

O logaritmo $\log_\kappa(\cdot)$ é o inverso de $\exp_\kappa(\cdot)$, sendo dado por:

$$\log_\kappa(u) = \begin{cases} \frac{u^\kappa - u^{-\kappa}}{2\kappa}, & \text{se } \kappa \neq 0, \\ \log(u), & \text{se } \kappa = 0. \end{cases}$$

Algumas propriedades algébricas da exponencial e do logaritmo convencionais são preservados respectivamente para $\exp_\kappa(\cdot)$ e $\log_\kappa(\cdot)$, como vemos a seguir

$$\exp_\kappa(u) \exp_\kappa(-u) = 1 \quad e \quad \log_\kappa(u) + \log_\kappa(u^{-1}) = 0.$$

Além disso, podemos verificar que a função $\exp_\kappa(\cdot)$ satisfaz diretamente as condições (a1) e (a2). Ademais, seja $u_0: T \rightarrow \mathbb{R}$ uma função mensurável para a qual $\int_T \exp_\kappa(u_0) d\mu < \infty$. Tal restrição tem por objetivo garantir a inserção de uma função $c: T \rightarrow \mathbb{R}$, que tornará a expressão uma função de probabilidade. Vamos mostrar que u_0 satisfaz a propriedade (a3) mostrada na equação (47). Com efeito, dados $u \in \mathbb{R}$ e $\alpha > 1$, podemos escrever que

$$\begin{aligned} \exp_\kappa(\alpha u) &= \alpha^{1/|\kappa|} (|\kappa|u + \sqrt{1/\alpha^2 + |\kappa|^2 u^2})^{1/|\kappa|} \\ &\leq \alpha^{1/|\kappa|} (|\kappa|u + \sqrt{1 + |\kappa|^2 u^2})^{1/|\kappa|} \\ &= \alpha^{1/|\kappa|} \exp_\kappa(u), \end{aligned}$$

em que $\exp_\kappa(\cdot) = \exp_{-\kappa}(\cdot)$. Então, concluímos que $\int_T \exp_\kappa(\alpha u_0) d\mu < \infty$ para todo $\alpha > 0$. Seja ainda qualquer função mensurável $c: T \rightarrow \mathbb{R}$, tal que $\int_T \varphi(c) d\mu = 1$. Para cada $\lambda > 0$, temos

$$\begin{aligned} \int_T \exp_\kappa(c + \lambda u_0) d\mu &\leq \frac{1}{2} \int_T \exp_\kappa(2c) d\mu + \frac{1}{2} \int_T \exp_\kappa(2\lambda u_0) d\mu \\ &\leq 2^{1/|\kappa|-1} \int_T \exp_\kappa(c) d\mu + 2^{1/|\kappa|-1} \int_T \exp_\kappa(\lambda u_0) d\mu \\ &< \infty, \end{aligned}$$

o que nos mostra que $\exp_\kappa(\cdot)$ satisfaz a condição mostrada em (47).

Uma generalização das famílias exponenciais é dada pelas famílias exponenciais deformadas MATSUZOE and HENMI (2013); MATSUZOE (2014). Com efeito, seja $\phi: (0, \infty) \rightarrow$

$(0, \infty)$ uma função crescente. A *função logaritmo deformado* (ou *função ϕ -logaritmo*) $\log_\phi : (0, \infty) \rightarrow (0, \infty)$ é definida por

$$\log_\phi(v) = \int_1^v \frac{1}{\phi(y)} dy, \quad \text{para cada } v > 0.$$

A função logaritmo deformado $\log_\phi(\cdot)$ é côncava e satisfaz $\log_\phi(1) = 0$. No intervalo $(-m, M)$, em que

$$m = \int_0^1 \frac{1}{\phi(y)} dy \quad \text{e} \quad M = \int_1^\infty \frac{1}{\phi(y)} dy,$$

a função $\log_\phi(\cdot)$ tem uma inversa. A *função deformada exponencial* (ou *função ϕ -exponencial*) $\exp_\phi : \mathbb{R} \rightarrow [0, \infty]$ é definida por

$$\exp_\phi(u) = \begin{cases} 0, & \text{se } u \in (-\infty, -m], \\ \log_\phi^{-1}(u), & \text{se } u \in (-m, M), \\ \infty, & \text{se } u \in [M, \infty). \end{cases}$$

A função exponencial deformada $\exp_\phi(\cdot)$ também pode ser expressa em termos de uma integral:

$$\exp_\phi(u) = \int_{-\infty}^u \lambda(x) dx, \quad \text{para todo } u \in \mathbb{R},$$

em que

$$\lambda(u) = \begin{cases} 0, & \text{se } u \in (-\infty, -m], \\ \phi(\exp_\phi(u)), & \text{se } u \in (-m, M), \\ \infty, & \text{se } u \in [M, \infty). \end{cases}$$

Com essas definições, temos que $\phi(v) = \lambda(\log_\phi(v))$ para todo $v > 0$. Quando a função ϕ é a identidade $\phi(u) = u$, as funções $\log_\phi(\cdot)$ e $\exp(\cdot)$ coincidem com as funções logaritmo e exponencial naturais. Vale ressaltar que nosso trabalho, de fato, nesse caso, nem toda exponencial deformada é uma φ -função. Isto é, a φ -função generaliza as famílias deformadas exponencias. Mais precisamente, se a φ -função é tal que $\varphi(0) = 1$, então $\varphi(\cdot)$ é uma exponencial deformada. Funções exponenciais deformadas cuja imagem não esteja contida em $(0, \infty)$ não são φ -funções. Um exemplo desse caso é encontrado em MATSUZOE and HENMI (2013); TSALLIS (1994).

Um fato extra da definição é que podemos relaxar a condição $\int_T \varphi(c) d\mu = 1$ na Definição 3.1, através do lema que se segue.

Lema 3.1 *Seja $\tilde{c} : T \rightarrow \mathbb{R}$ uma função mensurável tal que $\int_T \varphi(\tilde{c}) d\mu < \infty$. Então $\int_T \varphi(\tilde{c} + \lambda u_0) d\mu < \infty$, para todo $\lambda > 0$.*

Prova 1 *Caso $\int_T \varphi(\tilde{c}) d\mu \geq 1$, observe que pela condição (a1), para algum $\alpha > 1$, temos $\int_T \varphi(\tilde{c} - \alpha u_0) d\mu = 1$. Da definição de u_0 , podemos concluir que $\int_T \varphi(\tilde{c} + \lambda u_0) d\mu = \int_T \varphi(c + (\alpha + \lambda) u_0) d\mu < \infty$, em que $c = \tilde{c} - \alpha u_0$. Agora, assumamos que $\int_T \varphi(\tilde{c}) d\mu < 1$ e considere ainda*

cada conjunto mensurável $A \subseteq T$ com medida $0 < \mu(A) < \mu(T)$. Seja $u: T \rightarrow [0, \infty)$ uma função mensurável com suporte em A satisfazendo $\varphi(\tilde{c} + u)\mathbf{1}_A = [\varphi(\tilde{c}) + \alpha]\mathbf{1}_A$, em que $\mathbf{1}_A$ denota a função indicadora com respeito ao conjunto A e $\alpha = (1 - \int_T \varphi(\tilde{c})d\mu)/\mu(A)$. Definamos, para simplicidade de notação, $c = (\tilde{c} + u)\mathbf{1}_A + \tilde{c}\mathbf{1}_{T \setminus A}$, em que $T \setminus A$ denota o conjunto T menos os elementos do conjunto A , podemos ver que $\int_T \varphi(c)d\mu = 1$. Portanto, como $\int_T \varphi(\tilde{c})d\mu < \int_T \varphi(c)d\mu$ e, pela definição de u_0 $\int_T \varphi(c + \lambda u_0)d\mu$ é limitada, concluímos que

$$\int_T \varphi(\tilde{c} + \lambda u_0)d\mu \leq \int_T \varphi(c + \lambda u_0)d\mu < \infty, \quad \text{para qualquer } \lambda > 0,$$

o qual é o resultado desejado.

Nem sempre uma função $\varphi: \mathbb{R} \rightarrow (0, \infty)$ que satisfaça as condições (a1) e (a2) são φ -funções, isto é, também satisfaz a condição (a3). Para testificarmos essa afirmação, vejamos o exemplo que se segue:

Exemplo 9 Assuma que a uma medida μ é σ -finita e não atômica. Este é o caso da medida de Lebesgue. Vamos considerar a função

$$\varphi(u) = \begin{cases} e^{(u+1)^2/2}, & u \geq 0, \\ e^{(u+1/2)}, & u \leq 0, \end{cases}$$

a qual claramente é convexa e satisfaz os limites $\lim_{u \rightarrow -\infty} \varphi(u) = 0$ e $\lim_{u \rightarrow \infty} \varphi(u) = \infty$. Dada qualquer função mensurável $u_0: T \rightarrow (0, \infty)$, podemos encontrar uma função mensurável $c: T \rightarrow \mathbb{R}$ com $\int_T \varphi(c)d\mu < \infty$, para a qual a expressão (47) é não satisfeita.

Com efeito, para cada $m \geq 1$, definimos

$$v_m(t) := \left(m \frac{\log(2)}{u_0(t)} - \frac{u_0(t)}{2} - 1 \right) \mathbf{1}_{E_m}(t),$$

em que $E_m = \{t \in T : m \frac{\log(2)}{u_0(t)} - \frac{u_0(t)}{2} - 1 > 0\}$. Pelo fato de $v_m \uparrow \infty$, podemos encontrar uma subsequência $\{v_{m_n}\}$ tal que

$$\int_{E_{m_n}} e^{(v_{m_n} + u_0 + 1)^2/2} d\mu \geq 2^n.$$

De acordo com o Lema 8.3 de MUSIELAK (1983), existe uma subsequência $w_k = v_{m_{n_k}}$ e conjuntos disjuntos dois a dois $A_k \subseteq E_{m_{n_k}}$ para o qual

$$\int_{E_{m_n}} e^{(v_{m_n} + u_0 + 1)^2/2} d\mu = 1.$$

Vamos considerar o conjunto $c = \bar{c}\mathbf{1}_{T \setminus A} + \sum_{k=1}^{\infty} w_k \mathbf{1}_{A_k}$, em que $A = \bigcup_{k=1}^{\infty} A_k$ e \bar{c} é uma

função mensurável tal que $\varphi(c(t)) > 0$ para $t \in T \setminus A$ e $\int_{T \setminus A} \varphi(\bar{c}) d\mu < \infty$. Observando que

$$e^{(w_k(t)+u_0(t)+1)^2} d\mu = 1.$$

Vamos definir $c = \bar{c} \mathbf{1}_{T \setminus A} + \sum_{k=1}^{\infty} w_k \mathbf{1}_{A_k}$, em que $A = \bigcup_{k=1}^{\infty} A_k$ e \bar{c} é qualquer função mensurável tal que $\varphi(c(t)) > 0$ para $t \in T \setminus A$ e $\int_{T \setminus A} \varphi(\bar{c}) d\mu < \infty$. Observe que

$$e^{(w_k(t)+u_0(t)+1)^2/2} = 2^{m_{n_k}} e^{(w_k(t)+1)^2/2}, \quad \text{para todo } t \in A_k,$$

tomamos

$$\int_{A_k} e^{(w_k+1)^2/2} d\mu = \frac{1}{2^{m_{n_k}}}, \quad \text{para qualquer } m \geq 1.$$

Então, podemos escrever

$$\begin{aligned} \int_T \varphi(u) d\mu &= \int_{T \setminus A} \varphi(\bar{c}) d\mu + \sum_{k=1}^{\infty} \int_{A_k} e^{(w_k+1)^2/2} d\mu \\ &= \int_{T \setminus A} \varphi(\bar{c}) d\mu + \sum_{k=1}^{\infty} \frac{1}{2^{m_{n_k}}} < \infty. \end{aligned}$$

Por outro lado,

$$\begin{aligned} \int_T \varphi(c + u_0) d\mu &= \int_{T \setminus A} \varphi(\bar{c}) d\mu + \sum_{k=1}^{\infty} \int_{A_k} e^{(u_0+w_k+1)^2/2} d\mu \\ &= \int_{T \setminus A} \varphi(\bar{c}) d\mu + \sum_{k=1}^{\infty} 1 = \infty, \end{aligned}$$

a qual mostra que (47) é não satisfeita ■.

Ademais, podemos generalizar a formalização de variedade que é dada em uma família exponencial para uma família que será construída através da φ -função. É o que veremos na próxima seção.

3.2 Variedades Estatísticas Generalizadas

Mais uma vez, sejam (T, Σ, μ) um espaço de medida, L^0 o conjunto de todas as funções mensuráveis reais em T e \mathcal{P}_μ a família de distribuições adicionada de uma estrutura geométrica de variedade diferenciável, a qual é definida por VIGELIS and CAVALCANTE (2013b)

$$\mathcal{P}_\mu = \left\{ p \in L^0 : \int_T p d\mu = 1 \quad \text{e} \quad p > 0 \right\}. \quad (48)$$

Consideremos o modelo identificável $\mathcal{P}_G \subset \mathcal{P}_\mu$, em que

$$\mathcal{P}_G = \{ p_\theta : \theta \in \Theta \subseteq \mathbb{R}^n \}, \quad (49)$$

e p_θ é identificável com o parâmetro θ , se as seguintes condições abaixo são satisfeitas:

- (b1) Θ é um domínio em \mathbb{R}^n ;
- (b2) $p(t; \theta)$ é diferenciável em relação a θ ;
- (b3) a matriz (g_{ij}) da métrica generalizada é definida como

$$g_{ij} = -\frac{\int_T \frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j} \varphi'(f_\theta) d\mu}{\int_T u_0 \varphi'(f_\theta) d\mu}, \quad (50)$$

em que a função $f_\theta: T \rightarrow \mathbb{R}$ é definida por

$$f_\theta := \varphi^{-1}(p_\theta), \quad p_\theta \in \mathcal{P}_G. \quad (51)$$

Observe que f_θ tem a mesma motivação para sua definição em l_θ na família exponencial, como podemos comparar com a equação (6). É natural a expressão da métrica de Fisher, pois equivale-se à matriz de covariância. Ao longo desse trabalho, para efeito de simplificação, utilizaremos a seguinte notação que se segue:

$$E_\theta^{(n)}[\cdot] := \frac{\int_T (\cdot) \varphi^{(n)}(f_\theta) d\mu}{\int_T u_0 \varphi'(f_\theta) d\mu}, \quad (52)$$

em que $\varphi^{(n)}$ representa a derivada da φ função em relação a qualquer parâmetro θ^i . Particularmente, usaremos as notações $E'_\theta[\cdot]$, $E''_\theta[\cdot]$ e $E'''_\theta[\cdot]$ para denotar o caso particular de $E_\theta^{(n)}[\cdot]$, com $n = 1, 2, 3$, respectivamente. Isto é,

$$E'_\theta[\cdot] := \frac{\int_T (\cdot) \varphi'(f_\theta) d\mu}{\int_T u_0 \varphi'(f_\theta) d\mu}, \quad E''_\theta[\cdot] := \frac{\int_T (\cdot) \varphi''(f_\theta) d\mu}{\int_T u_0 \varphi'(f_\theta) d\mu} \quad (53)$$

e

$$E'''_\theta[\cdot] := \frac{\int_T (\cdot) \varphi'''(f_\theta) d\mu}{\int_T u_0 \varphi'(f_\theta) d\mu}. \quad (54)$$

Como consequência das notações mencionadas acima, podemos reescrever a expressão da métrica de Fisher dada em (50) por

$$g_{ij}(\theta) = -E'_\theta \left[\frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j} \right]. \quad (55)$$

É interessante observar que a expressão da métrica de Fisher para \mathcal{P}_G (55) tem a mesma forma da expressão da expressão da métrica de Fisher para as famílias exponenciais dada em (10). Além do mais, devido às condições de regularidade assumidas para a φ -família, pode-se permutar operador integral com respeito à medida μ e o operador derivada em relação ao parâmetro θ^j . Como a expressão (51) pode ser interpretada por $p_\theta = \varphi(f_\theta)$ e

$\int_T p_\theta d\mu = 1$, ao derivarmos essa expressão em relação a θ^i , obtemos

$$0 = \frac{\partial}{\partial \theta^i} \int_T p_\theta d\mu = \int_T \frac{\partial \varphi(f_\theta)}{\partial \theta^i} d\mu = \int_T \frac{\partial f_\theta}{\partial \theta^i} \varphi'(f_\theta) d\mu. \quad (56)$$

Uma interessante consequência é que ao derivarmos a equação (56) em relação ao parâmetro θ^j , temos que

$$0 = \int_T \frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j} \varphi'(f_\theta) d\mu + \int_T \frac{\partial f_\theta}{\partial \theta^i} \frac{\partial f_\theta}{\partial \theta^j} \varphi''(f_\theta) d\mu, \quad (57)$$

e assim, podemos exibir uma nova expressão para a métrica (55). De fato, com a equação (57), temos que

$$g_{ij}(\theta) = E_\theta'' \left[\frac{\partial f_\theta}{\partial \theta^i} \frac{\partial f_\theta}{\partial \theta^j} \right] = Cov_\theta \left[\frac{\partial f_\theta}{\partial \theta^i}, \frac{\partial f_\theta}{\partial \theta^j} \right]. \quad (58)$$

Mais uma vez, percebe-se a equivalência entre a expressão (58) dada pela parametrização através da função $\varphi(\cdot)$ e a que obtemos na família exponencial (5).

Uma família paramétrica de distribuições pode ser estruturada por uma métrica não riemanniana SALMON *et al.* (2014). Porém, considerando uma família paramétrica de distribuições, como as φ -família, a escolha da métrica de Fisher é devida a algumas propriedades que bem se ajustam as famílias de distribuições de probabilidade AMARI (2013). Em LEBANON (2005), os autores mostram que uma versão da métrica de Fisher aplicada à geometria da informação quântica. Para essa tese, a escolha como da métrica da informação para variedades estatísticas paramétricas tem as suas justificativas. No Lemma 11.3 de CENCOV (2000), prova-se que a métrica provinda da matriz de informação de Fisher é a única métrica que é invariante por morfismos de Markov. Isto é, o produto interno de campos de vetores em um ponto da variedade estatística podem ser preservados. Posteriormente, CAMPBELL (1986) estendeu os resultados de Čencov para modelos positivos não normalizados. Posteriormente, mais uma vez por Lebanon_2014 LI *et al.* (2014), obtém-se uma versão do teorema da caracterização geométrica para um conjunto de modelos condicionais não normalizados.

O teorema da caracterização é aplicado sobre um conjunto denotado por $\text{Caph}(\Omega, \mathbb{S}, \Theta)$ de todas as distribuições de probabilidade em uma álgebra \mathbb{S} finita, tal que os átomos A_j sejam probabilidades estritamente positivas, isto é

$$p_j = P\{A_j\} > 0, j = 1, \dots, m; \quad \sum_{i=1}^m p_j = 1.$$

Portanto, observa-se que os resultados de Čencov e os subsequentes ao seu, citados no parágrafo anterior, aplicam-se também à φ -família. Em especial, os dois que se seguem:

Lema 3.2 *Existe uma única métrica Riemanniana invariante nas distribuições como em*

$\text{Caph}(\Omega, \mathbb{S}, \Theta)$. Além disso, esta métrica é definida pela forma quadrática da forma

$$ds^2 = \int_{\Omega} \frac{(\delta P)\{d\omega\}(\delta P)\{d\omega\}}{P\{d\omega\}}, \quad (59)$$

em que $\delta P\{\cdot\}$ é a diferencial de P .

Dado o espaço paramétrico Θ e se consideramos as coordenadas (parâmetros) $\theta^1, \dots, \theta^n \in \Theta$ da φ -família, a expressão dada em (59) é reescrita como

$$ds^2 = \sum_{i,j=1}^k d\theta^i d\theta^j \mathbf{M}_P \frac{\partial \ln p}{\partial \theta^i} \frac{\partial \ln p}{\partial \theta^j}$$

com a matriz de informação de Fisher g dada por

$$g_{ij} = \mathbf{M}_P \frac{\partial \ln p}{\partial \theta^i} \frac{\partial \ln p}{\partial \theta^j} = \int_{\Omega} \frac{\partial p(\omega)}{\partial \theta^i} \frac{\partial p(\omega)}{\partial \theta^j} \frac{\mu\{d\omega\}}{p(\omega)},$$

em que $\mathbf{M}_P[\cdot] = E'_{\theta}[\cdot]$.

Definição 3.2 A família paramétrica \mathcal{P}_G munida da métrica g_{ij} dada em (50) é denominada variedade estatística generalizada.

A expressão da métrica de \mathcal{P}_G dada em (55) nos mostra que o espaço tangente $T_{p_{\theta}}\mathcal{P}_G$ também pode ser identificado com o espaço tangente $\tilde{T}_{p_{\theta}}\mathcal{P}_G$ VAILLANT *et al.* (2004). Mais precisamente, a segunda expressão da métrica (58) nos permite relacionar as funções $\partial f_{\theta}/\partial \theta^i$ com a base vetorial composta dos vetores $\partial/\partial \theta^i$. De fato, sejam os campos de vetores X e Y de $T_{p_{\theta}}\mathcal{P}$ dados por CARMO (1976a)

$$X = \sum_i a^i \frac{\partial}{\partial \theta^i} \quad \text{e} \quad Y = \sum_j b^j \frac{\partial}{\partial \theta^j}, \quad (60)$$

em que os campos (60) são as funções dadas por

$$\sum_i a^i \frac{\partial}{\partial \theta^i} \mapsto \sum_i a^i \frac{\partial f_{\theta}}{\partial \theta^i}$$

e

$$\sum_j b_j \frac{\partial}{\partial \theta^j} \mapsto \sum_j b_j \frac{\partial f_{\theta}}{\partial \theta^j}.$$

O produto interno dado pela métrica de Fisher (55) é definido por VAILLANT *et al.* (2004)

$$g(X, Y) = \sum_{i,j} g_{ij} a^i b^j = \sum_{i,j} E''_{\theta} \left[\frac{\partial f_{\theta}}{\partial \theta^i} \frac{\partial f_{\theta}}{\partial \theta^j} \right] a^i b^j = E''_{\theta}[\tilde{X}\tilde{Y}], \quad (61)$$

em que

$$\tilde{X} = \sum_i a^i \frac{\partial f_{\theta}}{\partial \theta^i} \quad \text{e} \quad \tilde{Y} = \sum_j b^j \frac{\partial f_{\theta}}{\partial \theta^j}. \quad (62)$$

com os quais podemos definir um produto interno dado por

$$\langle \tilde{X}, \tilde{Y} \rangle_\theta = E''_\theta[\tilde{X}\tilde{Y}].$$

Além disso, a expressão dada em (57) nos garante que cada campo \tilde{X} pertencente a $\tilde{T}_{p_\theta}\mathcal{P}_G$ satisfaz $E'_\theta[\tilde{X}] = 0$.

Sejam Γ_{ij}^k os símbolos de Christoffel da conexão de Levi-Civita ∇ associada com a métrica g_{ij} riemanniana. Com isso, os símbolos

$$\Gamma_{ijk} := \sum_m \Gamma_{ij}^m g_{mk} = \frac{1}{2} \left(\frac{\partial g_{ki}}{\partial \theta^j} + \frac{\partial g_{kj}}{\partial \theta^i} - \frac{\partial g_{ij}}{\partial \theta^k} \right) \quad (63)$$

são obtidos em termos de g_{ij} . Diferenciando com relação a θ^k a expressão para a métrica dada em (58), temos

$$\begin{aligned} \partial g_{ij} &= E''_\theta \left[\frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^k} \frac{\partial f_\theta}{\partial \theta^j} \right] + E''_\theta \left[\frac{\partial f_\theta}{\partial \theta^i} \frac{\partial^2 f_\theta}{\partial \theta^j \partial \theta^k} \right] \\ &+ E'''_\theta \left[\frac{\partial f_\theta}{\partial \theta^i} \frac{\partial f_\theta}{\partial \theta^j} \frac{\partial f_\theta}{\partial \theta^k} \right] - E''_\theta \left[\frac{\partial f_\theta}{\partial \theta^i} \frac{\partial f_\theta}{\partial \theta^j} \right] E''_\theta \left[u_0 \frac{\partial f_\theta}{\partial \theta^k} \right]. \end{aligned}$$

Portanto, usando os equivalentes para $\frac{\partial g_{ki}}{\partial \theta^j}$ e $\frac{\partial g_{kj}}{\partial \theta^i}$, temos que a conexão de Levi-Civita para \mathcal{P}_G é dada por

$$\begin{aligned} \Gamma_{ijk} &= E''_\theta \left[\frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j} \frac{\partial f_\theta}{\partial \theta^k} \right] + \frac{1}{2} E'''_\theta \left[\frac{\partial f_\theta}{\partial \theta^i} \frac{\partial f_\theta}{\partial \theta^j} \frac{\partial f_\theta}{\partial \theta^k} \right] \\ &- \frac{1}{2} E''_\theta \left[\frac{\partial f_\theta}{\partial \theta^k} \frac{\partial f_\theta}{\partial \theta^i} \right] E''_\theta \left[u_0 \frac{\partial f_\theta}{\partial \theta^j} \right] \\ &- \frac{1}{2} E''_\theta \left[\frac{\partial f_\theta}{\partial \theta^j} \frac{\partial f_\theta}{\partial \theta^k} \right] E''_\theta \left[u_0 \frac{\partial f_\theta}{\partial \theta^i} \right] \\ &+ \frac{1}{2} E''_\theta \left[\frac{\partial f_\theta}{\partial \theta^i} \frac{\partial f_\theta}{\partial \theta^j} \right] E''_\theta \left[u_0 \frac{\partial f_\theta}{\partial \theta^k} \right] \end{aligned} \quad (64)$$

Evidenciamos que para $f_\theta = \varphi^{-1}(p_\theta)$ e $u_0 = 1$, a expressão (63) é um caso particular de (64).

3.3 φ -Família

Nesta seção, iremos definir e expor alguns resultados de geometria da informação na variedade estatística que chamaremos de φ -família. Através da troca da função exponencial por uma φ -função como função de parametrização, vamos generalizar os resultados de cunho geométrico das famílias exponenciais discutidos no Capítulo 2. Seja $\varphi(\cdot)$ definida conforme a Definição 3.1 e $c: T \rightarrow \mathbb{R}$ uma função mensurável para a qual $p = \varphi(c)$ seja uma distribuição de probabilidade. Uma (paramétrica) φ -família $\mathcal{F}_p = \{p_\theta : \theta \in \Theta\}$ centrada em $p = \varphi(c)$ é o conjunto de todas as distribuições de probabilidade p_θ em \mathcal{P}

que podem ser escritas da forma VIGELIS, SOUZA, and CAVALCANTE (2015)

$$p_\theta := \varphi\left(c + \sum_{i=1}^n \theta^i u_i - \psi(\theta)u_0\right), \quad \text{para cada } \theta = (\theta^i) \in \Theta, \quad (65)$$

em que $\psi: \Theta \rightarrow \mathbb{R}$ é a *função de normalização* e as funções $u_1, \dots, u_n: T \rightarrow \mathbb{R}$ são mensuráveis, não triviais e linearmente independentes. Isto é, ψ é introduzida na expressão de p_θ para que esta seja uma distribuição de probabilidade em \mathcal{P} . Para o caso da família exponencial, a expressão de $\psi(\theta)$ é dada como na expressão (26).

As funções pertencentes à φ -família citadas em (65) satisfazem as seguintes propriedades VIGELIS and CAVALCANTE (2013b):

(c1) Para cada $1 \leq k \leq n$, existe $\varepsilon > 0$, tal que

$$\int_T \varphi(c + \lambda u_k) d\mu < \infty, \quad \text{para todo } \lambda \in (-\varepsilon, \varepsilon).$$

(c2) $\int_T u_k \varphi'_+(c) d\mu = 0$, em que $\varphi'_+(\cdot)$ denota a derivada à direita de $\varphi(\cdot)$.

(c3) $\int_T \varphi(c + \lambda \sum_{k=1}^n \theta^k u_k) d\mu < \infty$, para algum $\lambda > 1$, em que $\theta = (\theta^1, \dots, \theta^n) \in \Theta$.

Além disso, o domínio $\Theta \subseteq \mathbb{R}^n$ é definido como o conjunto de todos os vetores $\theta = (\theta^i)$ para os quais

$$\int_T \varphi\left(c + \lambda \sum_{i=1}^n \theta^i u_i\right) < \infty, \quad \text{para algum } \lambda > 1.$$

A condição (c2) garante que a função ψ seja não negativa. De fato, por conta da convexidade de $\varphi(\cdot)$, podemos escrever

$$\int_T \varphi(c) d\mu = \int_T \left[\varphi(c) + \left(\sum_{i=1}^n \theta^i u_i \right) \varphi'(c) \right] d\mu \leq \int_T \varphi\left(c + \sum_{i=1}^n \theta^i u_i\right) d\mu,$$

o que implica que $\psi(\theta) \geq 0$. Pela condição (c3), deduzimos que Θ é uma vizinhança aberta da origem. Pode-se mostrar que Θ é um domínio aberto e convexo FAN (1952). As condições (c1)–(c3) também são utilizadas para caracterizar φ -famílias não paramétricas VIGELIS and CAVALCANTE (2013b,a).

Observe que a condição (c3) sempre é satisfeita quando T é finito. Comparando-se as igualdades

$$\frac{\partial f_\theta}{\partial \theta^i} = u_i - \frac{\partial \psi}{\partial \theta^i}, \quad \text{e} \quad -\frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j} = -\frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j},$$

segue-se que

$$g_{ij} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}. \quad (66)$$

A expressão (66) é a Hessiana da função ψ . Portanto, concluímos que a φ -família \mathcal{F}_p é uma *variedade hessiana* SHIMA (2007). Da expressão (65), aplicando a inversa da φ -função, podemos escrever que

$$u_0\psi(\theta) = \sum_{i=1}^n \theta^i u_i + \varphi^{-1}(p_\theta) - \varphi^{-1}(\tilde{p}_\theta), \quad (67)$$

em que $\tilde{p}_\theta = \varphi(c)$. Com a condição (c2) satisfeita, escreve-se

$$\psi(\theta) = \int_T u_0 \varphi'(c) d\mu = \int_T [\varphi^{-1}(p) - \varphi^{-1}(\tilde{p})] \varphi'(c) d\mu.$$

Usando a derivada da função inversa LIMA (2004), $\varphi'(c) = \frac{1}{(\varphi^{-1})'(p)}$ e obtemos

$$\psi(\theta) = \frac{\int_T \frac{\varphi^{-1}(p) - \varphi^{-1}(\tilde{p})}{(\varphi^{-1})'(p)} d\mu}{\int_T \frac{u_0}{(\varphi^{-1})'(p)} d\mu}. \quad (68)$$

A expressão obtida acima será de extrema importância, pois com ela associaremos uma função de divergência. Para entender o porque, observe que para o caso evidenciado em (20), a expressão de $\psi(\theta)$ é dado por

$$\psi(\theta) = \int_T \log\left(\frac{p}{p_\theta}\right) p d\mu$$

provinda da família exponencial centrada em p (20), a qual coincide com a divergência de Kullback–Leibler, é um caso particular de (68).

Na seção seguinte, vamos falar sobre funções de divergência. Iremos generalizar o conceito de divergência de Rényi e, assim como nas famílias exponenciais, vamos construir a geometria induzida da φ -família.

3.4 Generalização da Divergência de Rényi

Nessa seção, essencialmente mostramos que podemos generalizar as divergências de Kullback–Leibler e de Rényi através da função normalizadora. Especialmente, trabalharemos com a divergência de Rényi que também é uma medida de informação que depende de um parâmetro α que é chamado de *ordem de divergência* KRISHNAMURTHY *et al.* (2014). A importância dessa função se concentra nos estudos em áreas de aplicação, como podemos citar os trabalhos em CSISZÁR (1995); HARREMOËS (2006). Para maiores detalhes sobre as divergências de Kullback–Leibler e Rényi, ver VAN ERVEN and HARREMOS (2014).

Desenvolveremos a generalização da divergência de Rényi citada a partir de uma φ -função. De igual modo no caso clássico VAN ERVEN and HARREMOËS (2010),

essa generalização dependerá do parâmetro $\alpha \in [-1, 1]$, em que precisamente em $\alpha = \pm 1$, faremos uso de limites.

Para T finito, a divergência de Rényi de ordem $\alpha > 0$, $\alpha \neq 1$, de uma distribuição de probabilidade $p = (p_1, \dots, p_n)$ em relação a outra $q = (q_1, \dots, q_n)$ é dada por VAN ERVEN and HARREMOËS (2010); VAN ERVEN and HARREMOS (2014)

$$\mathcal{D}_R^{(\alpha)}(p \parallel q) = \frac{1}{1 - \alpha} \log \left(\sum_{i=1}^n p_i^\alpha q_i^{1-\alpha} \right). \quad (69)$$

Para o caso T contínuo, quando $p, q \in \mathcal{P}_G$ forem duas distribuições de probabilidade, temos VAN ERVEN and HARREMOS (2014)

$$\mathcal{D}_R^{(\alpha)}(p \parallel q) = \frac{1}{\alpha(\alpha - 1)} \log \left(\int_T p^\alpha q^{1-\alpha} d\mu \right). \quad (70)$$

Para os casos de $\alpha = 0$ e $\alpha = 1$, a divergência de Kullback–Leibler é recuperada com o uso de limites. Supondo a existência dos limites, temos

$$\mathcal{D}_R^{(0)}(p \parallel q) = \lim_{\alpha \downarrow 0} \mathcal{D}_R^{(\alpha)}(p \parallel q) \quad (71)$$

e

$$\mathcal{D}_R^{(1)}(p \parallel q) = \lim_{\alpha \uparrow 1} \mathcal{D}_R^{(\alpha)}(p \parallel q), \quad (72)$$

definimos (ou recuperamos) a divergência de Kullback–Leibler que é dada por VAN ERVEN and HARREMOS (2014)

$$\mathcal{D}_{KL}(p \parallel q) = \int_T p \log \left(\frac{p}{q} \right) d\mu.$$

Costumeiramente, a divergência de Rényi é utilizada para todo valor $\alpha \in \mathbb{R}$. Porém, observa-se que utilizando um valor de $\alpha \notin [0, 1]$, temos uma inconsistência em (70) ao obtermos valores negativos para quaisquer distribuições de probabilidade p e q . Para evitarmos tais problemas, consideremos $\alpha \in [0, 1]$. Além disso, salientamos que a expressão (70) difere da versão clássica dada por VAN ERVEN and HARREMOS (2014), cuja expressão é dada por

$$\mathcal{D}^{(\alpha)}(p \parallel q) = \frac{1}{1 - \alpha} \log \left(\int_T p^\alpha q^{1-\alpha} d\mu \right) \quad (73)$$

com um parâmetro α . Tal mudança permite em se beneficiar de propriedades de simetria quando avaliarmos a divergência quando $\alpha \downarrow 0$ e $\alpha \uparrow 1$. Como estratégia de motivação para a generalização da divergência de Rényi adaptada em \mathcal{F}_p , vamos reescrever a

expressão (70) como

$$\mathcal{D}_R^{(\alpha)} = \frac{\kappa(\alpha)}{\alpha(1-\alpha)}, \quad \text{para todo } \alpha \in (0, 1), \quad (74)$$

em que

$$\kappa(\alpha) = -\log\left(\int_T p^\alpha q^{1-\alpha} d\mu\right), \quad \text{para todo } \alpha \in [0, 1]. \quad (75)$$

Ademais, a expressão (74) define implicitamente $\kappa(\alpha)$, pois podemos, dados p e q , escolher um único valor de $\kappa(\alpha)$ tal que a expressão

$$\int_T \exp(\alpha \log(p) + (1-\alpha) \log(q) + \kappa(\alpha)) d\mu = 1 \quad (76)$$

seja satisfeita, isto é, o termo do integrando seja uma função densidade de probabilidade. De maneira análoga, usaremos a expressão (74) para definir uma família de divergência entre duas distribuições de probabilidade p e q em \mathcal{F}_p . O detalhe que difere é que agora dados p e q , ao invés de $\kappa(\alpha)$ satisfazer (76), satisfaz

$$\int_T \varphi(\alpha \varphi^{-1}(p) + (1-\alpha) \varphi^{-1}(q) + \kappa(\alpha) u_0) d\mu = 1, \quad (77)$$

desde que uma φ -função tenha sido pré-fixada. Assim, definimos uma generalização para a divergência de Rényi por

$$\mathcal{D}_{R,\varphi}^{(\alpha)}(p \parallel q) = \frac{\kappa(\alpha)}{\alpha(1-\alpha)}, \quad \text{com } \alpha \in (0, 1), \quad (78)$$

desde que $\kappa(\alpha)$ satisfaça (77). Para $\alpha \in \{0, 1\}$, a generalização é definida pelos limites

$$\mathcal{D}_{R,\varphi}^{(0)}(p \parallel q) = \lim_{\alpha \downarrow 0} \mathcal{D}_{R,\varphi}^{(\alpha)}(p \parallel q) \quad (79)$$

e

$$\mathcal{D}_{R,\varphi}^{(1)}(p \parallel q) = \lim_{\alpha \uparrow 1} \mathcal{D}_{R,\varphi}^{(\alpha)}(p \parallel q). \quad (80)$$

Porém, ainda precisamos garantir que as expressões dos limites em (79) e (80) existam e sejam finitas. Além do mais, observamos que, assumindo a convergência nos limites, temos

$$\mathcal{D}_R^{(0)}(q \parallel p) = \mathcal{D}_R^{(1)}(p \parallel q) = \mathcal{D}_{\text{KL}}(p \parallel q) < \infty. \quad (81)$$

Nesse sentido, o lema abaixo será uma ferramenta útil.

Lema 3.3 *Suponha que $\varphi(\cdot)$ seja continuamente diferenciável. Se para $\alpha_0, \alpha_1 \in \mathbb{R}$, a expressão*

$$\int_T \varphi(\alpha \varphi^{-1}(p) + (1-\alpha) \varphi^{-1}(q) + \kappa(\alpha)) d\mu < \infty \quad (82)$$

é satisfeita para todo $\alpha \in [\alpha_0, \alpha_1]$, então a derivada de $\kappa(\alpha)$ existe para qualquer $\alpha \in (\alpha_0, \alpha_1)$, e é dada por

$$\frac{\partial \kappa}{\partial \alpha}(\alpha) = \frac{\int_T [\varphi^{-1}(q) - \varphi^{-1}(p)] \varphi'(c_\alpha) d\mu}{\int_T \varphi'(c_\alpha) u_0 d\mu}, \quad (83)$$

em que $c_\alpha = \alpha \varphi^{-1}(p) + (1 - \alpha) \varphi^{-1}(q) + \kappa(\alpha) u_0$.

Prova 2 Para $\alpha \in (\alpha_0, \alpha_1)$ e $\kappa > 0$, defina

$$g(\alpha, \kappa) = \int_T \varphi(\alpha \varphi^{-1}(p) + (1 - \alpha) \varphi^{-1}(q) + \kappa u_0) d\mu.$$

Como já citado anteriormente, podemos definir a função $\kappa(\alpha)$ de forma implícita por $g(\alpha, \kappa(\alpha)) = 1$. Se mostrarmos que:

- (i) a função $g(\alpha, \kappa)$ é contínua numa vizinhança de $(\alpha, \kappa(\alpha))$;
- (ii) as derivadas parciais $\frac{\partial g}{\partial \alpha}$ e $\frac{\partial g}{\partial \kappa}$ existem e são contínuas em $(\alpha, \kappa(\alpha))$;
- (iii) e $\frac{\partial g}{\partial \kappa}(\alpha, \kappa(\alpha)) > 0$,

então, pelo Teorema da Função Implícita LIMA (2004), $\kappa(\alpha)$ é diferenciável em $\alpha \in (\alpha_0, \alpha_1)$ e

$$\frac{\partial \kappa}{\partial \alpha}(\alpha) = -\frac{(\partial g / \partial \alpha)(\alpha, \kappa(\alpha))}{(\partial g / \partial \kappa)(\alpha, \kappa(\alpha))}. \quad (84)$$

Com efeito, mostremos a continuidade de $g(\alpha, \kappa)$. Fixado $\alpha \in (\alpha_0, \alpha_1)$ e $\kappa > 0$, seja $\kappa_0 = 2\kappa$. Denotando $A = \{t \in T : \varphi^{-1}(p(t)) > \varphi^{-1}(q(t))\}$, podemos escrever

$$\begin{aligned} \varphi(\beta \varphi^{-1}(p) + (1 - \beta) \varphi^{-1}(q) + \lambda u_0) &\leq \varphi(\varphi^{-1}(q) + \beta[\varphi^{-1}(p) - \varphi^{-1}(q)] + \kappa_0 u_0) \\ &\leq \varphi(\varphi^{-1}(q) + \alpha_1[\varphi^{-1}(p) - \varphi^{-1}(q)] + \kappa_0 u_0) \mathbf{1}_A \\ &\quad + \varphi(\varphi^{-1}(q) + \alpha_0[\varphi^{-1}(p) - \varphi^{-1}(q)] + \kappa_0 u_0) \mathbf{1}_{T \setminus A}, \end{aligned} \quad (85)$$

para cada $\beta \in (0, \alpha_0)$ e $\lambda \in (0, \kappa_0)$ em que $\mathbf{1}_A$ é a função indicadora de um conjunto A . Devido a função do lado direito de (85) ser integrável, pelo Teorema da Convergência Dominada SCHEFFÉ (1947), concluímos que

$$\lim_{(\beta, \lambda) \rightarrow (\alpha, \kappa)} g(\beta, \lambda) = g(\alpha, \kappa).$$

Agora, vamos mostrar que a derivada de $g(\alpha, \kappa)$ com respeito a α existe e é contínua. Considere a diferença

$$\frac{g(\gamma, \lambda) - g(\beta, \lambda)}{\gamma - \beta} = \int_T \frac{1}{\gamma - \beta} [\varphi(\tilde{c}_\beta + (\gamma - \beta)[\varphi^{-1}(p) - \varphi^{-1}(q)] + \lambda u_0) - \varphi(\tilde{c}_\beta + \lambda u_0)] d\mu,$$

em que $\tilde{c}_\beta = \beta \varphi^{-1}(p) + (1 - \beta) \varphi^{-1}(q)$. Represente por $f_{\beta, \gamma, \lambda}$ a função dentro da integral. Fixado $\alpha \in (\alpha_0, \alpha_1)$ e $\kappa > 0$, denote $\bar{\alpha}_0 = (\alpha_0 + \alpha)/2$, $\bar{\alpha}_1 = (\alpha + \alpha_1)/2$, e $\kappa_0 = 2\kappa$.

Porque $\varphi(\cdot)$ é convexo e crescente, segue que

$$|f_{\beta,\gamma,\lambda}| \leq f_{\bar{\alpha}_1,\alpha_1,\kappa_0} \mathbf{1}_A - f_{\bar{\alpha}_0,\alpha_0,\kappa_0} \mathbf{1}_{T \setminus A} =: f, \quad \text{para todo } \beta, \gamma \in (\bar{\alpha}_0, \bar{\alpha}_1) \text{ e } \lambda \in (0, \kappa_0),$$

em que $A = \{t \in T : \varphi^{-1}(p(t)) > \varphi^{-1}(q(t))\}$. Observando que f é integrável, nós podemos usar mais uma vez o Teorema da Convergência Dominada para obter

$$\lim_{\gamma \rightarrow \beta} \int_T f_{\beta,\gamma,\lambda} d\mu = \int_T \left(\lim_{\gamma \rightarrow \beta} f_{\beta,\gamma,\lambda} \right) d\mu,$$

e então,

$$\frac{\partial g}{\partial \alpha}(\beta, \lambda) = \int_T [\varphi^{-1}(p) - \varphi^{-1}(q)] \varphi'(\tilde{c}_\beta + \lambda u_0) d\mu. \quad (86)$$

Para $\beta \in (\bar{\alpha}_0, \bar{\alpha}_1)$ e $\lambda \in (0, \kappa_0)$, a função dentro do sinal da integral (86) é dominado por f . O Teorema da Convergência Dominada também mostra que $\frac{\partial g}{\partial \alpha}$ é contínua em (α, κ) :

$$\lim_{(\beta,\lambda) \rightarrow (\alpha,\kappa)} \frac{\partial g}{\partial \alpha}(\beta, \lambda) = \frac{\partial g}{\partial \alpha}(\alpha, \kappa).$$

Com argumentos similares, pode-se mostrar que $\frac{\partial g}{\partial \kappa}(\alpha, \kappa)$ existe e é contínua para qualquer $\alpha \in (\alpha_0, \alpha_1)$ e $\kappa > 0$, a qual é dada por

$$\frac{\partial g}{\partial \kappa}(\alpha, \kappa) = \int_T u_0 \varphi'(\tilde{c}_\alpha + \kappa u_0) d\mu. \quad (87)$$

Claramente, a expressão (87) implica que $\frac{\partial g}{\partial \kappa}(\alpha, \kappa) > 0$ para todo $\alpha \in (0, \alpha_0)$ e $\kappa > 0$.

Portanto, fazendo uso do lema 3.3, ratificamos que as expressões (79) e (80) estão bem definidas, ao consideramos a existência dos limites. Com efeito, suponha que $\varphi(\cdot)$ seja continuamente diferenciável.

(a) Se para algum $\alpha_0 > 1$ a expressão (82) é satisfeita para todo $\alpha \in (1, \alpha_0]$, então

$$\mathcal{D}_{R,\varphi}^{(1)}(p \parallel q) = \lim_{\alpha \uparrow 1} \mathcal{D}_{R,\varphi}^{(\alpha)}(p \parallel q) = -\frac{\partial \kappa}{\partial \alpha}(1) = \mathcal{D}_\varphi(p \parallel q) < \infty.$$

(b) Se para algum $\alpha_0 < 0$ a expressão (82) é satisfeita para $[\alpha_0, 0)$, então

$$\mathcal{D}_{R,\varphi}^{(0)}(p \parallel q) = \lim_{\alpha \downarrow 0} \mathcal{D}_{R,\varphi}^{(\alpha)}(p \parallel q) = \frac{\partial \kappa}{\partial \alpha}(0) = \mathcal{D}_\varphi(q \parallel p) < \infty.$$

Agora, vamos obter uma versão da divergência generalizada de Rényi para $\alpha \in [-1, 1]$. Essa nova versão está motivada no interesse de recuperar um par de conexões duais, cujos coeficientes são $\Gamma_{ijk}^{(1)}$ e $\Gamma_{ijk}^{(-1)}$. Com efeito, seja $\kappa(\alpha) = \kappa(\alpha; p, q) \geq 0$, tal que

$$\varphi\left(\frac{1-\alpha}{2}\varphi^{-1}(p) + \frac{1+\alpha}{2}\varphi^{-1}(q) + \kappa(\alpha)u_0\right),$$

seja uma distribuição de probabilidade em \mathcal{P} . Como já citado, temos que $\kappa(\alpha)$ é único devido a propriedade (a3) na definição de φ -família. Então, a α -divergência entre duas distribuições em \mathcal{P} é bem definida através de

$$\mathcal{D}_\varphi^{(\alpha)}(p \parallel q) = \frac{4}{1 - \alpha^2} \kappa(\alpha), \quad \alpha \in (-1, 1). \quad (88)$$

Para $\alpha = 1$ e $\alpha = -1$, definimos $\mathcal{D}_\varphi^{(1)}$ e $\mathcal{D}_\varphi^{(-1)}$ através dos limites abaixo

$$\mathcal{D}_\varphi^{(1)}(p \parallel q) = \lim_{\alpha \uparrow 1} \frac{4}{1 - \alpha^2} \kappa(\alpha) = -2\kappa'(1) \quad (89)$$

e

$$\mathcal{D}_\varphi^{(-1)}(p \parallel q) = \lim_{\alpha \downarrow -1} \frac{4}{1 - \alpha^2} \kappa(\alpha) = 2\kappa'(-1), \quad (90)$$

em que $\kappa'(1)$ e $\kappa'(-1)$ são as derivadas de $\kappa(\alpha)$ em $\alpha = 1$ e $\alpha = -1$, respectivamente. Podemos obter as expressões de $\kappa'(1)$ e $\kappa'(-1)$. Consideremos a seguinte equação

$$\int_T \varphi \left(\frac{1 - \alpha}{2} \varphi^{-1}(p) + \frac{1 + \alpha}{2} \varphi^{-1}(q) + \kappa(\alpha) u_0 \right) d\mu = 1. \quad (91)$$

Então,

$$\frac{\partial}{\partial \alpha} \int_T \varphi \left(\frac{1 - \alpha}{2} \varphi^{-1}(p) + \frac{1 + \alpha}{2} \varphi^{-1}(q) + \kappa(\alpha) u_0 \right) d\mu = 0.$$

Assumindo a convergência da integral, temos

$$\int_T \left[\frac{1}{2} \varphi^{-1}(q) - \frac{1}{2} \varphi^{-1}(p) + \kappa'(\alpha) u_0 \right] \varphi' \left(\frac{1 - \alpha}{2} \varphi^{-1}(p) + \frac{1 + \alpha}{2} \varphi^{-1}(q) + \kappa_\alpha u_0 \right) d\mu = 0. \quad (92)$$

Tomando-se os limites com $\alpha \rightarrow 1$ e $\alpha \rightarrow -1$, e usando o Lema 3.4 temos, respectivamente,

$$\kappa'(1) = -\frac{1}{2} \frac{\int_T [\varphi^{-1}(q) - \varphi^{-1}(p)] \varphi'(\varphi^{-1}(q)) d\mu}{\int_T u_0 \varphi'(\varphi^{-1}(q)) d\mu} \quad (93)$$

e

$$\kappa'(-1) = -\frac{1}{2} \frac{\int_T [\varphi^{-1}(q) - \varphi^{-1}(p)] \varphi'(\varphi^{-1}(p)) d\mu}{\int_T u_0 \varphi'(\varphi^{-1}(p)) d\mu}. \quad (94)$$

Portanto, as expressões (89) e (88) se reescrevem por

$$\mathcal{D}_\varphi^{(1)}(p \parallel q) = \frac{\int_T [\varphi^{-1}(q) - \varphi^{-1}(p)] \varphi'(\varphi^{-1}(q)) d\mu}{\int_T u_0 \varphi'(\varphi^{-1}(q)) d\mu} \quad (95)$$

e

$$\mathcal{D}_\varphi^{(-1)}(q \parallel p) = \frac{\int_T [-\varphi^{-1}(q) + \varphi^{-1}(p)] \varphi'(\varphi^{-1}(q)) d\mu}{\int_T u_0 \varphi'(\varphi^{-1}(p)) d\mu}. \quad (96)$$

Ademais, as expressões (95) e (96) são assimétricas, mas garantimos a dualidade na di-

vergência já que

$$\mathcal{D}_\varphi^{(1)}(p \parallel q) = \mathcal{D}_\varphi^{(-1)}(q \parallel p). \quad (97)$$

Uma outra forma de expressar uma generalização da divergência de Rényi é expressando-a em termos da função de normalização ψ da φ -família. Perceba que pela definição de convexidade e sendo ψ uma função convexa, temos

$$\psi\left(\frac{1-\alpha}{2}\theta + \frac{1+\alpha}{2}\vartheta\right) \leq \frac{1-\alpha}{2}\psi(\theta) + \frac{1+\alpha}{2}\psi(\vartheta). \quad (98)$$

Dessa forma, naturalmente podemos definir a divergência entre duas distribuições em uma φ -família, como vemos na proposição que se segue.

Proposição 3.1 *Dada uma φ -família \mathcal{F}_p , a divergência generaliza de Rényi para $\alpha \in (-1, 1)$ pode ser expressa em termos da função de normalização como segue:*

$$\mathcal{D}_\varphi^{(\alpha)}(p_\theta \parallel p_\vartheta) = \frac{2}{1+\alpha}\psi(\theta) + \frac{2}{1-\alpha}\psi(\vartheta) - \frac{4}{1-\alpha^2}\psi\left(\frac{1-\alpha}{2}\theta + \frac{1+\alpha}{2}\vartheta\right), \quad (99)$$

para todo $\theta, \vartheta \in \Theta$.

Prova 3 *Para tal importante constatação, observemos que a escolha de $\kappa(\alpha)$ em (74) é de tal modo que a expressão*

$$\int_T \varphi\left(\frac{1-\alpha}{2}\varphi^{-1}(p_\theta) + \frac{1+\alpha}{2}\varphi^{-1}(p_\vartheta) + \kappa(\alpha)u_0\right)d\mu = 1 \quad (100)$$

seja satisfeita. Isto é, $\varphi\left(\frac{1-\alpha}{2}\varphi^{-1}(p_\theta) + \frac{1+\alpha}{2}\varphi^{-1}(p_\vartheta) + \kappa(\alpha)u_0\right)$ é uma função densidade de probabilidade. Tomando as funções de densidade na φ -família, expressadas em (65), podemos escrever

$$\begin{aligned} & \frac{1-\alpha}{2}\varphi^{-1}(p_\theta) + \frac{1+\alpha}{2}\varphi^{-1}(p_\vartheta) + \kappa(\alpha)u_0 \\ &= c + \sum_{i=1}^n \left(\frac{1-\alpha}{2}\theta^i + \frac{1+\alpha}{2}\vartheta^i\right)u_i - \left(\frac{1-\alpha}{2}\psi(\theta) + \frac{1+\alpha}{2}\psi(\vartheta) - \kappa(\alpha)\right)u_0 \\ &= c + \sum_{i=1}^n \left(\frac{1-\alpha}{2}\theta^i + \frac{1+\alpha}{2}\vartheta^i\right)u_i - \psi\left(\frac{1-\alpha}{2}\theta + \frac{1+\alpha}{2}\vartheta\right)u_0. \end{aligned} \quad (101)$$

Observe que a primeira igualdade dos cálculos acima (101) é possível por conta da injetividade da função $\varphi(\cdot)$ (vide condição (a1) citada na Definição 3.1). Logo, isolando devidamente, segue que

$$\kappa(\alpha) = \frac{1-\alpha}{2}\psi(\theta) + \frac{1+\alpha}{2}\psi(\vartheta) - \psi\left(\frac{1-\alpha}{2}\theta + \frac{1+\alpha}{2}\vartheta\right).$$

De imediato, pela definição dada em (78), ratificamos a expressão (99)

Também podemos obter a conexão de Levi–Civita para as variedades estatísticas genera-

lizadas. Para tanto, consideremos antes o lema abaixo que estabelece a noção de derivada direcional para a função normalizadora. No Lema 10 que é enunciado e demonstrado em VIGELIS and CAVALCANTE (2013b), os autores expressam a derivada direcional da função normalizadora ψ . Considerando o nosso caso, que é paramétrico, o mesmo resultado pode ser adaptado segundo o lema que se segue.

Lema 3.4 *Assuma que $\varphi(\cdot)$ é continuamente diferenciável. Então, a função normalizadora $\psi: \Theta \rightarrow [0, \infty)$ é diferenciável e sua derivada em relação ao i -ésimo parâmetro θ^i pode ser escrita como*

$$\frac{\partial \psi}{\partial \theta^i}(\theta) = \frac{\mathbb{E}[u_i \varphi'(c + \theta^i u_i - \psi(\theta) u_0)]}{\mathbb{E}[u_0 \varphi'(c + \theta^i u_i - \psi(\theta) u_0)]}. \quad (102)$$

Esse lema será útil por dois motivos. Do ponto de vista geométrico, este nos proporciona recuperar uma expressão para a conexão métrica para as variedades estatísticas generalizadas. Além disso, é uma ferramenta para mostrar que a divergência de Rényi pode ser exibida através da função normalizadora ψ . Do ponto de vista de estimação, a expressão da derivada de ψ em relação a θ^i será parte de um algoritmo de estimação de parâmetros que citaremos no capítulo 4.

Com efeito, analogamente à expressão (2.34) de AMARI and NAGAOKA (2000), vamos recalculer a expressão para a métrica g_{ij} de \mathcal{P} . Da expressão (102), temos

$$\frac{\partial^2}{\partial \theta^i \partial \theta^j} \psi(\theta) = \frac{\partial}{\partial \lambda} \frac{\partial}{\partial \theta^i} (\theta^i u_i + \lambda u_j) = \frac{\partial}{\partial \lambda} \frac{\mathbb{E}[u_i \varphi'(c + \theta^i u_i + \lambda u_j - \psi(\theta^i u_i + \lambda u_j) u_0)]}{\mathbb{E}[u_0 \varphi'(c + \theta^i u_i + \lambda u_j - \psi(\theta^i u_i + \lambda u_j) u_0)]}. \quad (103)$$

Diferenciando a expressão (102) com respeito a θ^j , temos

$$\begin{aligned} \frac{\partial^2}{\partial \theta^i \partial \theta^j} \psi(\theta) &= \frac{\mathbb{E}[u_i \varphi''(c + \theta^i u_i - \psi(\theta) u_0) (u_j - \frac{\partial \psi}{\partial \theta^j}(\theta) u_0)]}{\mathbb{E}[u_0 \varphi'(c + \theta^i u_i - \psi(\theta) u_0)]} \\ &= \frac{\mathbb{E}[u_i \varphi'(c + \theta^i u_i - \psi(\theta) u_0)] \mathbb{E}[u_0 \varphi''(c + \theta^i u_i - \psi(\theta) u_0) (u_j - \frac{\partial \psi}{\partial \theta^j}(\theta) u_0)]}{\mathbb{E}[u_0 \varphi'(c + \theta^i u_i - \psi(\theta) u_0)] \mathbb{E}[u_0 \varphi'(c + \theta^i u_i - \psi(\theta) u_0)]} \\ &= \frac{\mathbb{E}[\varphi''(c + \theta^i u_i - \psi(\theta) u_0) (u_i - \frac{\partial \psi}{\partial \theta^i}(\theta) u_0) (u_j - \frac{\partial \psi}{\partial \theta^j}(\theta) u_0)]}{\mathbb{E}[u_0 \varphi'(c + \theta^i u_i - \psi(\theta) u_0)]}. \end{aligned} \quad (104)$$

Portanto, a partir da expressão (104) para calcular a conexão métrica combinada com

(104), temos:

$$\begin{aligned}
\Gamma_{ijk} &= \frac{1}{2} \left(\frac{\partial g_{ik}}{\partial \theta^j} + \frac{\partial g_{jk}}{\partial \theta^i} - \frac{\partial g_{ij}}{\partial \theta^k} \right) \\
&= \frac{1}{2} \frac{\partial}{\partial \theta^j} \frac{\mathbb{E}[\varphi''(c + \theta^i u_i - \psi(\theta) u_0)(u_i - \frac{\partial \psi}{\partial \theta^i}(\theta) u_0)(u_k - \frac{\partial \psi}{\partial \theta^k}(\theta) u_0)]}{\mathbb{E}[u_0 \varphi'(c + \theta^i u_i - \psi(\theta) u_0)]} \\
&\quad + \frac{1}{2} \frac{\partial}{\partial \theta^i} \frac{\mathbb{E}[\varphi''(c + \theta^i u_i - \psi(\theta) u_0)(u_j - \frac{\partial \psi}{\partial \theta^j}(\theta) u_0)(u_k - \frac{\partial \psi}{\partial \theta^k}(\theta) u_0)]}{\mathbb{E}[u_0 \varphi'(c + \theta^i u_i - \psi(\theta) u_0)]} \\
&\quad - \frac{1}{2} \frac{\partial}{\partial \theta^k} \frac{\mathbb{E}[\varphi''(c + \theta^i u_i - \psi(\theta) u_0)(u_i - \frac{\partial \psi}{\partial \theta^i}(\theta) u_0)(u_j - \frac{\partial \psi}{\partial \theta^j}(\theta) u_0)]}{\mathbb{E}[u_0 \varphi'(c + \theta^i u_i - \psi(\theta) u_0)]}. \tag{105}
\end{aligned}$$

Então, o tensor métrico é calculado obtendo-se a seguinte expressão abaixo

$$\Gamma_{ijk} = \frac{1}{2} \frac{\partial^3 \psi}{\partial \theta^i \partial \theta^j \partial \theta^k}. \tag{106}$$

A generalização da divergência de Rényi também pode ser calculada. Vejamos a proposição abaixo.

Proposição 3.2 *Dada uma φ -família \mathcal{F}_p , a divergência de Rényi é relacionada com a função de normalização ψ pela igualdade*

$$\mathcal{D}_\varphi(p_\theta \parallel p_\vartheta) = \psi(\vartheta) - \psi(\theta) - \nabla \psi(\theta)(\vartheta - \theta) \tag{107}$$

para todo $\theta, \vartheta \in \Theta$.

Prova 4 *Para mostrar (107), usamos*

$$\frac{\partial \psi}{\partial \theta^i}(\theta) = \frac{\int_T u_i \varphi'(f_\theta) d\mu}{\int_T u_0 \varphi'(f_\theta) d\mu},$$

a qual é uma consequência do Lema 10 de VIGELIS and CAVALCANTE (2013b). Pela derivada da função inversa LIMA (1976) $(\varphi^{-1})'(u) = \frac{1}{\varphi'(\varphi^{-1}(u))}$, expressão (68) com $p = p_\theta$ e $\tilde{p} = p_\vartheta$ resulta em

$$\mathcal{D}_\varphi(p_\theta \parallel p_\vartheta) = \frac{\int_T (f_\theta - f_\vartheta) \varphi'(f_\theta) d\mu}{\int_T u_0 \varphi'(f_\theta) d\mu}. \tag{108}$$

Inserindo a diferença

$$\begin{aligned}
\varphi^{-1}(p_\theta) - \varphi^{-1}(p_\vartheta) &= \left(c + \sum_{i=1}^n \theta^i u_i - \psi(\theta) u_0 \right) - \left(c + \sum_{i=1}^n \vartheta^i u_i - \psi(\vartheta) u_0 \right) \\
&= \psi(\vartheta) u_0 - \psi(\theta) u_0 - \sum_{i=1}^n (\vartheta^i - \theta^i) u_i,
\end{aligned}$$

em que obtemos a expressão (107)

Com esses conceitos e resultados estabelecidos, nosso próximo passo será estu-

dar a geometria da variedade estatística \mathcal{P} .

3.5 Geometria Induzida de $\mathcal{D}_\varphi(\cdot \parallel \cdot)$

Nesta seção, vamos caracterizar geometricamente a φ -família \mathcal{F}_p através da função de divergência, o que é fato possível em variedades estatísticas MATUMOTO (1993). Da definição 2.3, garantimos que

$$\mathcal{D}_\varphi^{(\alpha)}(p \parallel q) = 0 \quad \text{se e somente se} \quad p = q.$$

Portanto, utilizando-se da divergência de Rényi generalizada, uma métrica g_{ij} para \mathcal{P} pode ser induzida pela expressão

$$g_{ij} = - \left[\left(\frac{\partial}{\partial \theta^i} \right)_p \left(\frac{\partial}{\partial \theta^j} \right)_q \mathcal{D}_\varphi^{(\alpha)}(p \parallel q) \right]_{q=p}, \quad (109)$$

em que os índices p e q nas derivadas indicam em que ponto se calcula a derivada. Essa métrica induzida é invariante por mudanças de coordenadas pela regra da cadeia. Para se concluir que está bem definida, devemos mostrar que é positiva definida. Mostraremos isso sabendo que as métricas dadas em (55) e (58) são positiva definidas e, pela proposição abaixo, mostraremos que essas métricas citadas coincidem com a métrica induzida (109).

Proposição 3.3 *A métrica induzida por $\mathcal{D}_\varphi^{(\alpha)}(\cdot \parallel \cdot)$ coincide com as métricas dadas em (55) e (58).*

Prova 5 *Fixe $\alpha \in (-1, 1)$ e defina $c_\alpha = \frac{1-\alpha}{2} f_\theta + \frac{1+\alpha}{2} f_\vartheta + \kappa(\alpha) u_0$ tal que a expressão*

$$\int_T \varphi(c_\alpha) d\mu = 1 \quad (110)$$

seja satisfeita. Como já citado anteriormente, pelas convexidade da função φ , $\kappa(\alpha)$ é único. Aplicando o operador $\left(\frac{\partial}{\partial \theta^j} \right)_{p_\vartheta}$ em ambos os lados da expressão (110), obtemos

$$\int_T \left(\frac{1+\alpha}{2} \frac{\partial f_\vartheta}{\partial \theta^j} + \left(\frac{\partial}{\partial \theta^j} \right)_{p_\vartheta} \kappa(\alpha) u_0 \right) \varphi'(c_\alpha) d\mu = 0,$$

com a qual podemos expressar

$$\left(\frac{\partial}{\partial \theta^j} \right)_{p_\vartheta} \kappa(\alpha) = - \frac{1+\alpha}{2} \frac{\int_T \frac{\partial f_\vartheta}{\partial \theta^j} \varphi'(c_\alpha) d\mu}{\int_T u_0 \varphi'(c_\alpha) d\mu}. \quad (111)$$

Aplicando a derivada em $(\frac{\partial}{\partial \theta^i})_{p_\theta}$ em (111), obtemos

$$\begin{aligned} \left(\frac{\partial}{\partial \theta^i}\right)_{p_\theta} \left(\frac{\partial}{\partial \theta^j}\right)_{p_\vartheta} \kappa(\alpha) &= -\frac{1 + \alpha \int_T \left[\frac{1-\alpha}{2} \frac{\partial f_\theta}{\partial \theta^i} + (\frac{\partial}{\partial \theta^j})_{p_\theta} \kappa(\alpha) u_0\right] \frac{\partial f_\vartheta}{\partial \theta^j} \varphi''(c_\alpha) d\mu}{2 \int_T u_0 \varphi'(c_\alpha) d\mu} \\ &+ \frac{1 + \alpha \int_T \frac{\partial f_\vartheta}{\partial \theta^j} \varphi'(c_\alpha) d\mu}{2 \int_T u_0 \varphi'(c_\alpha) d\mu} \frac{\int_T u_0 \left[\frac{1-\alpha}{2} \frac{\partial f_\theta}{\partial \theta^i} + (\frac{\partial}{\partial \theta^j})_{p_\theta} \kappa(\alpha) u_0\right] \varphi''(c_\alpha) d\mu}{\int_T u_0 \varphi'(c_\alpha) d\mu}. \end{aligned} \quad (112)$$

Da expressão (56), concluímos que $\int_T \frac{\partial f_\vartheta}{\partial \theta^j} \varphi'(c_\alpha) d\mu = 0$ para $p_\vartheta = p_\theta$, o segundo termo do lado direito se anula e então escrevemos

$$\left[\left(\frac{\partial}{\partial \theta^i}\right)_{p_\theta} \left(\frac{\partial}{\partial \theta^j}\right)_{p_\vartheta} \kappa(\alpha)\right]_{p_\vartheta=p_\theta} = -\frac{1 - \alpha^2 \int_T \frac{\partial f_\theta}{\partial \theta^i} \frac{\partial f_\theta}{\partial \theta^j} \varphi'(f_\theta) d\mu}{4 \int_T u_0 \varphi'(f_\theta) d\mu}.$$

Portanto, usando as notações (53), podemos escrever

$$g_{ij} = -\left[\left(\frac{\partial}{\partial \theta^i}\right)_{p_\theta} \left(\frac{\partial}{\partial \theta^j}\right)_{p_\vartheta} \mathcal{D}_\varphi^{(\alpha)}(p_\theta \parallel p_\vartheta)\right]_{p_\vartheta=p_\theta} = E_\theta'' \left[\frac{\partial f_\theta}{\partial \theta^i} \frac{\partial f_\theta}{\partial \theta^j}\right].$$

Para o caso $\alpha = \pm 1$, ao compararmos as expressões (53) e (68), podemos escrever

$$\mathcal{D}_\varphi(p_\theta \parallel p_\vartheta) = E_\theta'[f_\theta - f_\vartheta]. \quad (113)$$

Usando as expressões

$$g_{ij} = \left[\left(\frac{\partial^2}{\partial \theta^i \partial \theta^j}\right)_p \mathcal{D}_\varphi^{(\alpha)}(p \parallel q)\right]_{q=p} = \left[\left(\frac{\partial^2}{\partial \theta^i \partial \theta^j}\right)_q \mathcal{D}_\varphi^{(\alpha)}(p \parallel q)\right]_{q=p},$$

segue que

$$g_{ij} = \left[\left(\frac{\partial^2}{\partial \theta^i \partial \theta^j}\right)_{p_\vartheta} \mathcal{D}_\varphi(p_\theta \parallel p_\vartheta)\right]_{p_\vartheta=p_\theta} = -E_\theta' \left[\frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j}\right].$$

Devido a $\mathcal{D}_\varphi^{(-1)}(p \parallel q) = \mathcal{D}_\varphi^{(1)}(q \parallel p) = \mathcal{D}_\varphi(p \parallel q)$, conclui-se que a expressão da métrica definida em (55) coincide com a métrica induzida por $\mathcal{D}_\varphi^{(-1)}(\cdot \parallel \cdot)$ e $\mathcal{D}_\varphi^{(1)}(\cdot \parallel \cdot)$.

Interessante destacar que a expressão (113) tem a mesma forma na família exponencial AMARI and NAGAOKA (2000). Vale ressaltar o exemplo abaixo.

Exemplo 10 Seja $\varphi(u) = \exp_\kappa(u)$ a exponencial de Kaniadakis cuja inversa é $\varphi^{-1}(u) = \log_\kappa(u)$, como citadas no Exemplo 8. Aplicando (68), a divergência aqui será dada por

$$\mathcal{D}_\kappa(p \parallel q) = \frac{\int_T \frac{\log_\kappa(p) - \log_\kappa(q)}{\log'_\kappa(p)} \mu}{\int_T \frac{u_0}{\log'_\kappa(p)} d\mu}. \quad (114)$$

Desde que o κ -logaritmo é definido como $\log_\kappa(u) = \frac{u^\kappa - u^{-\kappa}}{2\kappa}$ e sua derivada é dada por

$\log'_\kappa(u) = \frac{1}{u} \frac{u^\kappa + u^{-\kappa}}{2}$, reescreve-se (114) por

$$\mathcal{D}_\kappa(p \parallel q) = \frac{1}{\kappa} \frac{\int_T \left(\frac{p^\kappa - p^{-\kappa}}{p^\kappa + p^{-\kappa}} - \frac{q^\kappa - q^{-\kappa}}{p^\kappa + p^{-\kappa}} \right) d\mu}{\int_T \frac{2u_0}{p^\kappa + p^{-\kappa}} d\mu},$$

a qual é chamada de κ -divergência VIGELIS and CAVALCANTE (2013b).

Nas variedades estatísticas generalizadas, a divergência de Rényi generalizada $\mathcal{D}_\varphi^{(\alpha)}(\cdot \parallel \cdot)$ induz uma família de conexões $D^{(\alpha)}$, cujos símbolos de Christoffel $\Gamma_{ijk}^{(\alpha)}$ são dados por

$$\Gamma_{ijk}^{(\alpha)} = - \left[\left(\frac{\partial^2}{\partial \theta^i \partial \theta^j} \right)_p \left(\frac{\partial}{\partial \theta^k} \right)_q \mathcal{D}_\varphi^{(\alpha)}(p \parallel q) \right]_{q=p}. \quad (115)$$

Devido $\mathcal{D}_\varphi^{(\alpha)}(p \parallel q) = \mathcal{D}_\varphi^{(-\alpha)}(q \parallel p)$, temos que as conexões $D^{(\alpha)}$ e $D^{(-\alpha)}$ são mutuamente duais para todo $\alpha \in [-1, 1]$. Em outras palavras, $\Gamma_{ijk}^{(\alpha)}$ e $\Gamma_{ijk}^{(-\alpha)}$ satisfazem a relação $\frac{\partial g_{jk}}{\partial \theta^i} = \Gamma_{ijk}^{(\alpha)} + \Gamma_{ijk}^{(-\alpha)}$. Combinando as expressões (113) e (115), valem as seguintes expressões para as conexões.

$$\begin{aligned} \Gamma_{ijk}^{(-1)} &= \frac{\partial}{\partial \theta^k} \frac{\partial^2}{\partial \eta_i \partial \eta_j} E'_\theta[f_\theta - f_\vartheta] \\ &= \frac{\partial}{\partial \theta^k} \frac{\mathbb{E}[\frac{\partial^2 f_\vartheta}{\partial \eta_i \partial \eta_j} \varphi'(f_\theta)]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} \\ &= \frac{\mathbb{E}[\frac{\partial^2 f_\vartheta}{\partial \eta_i \partial \eta_j} \frac{\partial f_\theta}{\partial \theta^k} \varphi''(f_\theta)]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} - \frac{\mathbb{E}[\frac{\partial^2 f_\vartheta}{\partial \eta_i \partial \eta_j} \varphi'(f_\theta)]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} \frac{\mathbb{E}[u_0 \frac{\partial f_\theta}{\partial \theta^k} \varphi''(f_\theta)]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} \\ &= E''_\theta \left[\frac{\partial^2 f_\vartheta}{\partial \eta_i \partial \eta_j} \frac{\partial f_\theta}{\partial \theta^k} \right] - E'_\theta \left[\frac{\partial^2 f_\vartheta}{\partial \eta_i \partial \eta_j} \right] E''_\theta \left[u_0 \frac{\partial f_\theta}{\partial \theta^k} \right] \\ &\stackrel{\vartheta \rightarrow \theta}{=} E''_\theta \left[\frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j} \frac{\partial f_\theta}{\partial \theta^k} \right] - E'_\theta \left[\frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j} \right] E''_\theta \left[u_0 \frac{\partial f_\theta}{\partial \theta^k} \right]. \end{aligned} \quad (116)$$

Analogamente,

$$\begin{aligned} \Gamma_{ijk}^{(1)} &= \frac{\partial^2}{\partial \theta^i \partial \theta^j} \frac{\partial}{\partial \theta^k} E'_\theta[f_\theta - f_\vartheta] \\ &= \frac{\partial^2}{\partial \theta^i \partial \theta^j} \frac{\mathbb{E}[\frac{\partial f_\vartheta}{\partial \theta^k} \varphi'(f_\theta)]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} \\ &= \frac{\partial}{\partial \theta^j} \left(\frac{\mathbb{E}[\frac{\partial f_\vartheta}{\partial \theta^k} \frac{\partial f_\theta}{\partial \theta^i} \varphi''(f_\theta)]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} - \frac{\mathbb{E}[\frac{\partial f_\vartheta}{\partial \theta^k} \varphi'(f_\theta)]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} \frac{\mathbb{E}[u_0 \frac{\partial f_\theta}{\partial \theta^i} \varphi''(f_\theta)]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} \right). \end{aligned} \quad (117)$$

Considerando as notações (53) e (54), a expressão (117) pode ser reescrita como

$$\begin{aligned}
\Gamma_{ijk}^{(1)} &= \frac{\mathbb{E}\left[\frac{\partial f_\theta}{\partial \eta_k} \frac{\partial f_\theta}{\partial \theta^i} \frac{\partial f_\theta}{\partial \theta^j} \varphi'''(f_\theta)\right]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} + \frac{\mathbb{E}\left[\frac{\partial f_\theta}{\partial \eta_k} \frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j} \varphi''(f_\theta)\right]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} \\
&\quad - \frac{\mathbb{E}\left[\frac{\partial f_\theta}{\partial \eta_k} \frac{\partial f_\theta}{\partial \theta^i} \varphi''(f_\theta)\right]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} \frac{\mathbb{E}\left[u_0 \frac{\partial f_\theta}{\partial \theta^j} \varphi''(f_\theta)\right]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} \\
&\quad - \frac{\mathbb{E}\left[u_0 \frac{\partial f_\theta}{\partial \theta^i} \varphi''(f_\theta)\right]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} \frac{\mathbb{E}\left[\frac{\partial f_\theta}{\partial \eta_k} \frac{\partial f_\theta}{\partial \theta^j} \varphi''(f_\theta)\right]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} \\
&\stackrel{\vartheta \rightarrow \theta}{=} \frac{\mathbb{E}\left[\frac{\partial f_\theta}{\partial \theta^k} \frac{\partial f_\theta}{\partial \theta^i} \frac{\partial f_\theta}{\partial \theta^j} \varphi'''(f_\theta)\right]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} + \frac{\mathbb{E}\left[\frac{\partial f_\theta}{\partial \theta^k} \frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j} \varphi''(f_\theta)\right]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} \\
&\quad - \frac{\mathbb{E}\left[\frac{\partial f_\theta}{\partial \theta^k} \frac{\partial f_\theta}{\partial \theta^i} \varphi''(f_\theta)\right]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} \frac{\mathbb{E}\left[u_0 \frac{\partial f_\theta}{\partial \theta^j} \varphi''(f_\theta)\right]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} \\
&\quad - \frac{\mathbb{E}\left[u_0 \frac{\partial f_\theta}{\partial \theta^i} \varphi''(f_\theta)\right]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} \frac{\mathbb{E}\left[\frac{\partial f_\theta}{\partial \theta^k} \frac{\partial f_\theta}{\partial \theta^j} \varphi''(f_\theta)\right]}{\mathbb{E}[u_0 \varphi'(f_\theta)]} \\
&= E_\theta''' \left[\frac{\partial f_\theta}{\partial \theta^k} \frac{\partial f_\theta}{\partial \theta^i} \frac{\partial f_\theta}{\partial \theta^j} \right] + E_\theta'' \left[\frac{\partial f_\theta}{\partial \theta^k} \frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j} \right] \\
&\quad - E_\theta'' \left[\frac{\partial f_\theta}{\partial \theta^k} \frac{\partial f_\theta}{\partial \theta^i} \right] E_\theta'' \left[u_0 \frac{\partial f_\theta}{\partial \theta^j} \right] \\
&\quad - E_\theta'' \left[u_0 \frac{\partial f_\theta}{\partial \theta^i} \right] E_\theta'' \left[\frac{\partial f_\theta}{\partial \theta^k} \frac{\partial f_\theta}{\partial \theta^j} \right]. \tag{118}
\end{aligned}$$

Pode-se notar que as expressões (118) e (116) são uma generalização dos símbolos das conexões exponencial e mistura na variedade estatística exponencial, como já previamente mencionado. Observe ainda que

$$\int_T u_0 \varphi'(f_\theta) d\mu = \int_T p_\theta dx = 1$$

e

$$\begin{aligned}
E_\theta'' \left[u_0 \frac{\partial f_\theta}{\partial \theta^k} \right] &= \int_T \frac{\partial}{\partial \theta^k} (\log p_\theta) \exp''(\log p_\theta) dx \\
&= \frac{\partial}{\partial \theta^k} \int_T \frac{1}{p_\theta} p_\theta dx = 0. \tag{119}
\end{aligned}$$

Suponha que $f_\theta = l_\theta = \log(p_\theta)$ e $u_0 = 1$. Portanto, a conexão exponencial da família exponencial é obtida como

$$\begin{aligned}
\Gamma_{ijk}^{(e)} &:= \Gamma_{ijk}^{(1)} = E_\theta'' \left[\frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j} \frac{\partial f_\theta}{\partial \theta^k} \right] - E_\theta' \left[\frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j} \right] E_\theta'' \left[u_0 \frac{\partial f_\theta}{\partial \theta^k} \right] \\
&= E_\theta \left[\frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} \frac{\partial l_\theta}{\partial \theta^k} \right]
\end{aligned}$$

Analogamente, de forma mais direta, a conexão mistura da família exponencial é obtida

também como

$$\begin{aligned}\Gamma_{ijk}^{(m)} := \Gamma_{ijk}^{(-1)} &= E_\theta \left[\frac{\partial l_\theta}{\partial \theta^k} \frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} \right] + E_\theta \left[\frac{\partial l_\theta}{\partial \theta^k} \frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} \right] \\ &\quad - E_\theta \left[\frac{\partial l_\theta}{\partial \theta^k} \frac{\partial l_\theta}{\partial \theta^i} \right] E_\theta \left[u_0 \frac{\partial l_\theta}{\partial \theta^j} \right] \\ &\quad - E_\theta \left[u_0 \frac{\partial l_\theta}{\partial \theta^i} \right] E_\theta \left[\frac{\partial l_\theta}{\partial \theta^k} \frac{\partial l_\theta}{\partial \theta^j} \right].\end{aligned}$$

Usando a expressão (119), temos que

$$\Gamma_{ijk}^{(-1)} = E_\theta \left[\frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} \frac{\partial l_\theta}{\partial \theta^k} \right] + E_\theta \left[\frac{\partial l_\theta}{\partial \theta^k} \frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} \right],$$

como ocorre na conexão mistura citada em (16) quando especificamos $\alpha = -1$.

Um outro ponto importante é que derivando a expressão (58) em relação a θ^j , temos

$$\begin{aligned}\frac{\partial g_{ij}}{\partial \theta^k} &= E_\theta'' \left[\frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^k} \frac{\partial f_\theta}{\partial \theta^j} \right] + E_\theta'' \left[\frac{\partial f_\theta}{\partial \theta^i} \frac{\partial^2 f_\theta}{\partial \theta^j \partial \theta^k} \right] \\ &\quad + E_\theta''' \left[\frac{\partial f_\theta}{\partial \theta^k} \frac{\partial f_\theta}{\partial \theta^i} \frac{\partial f_\theta}{\partial \theta^j} \right] - E_\theta'' \left[\frac{\partial f_\theta}{\partial \theta^i} \frac{\partial f_\theta}{\partial \theta^i} \right] E_\theta'' \left[u_0 \frac{\partial f_\theta}{\partial \theta^k} \right].\end{aligned}$$

Então, concluímos que

$$\begin{aligned}\Gamma_{ijk} &= E_\theta'' \left[\frac{\partial f_\theta^2}{\partial \theta^i \partial \theta^j} \frac{\partial f_\theta}{\partial \theta^k} \right] + \frac{1}{2} E_\theta'' \left[\frac{\partial f_\theta}{\partial \theta^i} \frac{\partial f_\theta}{\partial \theta^j} \frac{\partial f_\theta}{\partial \theta^k} \right] \\ &\quad - \frac{1}{2} E_\theta'' \left[\frac{\partial f_\theta}{\partial \theta^i} \frac{\partial f_\theta}{\partial \theta^k} \right] E_\theta'' \left[u_0 \frac{\partial f_\theta}{\partial \theta^j} \right] - \frac{1}{2} E_\theta'' \left[\frac{\partial f_\theta}{\partial \theta^j} \frac{\partial f_\theta}{\partial \theta^k} \right] E_\theta'' \left[u_0 \frac{\partial f_\theta}{\partial \theta^i} \right] \\ &\quad + \frac{1}{2} E_\theta'' \left[\frac{\partial f_\theta}{\partial \theta^i} \frac{\partial f_\theta}{\partial \theta^j} \right] E_\theta'' \left[u_0 \frac{\partial f_\theta}{\partial \theta^i} \right].\end{aligned}\tag{120}$$

Mais uma vez, quando particularizamos $\varphi(\cdot) = \exp(\cdot)$ e $u_0 = 1$, obtemos a expressão da conexão métrica na família exponencial que é dada por

$$\Gamma_{ijk} = E_\theta \left[\frac{\partial^2 l_\theta}{\partial \theta^i \partial \theta^j} \frac{\partial l_\theta}{\partial \theta^k} \right] + \frac{1}{2} E_\theta \left[\frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^i} \right].$$

Além disso, com as expressões das conexões (116) e (118), podemos escrever a equação da α -conexão para a variedade estatística \mathcal{P} SHIMA (2007), como segue abaixo

$$\begin{aligned}
\Gamma_{ijk}^{(\alpha)} &= \frac{1-\alpha}{2}\Gamma_{ijk}^{(-1)} + \frac{1+\alpha}{2}\Gamma_{ijk}^{(1)} \\
&= E''_{\theta} \left[\frac{\partial^2 f_{\theta}}{\partial\theta^i \partial\theta^j} \frac{\partial f_{\theta}}{\partial\theta^k} \right] + \frac{1-\alpha}{2} E'''_{\theta} \left[\frac{\partial f_{\theta}}{\partial\theta^i} \frac{\partial f_{\theta}}{\partial\theta^j} \frac{\partial f_{\theta}}{\partial\theta^k} \right] \\
&\quad - \frac{1-\alpha}{2} E''_{\theta} \left[\frac{\partial f_{\theta}}{\partial\theta^j} \frac{\partial f_{\theta}}{\partial\theta^k} \right] E''_{\theta} \left[u_0 \frac{\partial f_{\theta}}{\partial\theta^i} \right] - \frac{1-\alpha}{2} E''_{\theta} \left[\frac{\partial f_{\theta}}{\partial\theta^i} \frac{\partial f_{\theta}}{\partial\theta^k} \right] E''_{\theta} \left[u_0 \frac{\partial f_{\theta}}{\partial\theta^j} \right] \\
&\quad - \frac{1+\alpha}{2} E''_{\theta} \left[\frac{\partial^2 f_{\theta}}{\partial\theta^i \partial\theta^j} \right] E''_{\theta} \left[u_0 \frac{\partial f_{\theta}}{\partial\theta^k} \right], \tag{121}
\end{aligned}$$

em que para $\alpha = 0$, recuperamos a conexão de Levi–Civita, como mostramos na expressão (64). Observe também que, como já foi citado na família exponencial, relacionamos a conexão métrica $\Gamma_{ijk}^{(\text{metric})} = \Gamma_{ijk}^{(0)}$ com a α -conexão $\Gamma_{ijk}^{(\alpha)}$ em (35). Obtemos equação análoga na φ -família quando relacionamos a conexão métrica (120) com

$$\begin{aligned}
T_{ijk} &= \frac{1}{2} E'''_{\theta} \left[\frac{\partial f_{\theta}}{\partial\theta^i} \frac{\partial f_{\theta}}{\partial\theta^j} \frac{\partial f_{\theta}}{\partial\theta^k} \right] - \frac{1}{2} E''_{\theta} \left[\frac{\partial f_{\theta}}{\partial\theta^k} \frac{\partial f_{\theta}}{\partial\theta^i} \right] E''_{\theta} \left[u_0 \frac{\partial f_{\theta}}{\partial\theta^j} \right] \\
&\quad - \frac{1}{2} E''_{\theta} \left[\frac{\partial f_{\theta}}{\partial\theta^j} \frac{\partial f_{\theta}}{\partial\theta^k} \right] E''_{\theta} \left[u_0 \frac{\partial f_{\theta}}{\partial\theta^i} \right] - \frac{1}{2} E''_{\theta} \left[\frac{\partial f_{\theta}}{\partial\theta^i} \frac{\partial f_{\theta}}{\partial\theta^j} \right] E''_{\theta} \left[u_0 \frac{\partial f_{\theta}}{\partial\theta^k} \right]. \tag{122}
\end{aligned}$$

Mais uma vez, se $\varphi(\cdot) = \exp(\cdot)$ e $u_0 = 1$, vale o caso da família exponencial, como citado na expressão (19).

Para $\alpha \in (-1, 1)$, os símbolos de Christoffel $\Gamma_{ijk}^{(\alpha)}$ podem ser escritos como a combinação convexa de $\Gamma_{ijk}^{(-1)}$ e $\Gamma_{ijk}^{(1)}$. Verifiquemos esse resultado através da proposição que se segue.

Proposição 3.4 *Os símbolos de Christoffel $\Gamma_{ijk}^{(\alpha)}$ induzidos pela função de divergência $\mathcal{D}_{\varphi}^{(\alpha)}(\cdot \parallel \cdot)$ satisfazem a relação*

$$\Gamma_{ijk}^{(\alpha)} = \frac{1-\alpha}{2}\Gamma_{ijk}^{(-1)} + \frac{1+\alpha}{2}\Gamma_{ijk}^{(1)}, \quad \text{para } \alpha \in [-1, 1]. \tag{123}$$

Prova 6 *Observe que para $\alpha = -1$ ou $\alpha = 1$, diretamente obtemos a relação (123). Para $\alpha \in (-1, 1)$, através da expressão (112), podemos escrever*

$$\begin{aligned}
\left(\frac{\partial}{\partial\theta^i} \right)_{p_{\theta}} \left(\frac{\partial}{\partial\theta^j} \right)_{p_{\theta}} \kappa(\alpha) &= - \frac{1+\alpha}{2} \frac{\int_T \left[\frac{1-\alpha}{2} \frac{\partial f_{\theta}}{\partial\theta^i} + \left(\frac{\partial}{\partial\theta^k} \right)_{p_{\theta}} \kappa(\alpha) u_0 \right] \frac{\partial f_{\theta}}{\partial\theta^j} \varphi''(c_{\alpha}) d\mu}{\int_T u_0 \varphi'(c_{\alpha}) d\mu} \\
&\quad + \frac{1+\alpha}{2} \frac{\int_T \frac{\partial f_{\theta}}{\partial\theta^k} \varphi'(c_{\alpha}) d\mu}{\int_T u_0 \varphi'(c_{\alpha}) d\mu} \frac{\int_T u_0 \left[\frac{1-\alpha}{2} \frac{\partial f_{\theta}}{\partial\theta^i} + \left(\frac{\partial}{\partial\theta^k} \right)_{p_{\theta}} \kappa(\alpha) u_0 \right] \varphi''(c_{\alpha}) d\mu}{\int_T u_0 \varphi'(c_{\alpha}) d\mu}. \tag{124}
\end{aligned}$$

Aplicando $(\frac{\partial}{\partial\theta^j})_{p_\theta}$ na primeira parcela da expressão (124) e então fazendo $p_\vartheta = p_\theta$, obtemos

$$\begin{aligned} & -\frac{1-\alpha^2}{4}E''_\theta\left[\frac{\partial^2 f_\theta}{\partial\theta^i\partial\theta^j}\frac{\partial f_\theta}{\partial\theta^k}\right] - \frac{1+\alpha}{2}\left(\frac{\partial^2}{\partial\theta^i\partial\theta^j}\right)_{p_\theta}\kappa(\alpha)E''_\theta\left[u_0\frac{\partial f_\theta}{\partial\theta^k}\right] \\ & - \frac{1-\alpha^2}{4}\frac{1-\alpha}{2}E'''_\theta\left[\frac{\partial f_\theta}{\partial\theta^i}\frac{\partial f_\theta}{\partial\theta^j}\frac{\partial f_\theta}{\partial\theta^k}\right] \\ & + \frac{1-\alpha^2}{4}\frac{1-\alpha}{2}E''_\theta\left[\frac{\partial f_\theta}{\partial\theta^i}\frac{\partial f_\theta}{\partial\theta^k}\right]E''_\theta\left[u_0\frac{\partial f_\theta}{\partial\theta^j}\right]. \end{aligned} \quad (125)$$

De igual modo, aplicando $(\frac{\partial}{\partial\theta^j})_{p_\theta}$ na segunda parcela da expressão (124) e então fazendo $p_\vartheta = p_\theta$, obtemos

$$\frac{1-\alpha^2}{2}\frac{1-\alpha}{2}E''_\theta\left[\frac{\partial f_\theta}{\partial\theta^j}\frac{\partial f_\theta}{\partial\theta^k}\right]E''_\theta\left[u_0\frac{\partial f_\theta}{\partial\theta^i}\right]. \quad (126)$$

Agrupando as expressões (125) e (126), podemos escrever

$$\begin{aligned} \Gamma_{ijk}^{(\alpha)} &= -\frac{4}{1-\alpha^2}\left[\left(\frac{\partial}{\partial\theta^i}\right)_{p_\theta}\left(\frac{\partial}{\partial\theta^k}\right)_{p_\theta}\kappa(\alpha)\right]_{p_\vartheta=p_\theta} \\ &= E''_\theta\left[\frac{\partial^2 f_\theta}{\partial\theta^i\partial\theta^j}\frac{\partial f_\theta}{\partial\theta^k}\right] + \frac{1-\alpha}{2}E'''_\theta\left[\frac{\partial f_\theta}{\partial\theta^i}\frac{\partial f_\theta}{\partial\theta^j}\frac{\partial f_\theta}{\partial\theta^k}\right] \\ & - \frac{1-\alpha}{2}E''_\theta\left[\frac{\partial f_\theta}{\partial\theta^j}\frac{\partial f_\theta}{\partial\theta^k}\right]E''_\theta\left[u_0\frac{\partial f_\theta}{\partial\theta^i}\right] \\ & - \frac{1-\alpha}{2}E''_\theta\left[\frac{\partial f_\theta}{\partial\theta^j}\frac{\partial f_\theta}{\partial\theta^k}\right]E''_\theta\left[u_0\frac{\partial f_\theta}{\partial\theta^i}\right] \\ & - \frac{1+\alpha}{2}E'_\theta\left[\frac{\partial^2 f_\theta}{\partial\theta^i\partial\theta^j}\right]E''_\theta\left[u_0\frac{\partial f_\theta}{\partial\theta^k}\right], \end{aligned} \quad (127)$$

em que usamos

$$\begin{aligned} \left(\frac{\partial^2}{\partial\theta^i\partial\theta^j}\right)_{p_\theta}\kappa(\alpha) &= \frac{1-\alpha^2}{4}\left[\left(\frac{\partial^2}{\partial\theta^i\partial\theta^j}\right)_{p_\theta}\mathcal{D}_\varphi^{(\alpha)}(p_\theta \parallel p_\vartheta)\right]_{p_\vartheta=p_\theta} \\ &= \frac{1-\alpha^2}{4}g_{ij} = -\frac{1-\alpha^2}{4}E'_\theta\left[\frac{\partial^2 f_\theta}{\partial\theta^i\partial\theta^j}\right]. \end{aligned}$$

Portanto, combinando as expressões (116), (118) e (127), garantimos a expressão (123). Através de uma conexão D , pode-se definir o transporte paralelo. Fixemos dois campos de vetores X e Y . Dado um ponto $p \in \mathcal{M}$, defina uma curva $\gamma : I \rightarrow \mathcal{M}$ tal que o campo X esteja definido ao longo de γ . Em outras palavras, $\gamma(t_0) = p$ e $\frac{d\gamma}{dt} = X(\gamma(t))$. Seja $P_{\gamma, t_0, t} : T_{\gamma(t_0)}\mathcal{M} \rightarrow T_{\gamma(t)}\mathcal{M}$ o transporte paralelo de um vetor ao longo de γ de t_0 para t . Então, podemos relacionar a conexão D como se segue PETERSEN (2006)

$$(D_X Y)(p) = \frac{d}{dt}P_{\gamma, t_0, t}^{-1}(Y(\gamma(t)))|_{t=t_0}.$$

Vamos supor que o conjunto T seja finito. Nesta caso, podemos considerar a variedade estatística generalizada $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ para o qual $\mathcal{P} = \mathcal{P}_\mu$. A conexão $D^{(1)}$ pode ser

obtida pelo transporte paralelo

$$P_{q,p} : \tilde{T}_q \mathcal{P} \rightarrow \tilde{T}_p \mathcal{P}$$

dado por

$$\tilde{X} \mapsto \tilde{X} - E'_\theta[\tilde{X}]u_0,$$

em que $p = p_\theta$. Destacamos que $P_{q,p}$ não depende da curva ligando p a q . Denote por $\gamma(t)$ a curva coordenada dada localmente por $\theta(t) = (\theta^1, \dots, \theta^i + t, \dots, \theta^n)$. Observando que $P_{\gamma(0),\gamma(t)}^{-1}$ aplica o vetor $\frac{\partial f_\theta}{\partial \theta^j}(t)$ para

$$\frac{\partial f_\theta}{\partial \theta^j}(t) - E'_{\theta(0)} \left[\frac{\partial f_\theta}{\partial \theta^j}(t) \right] u_0,$$

define-se a conexão

$$\begin{aligned} \tilde{D}_{\partial f_\theta / \partial \theta^i} \frac{\partial f_\theta}{\partial \theta^j} &= \frac{d}{dt} P_{\gamma(0),\gamma(t)}^{-1} \left(\frac{\partial f_\theta}{\partial \theta^j} \gamma(t) \right) \Big|_{t=0} \\ &= \frac{d}{dt} \left(\frac{\partial f_{\theta(t)}}{\partial \theta^j} - E'_{\theta(0)} \left[\frac{\partial f_{\theta(t)}}{\partial \theta^j} \right] u_0 \right) \Big|_{t=0} \\ &= \frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j} - E'_\theta \left[\frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j} \right] u_0. \end{aligned}$$

Denote por D a conexão correspondente para \tilde{D} , para o qual age em campos de vetores suaves em $T_p \mathcal{P}$. Com essa identificação, temos

$$\begin{aligned} g \left(D_{\partial / \partial \theta^i} \frac{\partial}{\partial \theta^j}, \frac{\partial}{\partial \theta^k} \right) &= \left\langle \tilde{D}_{\partial f_\theta / \partial \theta^i} \frac{\partial f_\theta}{\partial \theta^j}, \frac{\partial f_\theta}{\partial \theta^k} \right\rangle \\ &= E''_\theta \left[\frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j} \frac{\partial f_\theta}{\partial \theta^k} \right] - E'_\theta \left[\frac{\partial^2 f_\theta}{\partial \theta^i \partial \theta^j} \right] E''_\theta \left[u_0 \frac{\partial f_\theta}{\partial \theta^k} \right] \\ &= \Gamma_{ijk}^{(1)}, \end{aligned}$$

mostrando que $D = D^{(1)}$.

Nosso próximo passo será mostrar questões que envolvem dualidade entre as conexões e as coordenadas duais. Além disso, mostraremos uma adaptação para o caso das φ -família dos teoremas de Pitágoras e da Projeção.

3.6 Dualidade

Os conceitos de dualidade são naturalmente estudados em variedades estatísticas SHIMA (2007). Por exemplo, como podemos ver na expressão da α -conexão (16), as conexões $D^{(\alpha)}$ e $D^{(-\alpha)}$ são mutuamente duais. Em MATUMOTO (1993), mostra-se que em toda variedade estatística com uma métrica g_{ij} e com um par de conexões duais Γ_{ijk} e Γ_{ijk}^* , também podemos sempre definir uma classe de divergências.

3.6.1 Conexões Duais

Em variedades riemannianas, a conexão de Levi–Civita é a única conexão afim que preserva a métrica via transporte paralelo CARMO (1976a). Em variedades estatísticas, ser riemanniana não é sempre um fato, como vimos na seção anterior, em que as expressões para a conexão mistura (116) e exponencial (118) diferem uma em relação a outra. Em outras palavras, as conexões não são duais de si mesmas. Esse conceito de dualidade será formalizado nessa seção. A forma de preservar a métrica em variedades quaisquer é usar um par de conexões afins D e D^* duais. Denotamos o transporte paralelo associado às conexões D e D^* por Π e Π^* , respectivamente. Por definição, D e D^* são *conexões mutuamente duais* se AMARI (2013)

$$g(X, Y) = g(\Pi(X), \Pi^*(Y)). \quad (128)$$

Mais precisamente, para quaisquer campos de vetores $X, Y, Z \in \mathcal{T}(\mathcal{P})$, escolhidos nos espaços tangentes, D e D^* são mutuamente duais se a equação

$$Zg(X, Y) = g(D_Z X, Y) + g(X, D_Z^* Y) \quad (129)$$

é satisfeita e $\mathcal{T}(\mathcal{P})$ é o fibrado tangente da variedade estatística generalizada \mathcal{P} . Particularmente, considerando os vetores base do espaço tangente, podemos escrever a equação (129) como

$$\frac{\partial g_{jk}}{\partial \theta^i} = \Gamma_{ijk} + \Gamma_{ijk}^*. \quad (130)$$

Podemos estabelecer uma relação direta entre as equações (129) e (130), quando os campos de vetores descritos na equação (129) são, particularmente, os vetores base do plano tangente $\mathcal{T}(\mathcal{P})$. Portanto, podemos enunciar o teorema abaixo.

Teorema 3.1 *As definições de dualidade dadas nas equações (129) e (130) são equivalentes.*

Prova 7 *Caso exista uma conexão dual, a equação da dualidade*

$$g(X(t), Y^*(t)) = g(X(0), Y(0)), \quad \text{para todo } t,$$

é satisfeita, em que $X(t)$ é o transporte paralelo de $X(0)$ ao longo de t , com respeito ao símbolo Γ_{ijk} . Analogamente, $Y^(t)$ é o transporte paralelo de $Y(0)$ ao longo de t , com respeito ao símbolo Γ_{ijk}^* . Sejam $X = X^i \frac{\partial}{\partial \theta^i}$ e $Y = Y^{*j} \frac{\partial}{\partial \theta^j}$ dois campos vetoriais com as conexões D e D^* , respectivamente. Denotando por $\dot{\theta}$ como o campo vetorial para γ , podemos escrever*

$$\begin{aligned} 0 &= D_{\dot{\theta}} X = \left[\frac{d}{dt} X^i + \Gamma_{\lambda k}^i \dot{\theta}^\lambda X^k \right] \frac{\partial}{\partial \theta^i}, \\ 0 &= D_{\dot{\theta}} Y^* = \left[\frac{d}{dt} Y^{*j} + \Gamma_{\lambda k}^{*j} \dot{\theta}^\lambda Y^{*k} \right] \frac{\partial}{\partial \theta^j}, \end{aligned}$$

as quais nos proporcionam as seguintes expressões

$$\frac{d}{dt}X^i = -\Gamma_{\lambda k}^i \dot{\theta}^\lambda X^k, \quad (131)$$

$$\frac{d}{dt}Y^{*j} = -\Gamma_{\lambda k}^{*j} \dot{\theta}^\lambda Y^{*k}. \quad (132)$$

Usando as fórmulas (131) e (132), obtemos

$$\begin{aligned} 0 &= \frac{d}{dt}g(X, Y^*) \\ &= \frac{d}{dt}[X^i Y^{*j} g_{ij}] \\ &= -\Gamma_{\lambda k}^i \dot{\theta}^\lambda X^k Y^{*j} g_{ij} - X^i \Gamma_{\lambda k}^{*j} \dot{\theta}^\lambda Y^{*k} g_{ij} + X^i Y^j \frac{\partial}{\partial \theta^\lambda} g_{ij} \dot{\theta}^\lambda \\ &= \left(-\Gamma_{\lambda ij} - \Gamma_{\lambda ji}^* + \frac{\partial}{\partial \theta^\lambda} g_{ij} \right) X^i Y^{*j} \dot{\theta}^\lambda. \end{aligned} \quad (133)$$

Desde que $Y(0) = Y^*|_{\gamma(0)}$ e os vetores $X(0), Y(0)$ e $\dot{\theta}(0)$ possam assumir qualquer valor diferente de zero, concluímos que, afim de que a equação acima seja verdadeira, devemos ter $\partial_\lambda g_{ij} = \Gamma_{\lambda ij} + \Gamma_{\lambda ji}^*$, o qual é equivalente a (129).

Por outro lado, assumamos que (129) seja satisfeita. Segue-se que

$$\begin{aligned} \frac{d}{dt}g(X, Y^*) &= \dot{\theta}^i \frac{\partial}{\partial \theta^i} g(X, Y^*) \\ &= g(D_{\dot{\theta}} X, Y^*) + g(X, D_{\dot{\theta}}^* Y^*) \\ &= 0 + 0 = 0, \end{aligned}$$

o que implica em (130).

Nas variedades riemannianas, temos que a equação é sempre satisfeita, pois temos a conexão compatível com a métrica g , isto é, $D = D^*$. Em outras palavras, a conexão D é dual de si mesma. Além disso, sempre existe uma conexão dual.

Teorema 3.2 *A conexão dual D^* de uma conexão D qualquer sempre existe e é única.*

Prova 8 *Dada uma conexão D , podemos definir a sua conexão dual D^* através da expressão (130), em que Γ_{ijk} e Γ_{ik}^* são os símbolos de Christoffel das conexões D e D^* , respectivamente. A unicidade provém do fato da matriz da métrica ser tal que $\det(g_{ij}) \neq 0$.*

Exemplo 11 *Considere a expressão que relaciona duas conexões $\Gamma_{ijk}^{(\beta)}$ e $\Gamma_{ijk}^{(\alpha)}$ citada em (18). Fazendo $\alpha = 0$, obtemos que*

$$\Gamma_{ijk}^{(\beta)} = \Gamma_{ijk}^{(0)} + \frac{\beta}{2} T_{ijk},$$

em que $T_{ijk} = E_\theta[\frac{\partial l_\theta}{\partial \theta^i} \frac{\partial l_\theta}{\partial \theta^j} \frac{\partial l_\theta}{\partial \theta^k}]$. Dessa forma,

$$\begin{aligned} \Gamma_{ijk}^{(\beta)} + \Gamma_{ijk}^{(-\beta)} &= \Gamma_{ijk}^{(0)} + \frac{\beta}{2} T_{ijk} + \Gamma_{ijk}^{(0)} - \frac{\beta}{2} T_{ijk} \\ &= 2\Gamma_{ijk}^{(0)} \end{aligned}$$

o que pela (123) garantimos a dualidade entre $D^{(\beta)}$ e $D^{(-\beta)}$.

Consideremos dois campos de vetores X e Y ao longo de uma curva $\gamma : I \rightarrow \mathcal{M}$, \mathcal{M} uma variedade, que se comportam como o transporte paralelo dos vetores $X(t_0)$ e $Y(t_0)$, $t_0 \in I$, relativos a alguma conexão D . De um modo geral, transporte paralelo não preserva o produto interno. Em outras palavras,

$$g(X(t), Y(t)) \neq g(X(t_0), Y(t_0)), \quad \text{para algum } t \in I.$$

Vejamos a figura que se segue que mostra melhor essa restrição.

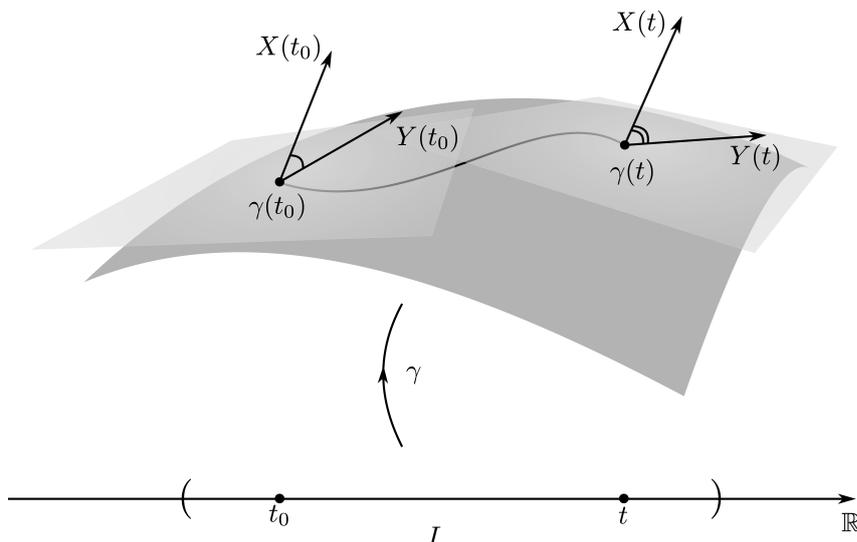


Figura 2: Transporte paralelo sem conexão dual.

Quando esse produto interno satisfizer a igualdade, é porque definimos que a conexão D é compatível com a métrica CARMO (1976a). Para o caso que a desigualdade ocorra, podemos garantir a igualdade ajustando o ângulo com um vetor Y^* de tal modo que preserve o ângulo formado com X seja o mesmo em t_0 e um t qualquer. Em outras palavras, dado uma conexão D , fazemos uso teorema 3.2 e tomamos uma outra conexão denotada D^* de tal modo que

$$g(X(t), Y^*(t)) = g(X(t_0), Y(t_0)), \quad \text{para todo } t \in I. \quad (134)$$

O campo vetorial Y^* é definido como o transporte paralelo de Y ao longo de γ . Portanto, podemos definir a dualidade entre duas conexões D e D^* quando o produto interno entre quaisquer vetores $X(t_0)$ e $Y(t_0)$ são preservados. Vejamos essa observação na figura que

se segue.

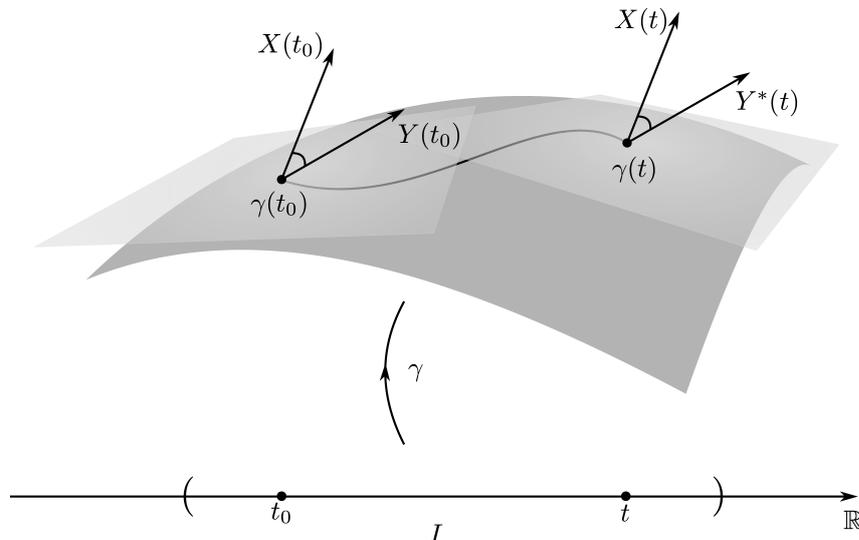


Figura 3: Transporte paralelo com conexão dual.

Um importante resultado é que podemos mostrar que as equações (129) e (134) são equivalentes. Vejamos isso no teorema que se segue.

Teorema 3.3 *As expressões que definem dualidade em (129) e (134) são equivalentes.*

Prova 9 *Assumamos como verdade (134). Dada uma curva $\gamma : (-\varepsilon, \varepsilon) \rightarrow \mathcal{M}$ com $\gamma(0) = p \in \mathcal{M}$, com a qual, devido a parametrização da variedade, podemos expressar o ponto p localmente como $(\gamma^1(t), \dots, \gamma^n(t))$ em algum sistema de coordenadas (x^1, \dots, x^n) . Ao longo da curva γ , denotemos o transporte paralelo dos vetores $X(0)$ e $Y(0)$ por X e Y^* , relativos às conexões D e D^* , duais entre si, respectivamente. Baseado em (60), escrevamos os vetores X e Y^* por*

$$X = \sum_i a^i \frac{\partial}{\partial x^i} \quad e \quad Y = \sum_j b^j \frac{\partial}{\partial x^j}.$$

Denotando os símbolos de Christoffel com respeito às conexões D e D^* por Γ_{ij}^k e Γ_{ij}^{*k} , respectivamente, temos

$$0 = D_{d\gamma/dt} X = \sum_m \left(\frac{da^m}{dt} + \sum_{i,j} \frac{d\gamma^i}{dt} a^j \Gamma_{ij}^m \right) \frac{\partial}{\partial x^m}$$

e

$$0 = D_{d\gamma/dt}^* Y^* = \sum_m \left(\frac{db^m}{dt} + \sum_{i,k} \frac{d\gamma^i}{dt} b^k \Gamma_{ik}^{*m} \right) \frac{\partial}{\partial x^m},$$

com os quais, lembrando que $\frac{\partial}{\partial x^i}$ é uma base para o espaço tangente em todo ponto de γ , respectivamente, temos

$$\frac{da^m}{dt} = - \sum_{i,j} \frac{d\gamma^i}{dt} a^j \Gamma_{ij}^m$$

e

$$\frac{db^m}{dt} = - \sum_{i,k} \frac{d\gamma^i}{dt} b^k \Gamma_{ik}^{*m},$$

para cada m . Portanto, usando essas fórmulas, obtemos

$$\begin{aligned} 0 &= \frac{d}{dt}g(X, Y^*) = \frac{d}{dt} \sum_{j,k} g_{jk} a^j b^k \\ &= \sum_{j,k} \frac{dg_{jk}}{dt} a^j b^k + \sum_{m,k} g_{mk} \frac{da^m}{dt} b^k + \sum_{j,m} g_{jm} a^j \frac{db^m}{dt} \\ &= \sum_{j,k} \frac{dg_{jk}}{dt} a^j b^k - \sum_{m,k} g_{m,k} \left(\sum_{i,j} \frac{d\gamma^i}{dt} a^j \Gamma_{ij}^m \right) b^k - \sum_{j,m} g_{jm} a^j \left(\sum_{i,k} \frac{d\gamma^i}{dt} b^k \Gamma_{ik}^{*m} \right) \\ &= \sum_{i,j,k} \frac{\partial g_{jk}}{\partial x^i} \frac{d\gamma^i}{dt} a^j b^k - \sum_{i,j,k} \Gamma_{ijk} \frac{d\gamma^i}{dt} a^j b^k - \sum_{i,j,k} \Gamma_{ijk}^* \frac{d\gamma^i}{dt} a^j b^k \\ &= \sum_{i,j,k} \left(\frac{\partial g_{jk}}{\partial x^i} - \Gamma_{ijk} - \Gamma_{ijk}^* \right) \frac{d\gamma^i}{dt} a^j b^k. \end{aligned}$$

Como os vetores $X(0)$, $Y(0)$ e $(d\gamma/dt)(0)$ podem assumir quaisquer valores, concluímos que

$$\frac{\partial g_{jk}}{\partial x^i} - \Gamma_{ijk} - \Gamma_{ijk}^* = 0.$$

Isto é,

$$\frac{\partial g_{jk}}{\partial x^i} = \Gamma_{ijk} + \Gamma_{ijk}^*,$$

o que como já vimos, equivale-se a (129).

Analogamente, assumamos como verdade (129). Podemos escrever

$$\begin{aligned} \frac{d}{dt}g(X, Y^*) &= g(D_{d\gamma/dt}X, Y^*) + g(X, D_{d\gamma/dt}^*Y^*) \\ &= g(0, Y^*) + g(X, 0) \\ &= 0. \end{aligned}$$

Logo, a variação do produto internos dos campos é zero, o que implica que eles preservam o ângulo entre eles. Isto é, vale (134).

Teorema 3.4 *Se uma variedade M admite uma conexão D , então existe um sistema de coordenadas locais em M tal que $\Gamma_{ij}^k = 0$. Além disso, as funções de transição entre os sistemas de coordenadas são transformações afins CARMO (1976a).*

Prova 10 *A prova desse teorema pode ser encontrada em SHIMA (2007).*

Vale salientar que um sistema de coordenadas (x^j) para o qual vale $\Gamma_{ij}^k = 0$ com respeito à conexão D é chamado de afim. Um outro resultado importante relaciona conexão *flat* com conexão dual *flat*, o que é uma consequência direta do teorema que se segue.

Teorema 3.5 *Os símbolos dos tensores curvaturas R_{ijkl} e R_{ijkl}^* das respectivas conexões duais D e D^* satisfazem*

$$R_{ijkl} = -R_{ijlk}^*.$$

Prova 11 *Considere a equação da dualidade dada em (129). Podemos escrever*

$$\begin{aligned} g\left(D_{\partial/\partial x^i} D_{\partial/\partial x^j} \frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^l}\right) &= \frac{\partial}{\partial x^i} g\left(D_{\partial/\partial x^j} \frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^l}\right) \\ &\quad - g\left(D_{\partial/\partial x^j} \frac{\partial}{\partial x^k}, D_{\partial/\partial x^i}^* \frac{\partial}{\partial x^l}\right) \\ &= \frac{\partial^2}{\partial x^i \partial x^j} g\left(\frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^l}\right) - \frac{\partial}{\partial x^i} g\left(\frac{\partial}{\partial x^k}, D_{\partial/\partial x^j}^* \frac{\partial}{\partial x^l}\right) \\ &\quad - \frac{\partial}{\partial x^j} g\left(\frac{\partial}{\partial x^k}, D_{\partial/\partial x^i}^* \frac{\partial}{\partial x^l}\right) \\ &\quad + g\left(\frac{\partial}{\partial x^k}, D_{\partial/\partial x^j}^* D_{\partial/\partial x^i}^* \frac{\partial}{\partial x^l}\right) \end{aligned}$$

e

$$\begin{aligned} g\left(D_{\partial/\partial x^i} D_{\partial/\partial x^j} \frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^l}\right) &= \frac{\partial}{\partial x^j} g\left(D_{\partial/\partial x^i} \frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^l}\right) \\ &\quad - g\left(D_{\partial/\partial x^i} \frac{\partial}{\partial x^k}, D_{\partial/\partial x^j}^* \frac{\partial}{\partial x^l}\right) \\ &= \frac{\partial^2}{\partial x^i \partial x^j} g\left(\frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^l}\right) - \frac{\partial}{\partial x^j} g\left(\frac{\partial}{\partial x^k}, D_{\partial/\partial x^i}^* \frac{\partial}{\partial x^l}\right) \\ &\quad - \frac{\partial}{\partial x^i} g\left(\frac{\partial}{\partial x^k}, D_{\partial/\partial x^j}^* \frac{\partial}{\partial x^l}\right) \\ &\quad + g\left(\frac{\partial}{\partial x^k}, D_{\partial/\partial x^i}^* D_{\partial/\partial x^j}^* \frac{\partial}{\partial x^l}\right). \end{aligned}$$

Segue-se que

$$\begin{aligned} R_{ijkl} &= g\left(D_{\partial/\partial x^i} D_{\partial/\partial x^j} \frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^l}\right) - g\left(D_{\partial/\partial x^j} D_{\partial/\partial x^i} \frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^l}\right) \\ &= g\left(D_{\partial/\partial x^k}^* D_{\partial/\partial x^j}^* \frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^l}\right) - g\left(D_{\partial/\partial x^k} D_{\partial/\partial x^i} \frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^l}\right) \\ &= -R_{ijlk}^*, \end{aligned}$$

a qual é a expressão desejada.

O teorema anterior nos permite formalizar a seguinte definição.

Definição 3.3 *Uma variedade \mathcal{M} que seja flat com respeito a um par de conexões duais é chamada de flat dual.*

Embora de maneira implícita, exemplificamos variedades flat em (28) e (30).

3.6.2 Coordenadas Duais

Para uma variedade *flat* dual com conexões D e D^* , podemos adicionar dois sistemas de coordenadas, duais entre si, que se relacionam com D e D^* . Definimos esses sistemas de coordenadas duais como se segue:

Definição 3.4 *Dois sistemas de coordenadas (θ^i) e (η_j) são ditos duais entre si se seus vetores bases coordenados satisfazem a condição*

$$g\left(\frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\eta_j}\right) = \delta_i^j,$$

em que $\delta_i^j = 1$ para $i = j$, e $\delta_i^j = 0$ para $i \neq j$.

De um modo geral, não temos garantia da existência de dois sistemas de coordenadas *flat* duais em variedades quaisquer, mas para as variedades estatísticas, temos essa garantia.

Sejam (θ^i) e (η_j) dois sistema de coordenadas duais. Com esse raciocínio, expressamos cada coordenada por

$$\theta^i = \theta^i(\eta_1, \dots, \eta_n) \quad \text{e} \quad \eta_j = \eta_j(\theta^1, \dots, \theta^n),$$

na qual recuperamos pela regra da cadeia que

$$\frac{\partial}{\partial\theta^i} = \sum_j \frac{\partial\eta_j}{\partial\theta^i} \frac{\partial}{\partial\eta_j} \quad \text{e} \quad \frac{\partial}{\partial\eta_j} = \sum_i \frac{\partial\theta^i}{\partial\eta_j} \frac{\partial}{\partial\theta^i}. \quad (135)$$

As matrizes $g_{ij}(\theta) = g\left(\frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\theta^j}\right)$ e $h_{ij}(\eta) = g\left(\frac{\partial}{\partial\eta_i}, \frac{\partial}{\partial\eta_j}\right)$ são induzidas pelos vetores das respectivas bases coordenadas duais (135). Garante-se que elas são inversas uma da outra MATSUZOE and HENMI (2013); MATSUZOE (2014), isto é, $h_{ji} = g_{ij}$. Este resultado segue de

$$g_{ij}(\theta) = g\left(\frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\theta^j}\right) = \sum_k \frac{\partial\eta_k}{\partial\theta^i} g\left(\frac{\partial}{\partial\eta_k}, \frac{\partial}{\partial\theta^j}\right) = \frac{\partial\eta_j}{\partial\theta^i}, \quad (136)$$

e

$$h_{ij}(\eta) = g\left(\frac{\partial}{\partial\eta_i}, \frac{\partial}{\partial\eta_j}\right) = \sum_k \frac{\partial\theta^k}{\partial\eta_j} g\left(\frac{\partial}{\partial\eta_i}, \frac{\partial}{\partial\theta^k}\right) = \frac{\partial\theta^i}{\partial\eta_j}, \quad (137)$$

desde que as matrizes jacobianas $\left(\frac{\partial\eta_j}{\partial\theta^i}\right)$ e $\left(\frac{\partial\theta^i}{\partial\eta_j}\right)$ são matrizes inversas uma da outra. A expressão (135) pode ser reescrita como

$$\frac{\partial}{\partial\theta^i} = \sum_j g_{ij} \frac{\partial}{\partial\eta_j}, \quad \frac{\partial}{\partial\eta_j} = \sum_i g^{ij} \frac{\partial}{\partial\theta^i}. \quad (138)$$

No caso do sistema de coordenadas euclidiano, as coordenadas duais coincidem com as convencionais.

O próximo resultado mostra que um sistema de coordenadas duais podem ser expressos em termos das funções potenciais SHIMA (2007).

Teorema 3.6 *Sejam (θ^i) e (η_j) dois sistemas de coordenadas cujas respectivas funções de Legendre $\psi(\theta)$ e $\psi^*(\eta)$ são tais que MATSUZOE and HENMI (2013); MATSUZOE (2014)*

$$\theta^i = \frac{\partial \psi^*}{\partial \eta_i} \quad e \quad \eta_j = \frac{\partial \psi}{\partial \theta^j}. \quad (139)$$

Portanto, temos

$$g_{ij} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j} \quad e \quad g^{ij} = \frac{\partial^2 \psi^*}{\partial \eta_i \partial \eta_j}. \quad (140)$$

Adicionalmente,

$$\psi(\theta) + \psi^*(\eta) = \theta^i \eta_i. \quad (141)$$

De modo inverso, quando a função potencial $\psi(\theta)$ existe tal que $g_{ij} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}$, temos um sistema de coordenadas (η_j) cujo dual é o (θ^i) e a equação (141) nos proporciona a outra função de Legendre $\psi^*(\eta)$ respectiva para as coordenadas duais.

Prova 12 *Esta prova pode ser encontrada em COELHO et al. (2015). Outras referências como AMARI and NAGAOKA (2000); MATSUZOE and HENMI (2013) também falam desse teorema. Mesmo assim, escreveremos a prova, pois ela explicita relação entre as coordenadas convencionais (θ^i) e as duais (η_j) . Sejam $(\theta)^i$ e (η_j) um sistema de coordenadas duais entre si. A expressão (136) garante que sempre temos $g_{ij} = \frac{\partial \eta_j}{\partial \theta^i}$, cuja simetria resulta em $\frac{\partial \eta_j}{\partial \theta^i} = \frac{\partial \eta_i}{\partial \theta^j}$, o que implica em $\eta_j = \frac{\partial \psi}{\partial \theta^j}$. Analogamente, a expressão (137), pelos mesmos argumentos implica que $\theta^i = \frac{\partial \psi^*}{\partial \eta^i}$. As expressões (140) são obtidas como em (66). Para mostrar a equação (141), vejamos que*

$$\begin{aligned} \frac{\partial^2}{\partial \eta_i \partial \eta_j} \left(\sum_l \theta^l \eta_l \right) &= \frac{\partial}{\partial \eta_i} \left(\sum_l \frac{\partial \theta^l}{\partial \eta_j} \eta_l + \theta^j \right) \\ &= \sum_l \frac{\partial^2 \theta^l}{\partial \eta_i \partial \eta_j} \eta_l + \frac{\partial \theta^i}{\partial \eta_j} + \frac{\partial \theta^j}{\partial \eta_i} \\ &= \left(\sum_l \frac{\partial^2 \theta^l}{\partial \eta_i \partial \eta_j} \eta_l \right) + \frac{\partial \psi^*}{\partial \eta_i \partial \eta_j}. \end{aligned} \quad (142)$$

Considerando o fato de que

$$\begin{aligned}
\frac{\partial^2 \psi^*}{\partial \eta_i \partial \eta_j} &= g\left(\frac{\partial}{\partial \eta_i}, \frac{\partial}{\partial \eta_j}\right) \\
&= \sum_{k,l} g\left(\frac{\partial}{\partial \theta^k}, \frac{\partial}{\partial \theta^l}\right) \frac{\partial \theta^k}{\partial \eta_i} \frac{\partial \theta^l}{\partial \eta_j} \\
&= \sum_{k,l} g_{kl} \frac{\partial \theta^k}{\partial \eta_i} \frac{\partial \theta^l}{\partial \eta_j} \\
&= \sum_{k,l} \frac{\partial^2 \psi}{\partial \theta^k \partial \theta^l} \frac{\partial \theta^k}{\partial \eta_i} \frac{\partial \theta^l}{\partial \eta_j},
\end{aligned}$$

podemos escrever

$$\begin{aligned}
\frac{\partial^2 \psi^*}{\partial \eta_i \partial \eta_j} &= \sum_l \frac{\partial}{\partial \eta_i} \frac{\partial \psi}{\partial \theta^l} \frac{\partial \theta^l}{\partial \eta_j} \\
&= \frac{\partial}{\partial \eta_i} \sum_l \left(\frac{\partial \psi}{\partial \theta^l} \frac{\partial \theta^l}{\partial \eta_j} \right) - \sum_l \frac{\partial \psi}{\partial \theta^l} \frac{\partial^2 \theta^l}{\partial \eta_i \partial \eta_j} \\
&= \frac{\partial^2 \psi}{\partial \eta_i \partial \eta_j} - \sum_l \frac{\partial^2 \theta^l}{\partial \eta_i \partial \eta_j} \eta_l
\end{aligned} \tag{143}$$

Usando (143) em (142), temos

$$\frac{\partial^2 \psi}{\partial \eta_i \partial \eta_j} + \frac{\partial^2 \psi^*}{\partial \eta_i \partial \eta_j} = \frac{\partial^2}{\partial \eta_i \partial \eta_j} \left(\sum_l \theta^l \eta_l \right),$$

a qual resulta em

$$\psi + \psi^* = \sum_i \theta^i \eta_i + \sum_i a^i \eta_i + a, \tag{144}$$

para algumas constantes a^i e a . Diferenciando ambos os lados de (144) com respeito a η_i , obtemos $a^i = 0$. Analogamente, assumamos que ψ seja uma função normalizadora (Legendre) em \mathcal{P} tal que $g_{ij} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}$. O sistema de coordenadas $(\eta_j) = \left(\frac{\partial \psi}{\partial \theta^j}\right)$ satisfaz

$$\begin{aligned}
g\left(\frac{\partial}{\partial \theta^i}, \frac{\partial}{\partial \eta_j}\right) &= g\left(\frac{\partial}{\partial \theta^i}, \sum_k \frac{\partial \theta^k}{\partial \eta_j} \frac{\partial}{\partial \theta^k}\right) \\
&= \sum_k g\left(\frac{\partial}{\partial \theta^i}, \frac{\partial}{\partial \theta^k}\right) \frac{\partial \theta^k}{\partial \eta_j} \\
&= \sum_k g_{ik} \frac{\partial \theta^k}{\partial \eta_j}.
\end{aligned}$$

Além do mais, observe que $\left(\frac{\partial \theta^k}{\partial \eta_j}\right)$ é a inversa de $(g_{kj}) = \left(\frac{\partial \eta_j}{\partial \theta^k}\right)$. Então, concluímos que $g\left(\frac{\partial}{\partial \theta^i}, \frac{\partial}{\partial \eta_j}\right) = \delta_i^j$, mostrando que o sistema de coordenadas (θ^i) e (η_j) são duais entre si.

O exemplo 3 exhibe para uma família de gaussianas as coordenadas convencionais.

Exemplo 12 *Sejam T um conjunto finito e $p \in \mathcal{P}_\mu$. Mais uma vez, tomemos a família centrada em p $\mathcal{E}_p = \{p_\theta : \theta \in \Theta\}$, cujas distribuições são descritas através de*

$$p_\theta = \exp\left(\sum_{i=1}^n \theta^i u_i - \psi(\theta)\right)p,$$

para as funções exponenciais $u_1, \dots, u_n : T \rightarrow \mathbb{R}$ satisfazendo as condições (U1) e (U2), citadas na seção 2.2.2, quando se define a família exponencial. Dadas as distribuições de probabilidade p_1, \dots, p_n , consideramos a família mistura $\mathcal{M}_p = \{q_\eta : \eta \in H\} \subseteq \mathcal{P}_\mu$ de todas as distribuições de probabilidades da forma

$$q_\eta = \sum_{i=1}^n \eta_i p_i + \left(1 - \sum_{i=1}^n \eta_i\right)p.$$

Devido ao fato de T ser finito, temos que $\mathcal{P}_\mu = \mathcal{E}_p = \mathcal{M}_p$. Então, a fim de garantir que os sistemas de coordenadas (θ^i) e (η_j) , imporemos condições sobre as funções u_1, \dots, u_n e as distribuições de probabilidades p_1, \dots, p_n . Com efeito, suponha que os sistemas de coordenadas (θ^i) e (η_j) sejam mutuamente duais. Da igualdade entre os conjuntos $\mathcal{P}_\mu = \mathcal{M}_p$, derivemos em relação ao parâmetro θ^j a equação $p_\theta(x) = q_\eta(x)$. Obtemos

$$\sum_{i=1}^n \frac{\partial \eta_i}{\partial \theta^j} (p_i - p) = p_\theta \left(u_j - \frac{\partial \psi}{\partial \theta^j} \right).$$

Desde que $\frac{\partial \psi}{\partial \theta^j}(0) = 0$ e $g_{ij} = \frac{\partial \eta_i}{\partial \theta^j}$, segue-se que

$$\sum_{i=1}^n \frac{p_i - p}{p} g_{ij}(0) = u_j. \quad (145)$$

Por outro lado, assumamos que as funções p_1, \dots, p_n e u_1, \dots, u_n satisfazem a condição (145). Sejam $\theta = (\theta^i)$ e $\eta = (\eta_j)$ tais que $p_\theta(x) = q_\eta(x)$. Da propriedade (U1) $E_\theta[u_j] = 0$, para todo j , temos

$$E_\theta[u_j] = \sum_{i=1}^n \eta_i E_{p_i}[u_j]. \quad (146)$$

Aplicando a derivada em relação ao parâmetro θ^i em $\int_T p_\theta(x) d\mu = 1$, temos

$$0 = \int_T \frac{\partial p_\theta}{\partial \theta^i} d\mu = \int_T p_\theta \left(u_i(x) - \frac{\partial \psi}{\partial \theta^i} \right) d\mu,$$

na qual implica que

$$\frac{\partial \psi}{\partial \theta^i} = E_p[u_i]. \quad (147)$$

Multiplicando ambos os lados de (145) por u_k e aplicando o operador esperança $E_p[\cdot]$, obtemos

$$\sum_{i=1}^n E_{p_i}[u_k] g_{ij}(0) = E_p[u_k u_j] = g_{kj}(0).$$

Devido a matriz (g_{ij}) ser positiva definida, segue que $E_{p_i}[u_k] = \delta_i^k$. Este resultado junto com as expressões (146) e (147) implica que $\frac{\partial \psi}{\partial \theta^i} = \eta_i$. Logo, o teorema (3.6) garante que os sistemas de coordenadas (η_j) e (θ^i) são mutuamente duais.

Teorema 3.7 *Se uma variedade \mathcal{M} é flat com respeito a um par de conexões duais entre si D e D^* , então temos um sistema de coordenadas dual (θ^i) e (η_j) , respectivamente.*

Prova 13 *Seja (θ^i) um sistema de coordenadas afim, em que $\Gamma_{ijk} = 0$ com respeito à conexão D . Da equação da dualidade (129), temos que $\Gamma_{ijk}^* = \frac{\partial g_{jk}}{\partial \theta^i}$. Do fato de D^* ser flat AMARI and NAGAOKA (2000), escrevemos $\frac{\partial g_{jk}}{\partial \theta^i} = \Gamma_{ijk}^* = \Gamma_{jik}^* = \frac{\partial g_{ik}}{\partial \theta^j}$, mostrando a existência da função η_k tal que $g_{jk} = \frac{\partial \eta_k}{\partial \theta^j}$. Pela simetria, $g_{ij} = g_{ji}$, segue-se que $\frac{\partial \eta_i}{\partial \theta^j} = \frac{\partial \eta_j}{\partial \theta^i}$. Como consequência, existe uma função potencial ψ em \mathcal{M} , tal que $g_{ij} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}$. Do Teorema Proposição 3.6, garantimos a existência de e um sistema de coordenadas (η_j) dual com respeito a (θ^i) . Resta-nos mostrar que (η_j) é afim com respeito a D^* . Pela definição de dualidade,*

$$\frac{\partial}{\partial \theta^i} g \left(\frac{\partial}{\partial \theta^j}, \frac{\partial}{\partial \eta_k} \right) = \frac{\partial}{\partial \theta^i} \delta_j^k = 0.$$

Usando as expressões (138), podemos escrever

$$\begin{aligned} \frac{\partial}{\partial \theta^i} g \left(\frac{\partial}{\partial \theta^j}, \frac{\partial}{\partial \eta_k} \right) &= g \left(D_{\frac{\partial}{\partial \theta^i}} \frac{\partial}{\partial \theta^j}, \frac{\partial}{\partial \eta_k} \right) + g \left(\frac{\partial}{\partial \theta^j}, D_{\frac{\partial}{\partial \theta^i}}^* \frac{\partial}{\partial \eta_k} \right) \\ &= \sum_l g^{lk} g \left(D_{\frac{\partial}{\partial \theta^i}} \frac{\partial}{\partial \theta^j}, \frac{\partial}{\partial \theta^l} \right) + \sum_{l,m} g_{jl} g_{im} g \left(\frac{\partial}{\partial \eta_l}, D_{\frac{\partial}{\partial \theta^i}}^* \frac{\partial}{\partial \eta_m} \right) \\ &= \sum_{l,m} g_{jl} g_{im} g \left(\frac{\partial}{\partial \eta_l}, D_{\frac{\partial}{\partial \theta^i}}^* \frac{\partial}{\partial \eta_m} \right), \end{aligned}$$

o qual mostra que os símbolos de Christoffel de D^* com respeito a (η_j) são identicamente nulos. Logo, garantimos que (η_j) é um sistema de coordenadas dual para a conexão D^* .

Dois outros resultados importantes envolvendo as conexões duais são o Teorema de Pitágoras e o Teorema da Projecção, ambos aplicados em variedades duais AMARI and NAGAOKA (2000). Seja \mathcal{M} uma variedade flat equipada com um par de conexões D e D^* duais entre si. Ainda de acordo com a notação já citada, sejam mais uma vez (θ^i) e (η_j) sistemas de coordenadas duais com respeito às conexões D e D^* , respectivamente, cujas funções potenciais são ψ e ψ^* , também respectivamente. Conforme visto em SHIMA (2007), dada uma variedade flat dual, a divergência entre dois pontos p e q em

\mathcal{M} é definida como

$$\mathcal{D}(p \parallel q) = \psi(p) + \psi^*(q) - \sum_i \theta^i(p) \eta_i(q). \quad (148)$$

Mais uma vez cito que pode-se recuperar a métrica de \mathcal{M} usando a divergência (148). Precisamente,

$$\left[\left(\frac{\partial^2}{\partial \theta^i \partial \theta^j} \right)_p \mathcal{D}(p \parallel q) \right]_{q=p} = g_{ij}(p) \quad \text{e} \quad \left[\left(\frac{\partial^2}{\partial \eta^i \partial \eta^j} \right)_q \mathcal{D}(p \parallel q) \right]_{q=p} = g^{ij}(q).$$

No sistema de coordenadas (θ^i) , os coeficientes das conexões são dados por

$$\Gamma_{ijk} = 0 \quad \text{e} \quad \Gamma_{ijk}^* = \frac{\partial^3 \psi}{\partial \theta^i \partial \theta^j \partial \theta^k}.$$

Analogamente em ao sistema de coordenadas (η_j) ,

$$\Gamma_{ijk} = \frac{\partial^3 \psi^*}{\partial \eta_i \partial \eta_j \partial \eta_k} \quad \text{e} \quad \Gamma_{ijk}^* = 0.$$

Considere as seguintes expressões abaixo, nas quais $p, q, r \in \mathcal{M}$,

$$\mathcal{D}(p \parallel q) = \psi(p) + \psi^*(q) - \sum_{i=1}^n \theta^i(p) \eta_i(q)$$

e

$$\mathcal{D}(q \parallel r) = \psi(q) + \psi^*(r) - \sum_{i=1}^n \theta^i(q) \eta_i(r).$$

Com algumas manipulações, obtemos a expressão

$$\begin{aligned} \mathcal{D}(p \parallel q) + \mathcal{D}(q \parallel r) - \mathcal{D}(p \parallel r) &= \sum_{i=1}^n \theta^i(p) \eta_i(r) + \sum_{i=1}^n \theta^i(q) \eta_i(q) \\ &\quad - \sum_{i=1}^n \theta^i(p) \eta_i(r) - \sum_{i=1}^n \theta^i(q) \eta_i(r) \\ &= \sum_{i=1}^n [\theta^i(p) - \theta^i(q)] \cdot [\eta_i(r) - \eta_i(q)]. \end{aligned} \quad (149)$$

A expressão (149) é chamada de *desigualdade triangular*. Através dessa relação, podemos estender o Teorema de Pitágoras para \mathcal{P} . Antes de enunciá-lo, vejamos a seguinte definição auxiliar abaixo.

Definição 3.5 Para duas distribuições p_{θ_1} e p_{θ_2} em \mathcal{P} , a curva

$$\theta(t) = t\theta^1 + (1-t)\theta^2 \quad (150)$$

é chamada de φ -geodésica, em que $t \in [0, 1]$ é o parâmetro da curva e (θ^i) é o sistema de

coordenadas convencional. Analogamente, a φ -geodésica dual é a curva dada por

$$\eta(t) = t\eta_1 + (1 - t)\eta_2, \quad (151)$$

em que $t \in [0, 1]$ é o parâmetro da curva e (η_i) é o sistema de coordenadas dual.

Agora, podemos generalizar o Teorema de Pitágoras.

Teorema 3.8 *Sejam p, q e r três pontos em \mathcal{P} . Considere que a geodésica γ_1 , relacionada à conexão D , conecta p a q e a geodésica γ_2 , relacionada à conexão D^* , conecta q a r . Então, a relação pitagórica*

$$\mathcal{D}(p \parallel q) + \mathcal{D}(q \parallel r) = \mathcal{D}(p \parallel r) \quad (152)$$

é satisfeita.

Prova 14 *Sejam (θ^i) e (η_j) dois sistemas de coordenadas afins dual de \mathcal{P} com respeito às conexões D e D^* , respectivamente. Considere as geodésicas $\gamma_1(t)$ e $\gamma_2(t)$ como em (150) e (151), respectivamente. Além disso, sejam os vetores tangentes ao longo das geodésicas dados por*

$$\frac{d\gamma_1}{dt} = \sum_i [\theta^i(p) - \theta^i(q)] \frac{\partial}{\partial \theta^i} \quad (153)$$

e

$$\frac{d\gamma_2}{dt} = \sum_i [\eta_i(r) - \eta_i(q)] \frac{\partial}{\partial \eta_i}. \quad (154)$$

Usando a dualidade das coordenadas (θ^i) e (η_j) , as curvas são ortogonais em q . Nessa particularidade, a expressão (149) se resume ao Teorema de Pitágoras.

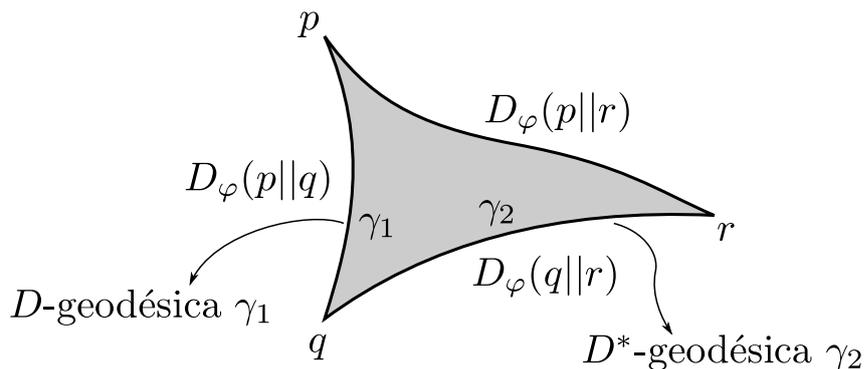


Figura 4: Teorema de Pitágoras.

Uma consequência direta do teorema acima é o que denominamos por Teorema da Projeção.

Corolário 3.1 *Sejam \mathcal{P} uma variedade flat dual e \mathcal{N} uma subvariedade auto-paralela de \mathcal{P} , com respeito à conexão dual D^* . Fixado um ponto $\mathcal{P} \setminus \mathcal{N}$, o mínimo da função $r \mapsto \mathcal{D}_\varphi(p \parallel r)$ é atingido no ponto q se, e somente se, a D -geodésica conectando p a q é ortogonal a \mathcal{N} em q .*

Prova 15 *Sejam (ϑ^K) um sistema de coordenadas de $r \in \mathcal{N}$ e $\gamma_1(t)$ a geodésica associada à conexão D ligando os pontos p e r por $t\theta(p) + (1-t)\theta(r)$. De (148), podemos escrever*

$$\begin{aligned} \frac{\partial}{\partial \vartheta^k} \mathcal{D}(p \parallel r) &= \sum_{i,j} \frac{\partial \theta^i}{\partial \vartheta^k} \frac{\partial \eta_j}{\partial \theta^i} \frac{\partial}{\partial \eta_j} \mathcal{D}(p \parallel r) \\ &= - \sum_{i,j} [\theta^j(p) - \theta^j(r)] \frac{\partial \theta^i}{\partial \vartheta^k} g_{ij} \\ &= -g \left(\sum_j [\theta^j(p) - \theta^j(r)] \frac{\partial}{\partial \theta^j}, \sum_i \frac{\partial \theta^i}{\partial \vartheta^k} \frac{\partial}{\partial \theta^i} \right) \\ &= -g \left(\frac{d\gamma_1}{dt}, \frac{\partial}{\partial \vartheta^k} \right) \end{aligned}$$

Portanto, $\frac{\partial}{\partial \vartheta^k} \mathcal{D}(p \parallel q) = 0$ se, e somente se, $\gamma_1(t)$ é ortogonal em $q \in \mathcal{N}$.

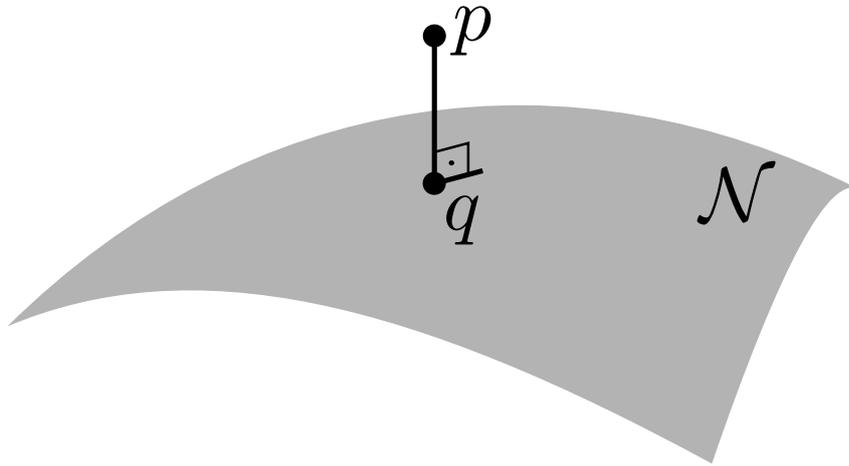


Figura 5: Teorema da Projeção.

Nossas contribuições ao longo desse capítulo focaram-se na parte de geometria. No capítulo que se segue, vamos mostrar alguns resultados de aplicabilidade. Especialmente, abordaremos como se estimar parâmetros que especificam as distribuições de probabilidade p_θ . Para tal, algumas definições e resultados preliminares de inferência serão abordados, para que apliquemos ao contexto da φ -família.

3.7 Síntese

Neste capítulo se concentra-se a maior parte dos resultados dessa tese. Motivado nas pesquisas em Geometria da Informação que são amplamente difundidas, como podemos ver nos trabalhos indicados em AMARI and NAGAOKA (2000); AMARI (2013); KANIADAKIS, LISSIA, and SCARFONE (2004); NAUDTS (2011); CENA and PISTONE (2007), dentre muitas outras, apresentamos a geometria de uma nova família paramétrica, denominada φ -família. Esta família generaliza as famílias exponenciais deformadas, logo também a família exponencial, esta última amplamente estudada por AMARI and NA-

GAOKA (2000); AMARI, OHARA, and MATSUZOE (2012); AMARI (2013). Definimos uma função chamada φ -função que possui propriedades similares à função exponencial, que foi introduzida em VIGELIS and CAVALCANTE (2013b). Ao trocarmos a função exponencial por uma φ -função, construímos uma função de divergência com a qual induzimos a estrutura de geometria da φ -família.

Vimos que a φ -divergência generaliza a divergência de Rényi e considerando uma família exponencial centrada em uma função de probabilidade p , a φ -divergência coincide com a de Kullback–Leibler. Recuperamos um par de conexões duais, com as quais calculamos os seus respectivos símbolos de Christoffel. Por fim, estudamos as coordenadas duais com as quais exibimos uma versão do Teorema de Pitágoras e o Teorema da Projeção para as φ -família.

4 ESTIMADOR DE MÁXIMA VEROSSIMILHANÇA MODIFICADO

Neste capítulo propomos uma modificação para o *Maximum Likelihood Estimator* (do inglês MLE) desenvolvido por R. A. Fisher ALDRICH (1997) para uma amostra segundo uma distribuição na φ -família, como definido em 3.1. A importância do MLE está nas aplicações possíveis, como ocorre em sistemas de comunicação, inferência estatística, processamento de sinais dentre outros, como podemos citar KAY (1993); Trees, Bell, and Tian (2013). O fundamento geral do MLE está em tomar uma amostra de uma variável aleatória X e estimar qual a distribuição de probabilidade que melhor caracteriza essa amostra. Essa ideia será estendida para as distribuições definidas na φ -família. Será mostrada uma versão generalizada do método de estimação de máxima verossimilhança. Este método depende fortemente do uso da φ -função para estimar os parâmetros dada uma amostra. Do fato que uma φ -família é paramétrica e dada uma amostra distribuída segundo uma φ -função p_θ como em (65), identificamos o parâmetro θ da distribuição p_θ .

Mostramos uma condição necessária para que um valor θ^* maximize a função de verossimilhança desenvolvida para a φ -família. Mostraremos que esse fato está intrinsecamente ligado com a minimização da φ -divergência, citada no capítulo anterior. Definimos uma classe especial de φ -função denominada de φ -gaussiana, a qual será útil para efeitos de simulação. Além do mais, proporemos um algoritmo de busca de parâmetro nas φ -gaussianas. Além disso, como é natural analisar em problemas de estimação, estudaremos questões de convergência assintótica do método estabelecido, em que algumas simulações dão indicativos de sua eficácia e exibiremos uma versão mais geral da Desigualdade de Cramér–Rao. Por fim, algumas propriedades da φ -família nos permitirão inferir sobre critérios estatísticos intrínsecos a \mathcal{P} .

4.1 Estimador de Máxima Verossimilhança Generalizado

Nessa seção, apresenta-se um resumo dos principais conceitos e resultados que tratam de estimação. Com base neles e na caracterização da φ -família dada no capítulo anterior, vamos mostrar uma generalização do método clássico de máxima verossimilhança. Alguns teoremas e proposições já bem conhecidos não serão necessariamente demonstrados, mas as devidas referências dos teoremas são citadas.

4.1.1 Preliminares de Teoria da Estimação e Análise de Convergência

Para a teoria de estimação de parâmetros, consideremos uma sequência de variáveis aleatórias X_1, \dots, X_n independentes e identicamente distribuídas a uma função densidade de probabilidade p_θ . Por questões de abreviação, ao termo independente e identicamente

distribuídas segundo uma função de densidade p_θ usaremos iid Mais uma vez citamos, o parâmetro $\theta \in \Theta$ associa-se injetivamente a uma função densidade de probabilidade p_θ . Isto é, p_θ pertence a uma família paramétrica como mostrado em (49). Um dos objetivos dessa seção é desenvolver uma *função estatística* $E_n = E_n(X_1, \dots, X_n)$ KAY (1993) em que dadas n observações de uma amostra, possamos estimar o parâmetro θ que caracterize p_θ . Em outras palavras, a função E_n é um *estimador* de θ .

Definição 4.1 *Seja $\{X_1, \dots, X_n\}$ uma amostra de uma distribuição parametrizada por θ . O erro quadrado médio (MSE proveniente do inglês) de um estimativa $\hat{\theta}$ de θ é definido como WALPOLE et al. (1993)*

$$EQM(\hat{\theta}) = E_\theta[\hat{\theta} - \theta]^2. \quad (155)$$

A expressão (155) pode ser reescrita por

$$\begin{aligned} EQM(\hat{\theta}) &= E_\theta[(\hat{\theta} - E_\theta[\hat{\theta}]) + (E_\theta[\hat{\theta}] - \theta)]^2 \\ &= E_\theta[\hat{\theta} - E_\theta[\hat{\theta}]]^2 + [E_\theta(\hat{\theta}) - \theta]^2 \\ &= Var[\hat{\theta}] + [E_\theta(\hat{\theta}) - \theta]^2, \end{aligned} \quad (156)$$

na qual o termo $E_\theta(\hat{\theta}) - \theta$ é chamado de *viés* do estimador LEON-GARCIA (2008).

Quando a função de densidade conjunta $L_\theta(X_1, \dots, X_n)$ é vista como uma função de θ , dada uma amostra $\{X_1, \dots, X_n\}$, esta é chamada de *função de verossimilhança* LEON-GARCIA (2008). Em algumas situações, ao invés de se trabalhar com a função de verossimilhança L_θ , trabalha-se com o logaritmo de L_θ , resultando na função $l_\theta = \log(L_\theta)$ LEON-GARCIA (2008). Especialmente quando a amostra é i.i.d., essa mudança torna-se mais útil, devido a propriedades do logaritmos e da definição de variável aleatória independente LIPTSER and SHIRYAEV (2013), o que torna a função l_θ linearizada. Esse ponto será de extrema importância para definirmos a versão com a qual generalizaremos o método de estimação de máxima verossimilhança generalizado.

Das funções estatísticas mais conhecidas, destaca-se a *função score* KAY (2013)

$$U_n(\theta) = \frac{\partial}{\partial \theta} l_\theta(X_1, \dots, X_n) \quad (157)$$

devido à propriedades intrínsecas, em especial a que em média ela é identicamente nula para todo parâmetro. Analisando a expressão (157), percebe-se que ela mede a taxa com a qual a função de verossimilhança varia de acordo com a variação do parâmetro θ . Uma particularidade que decorre de (157) é que quando a função de probabilidade conjunta da amostra se fatora no produto de suas marginais, temos

$$U_n(\theta) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \log p_\theta(x_i), \quad (158)$$

em que $p_\theta(x_i)$ é a função densidade de probabilidade da variável aleatória X_i e desde que X_1, \dots, X_n sejam iid. Além disso, a expressão (158) tem parcelas iguais. Aplicando o operador esperança em (158), obtemos

$$\mathbb{E}_\theta[U_n(\theta)] = n\mathbb{E}_\theta\left[\frac{p'(X_1)}{p_\theta(X_1)}\right] = n \int_T \frac{p'(X_1)}{p_\theta(X_1)} p_\theta dx.$$

Assumindo mais uma vez que o sinal de integração e a derivada em relação ao parâmetro θ comutam entre si e que o conjunto

$$\{x : p_\theta(x) > 0\}$$

não depende de θ , podemos escrever

$$\mathbb{E}_\theta\left[\left|\frac{\partial}{\partial\theta} \log p_\theta(X)\right|\right] = \int_T |f'_\theta| dx < \infty, \quad \forall \theta \in \Theta. \quad (159)$$

A expressão da função score dada em (158) é base para definirmos uma outra expressão, em termos de notação, da matriz de *Informação de Fisher* $I(\theta)$ que é dada por

$$I(\theta) = \mathbb{E}_\theta\left[\left|\frac{\partial^2}{\partial\theta^2} \log p_\theta(X)\right|\right] < \infty, \quad \theta \in \Theta, \quad (160)$$

tal que

$$\mathbb{E}_\theta[U_n^2(\theta)] = nI(\theta).$$

Observe que em (5), a expressão do elemento da matriz g_{ij} coincide com o da expressão (160) o que mostra uma estreita relação entre os estudos de estimação e de variedades estatísticas. Além do mais, a matriz $I(\theta)$ é uma matriz de momentos de ordem 2. Também podemos interpretar $I(\theta)$ como uma medida de informação de um parâmetro θ extraído de uma amostra.

A equação (157) nos descreve a função *score* como uma função do vetor aleatório $\mathbf{X} = (X_1, \dots, X_n)$, em que as variáveis aleatórias X_i são i.i.d. Com esta notação, percebe-se que

$$\begin{aligned} \mathbb{E}\left[\frac{\partial \log p_\theta(\mathbf{X})}{\partial\theta}\right] &= \int \frac{1}{f_\theta(\mathbf{X})} \frac{\partial p_\theta(\mathbf{X})}{\partial\theta} f_\theta(\mathbf{X}) d\mathbf{X} \\ &= \int \frac{\partial p_\theta(\mathbf{X})}{\partial\theta} d\mathbf{X} \\ &= \frac{\partial}{\partial\theta} \int p_\theta(\mathbf{X}) d\mathbf{X} \\ &= \prod_{i=1}^n \int p_\theta(X_i) dX_i \\ &= 0, \end{aligned}$$

assumindo aqui mais uma vez a permutação entre integral e derivada. Em outras palavras, a matriz de informação de Fisher $I(\theta)$ é igual a matriz de covariância da função *score*.

Intuitivamente, quanto maior o número de elementos de uma amostra, mais verossímil é o estimador. Podemos formalizar melhor essa ideia de *convergência assintótica* através das próximas duas definições que se seguem.

Definição 4.2 (LEON-GARCIA (2008)) *Uma sequência de estimadores $E_n = E_n(X_1, \dots, X_n)$ é fracamente consistente para o parâmetro θ se, para cada $\varepsilon > 0$ e para todo $\theta \in \Theta$,*

$$\lim_{n \rightarrow \infty} P(|E_n - \theta| < \varepsilon) = 1.$$

Em outras palavras, para uma amostra de tamanho extremamente grande, o estimador converge para o parâmetro θ , em termos de probabilidade. Vejamos o exemplo que se segue para melhor ilustrar essa definição.

Exemplo 13 *Seja $\{X_n\}_{\mathbb{N} \ni n \geq 1}$ uma sequência de variáveis aleatórias i.i.d., distribuídas segundo a normal $N(0, 1)$ e consideremos a sequência de estimadores que associa uma amostra de tamanho n com a média amostral \bar{X}_n dada por*

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Então,

$$\begin{aligned} P_\theta(|\bar{X}_n - \theta| < \varepsilon) &= \int_{\theta - \varepsilon}^{\theta + \varepsilon} \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} e^{-(n/2)(\bar{x}_n - \theta)^2} d\bar{x}_n \\ &= \int_{-\varepsilon}^{\varepsilon} \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} e^{-(n/2)y^2} dy \\ &= \int_{-\varepsilon\sqrt{n}}^{\varepsilon\sqrt{n}} \left(\frac{1}{2\pi}\right)^{\frac{1}{2}} e^{-(1/2)t^2} dt \\ &= P(-\varepsilon\sqrt{n} < Z < \varepsilon\sqrt{n}) \\ &\rightarrow 1 \quad \text{com} \quad n \rightarrow \infty, \end{aligned}$$

em que $Z \sim N(0, 1)$. Portanto, \bar{X}_n é um estimador consistente para o parâmetro θ .

e foma equivalente, podemos expressar a definição acima com outra conotação.

Definição 4.3 (LEON-GARCIA (2008)) *Uma sequência de estimadores $E_n(X)$ de um parâmetro θ converge fracamente para θ se $\forall \varepsilon > 0$,*

$$P(|E_n(X) - \theta| > \varepsilon) \rightarrow 0, \quad n \rightarrow \infty.$$

Em outras palavras, esse estimador é consistente.

Uma outra forma de se classificar um estimador como eficiente está relacionada

com a sua função de densidade conjunta dado um parâmetro θ . Segue-se no teorema abaixo.

Teorema 4.1 (Teorema da Fatoração, ZACKS (1971)) *Uma estatística E_n é suficiente para uma parâmetro θ com respeito a uma família paramétrica se, e somente se, a função de densidade conjunta $L_\theta(X_1, \dots, X_n; \theta)$ pode ser fatorada como*

$$L_\theta(X_1, \dots, X_n; \theta) = g_1(T_n; \theta)g_2(X_1, \dots, X_n), \quad \theta \in \Theta.$$

Para que possamos focar na comparação de nosso método de estimação proposto com o MLE, faz-se necessário uma breve apresentação do MLE. Existem inúmeros outros métodos de estimação como o Método dos Momentos e o Método dos Mínimos Quadrados ZACKS (1971). Porém, esses não são foco dessa tese e por isso não comentaremos sobre eles.

Com efeito, consideremos uma amostra $\{X_1, \dots, X_n\}$ de uma variável aleatória iid, distribuída segundo uma função densidade de probabilidade p_θ . Como já citado, a função de verossimilhança de θ segundo a amostra é dada por HAMILTON (1994)

$$l_\theta(x) = \prod_{i=1}^n p_\theta(x_i), \quad (161)$$

na qual $p_\theta(x_i)$ é a função densidade de probabilidade da variável aleatória X_i .

Definição 4.4 *O estimador de máxima verossimilhança de θ é denominado por $\hat{\theta}_{MLE}$ tal que*

$$\hat{\theta}_{MLE} = \arg \max_{\theta \in \Theta} L_\theta(x).$$

Do fato da função logaritmo ser estritamente crescente, o $\hat{\theta}_{MLE}$ é de igual modo calculado quando o definimos por

$$\hat{\theta}_{MLE} = \arg \max_{\theta \in \Theta} l_\theta(x). \quad (162)$$

Na equação (162), o estimador $\hat{\theta}_{MLE}$ pode ser calculado pelas condições de primeira ordem para derivadas, caso o parâmetro θ seja unidimensional. Isto é, dentre os pontos críticos de l_θ , estabeleceremos o $\hat{\theta}_{MLE}$. Mais precisamente, os estimadores serão dados pelas raízes da equação

$$\frac{\partial l_\theta}{\partial \theta} = 0. \quad (163)$$

Por questão da convexidade da função de verossimilhança, temos a unicidade desse estimador. A equação (163) é chamada de *equação de verossimilhança* WALD (1949). Paralelamente, para garantirmos que $\hat{\theta}_{MLE}$ seja de máximo, devemos ter a condição de

segunda ordem abaixo satisfeita

$$\left(\frac{\partial^2 l_\theta}{\partial \theta^2}\right)_{\theta=\hat{\theta}} < 0.$$

Destaca-se que quando usamos a equação (163) equipada com a função logaritmo, a condição de primeira ordem se reescreve utilizando a função *score*

$$U_n(\theta) = \frac{\partial \log L_\theta}{\partial \theta} = 0 \quad (164)$$

baseado na expressão da função *score* expressa em (157). Contextualizemos esses conceitos com o exemplo abaixo.

Exemplo 14 *Seja $\{X_1, \dots, X_n\}$ uma amostra de uma variável aleatória distribuída segundo à distribuição gaussiana padrão $N(\theta_1, \theta_2)$. O logaritmo da função de verossimilhança é dada por*

$$\log L_\theta(x) = -\frac{n}{2} \log(2\pi\theta_2^2) - \sum_{i=1}^n (x_i - \theta_1)^2/2.$$

Algumas manipulações resultam em $\hat{\theta}_1 = \sum_{i=1}^n x_i/n$, a qual por definição é a média amostral \bar{X} e $\hat{\theta}_2 = \sum_{i=1}^n (x_i - \bar{x})^2/n$. Isto é, a estimativa de θ_1 que maximiza a equação de verossimilhança é obtida com a média amostral. Além disso, $\frac{\partial^2 \log l_\theta}{\partial \theta_1^2} = -n < 0$ e $\frac{\partial^2 \log l_\theta}{\partial \theta_2^2} = -\frac{n}{\theta_2^2} < 0$. Por fim, $\mathbb{E}_\theta[\hat{\theta}_1] = \theta_1$, o que nos garante que o estimador é não enviesado. Em contra partida, a média do estimador $\hat{\theta}_2$ é $\mathbb{E}_\theta[\hat{\theta}_2] = \frac{n-1}{n}\theta_2$, o que implica que se tem viés. Porém,

$$\lim_{n \rightarrow \infty} \frac{n-1}{n} \theta_2 = \theta_2, \quad (165)$$

portanto a expressão (165) nos dá a consistência da estimativa de θ_2 .

Um fato a se destacar é que podemos estabelecer critérios para que uma função de um estimador consistente também seja consistente. Mais formalmente, vejamos o teorema abaixo.

Teorema 4.2 *Seja $\{X_n\}_{\mathbb{N} \ni n \geq 1}$ uma sequência iid distribuída segundo uma função densidade de probabilidade conjunta p_θ é tal que $l_\theta(x) = \prod_{i=1}^n p_\theta(x_i)$ a respectiva função de verossimilhança, em que $p_\theta(x_i)$ é a função densidade de probabilidade da variável aleatória X_i . Suponhamos que uma função $g(\theta)$ seja contínua em Θ e que $\hat{\theta}_{MLE}$ seja o estimador de máxima verossimilhança de θ . Então, para todo $\varepsilon > 0$ e para todo $\theta \in \Theta$,*

$$\lim_{n \rightarrow \infty} P(|g(\hat{\theta}_{MLE}) - g(\theta)| \geq \varepsilon) = 0.$$

Em outras palavras, $g(\hat{\theta}_{MLE})$ é um estimador consistente de $g(\theta)$.

Podemos ver com detalhes a demonstração em KENDALL, STUART, and ORD (1968), em que basicamente o fato da continuidade da função g permite que esta seja aplicada dentro da função de probabilidade P , ao permutarmos com o limite. Outros

pontos a serem abordados são aqueles que se referem à convergência de sequência de variáveis aleatórias. Nesse sentido, citamos os principais resultados e definições que abordam a convergência estocástica, omitindo alguns preliminares como sequência de números reais e álgebra de eventos.

Existem alguns tipos de convergência envolvendo variável aleatória como convergência quase certa, convergência na r -ésima média dentre outras SCHMETTERER (2012). Porém, vamos citar apenas aquelas que serão utilizadas para análise do método de convergência proposto no final dessa seção. Com efeito, vejamos a definição abaixo.

Definição 4.5 *Uma sequência de variáveis aleatórias $(X_n)_{n \geq 1}$ converge em probabilidade para uma variável aleatória X se para todo $\varepsilon > 0$, temos CASELLA and BERGER (2002)*

$$\lim_{n \rightarrow \infty} P(|X_n - X| > \varepsilon) = 0. \quad (166)$$

Com efeito de simplificação, quando a sequência de variáveis aleatórias $(X_n)_{n \geq 1}$ satisfizer a condição (166), denotaremos por $X_n \xrightarrow{P} X$. Para melhor entendimento da definição, consideremos o exemplo abaixo.

Exemplo 15 *Seja a sequência $(X_n)_{n \geq 3}$ de variáveis aleatórias independentes distribuídas segundo uma a função densidade de probabilidade dada por*

$$P(X_n = 0) = 1 - \frac{1}{\log n} \quad e \quad P(X_n = n) = \frac{1}{\log n}, \quad n \geq 3.$$

Temos que para $\varepsilon \in (0, 1)$, $P(|X_n| > \varepsilon) = P(X_n = n)$ e para $\varepsilon \geq 1$, $P(|X_n| > \varepsilon) \leq P(X_n = n)$. Como $\lim_{n \rightarrow \infty} \frac{1}{\log n} = 0$, então para todo $\varepsilon > 0$, $\lim_{n \rightarrow \infty} P(|X_n| > \varepsilon) = 0$.

Um outro tipo de convergência é a ser comentada é *Convergência em Distribuição*. Especificamente ao longo dessa tese, a convergência em distribuição será de suma importância por causa de vários resultados já conhecidos que serão aplicados para garantir a convergência da versão mais geral do MLE.

Definição 4.6 *Uma sequência de variáveis aleatórias $\{X_n\}_{\mathbb{N} \ni n \geq 1}$ converge em distribuição para uma variável aleatória X se, para todo ponto de continuidade x da função de distribuição de probabilidade F_X da variável aleatória X , temos CASELLA and BERGER (2002)*

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x). \quad (167)$$

Com efeito de simplificação, se a sequência de variáveis aleatórias $(X_n)_{n \geq 1}$ satisfizer a condição (167), denotaremos por $X_n \xrightarrow{D} X$. Um importante resultado nos garante que temos convergência em distribuição mediante a convergência em probabilidade. Mais formalmente, vejamos o teorema abaixo que pode ser encontrado em BILLINGSLEY (2013).

Teorema 4.3 *Se a sequência de variáveis aleatórias $\{X_n\}_{\mathbb{N} \ni n \geq 1}$ converge em probabilidade para X , então também converge em distribuição.*

Prova 16 Suponha que $X_n \xrightarrow{P} X$ e que X tenha função de distribuição F_X . Seja x um ponto de continuidade de F e $\varepsilon > 0$ tão pequeno quanto se queira. Então,

$$\begin{aligned} F(x - \varepsilon) &= P(X \leq x - \varepsilon) \\ &= P(X \leq x - \varepsilon, |X_n - X| \leq \varepsilon) + P(X \leq x - \varepsilon, |X_n - X| > \varepsilon) \\ &= P(X_n \leq x) + P(|X_n - X| > \varepsilon) \end{aligned} \tag{168}$$

e

$$F(x + \varepsilon) \geq P(X_n \leq x) - P(|X_n - X| > \varepsilon).$$

As definições e resultados citados nessa subseção até aqui são pré-requisitos para a análise de convergência assintótica propriamente dita. Nesse sentido, esses preliminares serão citados nas definições e resultados de convergência que se seguem. Abordaremos um conceito essencial para darmos prosseguimento: *funções de estimação* CHANDRASEKAR and KALE (1984); SEN, SINGER, and DE LIMA (2010). Veremos uma condição necessária para se garantir convergência envolvendo as funções de estimação, usando as propriedades de *linearidade assintótica uniforme* de estimadores gerados pelas *equações de estimação*. Esses resultados serão importantes, pois nem sempre conseguimos exibir a função de estimação de forma explícita como função das observações (ou amostra) ou podemos ter problemas relacionados à unicidade de parâmetro que maximiza uma função de verossimilhança de um estimador. Tal observação tem destaque, pois nos deparamos com esse problema de uma condição implícita para a maximização da função de verossimilhança em nosso problema aplicado. Discutiremos isso no final dessa subseção. Com efeito, consideremos inicialmente a definição que se segue.

Definição 4.7 Uma função de estimação é uma função $\phi: T \times \Theta \rightarrow \mathbb{R}^p$ tal que para cada $\theta \in \Theta$, $\phi(\cdot, \theta) \rightarrow \mathbb{R}^p$ é uma variável aleatória.

Seja (T, Σ, μ) um espaço de medida, em que parametrizamos a variedade estatística \mathcal{P} . Suponhamos a existência de uma amostra de n vetores aleatórios independentes $\mathbf{x}_i = (x_{i1}, \dots, x_{in})^T$, com $i = 1, \dots, n$, em que cada índice i associa-se com uma função de estimação ϕ_i . Estendemos, através de ϕ_i , a função de estimação para as amostras por

$$\Phi_n(\mathbf{x}, \theta) = \sum_{i=1}^n \phi_i(\mathbf{x}_i, \theta),$$

em que $\mathbf{x} = (x_1, \dots, x_n)$.

Motivado pela estimação de parâmetros, focaremos nas funções de estimação cujas raízes coincidem com os estimadores dos parâmetros. Isto é,

$$\Phi_n(\mathbf{x}, \hat{\theta}_n) = 0. \tag{169}$$

A expressão (169) é chamada de *equação de estimação*. Além disso, a função *score* da família exponencial (157) é um exemplo de função de estimação.

Podemos, através de uma função de estimação $\Phi(\theta)$, definir várias outras funções de estimação através da equivalência definida abaixo:

Definição 4.8 *As funções de estimação $\Phi(\theta)$ e $\Psi(\theta)$ são equivalentes se, e somente se,*

$$\Psi(\theta) = \mathbf{C}(\theta)\Phi(\theta), \quad (170)$$

sendo $\mathbf{C}(\theta)$ uma matriz de ordem p de posto completo BASAWA, GODAMBE, and TAYLOR (1997).

A equivalência (170) indica que existem infinitas funções de equivalência. Ademais, por conta de $\mathbf{C}(\theta)$ ser de posto completo, uma raiz de uma função de estimação qualquer será também raiz de todas as funções equivalente.

Um outro conceito extremamente importante é o de função de estimação não viesada que será citado abaixo e pode ser encontrada em BASAWA, GODAMBE, and TAYLOR (1997).

Definição 4.9 *Uma função de estimação Φ será denominada não viesada se a equação de estimação abaixo*

$$E_{\theta}[\Phi(\theta)] = 0 \quad (171)$$

é satisfeita para todo $\theta \in \Theta$.

Para melhor entendimento da definição acima, considere o exemplo de aplicação abaixo.

Exemplo 16 *Seja $p(\mathbf{x}_i, \theta) = \frac{1}{\theta} \exp(-\frac{x_i}{\theta})$ a função de distribuição exponencial. A função de estimação associada a essa distribuição é dada por*

$$\Phi(\theta) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \log\left(\frac{1}{\theta} \exp\left(-\frac{x_i}{\theta}\right)\right) = 0 \iff n - \sum_{i=1}^n \frac{x_i}{\theta} = 0.$$

Além disso,

$$\begin{aligned} E_{\theta}[\Phi(\theta)] &= \int_T \left(n - \sum_{i=1}^n \frac{x_i}{\theta} \right) p(x_i, \theta) dx_i \\ &= n - \frac{n\theta}{\theta} \\ &= 0. \end{aligned}$$

As definições dadas acima estabelecem restrições às funções de estimação para que se tenha propriedades assintóticas satisfatórias. Vale destacar que no Capítulo 2, vimos que $\mathbb{E}_{\theta}[\frac{\partial l_{\theta}}{\partial \theta^i}] = 0$, em que $\frac{\partial l_{\theta}}{\partial \theta^i}$ é a função score. Portanto, dados os conceitos de função de estimação e equação de estimação, percebemos que implicitamente estava se garantindo

as devidas condições para convergência assintótica do Estimador de Máxima Verossimilhança. A ideia será explorar as propriedades assintóticas das funções de estimação e, em seguida, estendê-las aos estimadores. A propriedade de maior relevância a ser trabalhada é a de regularidade.

Definição 4.10 *Uma função de estimação $\Phi(\mathbf{x}, \theta)$ é dita regular se para qualquer parâmetro $\theta \in \Theta$, as seguintes condições são satisfeitas GODAMBE (1960):*

- 1 $\Phi(\mathbf{x}, \theta)$ é não viciada.
- 2 As derivadas parciais de $\Phi(\mathbf{x}, \theta)$ com respeito ao parâmetro θ^i existem, para qualquer $x \in T$.
- 3 $\frac{\partial}{\partial \theta^i} E_\theta[\Phi(\mathbf{x}, \theta)] = E_\theta[\frac{\partial \Phi}{\partial \theta^i}]$, isto é, os operadores derivada e integral permutam.
- 4 A matriz

$$\mathbf{V}_\Phi(\theta) = E_\theta[\Phi(\mathbf{x}, \theta)\Phi^T(\mathbf{x}, \theta)]$$

é positiva definida, em que as entradas da matriz são dadas por

$$v_{ij} = E_\theta[\Phi_i(\mathbf{x}, \theta)\Phi_j(\mathbf{x}, \theta)] \in \mathbb{R}.$$

- 5 $E_\theta[\frac{\partial \Phi_i}{\partial \theta^r}(\mathbf{x}, \theta)\frac{\partial \Phi_j}{\partial \theta^s}(\mathbf{x}, \theta)] \in \mathbb{R}$ e $\mathbf{S}_\Phi(\theta) = E_\theta[\nabla_\theta \Phi(\mathbf{x}, \theta)]$ é não identicamente nula, em que ∇ é o operador gradiente em relação ao parâmetro θ .

Observemos que para o caso uniparamétrico, as condições 4 e 5 da função de estimação convertem-se, respectivamente em

$$0 < E_\theta[\Phi^2(\theta)] < \infty \quad \text{e} \quad -\infty < E_\theta\left[\frac{\partial \Phi}{\partial \theta}\right] < \infty.$$

Outro conceito de extrema relevância é o de *Matriz de Informação de Godambe*, cuja equivalente nas famílias exponenciais é a matriz de informação de Fisher e já destacamos que estabelece uma referência da quantidade de informação que se pode obter de um parâmetro dada uma amostra de uma variável aleatória.

Definição 4.11 *Seja $\Phi(\theta)$ uma função de estimação regular. A matriz*

$$\mathbf{J}_\Phi(\theta) = \mathbf{S}_\Phi^T(\theta)\mathbf{V}_\Phi^{-1}(\theta)\mathbf{S}_\Phi(\theta) \tag{172}$$

é denominada matriz de informação de Godambe de θ associada a Φ .

Um fato importante que decorre de (172) e da função *score* que provém de uma distribuição de θ é uma função regular e, além disso, $\mathbf{S}_\Phi(\theta) = -\mathbf{V}_\Phi(\theta)$, o que implica que a matriz de Godambe realmente coincide com a matriz de informação de Fisher, como afirmado antes, isto é,

$$\mathbf{J}_\Phi(\theta) = -\mathbf{S}_\Phi^T(\theta)\mathbf{V}_\Phi^{-1}(\theta)\mathbf{V}_\Phi(\theta) = -\mathbf{S}_\Phi^T(\theta). \tag{173}$$

Completa-se a esses fatos de regularidade explorar as propriedades de linearidade das

equações de estimação, para então obter resultados de convergência assintótica dos estimadores de fato. Consideremos a seguinte equação de estimação abaixo, para o caso uniparamétrico e com variáveis aleatórias também unidimensionais,

$$M_n(\hat{\theta}_n) = n^{-\frac{1}{2}} \sum_{i=1}^n \phi(X_i, \hat{\theta}_n) = 0, \quad (174)$$

em que função ϕ será escolhida de tal forma que

$$\int_{\mathbb{R}} \phi(x, \theta) dF_X(x, \theta) = 0, \quad (175)$$

e

$$\int_{\mathbb{R}} \phi^2(x, \theta) dF_X(x, \theta) = \sigma^2(\psi, \theta) < \infty, \quad (176)$$

com F_X denotando a função de distribuição da variável aleatória X . Observe que as expressões (171) e (175) são equivalentes. Portanto, vamos avaliar o parâmetro $\hat{\theta}_n$ que é descrito na expressão de (174). A seguir, apresentaremos algumas notações que serão utilizadas no principal teorema desta seção, em que estabeleceremos sob algumas hipóteses, a convergência em distribuição de estimadores provenientes de equação de estimação. Defina

$$\lambda(\phi, \theta, \delta) = \mathbb{E}_{\theta}[\phi(X, \theta + \delta)] \quad (177)$$

para todo δ em uma vizinhança de 0 conveniente. É sensato definirmos (177), considerando a continuidade da função ϕ . Observe que, pela equação (175), $\lambda(\phi, \theta, 0) = \mathbb{E}_{\theta}[\phi(X, \theta)] = 0$. Portanto, temos

$$\begin{aligned} \lambda(\phi, \theta, \delta) &= \int_{\mathbb{R}} [\phi(x, \theta + \delta) - \phi(x, \theta)] dF_X(x; \theta) \\ &= \delta \int_{\mathbb{R}} \frac{1}{\delta} [\phi(x, \theta + \delta) - \phi(x, \theta)] dF_X(x; \theta). \end{aligned}$$

Considerando que ϕ é absolutamente contínua, ao tomarmos o limite de $\lambda(\phi, \theta, \delta)$ quando $\delta \rightarrow 0$, temos

$$\nu(\phi, \theta) = \lim_{\delta \rightarrow 0} \frac{1}{\delta} \lambda(\phi, \theta, \delta) = \int_{\mathbb{R}} \phi'(x, \theta) dF_X(x; \theta), \quad (178)$$

em que

$$\phi'(x, \theta) = \lim_{\delta \rightarrow 0} [\phi(x, \theta + \delta) - \phi(x, \theta)] / \delta.$$

Denote por f_X a função densidade de probabilidade associada à variável aleatória X . Com algumas manipulações, a expressão de (178) pode ser reescrita como

$$\nu(\phi, \theta) = \int_{\mathbb{R}} \phi(x, \theta) f'_{\theta}(x; \theta) dx,$$

em que $f'_\theta(x; \theta) = \lim_{\delta \rightarrow 0} \frac{1}{\delta} [f_X(x; \theta + \delta) - f_X(x; \theta)]$. A função *score* associada à distribuição de probabilidade f_X será dada por

$$\phi_f(x, \theta) = f'_X(x; \theta) / f_X(x; \theta),$$

enquanto a informação de Fisher em θ é

$$I_f(\theta) = \int_{\mathbb{R}} [f'_X(x; \theta) / f_X(x; \theta)]^2 f(x; \theta) dx.$$

Dessa forma, podemos reescrever (178) por

$$\nu(\phi, \theta) = \sigma(\phi, \theta) [I_f(\theta)]^{\frac{1}{2}} \rho[\phi(\cdot, \theta), \phi_f(\cdot, \theta)] \quad (179)$$

em que ρ é o coeficiente de correlação entre as funções *score* ϕ e ϕ_f e σ é a função de variância. Com essas notações dadas, citemos o teorema que se segue:

Teorema 4.4 *Se o estimador $\hat{\theta}_n$ satisfaz a equação (174), respeitando as condições de regularidade descritas na Definição 4.10, então*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, [I_f(\theta)\rho^2(\phi, \phi_f)]^{-1}).$$

Prova 17 *Mais uma vez, embora a prova esteja em SEN, SINGER, and DE LIMA (2010), apresentaremos um resumo da mesma. Considerando*

$$\hat{u} = \sqrt{n}(\hat{\theta}_n - \theta), \quad (180)$$

podemos estabelecer que $\hat{\theta}_n = \theta + n^{-\frac{1}{2}}\hat{u}$. Portanto, $M_n(\hat{\theta}_n) = M_n(\theta + n^{-\frac{1}{2}}\hat{u}) = 0$, em que M_n satisfaz (174) e, considerando (175) e (175), o Teorema Centraldo Limite CASELLA and BERGER (2002) garante que

$$M_n(\theta) \xrightarrow{D} N(0, \sigma^2(\phi, \theta)).$$

Caminhando para os resultados finais, falaremos sobre o Método Delta, que tem a sua versão linear e multivariada. Este método é um resultado que aumenta significativamente o Teorema Central do Limite CASELLA and BERGER (2002).

Teorema 4.5 (Método Delta) *Suponha que $\sqrt{n}(T_n - \theta)/\sigma \xrightarrow{D} N(0, 1)$ e seja g uma função contínua tal que $g'(\theta) \neq 0$. Então, segue que*

$$\sqrt{n}[g(T_n) - g(\theta)]/\sigma g'(\theta) \xrightarrow{D} N(0, 1). \quad (181)$$

O Teorema 4.5 pode ser generalizado para o caso multivariado. Seja $\{\mathbf{T}_n\}$ uma sequência de variáveis vetoriais aleatórias e suponha que $\sqrt{n}(\mathbf{T}_n - \theta) \xrightarrow{D} N(\mathbf{0}, \mathbf{\Sigma})$. Essencialmente,

estamos interessados na distribuição assintótica de $\sqrt{n}[g(\mathbf{T}_n) - g(\theta)]$, em que $g(\cdot)$ é uma função real de \mathbf{T}_n . Isto é o que se mostra no seguinte teorema.

Teorema 4.6 *Seja $\{\mathbf{T}_n\}$ uma sequência de variáveis p -vetoriais de tal modo que $\sqrt{n}(\mathbf{T}_n - \theta) \xrightarrow{D} N(\mathbf{0}, \Sigma)$ e considere uma função real $g(\mathbf{T}_n)$ tal que $\dot{\mathbf{g}}(\theta) = \partial g / \partial \mathbf{x}|_{\theta}$ é não nulo e contínua em uma vizinhança de θ . Então,*

$$\sqrt{n}[g(\mathbf{T}_n) - g(\theta)] \xrightarrow{D} N(0, \gamma^2) \quad \text{com} \quad \gamma^2 = [\dot{\mathbf{g}}(\theta)]' \Sigma [\dot{\mathbf{g}}(\theta)].$$

Na literatura, vários métodos possibilitam estimar parâmetros, como o MLE já citado, o método dos momentos e outros LEHMANN and CASELLA (2006). É de natural interesse que, dado estimadores, precisamos compará-los entre si em termos de eficiência. Uma das maneiras de se analisar a eficiência é pela desigualdade de Cramér–Rao, em que obtemos um limite inferior da matriz de variância de um estimador pela inversa da matriz de informação RAO (1992). Especificamente ao longo dessa seção, mostraremos uma versão dessa desigualdade aplicada no contexto da φ -família.

Para efeitos de simulação, iremos considerar o caso particular das φ -famílias das distribuições para as quais chamaremos de φ -gaussiana, que identifica-se com a distribuição gaussiana padrão. Com efeito, seja X_1, \dots, X_n uma sequência de n variáveis aleatórias identicamente distribuídas, cuja densidade de probabilidade dependa de um parâmetro $\theta \in \Theta$, em que Θ aqui é um conjunto compacto e convexo, o que nos garante encontrar um ponto de máximo. Com essas considerações, reescrevemos a expressão (162) como

$$\hat{\theta}_{MLE} = \arg \max_{\theta \in \Theta} \sum_{i=1}^n \log p_{\theta}(x_i), \quad (182)$$

em que $\{x_1, \dots, x_n\}$ é uma amostra das variáveis aleatórias $\{X_1, \dots, X_n\}$, respectivamente. Uma hipótese necessária a se impor é garantir que a derivada de segunda ordem seja negativa, dessa forma, temos localmente a unicidade do $\hat{\theta}_{MLE}$ LEON-GARCIA (2008).

De maneira análoga à equação (182), propomos um método de estimação adaptado a \mathcal{P} . Alguns trabalhos MARTINS and STEDINGER (2000); ZHANG (2009) já fizeram adaptações específicas ao MLE. Em nosso caso, porém, perceberemos que o MLE aplicado à famílias exponenciais se comporta como um caso particular do que iremos definir abaixo. De fato, seja φ uma φ -função como na definição 3.1. Vamos definir o estimador de máxima verossimilhança modificado como

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \sum_{i=1}^n \varphi^{-1}(p_{\theta}(x_i)). \quad (183)$$

Um relevante resultado dessa tese é afirmar que existe uma relação direta entre a maximização do método de estimação (183) proposto e a minimização da função de divergência $\mathcal{D}_{\varphi}(\cdot \parallel \cdot)$. Para tal, aproximemos a função de distribuição empírica $p_n(x) =$

$\frac{1}{n} \sum_{i=1}^n \delta(x - x_i)$ EVANS, HASTINGS, and PEACOCK (2000) por

$$p_{n,\varepsilon}(x) = \frac{1}{n} \sum_{i=1}^n u_\varepsilon(x - x_i), \quad (184)$$

na qual

$$u_\varepsilon(t) = \begin{cases} 1/\varepsilon, & \text{para } -\varepsilon/2 \leq t \leq \varepsilon/2, \\ 0, & \text{caso contrário.} \end{cases} \quad (185)$$

Proposição 4.1 *Suponha que as funções $\varphi(\cdot)$, u_0 e p_θ sejam contínuas. Então,*

$$\mathcal{D}_\varphi(p_{n,\varepsilon} \parallel p_\theta) - \frac{n\varphi^{-1}(\frac{1}{n\varepsilon})}{\sum_{i=1}^n \frac{1}{\varepsilon} \int_{x_i-\varepsilon/2}^{x_i+\varepsilon/2} u_0 dx} \xrightarrow{\varepsilon \rightarrow 0} - \frac{\sum_{i=1}^n \varphi^{-1}(p_\theta(x_i))}{\sum_{i=1}^n u_0(x_i)}.$$

Prova 18 *Através das equações (184) e (185), podemos escrever*

$$\begin{aligned} \int_{\mathbb{R}} \frac{\varphi^{-1}(p_{n,\varepsilon})}{(\varphi^{-1})'(p_{n,\varepsilon})} dx &= \sum_{i=1}^n \int_{x_i-\varepsilon/2}^{x_i+\varepsilon/2} \frac{\varphi^{-1}(\frac{1}{n\varepsilon})}{(\varphi^{-1})'(\frac{1}{n\varepsilon})} dx \\ &= \frac{\varepsilon}{(\varphi^{-1})'(\frac{1}{n\varepsilon})} n\varphi^{-1}\left(\frac{1}{n\varepsilon}\right) \end{aligned} \quad (186)$$

e, analogamente,

$$\int_{\mathbb{R}} \frac{\varphi^{-1}(p_\theta)}{(\varphi^{-1})'(p_{n,\varepsilon})} dx = \sum_{i=1}^n \int_{x_i-\varepsilon/2}^{x_i+\varepsilon/2} \frac{\varphi^{-1}(p_\theta)}{(\varphi^{-1})'(\frac{1}{n\varepsilon})} dx = \frac{\varepsilon}{(\varphi^{-1})'(\frac{1}{n\varepsilon})} \sum_{i=1}^n \frac{1}{\varepsilon} \int_{x_i-\varepsilon/2}^{x_i+\varepsilon/2} \varphi^{-1}(p_\theta) dx. \quad (187)$$

Além disso, temos

$$\begin{aligned} \int_{\mathbb{R}} \frac{u_0}{(\varphi^{-1})'(p_{n,\varepsilon})} dx &= \sum_{i=1}^n \int_{x_i-\varepsilon/2}^{x_i+\varepsilon/2} \frac{u_0}{(\varphi^{-1})'(\frac{1}{n\varepsilon})} dx \\ &= \frac{\varepsilon}{(\varphi^{-1})'(\frac{1}{n\varepsilon})} \sum_{i=1}^n \frac{1}{\varepsilon} \int_{x_i-\varepsilon/2}^{x_i+\varepsilon/2} u_0 dx. \end{aligned} \quad (188)$$

Usando as expressões (186), (187) e (188), concluímos que

$$\begin{aligned} \mathcal{D}_\varphi(p_{n,\varepsilon} \parallel p_\theta) &- \frac{n\varphi^{-1}(\frac{1}{n\varepsilon})}{\sum_{i=1}^n \frac{1}{\varepsilon} \int_{x_i-\varepsilon/2}^{x_i+\varepsilon/2} u_0 dx} \\ &= - \frac{\sum_{i=1}^n \frac{1}{\varepsilon} \int_{x_i-\varepsilon/2}^{x_i+\varepsilon/2} \varphi^{-1}(p_\theta) dx}{\sum_{i=1}^n \frac{1}{\varepsilon} \int_{x_i-\varepsilon/2}^{x_i+\varepsilon/2} u_0 dx}. \end{aligned} \quad (189)$$

Pela hipótese da continuidade das funções, a expressão (189) tende para

$$-\frac{\sum_{i=1}^n \varphi^{-1}(p_{\theta}(x_i))}{\sum_{i=1}^n u_0(x_i)}, \quad (190)$$

quando $\varepsilon \rightarrow 0$.

A expressão obtida em (190) é de extrema importância. Perceba que o numerador dela é o mesmo da expressão (183). Em outras palavras, a função φ -divergência tem uma outra significativa contribuição, pois ela nos possibilita relacionar a maximização do método de estimação de verossimilhança generalizado (183) pela minimização da função de divergência dada em (190). Esse tipo de desenvolvimento não é incomum, pois como sabemos COVER and THOMAS (2006), podemos maximizar a entropia de Shannon, através na minimização da divergência de Kullback–Leibler. Então, o método proposto em (183) é análogo quando minimizarmos a divergência. Usando a Proposição 4.1 em (65), definimos a *função de verossimilhança generalizada* aplicada a φ -família por

$$\mathcal{L}_{\varphi^{-1}}(\theta) = \sum_{i=1}^n \{c(x_i) + \theta^k u_k(x_i) - \psi(\theta) u_0(x_i)\}. \quad (191)$$

Derivando (191) em relação a θ^k , $k = 1, \dots, n$, temos

$$\frac{\partial}{\partial \theta^k} \mathcal{L}_{\varphi^{-1}}(\theta) = \sum_{i=1}^n \left\{ u_k(x_i) - \frac{\partial \psi}{\partial \theta^k} u_0(x_i) \right\} = 0,$$

o que equivalentemente resulta em

$$\frac{\partial \psi}{\partial \theta^k} = \frac{\sum_{i=1}^n u_k(x_i)}{\sum_{i=1}^n u_0(x_i)},$$

na qual se torna uma condição necessária para a maximização da função de verossimilhança dada em (191).

Em trabalho anterior SOUZA, CAVALCANTE, and VIGELIS (2016), apresentamos uma classe de funções de densidade de probabilidade denominada por φ -gaussiana, a qual está contida em uma φ -família. Dada uma φ -função, definimos uma φ -gaussiana como a função definida por

$$p_{\mu, \sigma} = \varphi \left(-\frac{(x - \mu)^2}{2\sigma^2} - \psi(\sigma) \right), \quad (192)$$

em que os parâmetros μ e σ representam, respectivamente, a média amostral e uma medida de dispersão similar ao desvio padrão. Além do mais, a função $\psi(\sigma)$ tem a mesma utilidade como em (65), isto é, é uma função normalizadora. Observe também que esse nome é sugestivo, pois a φ -gaussiana tem como caso particular a distribuição gaussiana

padrão, definida por

$$f_{\mu,\sigma} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty, \mu \in \mathbb{R} \quad \text{e} \quad \sigma > 0. \quad (193)$$

Para tal afirmação se fazer valer, basta definir $\varphi = \exp$ e considerar a função $\psi(\sigma)$ dada em (26).

Como citamos no parágrafo anterior, o parâmetro σ é uma espécie de medida de dispersão. Mostremos o que isso significa. Considerando $T = \mathbb{R}$ e μ uma medida de Lebesgue e do fato de (192) ser uma distribuição de probabilidade, derivando a expressão abaixo

$$\int_{\mathbb{R}} \varphi\left(-\frac{(x-\mu)^2}{2\sigma^2} - \psi(\sigma)\right) dx = 1, \quad (194)$$

em relação a μ , obtemos

$$\int_{\mathbb{R}} \frac{(x-\mu)}{\sigma^2} \varphi'\left(-\frac{(x-\mu)^2}{2\sigma^2} - \psi(\sigma)\right) dx = 0. \quad (195)$$

Permutando integral com derivada, já que temos as condições de regularidade, e derivando-se mais uma vez em relação a μ , obtemos

$$-\int_{\mathbb{R}} \frac{1}{\sigma^2} \varphi'\left(-\frac{(x-\mu)^2}{2\sigma^2} - \psi(\sigma)\right) dx + \int_{\mathbb{R}} \frac{(x-\mu)^2}{\sigma^4} \varphi''\left(-\frac{(x-\mu)^2}{2\sigma^2} - \psi(\sigma)\right) dx = 0. \quad (196)$$

Reorganizando a equação (196), temos uma medida que se assemelha com variância

$$\sigma^2 = \frac{\int_T (x-\mu)^2 \varphi''\left(-\frac{(x-\mu)^2}{2\sigma^2} - \psi(\sigma)\right) dx}{\int_T \varphi'\left(-\frac{(x-\mu)^2}{2\sigma^2} - \psi(\sigma)\right)}. \quad (197)$$

Considerando a notação

$$E''_{\sigma,\mu}[\cdot] = \frac{\int_{\mathbb{R}} (\cdot) \varphi''(f_{\sigma,\mu}) dx}{\int_{\mathbb{R}} \varphi'(f_{\sigma,\mu}) dx},$$

na qual $f_{\sigma,\mu} = \varphi^{-1}(p_{\sigma,\mu})$, dada em (54), obtemos

$$\sigma^2 = E''_{\sigma,\mu}[(X-\mu)^2]. \quad (198)$$

Iremos agora focar na aplicação do método de estimação proposto em (183), aplicado na expressão de (191) para uma família de distribuições como em (192) é

$$\mathcal{L}_{\varphi^{-1}}(\sigma) = \sum_{i=1}^n \left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} - \psi(\sigma) \right\}. \quad (199)$$

Portanto, um possível estimador de σ é obtido quando o mesmo é ponto crítico de (199),

isto é, $\frac{\partial \mathcal{L}_{\varphi^{-1}}(\sigma)}{\partial \sigma} = 0$, no que implica

$$\frac{\partial \mathcal{L}_{\varphi^{-1}}}{\partial \sigma} = \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^3} - n\psi'(\sigma) = 0. \quad (200)$$

De maneira equivalente, a equação (200) pode ser reescrita por

$$\sigma^3 \psi'(\sigma) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2, \quad (201)$$

em que se torna uma condição necessária para maximizar $\mathcal{L}_{\varphi^{-1}}(\sigma)$. Isto é, a função (191) possui um estimador para o parâmetro σ quando este maximiza $\mathcal{L}_{\varphi^{-1}}$. Em outras palavras, σ é raiz da equação (201). Analogamente, derivando a expressão (199) com respeito ao parâmetro μ , obtemos a seguinte equação

$$\frac{\partial \mathcal{L}_{\varphi^{-1}}}{\partial \mu} = \sum_{i=1}^n (x_i - \mu) = 0.$$

Algumas manipulações nos mostram que o estimador para μ é

$$\hat{\mu} = \sum_{i=1}^n \frac{X_i}{n} = \bar{X}_n, \quad (202)$$

isto é, a própria média amostral. Essa medida de estimação é importante, pois não causa problema na expressão (201), já que temos dois parâmetros envolvidos na equação. Isto é, usando a expressão (202) em (201), fica possível sim estimar o parâmetro σ . Caso tenhamos uma variável aleatória contínua, podemos deduzir a equivalente da equação (201), ainda no mesmo intuito de maximizar a função de verossimilhança generalizada. Nesse sentido, derivemos mais uma vez a equação (194), agora em relação ao parâmetro σ . Obtemos então

$$\int_{\mathbb{R}} \left(\frac{(x - \mu)^2}{\sigma^3} - \psi'(\sigma) \right) \varphi' \left(-\frac{(x - \mu)^2}{2\sigma^2} - \psi(\sigma) \right) dx = 0 \iff$$

$$\int_{\mathbb{R}} (x - \mu)^2 \varphi' \left(-\frac{(x - \mu)^2}{2\sigma^2} - \psi(\sigma) \right) dx = \sigma^3 \psi'(\sigma) \int_{\mathbb{R}} \varphi' \left(-\frac{(x - \mu)^2}{2\sigma^2} - \psi(\sigma) \right) dx.$$

Logo,

$$\sigma^3 \psi'(\sigma) = \frac{\int_{\mathbb{R}} (x - \mu)^2 \varphi' \left(-\frac{(x - \mu)^2}{2\sigma^2} - \psi(\sigma) \right) dx}{\int_{\mathbb{R}} \varphi' \left(-\frac{(x - \mu)^2}{2\sigma^2} - \psi(\sigma) \right) dx} dx. \quad (203)$$

Usando a notação $E'_{\sigma, \mu}[\cdot] = \frac{\int_{\mathbb{R}} (\cdot) \varphi'(f_{\sigma, \mu}) dx}{\int_{\mathbb{R}} \varphi'(f_{\sigma, \mu}) dx}$, observamos que podemos reescrever (201) na sua respectiva versão contínua por

$$\sigma^3 \psi'(\sigma) = E'_{\sigma, \mu}[(X - \mu)^2]. \quad (204)$$

A expressão (204) acima, ao mesmo tempo que impõe uma condição necessária e suficiente para encontrar o estimador do parâmetro σ , também evidencia um problema: o parâmetro citado está escrito de forma implícita na equação. Ao mesmo tempo que se torna um problema analítico nos motiva a desenvolver um método numérico para a solução da equação (204). Desta feita, desenvolvemos um algoritmo baseado nas características geométricas da φ -família e que resultará numa solução em σ para a equação (204). Para tal, utilizaremos métodos numéricos para se obter integrais e derivadas. Vejamos alguns pontos que devem ser observados.

- Observação 1) Dada uma φ -função, nem sempre podemos obter a função normalizadora ψ , como visto em (26), embora pelas propriedades da φ -função, seja possível obter uma única ψ como explicado na após a definição de φ -função. Portanto, essa unicidade nos permite obter dado um valor de σ atribuído aleatoriamente, usar a equação (194), para se obter $\psi(\sigma)$, sem ter a sua respectiva função explícita. Mais precisamente, supondo a função

$$I(k) := \int_{\mathbb{R}} \varphi\left(-\frac{(x-\mu)^2}{2\sigma^2} - k\right) dx, \quad (205)$$

o valor de $\psi(\sigma)$ coincide com a raiz da função $I(k)$.

- Observação 2) Não termos a expressão explícita de ψ não será problema para obtermos a expressão de ψ' , fato este devido à expressão (102);
- Observação 3) Obtendo-se o valor de $\psi'(\sigma)$, diretamente obtemos o primeiro membro da equação (204);

Com essas observações, apresentamos o **Algoritmo do MLE Modificado**.

-
1. Supor um valor inicial para σ ;
 2. Obter o valor de $\psi(\sigma)$ através da equação $\int_{\mathbb{R}} \varphi\left(-\frac{(x-\mu)^2}{2\sigma^2} - \psi(\sigma)\right) dx = 1$;
 3. Obter o valor de $\psi'(\sigma)$ através da expressão $\frac{\partial \psi}{\partial \theta^i}(\theta) = \frac{\mathbb{E}[u_i \varphi'(c + \theta^i u_i - \psi(\theta) u_0)]}{\mathbb{E}[u_0 \varphi'(c + \theta^i u_i - \psi(\theta) u_0)]}$;
 4. Obter um novo valor de σ , ao substituir $\psi'(\sigma)$ na equação $\sigma^3 \psi'(\sigma) = E'_{\sigma, \mu}[(x - \mu)^2]$.
-

Para analisarmos a performance do algoritmo, vamos atuar com dois exemplos de φ -função, a saber $\varphi(u) = \log(\exp(u) + 1)$ e função de exponencial de Kaniadakis $\varphi(u) = \exp_{\kappa}(u)$, com $\kappa = 0,2$. Portanto, os gráficos abaixo revelam, para cada uma das funções, como a sequência dos valores de estimação $\hat{\sigma}$ convergem para o parâmetro real σ .

Gráfico 1 – Estimador de máxima verossimilhança generalizado para $\varphi(u) = \log(\exp(u) + 1)$

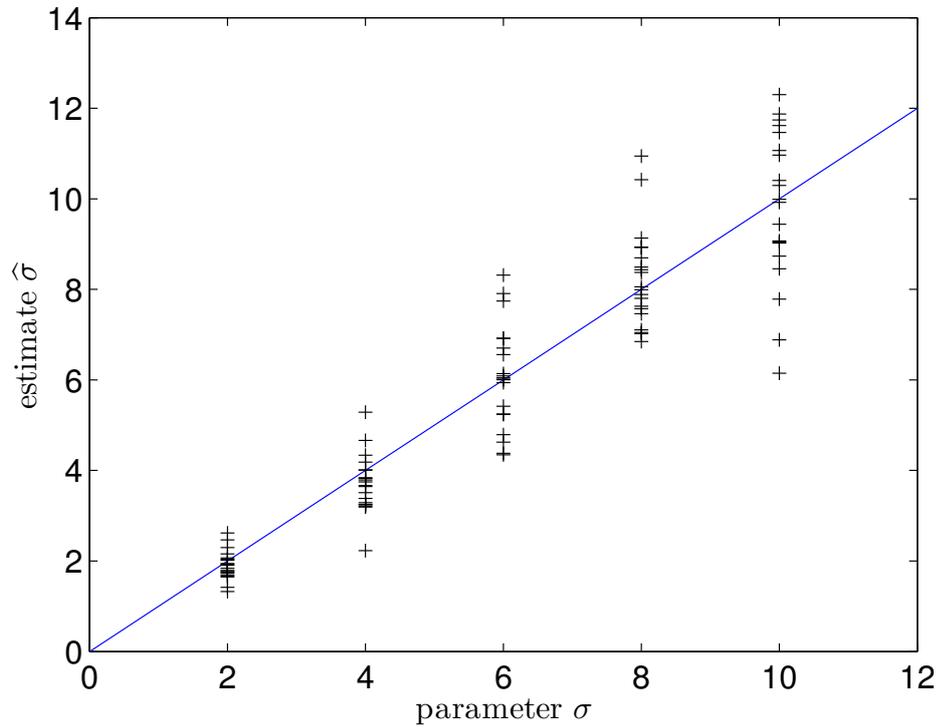
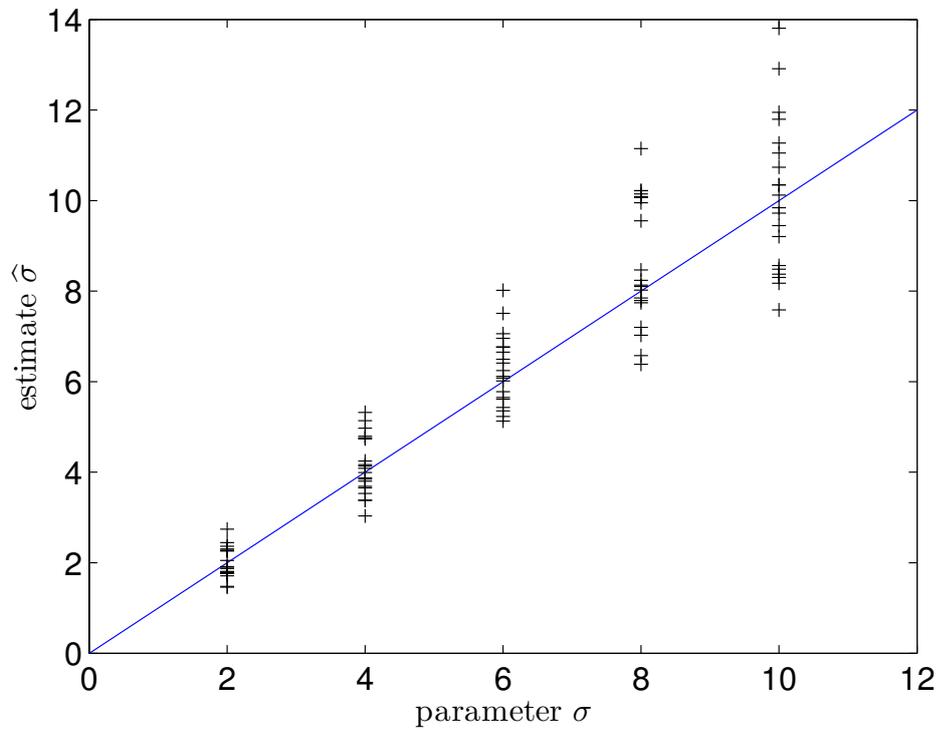


Gráfico 2 – Estimador de máxima verossimilhança generalizado para $\varphi(u) = \exp_{\kappa}(u)$, com $\kappa = 0.2$.



Em ambas os gráficos acima, para cada valor real de σ , faz-se uso de uma

sequência de 20 estimadores observados. Por algum motivo ainda a ser estudado através da análise de erro quadrado médio para valores de σ menores, percebemos uma menor dispersão das observações. Porém, mesmo assim, temos que o estimador é não viesado, pois na média, os estimadores se concentram sobre a linha reta. No caso da função exponencial, percebemos que a dispersão das observações foi maior.

Uma segunda forma de análise, a analítica, está baseada no Teorema 4.5. Iremos formalizar alguns de nossos resultados para que se enquadrem nas hipóteses do teorema citado. A expressão dada em (201) nos permite observar que o seu primeiro membro é escrito exclusivamente pelo parâmetro σ , este sendo o foco de interesse da estimação. No segundo membro, a expressão mostra uma relação entre os valores da amostra. Dessa forma, definamos uma função $f: (0, \infty) \rightarrow (0, \infty)$ dada por

$$f(x) = x^3\psi'(x) \quad (206)$$

e consideremos a sua inversa $g = f^{-1}$. Com essa notação, definimos a função de estimação $\hat{\sigma}_n : T \rightarrow \mathbb{R}$ um $(\sigma_n)_{n \geq 1}$ dada por

$$\sigma_n = g(T_n), \quad (207)$$

em que $T_n = \frac{(X_i - \mu)^2}{n}$. Dessa forma, temos como estimar o parâmetro σ que maximiza a função de verossimilhança expressa em (199). Pela derivada da inversa LIMA (2004), $g' = \frac{1}{f'}$, temos que

$$g' = \frac{1}{3x^2\psi'(x) + x^3\psi''(x)}. \quad (208)$$

A expressão (208) é diferente de zero para todo x . Considerando que a expressão (208) esteja bem definida, isto é, $3x^2\psi'(x) + x^3\psi''(x) \neq 0$, podemos usar o Teorema 4.5 e concluir que o método de estimação proposto para o parâmetro σ converge em distribuição. Como detalhe final, vamos avaliar a variância desse estimador.

4.2 Desigualdade de Cramér–Rao

Um importante resultado que se segue estabelece como limitante inferior para uma função de estimação não viesada a inversa da matriz de informação de Fisher $I(\theta)$. Quando esse limite é atingido por algum estimador não enviesado, temos a menor variância possível do estimador, o que é diretamente percebido através da expressão (156), já que $\mathbb{E}_\theta(\hat{\theta}) = \theta$ minimiza o erro. Isto é, o *EQM* coincide com a variância.

Teorema 4.7 *Sejam $\{X_1, \dots, X_n\}$ uma amostra de uma variável aleatória X distribuídas conforme uma função de densidade de probabilidade p_θ e um estimador $E_n(X)$ não viesado e de variância finita do parâmetro θ . Então,*

$$\text{Var}[E_n(X)] \geq \frac{1}{I_F(\theta)}. \quad (209)$$

A demonstração do Teorema 4.7, denominado por *Teorema da Desigualdade de Cramér–Rao*, é encontrada em KAY (1993). Vale destacar que o denominador do lado direito da expressão (209) é a matriz de informação de Fisher $I(\theta)$, exibida previamente em (5). Esse resultado será generalizado a logo a seguir. Além disso, um estimador de θ que satisfaz o mínimo de (209) é chamado de *eficiente* DAVID and NAGAJARA (1981). Com esta definição, calcula-se a eficiência de um estimador $\hat{\theta}$ através da razão abaixo que é limitada devido ao Teorema 4.7 DAVID and NAGAJARA (1981).

$$\text{eficiência}(\hat{\theta}) = \frac{1/I(\theta)}{\text{Var}(\hat{\theta})} \leq 1.$$

Dando continuidade a alguns novos resultados, vamos expor uma versão mais geral do Teorema 4.7. A análise de desempenho de estimadores por limitantes da variância é objeto de estudos há décadas como podemos observar em alguns trabalhos PUNTANEN and STYAN (1996); RAO (1992); VAN DEN BOS (2007). Algumas versões mais gerais do que essas últimas referenciadas tem sido discutidas para casos, por exemplo, em que os parâmetros são vetoriais GINI, REGGIANNINI, and MENGALI (1998). Em BATALAMA and KAZAKOS (1997), o autor mostra que a distribuição gaussiana generalizada maximiza uma generalização da desigualdade de Cramér–Rao para o p -ésimo momento central. Seguindo essa mesma linha, podemos adaptar à φ -família uma versão do limitante de Cramér–Rao.

Com efeito, consideremos a família de distribuições parametrizadas pela φ -família (65). Com a notação (52), seja

$$E'_\theta[\hat{\theta}] = \frac{\int_T \hat{\theta} \varphi'(f_\theta) d\mu}{\int_T u_0 \varphi'(f_\theta) d\mu}. \quad (210)$$

Derivando a expressão acima com respeito a θ , obtemos

$$\begin{aligned} \frac{\partial}{\partial \theta} E'_\theta[\hat{\theta}] &= \frac{(\int_T \hat{\theta} \varphi''(f_\theta) \frac{\partial f_\theta}{\partial \theta} d\mu)(\int_T u_0 \varphi'(f_\theta) d\mu)}{(\int_T u_0 \varphi'(f_\theta) d\mu)^2} - \frac{(\int_T \hat{\theta} \varphi'(f_\theta) d\mu)(\int_T \hat{\theta} \varphi''(f_\theta) \frac{\partial f_\theta}{\partial \theta} d\mu)}{(\int_T u_0 \varphi'(f_\theta) d\mu)^2} \\ &= \frac{\int_T \hat{\theta} \varphi''(f_\theta) \frac{\partial f_\theta}{\partial \theta} d\mu}{\int_T u_0 \varphi'(f_\theta) d\mu} - \frac{\int_T \hat{\theta} \varphi'(f_\theta) d\mu}{\int_T u_0 \varphi'(f_\theta) d\mu} \frac{\int_T \hat{\theta} \varphi''(f_\theta) \frac{\partial f_\theta}{\partial \theta} d\mu}{\int_T u_0 \varphi'(f_\theta) d\mu} \\ &= E''_\theta \left[\hat{\theta} \frac{\partial f_\theta}{\partial \theta} - E'_\theta[\hat{\theta}] u_0 \frac{\partial f_\theta}{\partial \theta} \right] \\ &= E''_\theta \left[(\hat{\theta} - E'_\theta[\hat{\theta}] u_0) \frac{\partial}{\partial \theta} f_\theta \right]. \end{aligned} \quad (211)$$

Sabemos que se X e Y são duas variáveis aleatórias com variância finita. A desigualdade de Cauchy–Schwartz é dada por

$$E[XY] \leq \sqrt{E[X^2]E[Y^2]}. \quad (212)$$

Portanto, usando a desigualdade (212), temos

$$E''_{\theta} \left[(\hat{\theta} - E'_{\theta}[\hat{\theta}]) u_0 \frac{\partial f_{\theta}}{\partial \theta} \right]^2 \leq E''_{\theta} [(\hat{\theta} - E'_{\theta}[\hat{\theta}]) u_0]^2 E''_{\theta} \left[\left(\frac{\partial f_{\theta}}{\partial \theta} \right)^2 \right]. \quad (213)$$

Portanto, usando (211) em (213), temos

$$\left(\frac{\partial}{\partial \theta} E'_{\theta}[\hat{\theta}] \right)^2 \leq E''_{\theta} [(\hat{\theta} - E'_{\theta}[\hat{\theta}]) u_0]^2 E''_{\theta} \left[\left(\frac{\partial f_{\theta}}{\partial \theta} \right)^2 \right],$$

a qual analogamente resulta na versão generalizada da Desigualdade de Cramér–Rao

$$E''_{\theta} [(\hat{\theta} - E'_{\theta}[\hat{\theta}]) u_0]^2 \geq \frac{\left(\frac{\partial}{\partial \theta} E'_{\theta}[\hat{\theta}] \right)^2}{E''_{\theta} \left[\left(\frac{\partial f_{\theta}}{\partial \theta} \right)^2 \right]}. \quad (214)$$

De fato, a desigualdade (214) é uma versão mais geral do teorema 4.7. Com efeito, suponhamos que $\varphi^{-1}(p_{\theta}) = f_{\theta} = l_{\theta} = \log p_{\theta}$ e $u_0(x) \equiv 1$. Consideremos as notações pré-definidas em (52). Observe que

$$\begin{aligned} E'_{\theta}[\hat{\theta}] &= \frac{\int_T \hat{\theta} \varphi'(f_{\theta}) d\mu}{\int_T u_0 \varphi'(f_{\theta}) d\mu} \\ &= \frac{\int_T \hat{\theta} \exp'(\log(p_{\theta})) d\mu}{\int_T \exp'(\log(p_{\theta})) d\mu} \\ &= \int_T \hat{\theta} p_{\theta} d\mu. \end{aligned}$$

Logo, implica-se que

$$\left(\frac{\partial}{\partial \theta} E'_{\theta}[\hat{\theta}] \right)^2 = \left(\frac{\partial}{\partial \theta} \int_T \hat{\theta} p_{\theta} d\mu \right)^2 = 1. \quad (215)$$

Além disso, de maneira mais direta

$$\begin{aligned} E''_{\theta} \left[\left(\frac{\partial f_{\theta}}{\partial \theta} \right)^2 \right] &\stackrel{f_{\theta} \equiv l_{\theta}}{=} E''_{\theta} \left[\left(\frac{\partial l_{\theta}}{\partial \theta} \right)^2 \right] \\ &= E_{\theta} \left[\left(\frac{\partial l_{\theta}}{\partial \theta} \right)^2 \right] \\ &= -E_{\theta} \left[\frac{\partial^2 l_{\theta}}{\partial \theta^2} \right], \end{aligned} \quad (216)$$

em que essa última igualdade vem de (9). Por fim,

$$E''_{\theta} [(\hat{\theta} - E'_{\theta}[\hat{\theta}]) u_0]^2 = E_{\theta} [(\hat{\theta} - E_{\theta}[\hat{\theta}])^2] = \text{Var}[\hat{\theta}]. \quad (217)$$

Dessa forma, utilizando-se das expressões (215), (216) e (217) e as substituindo em (214), obtemos a desigualdade clássica da desigualdade de Cramér–Rao expressada dada em (209).

4.3 Síntese

Neste capítulo, apresentamos uma rápida revisão do método de estimação de máxima verossimilhança e de conceitos e resultados que versam sobre convergência. Com essas preliminares, propomos um método de estimação denominado por máxima verossimilhança generalizado que atuará sobre as φ -família. Isto é, dado que a φ -família é uma família paramétrica, dada uma amostra, estima-se o parâmetro θ que identifica injetivamente uma função densidade de probabilidade p_θ .

Caracterizamos uma condição necessária para que um estimador de um parâmetro θ maximize a função de verossimilhança definida em um espaço paramétrico. Mostra-se que a minimização da função de divergência $\mathcal{D}_\varphi(\cdot \| \cdot)$ é equivalente a maximizar a função de verossimilhança. Discutimos sobre a família φ -gaussiana, um subfamília da φ -família, em que mostramos que o método de estimação proposto pode não implicar em uma solução analítica. Fato este nos motivou a propor um algoritmo de busca de parâmetros, específico para a φ -gaussiana, na qual vimos que tem boa eficiência por simulações numéricas e análise analítica.

5 CONCLUSÕES, PERSPECTIVAS e APLICAÇÕES

5.1 Conclusões

Nessa tese, as principais contribuições se destacam para as áreas de geometria da informação e inferência. Para a Geometria da Informação, a partir das propriedades de uma φ -função, descrevemos os principais elementos de geometria que caracterizam uma nova família de distribuições de probabilidade denominada por φ -família. Para questões de inferência, propomos um método de estimação baseado no de verossimilhança, com a qual percebemos sua eficiência em identificar parâmetros de distribuições numa φ -família. Além do mais, esse método pode ser aplicado, através de simulações numéricas, para uma classe geral similar à família de gaussianas.

Mais precisamente, em geometria da informação, a mudança da função exponencial para uma φ -função como função de parametrização de uma variedade estatística, possibilita-nos estender resultados que são aplicados em famílias exponenciais e famílias exponenciais deformadas AMARI and NAGAOKA (2000); MATSUZOE and HENMI (2013). Introduzimos um par de conexões $D^{(1)}$ e $D^{(-1)}$ induzidos por uma chamada φ -divergência. Generalizamos a divergência de Rényi com a qual propomos uma classe de funções de divergência $\mathcal{D}_\varphi(\cdot \parallel \cdot)$. Essa classe de divergência nos garante uma classe de α -conexões $D^{(\alpha)}$, nas quais a combinação convexa $D^{(\alpha)} = \frac{1-\alpha}{2}D^{(-1)} + \frac{1+\alpha}{2}D^{(1)}$ é satisfeita. Observe que a α -conexão é a mesma quando a α -divergência de Amari. Além disso, a generalização da divergência de Rényi nos mostra a importância das φ -funções na geometria de modelos não padrões.

Referindo-se agora à inferência, o fato da φ -família ser paramétrica nos sugere o problema de estimação de parâmetros. Portanto, definimos um método de estimação denominado Estimação de Verossimilhança Generalizado, o qual se mostrou útil para as φ -famílias. Mostramos que maximizar a função de verossimilhança equivale à minimizar a φ -divergência e propomos um algoritmo de busca numérica para estimação de parâmetros para a φ -gaussiana, o qual se mostrou eficiente.

Essa generalização tornou-se útil para problemas em que uma amostra não seja distribuída segundo funções densidade de probabilidade convencionais.

5.2 Perspectivas

Devido ao fato de pesquisas em Geometria da Informação são constantemente desenvolvidas nas mais diversas áreas da engenharia, estatística e matemática, é natural abordarmos ainda algumas questões que surgem naturalmente. Diante de todo o trabalho proposto, podemos citar algumas sugestões.

- Sabe-se que é possível definir uma função de divergência em uma variedade estatística dual MATUMOTO (1993), fato este que nos induziu a definir uma φ -divergência

na φ -família \mathcal{P} . Porém quando não temos uma conexão dual, Amari e Nihat AY and AMARI (2015) mostram como se pode definir a divergência canônica utilizando integração geodésica da aplicação exponencial inversa. Prova-se que quando supõe-se a dualidade, esta divergência coincide conforme se propõe em MATUMOTO (1993). Pode-se investigar como essa propriedade pode ser adaptada na φ -família.

- Uma generalização da família exponencial é dada pela família exponencial deformada MATSUZOE (2014). Um exemplo dessa família é dado através da uso da função q -exponencial que resulta na da família q -exponencial, a qual é uma variedade de curvatura constante. Pode-se classificar uma φ -família com respeito a algumas particularidade de curvatura.
- Desenvolver um algoritmo de busca de parâmetro mais geral do que o citado na seção 4.2, no sentido de se estender para qualquer φ -família e analisar a sua eficiência por simulações computacionais e analiticamente. Como fazer o estudo de convergência analítica pode ser citado na sugestão que se segue;
- Na seção 4.2, mostramos analiticamente a convergência do método de estimação generalizado para as chamadas φ -gaussianas, através da equivalência entre maximizar a função de verossimilhança em relação ao parâmetro desejado e uma solução da equação $\sigma^3\psi'(\sigma) = E'_{\sigma,\mu}[(x - \mu)^2]$, na qual nos permite encontrar uma função que retorna σ em função da amostra, como descrito nas expressões (206) e (207). Tais expressões nos permitem usar o teorema 4.5 que garante a convergência em distribuição. Tal equação foi possível ser obtida por propriedades intrínsecas da φ -gaussiana. Para uma amostra distribuída segunda uma φ -função qualquer, como em (65), não conseguimos escrever o problema nas hipóteses do teorema 4.5. Porém, o teorema 4.4 garante a convergência caso mostrássemos as condições de regularidades descritas em 4.10. Mais precisamente, precisaríamos mostrar que

$$E_{\theta} \left[\frac{\partial}{\partial \theta^i} \varphi^{-1}(p_{\theta}) \right] = 0,$$

para todo θ . Porém, temos uma versão mais geral e equivalente

$$E'_{\theta} \left[\frac{\partial}{\partial \theta^i} \varphi^{-1}(p_{\theta}) \right] = 0.$$

A ideia é ver quais as conclusões do teorema 4.4 para essa hipótese mais enfraquecida;

- Fazer uma análise do erro quadrado médio mostrado nos gráficos 1 e 2. Percebe-se que nessa figuras, o estimador de parâmetros em ambas as φ -funções começa a ter uma maior dispersão, o que nos induz a fazer esse estudo;
- Apresentamos uma versão mais geral da Desigualdade de Cramér–Rao apresentada em sua forma na expressão (214). Uma sequência natural a ser estudada é mostrar

que existe um estimador que $\hat{\theta}$ que atinge o limitante inferior. A estratégia é conhecida e pode ser encontrada na equação (3.7) de KAY (2013). Em outras palavras, um estimador não viesado $\hat{\theta}$ atinge o limite inferior para todo θ se, e somente se, podemos fatorar a função score por

$$\frac{\partial \ln p_{\theta}(x)}{\partial \theta} = I(\theta)(g(x) - \theta)$$

para algumas funções g e I . Dessa forma, ao definirmos $\hat{\theta} = g(x)$, definimos um estimador que atinge o mínimo do limitante de Cramér–Rao;

- Destacar a geometria afim de uma variedade, sujeita à uma φ -função, para que se possa fazer cálculos de curvaturas. Isto é, fazer um estudo da geometria de segunda ordem;

5.3 Aplicações

Geometria da Informação é um campo de pesquisa que nos proporciona investigações em várias direções, dentre as quais podemos destacar aplicações em inferência estatística, teoria da informação, neurocomputação, dentre outras. A estrutura geométrica que equipara as distribuições de probabilidade como uma variedade riemanniana, ou variedade estatística, como já definimos, é o modelo mais usual para extrair propriedades que possam ter aplicações AMARI and NAGAOKA (2000). Dentro do contexto das φ -família, é um interesse analisar as possíveis interpretações em geometria da informação em campos variacionais. Mais precisamente, quando consideramos a φ -gaussiana, podemos focar na avaliação da matriz de Fisher multivariacional, como feito no caso para a gaussiana multivariacional em COSTA, SANTOS, and STRAPASSON (2015).

REFERÊNCIAS

- ABBENA, Elsa; SALAMON, Simon; GRAY, Alfred. *Modern differential geometry of curves and surfaces with Mathematica*. CRC press, 2006.
- ALDRICH, John. R. A. Fisher and the Making of Maximum Likelihood 1912-1922. *Statistical Science*, v. 12, n. 3, p. 162–176, 1997.
- AMARI, Shun-Ichi. Information geometry and its applications: survey. *Geometric science of information*, Springer, Heidelberg, v. 8085 of *Lecture Notes in Comput. Sci.*, p. 3. 2013. URL http://dx.doi.org/10.1007/978-3-642-40020-9_1.
- AMARI, Shun-Ichi; KUMON, Masayuki. Estimation in the presence of infinitely many nuisance parameters—geometry of estimating functions. *Ann. Statist.*, v. 16, n. 3, p. 1044–1068, 1988. URL <http://dx.doi.org/10.1214/aos/1176350947>.
- AMARI, Shun-Ichi; NAGAOKA, Hiroshi. *Methods of information geometry*, v. 191 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI; Oxford University Press, Oxford, 2000, x+206 p. Translated from the 1993 Japanese original by Daishi Harada.
- AMARI, Shun-Ichi; OHARA, Atsumi. Geometry of q -exponential family of probability distributions. *Entropy*, v. 13, n. 6, p. 1170–1185, 2011. URL <http://dx.doi.org/10.3390/e13061170>.
- AMARI, Shun-Ichi; OHARA, Atsumi; MATSUZOE, Hiroshi. Geometry of deformed exponential families: invariant, dually-flat and conformal geometries. *Phys. A*, v. 391, n. 18, p. 4308–4319, 2012. URL <http://dx.doi.org/10.1016/j.physa.2012.04.016>.
- AY, Nihat; AMARI, Shun-Ichi. A novel approach to canonical divergences within information geometry. *Entropy*, v. 17, n. 12, p. 8111–8129, 2015.
- BASAWA, Ishwar V; GODAMBE, Vp; TAYLOR, Robert Lee. Selected proceedings of the Symposium on Estimating Functions. IMS, 1997.
- BATALAMA, Stella n; KAZAKOS, Dimitri. On the generalized Cramer-Rao bound for the estimation of the location. *IEEE transactions on signal processing*, v. 45, n. 2, p. 487–492, 1997.
- BILLINGSLEY, Patrick. *Convergence of probability measures*. John Wiley & Sons, 2013.
- CALIN, Ovidiu; UDRISTE, Constantin. *Geometric modeling in probability and statistics*. Springer, Cham, 2014, xxiv+375 p. URL <http://dx.doi.org/10.1007/978-3-319-07779-6>.

- CAMERON, A Colin; TRIVEDI, Pravin K. Econometric models based on count data. Comparisons and applications of some estimators and tests. *Journal of applied econometrics*, v. 1, n. 1, p. 29–53, 1986.
- CAMPBELL, LL. An extended Čencov characterization of the information metric. *Proceedings of the American Mathematical Society*, v. 98, n. 1, p. 135–141, 1986.
- CARMO, Manfredo P. do. *Differential geometry of curves and surfaces*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1976a, viii+503 p. Translated from the Portuguese.
- CARMO, Manfredo Perdigao do. *Differential geometry of curves and surfaces*, v. 2. Prentice-hall Englewood Cliffs, 1976b.
- CASELLA, George; BERGER, Roger L. *Statistical inference*, v. 2. Duxbury Pacific Grove, CA, 2002.
- CENA, Alberto; PISTONE, Giovanni. Exponential statistical manifold. *Ann. Inst. Statist. Math.*, v. 59, n. 1, p. 27–56, 2007. URL <http://dx.doi.org/10.1007/s10463-006-0096-y>.
- CENCOV, Nikolai Nikolaevich. *Statistical decision rules and optimal inference*. 53. American Mathematical Soc., 2000.
- CHANDRASEKAR, B; KALE, BK. Unbiased statistical estimation functions for parameters in presence of nuisance parameters. *Journal of statistical planning and inference*, v. 9, n. 1, p. 45–54, 1984.
- COELHO, Rosângela Fernandes; NASCIMENTO, Vitor Heloiz; de QUEIROZ, Ricardo Lopes; ROMANO, João Marcos Travassos; CAVALCANTE, Charles Casimiro; VIGELIS, Rui Facundo. Information Geometry: An Introduction to New Models for Signal Processing. *Signals and Images: Advances and Results in Speech, Estimation, Compression, Recognition, Filtering, and Processing*, CRC Press, p. 455–491. 2015.
- COSTA, Sueli IR; SANTOS, Sandra A; STRAPASSON, João E. Fisher information distance: a geometrical reading. *Discrete Applied Mathematics*, v. 197, p. 59–69, 2015.
- COVER, Thomas M.; THOMAS, Joy A. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2. ed., 2006, xxiv+748 p.
- COVER, Thomas M; THOMAS, Joy A. *Elements of information theory*. John Wiley & Sons, 2012.
- CSISZÁR, Imre. Generalized cutoff rates and Renyi's information measures. *IEEE Transactions on information theory*, v. 41, n. 1, p. 26–34, 1995.

- DAVID, Herbert Aron; NAGAJARA, Haikady Navada. *Order statistics*. Wiley Online Library, 1981.
- EVANS, Merran; HASTINGS, Nicholas; PEACOCK, Brian. *Statistical distributions*. 2000.
- FAN, Ky. Fixed-point and minimax theorems in locally convex topological linear spaces. *Proceedings of the National Academy of Sciences of the United States of America*, v. 38, n. 2, p. 121, 1952.
- FISHER, Ronald Aylmer. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, v. 222, p. 309–368, 1922.
- GINI, Fulvio; REGGIANNINI, Ruggero; MENGALI, Umberto. The modified Cramer-Rao bound in vector parameter estimation. *IEEE Transactions on Communications*, v. 46, n. 1, p. 52–60, 1998.
- GODAMBE, Vidyadhar P. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, v. 31, n. 4, p. 1208–1211, 1960.
- GRASSELLI, MR. Dual connections in nonparametric classical information geometry. *Annals of the Institute of Statistical Mathematics*, v. 62, n. 5, p. 873–896, 2010.
- HAMILTON, James Douglas. *Time series analysis*, v. 2. Princeton university press Princeton, 1994.
- HARREMOËS, Peter. Interpretations of Rényi entropies and divergences. *Physica A: Statistical Mechanics and its Applications*, v. 365, n. 1, p. 57–62, 2006.
- HIRIART-URRUTY, Jean-Baptiste; Strodiot, Jean-Jacques; Nguyen, V Hien. Generalized Hessian matrix and second-order optimality conditions for problems with $C^1, 1$ data. *Applied mathematics and optimization*, v. 11, n. 1, p. 43–56, 1984.
- JAMES, Barry R Barry R. *Probabilidade: um curso em nível intermediário*. 519.2. 1996.
- KANIADAKIS, G.; LISSIA, M.; SCARFONE, A. M. Deformed logarithms and entropies. *Phys. A*, v. 340, n. 1-3, p. 41–49, 2004. URL <http://dx.doi.org/10.1016/j.physa.2004.03.075>. News and expectations in thermostatics.
- KANIADAKIS, Giorgio; SCARFONE, AM. A new one-parameter deformation of the exponential function. *Physica A: Statistical Mechanics and its Applications*, v. 305, n. 1, p. 69–75, 2002.

KAY, Steven M. *Fundamentals of Statistical Signal Processing: Estimation Theory*, v. 1 of *Prentice Hall Signal Processing Series*. Prentice-Hall, 1993.

KAY, Steven M. *Fundamentals of Statistical Signal Processing: Practical Algorithm Development*, v. 3. Pearson Education, 2013.

KENDALL, Maurice G; STUART, Alan; ORD, JK. *The advanced theory of statistics*, v. 3. London, 1968.

KRISHNAMURTHY, Akshay; KANDASAMI, Kirthevasan; POCZOS, Barnabas; Wasserman, Larry A. Nonparametric Estimation of Renyi Divergence and Friends. *ICML*. 2014, p. 919–927.

LAURITZEN, Steffen Lilholt. *Statistical manifolds*. JSTOR, 1984.

LEBANON, Guy. Axiomatic geometry of conditional models. *IEEE Transactions on Information Theory*, v. 51, n. 4, p. 1283–1294, 2005.

LEE, John M. Smooth manifolds. *Introduction to Smooth Manifolds*, Springer, p. 1–29. 2003.

LEHMANN, Erich Leo; CASELLA, George. *Theory of point estimation*. Springer Science & Business Media, 2006.

LEON-GARCIA, Alberto. *Probability, statistics, and random processes for electrical engineering*. Pearson/Prentice Hall Upper Saddle River, NJ, 2008.

LI, Fuxin; CARREIRA, Joao; LEBANON, Guy; SMINCHISESCU, Cristian. Composite Statistical Learning and Inference for Semantic Segmentation. 2014.

LIMA, Elon Lages. *Curso de Análise, vol. 2*. 1976.

LIMA, Elon Lages. *Análise real*. Impa, 2004.

LINDSAY, Bruce G. Exponential family mixture models (with least-squares estimators). *The Annals of Statistics*, p. 124–137, 1986.

LINDSAY, Bruce G; *et al.* The geometry of mixture likelihoods, part II: the exponential family. *The Annals of Statistics*, v. 11, n. 3, p. 783–792, 1983.

LIPTSER, Robert; SHIRYAEV, Albert N. *Statistics of random Processes: I. general Theory*, v. 5. Springer Science & Business Media, 2013.

MARTINS, Eduardo S.; STEDINGER, Jery R. Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resour. Res.*,

v. 36, n. 3, p. 737–744, 2000. URL <http://dx.doi.org/10.1029/1999WR900330>.

MATSUZOE, Hiroshi. Hessian structures on deformed exponential families and their conformal structures. *Differential Geom. Appl.*, v. 35, n. suppl., p. 323–333, 2014. URL <http://dx.doi.org/10.1016/j.difgeo.2014.06.003>.

MATSUZOE, Hiroshi; HENMI, Masayuki. Hessian structures on deformed exponential families. *Geometric science of information*, Springer, Heidelberg, v. 8085 of *Lecture Notes in Comput. Sci.*, p. 275–282. 2013. URL http://dx.doi.org/10.1007/978-3-642-40020-9_29.

MATUMOTO, Takao. Any statistical manifold has a contrast function—on the C^3 -functions taking the minimum at the diagonal of the product manifold. *Hiroshima Math. J.*, v. 23, n. 2, p. 327–332, 1993. URL <http://projecteuclid.org/euclid.hmj/1206128255>.

MOTA, João César Moura; CAVALCANTE, Charles Casimiro; LUCENA, Antonio Macilio Pereira de. Detecção de sinais não-ortogonais com superposição espectral usando o critério MV. Simpósio Brasileiro de Telecomunicações, 2005.

MURRAY, Michael K.; RICE, John W. *Differential geometry and statistics*. Monographs on statistics and applied probability. London, Glasgow, New York: Chapman & Hall, 1993. URL <http://opac.inria.fr/record=b1083593>.

MUSIELAK, Julian. *Orlicz spaces and modular spaces*. 1983.

NAUDTS, Jan. Estimators, escort probabilities, and ϕ -exponential families in statistical physics. *JIPAM. J. Inequal. Pure Appl. Math.*, v. 5, n. 4, p. Article 102, 15 pp. (electronic), 2004.

NAUDTS, Jan. *Generalised thermostatics*. Springer-Verlag London, Ltd., London, 2011, x+201 p. URL <http://dx.doi.org/10.1007/978-0-85729-355-8>.

O'NEILL, Barrett. *Semi-Riemannian Geometry With Applications to Relativity*, 103, v. 103. Academic press, 1983.

PETERSEN, Peter. *Riemannian geometry*, v. 171. Springer, 2006.

PISTONE, Giovanni. κ -exponential models from the geometrical viewpoint. *The European Physical Journal B*, v. 70, n. 1, p. 29–37, 2009.

PISTONE, Giovanni; ROGANTIN, Maria Piera. The exponential statistical manifold: mean parameters, orthogonality and space transformations. *Bernoulli*, v. 5, n. 4, p. 721–760, 1999. URL <http://dx.doi.org/10.2307/3318699>.

PISTONE, Giovanni; SEMPI, Carlo. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Statist.*, v. 23, n. 5, p. 1543–1561, 1995. URL <http://dx.doi.org/10.1214/aos/11176324311>.

PUNTANEN, Simo; STYAN, George PH. A brief biography and appreciation of Calyampudi Radhakrishna Rao, with a bibliography of his books and papers. *Linear Algebra and its Applications*, v. 237, p. 1–40, 1996.

RAO, C Radhakrishna. Information and the accuracy attainable in the estimation of statistical parameters. *Breakthroughs in statistics*, Springer, p. 235–247. 1992.

ROGERS, Hartley. *Theory of recursive functions and effective computability*, v. 126. McGraw-Hill New York, 1967.

SALMON, Joseph; HARMANY, Zachary; DELEDALLE, Charles-Alban; WILLETT, Rebecca. Poisson noise reduction with non-local PCA. *Journal of mathematical imaging and vision*, v. 48, n. 2, p. 279–294, 2014.

SASAKI, Shigeo. ON THE DIFFERENTIAL GEOMETRY OF TANGENT BUNDLES OF RIEWANNIAN MANIFOLDS. *Tohoku Mathematical Journal, Second Series*, v. 10, n. 3, p. 338–354, 1958.

SCHEFFÉ, Henry. A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, v. 18, n. 3, p. 434–438, 1947.

SCHMETTERER, Leopold. *Introduction to mathematical statistics*, v. 202. Springer Science & Business Media, 2012.

SEN, Pranab K; SINGER, Julio M; DE LIMA, Antonio C Pedroso. *From finite sample to asymptotic methods in statistics*. Cambridge University Press, 2010.

SHANKS, ME; PURSELL, Lyle E. The Lie algebra of a smooth manifold. *Proceedings of the American Mathematical Society*, v. 5, n. 3, p. 468–472, 1954.

SHANNON, Claude Elwood. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, v. 5, n. 1, p. 3–55, 2001.

SHIMA, Hirohiko. *The geometry of Hessian structures*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2007, xiv+246 p. URL <http://dx.doi.org/10.1142/9789812707536>.

SOUZA, David C de; CAVALCANTE, Charles C; VIGELIS, Rui F. Modified maximum likelihood estimator. *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2016 IEEE*. IEEE, 2016, p. 1–5.

SUN, Hua-fei; PENG, Lin-yu; ZHANG, Zhen-ning; ZHANG, Liang; HE, Lang; LIU, Bingchen; LI, Fengjie; ZHENG, Sining; ZHAO, Liqin; AN, Li-Jian; *et al.* Information geometry and its applications. *Adv. Math.(China)*, v. 40, p. 257–269, 2011.

Trees, HARRY L. Van; Bell, KRISTINE L.; Tian, ZHI. *Detection, Estimation and Modulation Theory*, v. 1. Wiley, 2. ed., 2013.

TSALLIS, Constantino. What are the numbers that experiments provide. *Quimica Nova*, v. 17, n. 6, p. 468–471, 1994.

VAILLANT, Marc; MILLER, Michael I; YOUNES, Laurent; TROUVÉ, Alain. Statistics on diffeomorphisms via tangent space representations. *NeuroImage*, v. 23, p. S161–S169, 2004.

VAN DEN BOS, Adriaan. *Parameter estimation for scientists and engineers*. John Wiley & Sons, 2007.

VAN ERVEN, Tim; HARREMOËS, Peter. Rényi divergence and majorization. *2010 IEEE International Symposium on Information Theory*. IEEE, 2010, p. 1335–1339.

VAN ERVEN, Tim; HARREMOËS, Peter. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inform. Theory*, v. 60, n. 7, p. 3797–3820, 2014. URL <http://dx.doi.org/10.1109/TIT.2014.2320500>.

VAN ERVEN, Tim; HARREMOS, Peter. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, v. 60, n. 7, p. 3797–3820, 2014.

VAN TREES, Harry L. *Detection, estimation, and modulation theory*. John Wiley & Sons, 2004.

VIGELIS, Rui F.; CAVALCANTE, Charles C. The Δ_2 -condition and ϕ -families of probability distributions. *Geometric science of information*, Springer, Heidelberg, v. 8085 of *Lecture Notes in Comput. Sci.*, p. 729–736. 2013a. URL http://dx.doi.org/10.1007/978-3-642-40020-9_81.

VIGELIS, Rui F.; CAVALCANTE, Charles C. On ϕ -families of probability distributions. *J. Theoret. Probab.*, v. 26, n. 3, p. 870–884, 2013b. URL <http://dx.doi.org/10.1007/s10959-011-0400-5>.

VIGELIS, Rui F.; SOUZA, David C de.; CAVALCANTE, Charles Casimiro. New Metric and Connections in Statistical Manifolds. Frank Nielsen; Frédéric Barbaresco (Ed.), *GSI*. Springer, 2015, v. 9389 of *Lecture Notes in Computer Science*, p. 222–229. URL <http://dblp.uni-trier.de/db/conf/gsi/gsi2015.html#VigelisSC15>.

WALD, Abraham. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, v. 20, n. 4, p. 595–601, 1949.

WALPOLE, Ronald E; MYERS, Raymond H; MYERS, Sharon L; YE, Keying. *Probability and statistics for engineers and scientists*, v. 5. Macmillan New York, 1993.

ZACKS, Shelemyahu. *The theory of statistical inference*, v. 34. Wiley New York, 1971.

ZHANG, Cun-Hui. Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica*, p. 1297–1318, 2009.

6 APÊNDICE-INTRODUÇÃO À GEOMETRIA RIEMANNIANA

Ao longo desta tese, alguns conceitos e resultados de geometria riemanniana de extrema importância foram citados e utilizados. Neste apêndice, iremos formalizá-los e detalhá-los de maneira rápida a fim de melhor compreensão da leitura e como forma de consulta tendo em vista requisitos mínimos.

6.1 Variedades Diferenciáveis

As *variedades diferenciáveis* são estruturas geométricas que nos permite estudar formular conceitos de caráter local em espaços mais gerais que o \mathbb{R}^n . Isto é, dadas algumas restrições de definição, qualquer conjunto que se defina como variável é localmente semelhante ao \mathbb{R}^n . Podemos formalizar essa ideia através da definição que se segue.

Definição 6.1 *Uma variedade diferenciável \mathcal{M} é um conjunto não vazio munido de um conjunto de aplicações biunívocas $\varphi_\alpha: U_\alpha \subset \mathbb{R}^n \rightarrow \mathcal{M}$ de abertos U_α de \mathbb{R}^n em \mathcal{M} satisfazendo as seguintes condições que se seguem CARMO (1976a):*

- (i) $\bigcup_\alpha \varphi_\alpha(U_\alpha) = \mathcal{M}$;
- (ii) Para dois abertos U_α e U_β de \mathbb{R}^n tal que $\varphi_\alpha(U_\alpha) \cap \varphi_\beta(U_\beta) = W \neq \emptyset$, os conjuntos $\varphi_\alpha^{-1}(W)$ e $\varphi_\beta^{-1}(W)$ são conjuntos abertos em \mathbb{R}^n e a aplicação $\varphi_\beta^{-1} \circ \varphi_\alpha$ é diferenciável;

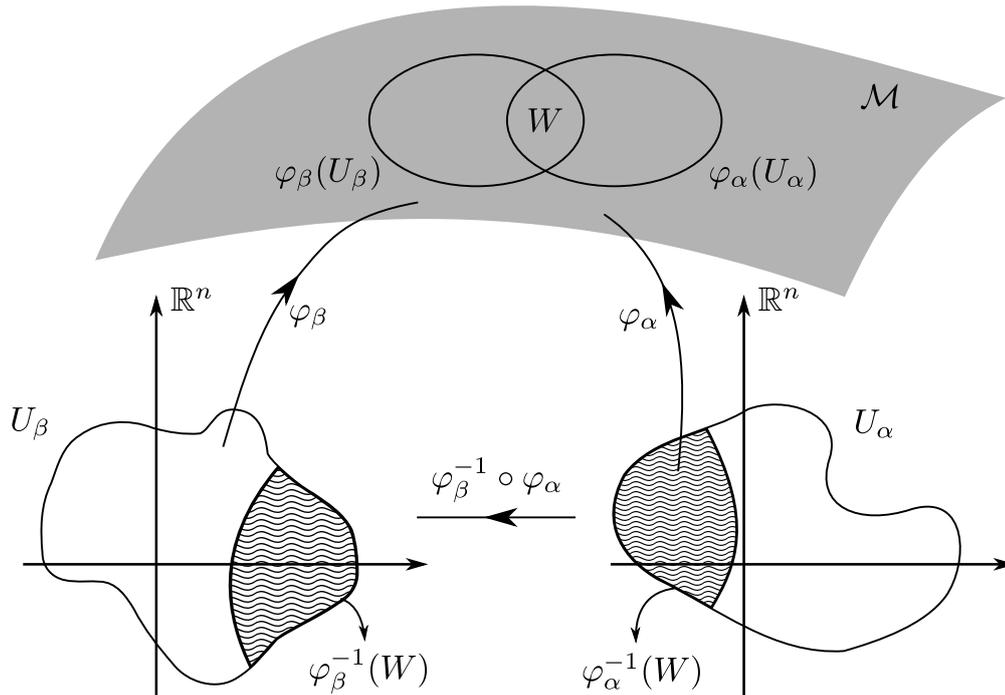


Figura 6: Parametrização diferenciável.

- (iii) Dada qualquer outra família $\{(U_{\alpha'}, \varphi_{\alpha'})\}$ que satisfaça as condições (i) e (ii), admitimos que $\{(U_{\alpha'}, \varphi_{\alpha'})\} \subseteq \{(U_{\alpha}, \varphi_{\alpha})\}$. Em outras palavras, a família $\{(U_{\alpha}, \varphi_{\alpha})\}$ é máxima relativamente às condições (i) e (ii).

Ademais, o par $(U_{\alpha}, \varphi_{\alpha})$ com $p \in \varphi_{\alpha}(U_{\alpha})$ é uma parametrização de \mathcal{M} em p . Em outras palavras, $\varphi_{\alpha}(U_{\alpha})$ é chamada de *vizinhança coordenada* em p . Um exemplo natural de variedade diferenciável é a φ -família definida em (65), em que a função φ se comporta como a função de parametrização citada na Definição 6.1. Uma consequência que se segue da definição de variedade é que podemos estabelecer a noção de função diferenciável definida em uma variedade.

Definição 6.2 *Sejam \mathcal{M}_1 e \mathcal{M}_2 duas variedades diferenciáveis de dimensão n e m , respectivamente. Uma função $f: \mathcal{M}_1 \rightarrow \mathcal{M}_2$ é diferenciável no ponto $p \in \mathcal{M}_1$ se dada uma parametrização $\varphi_2: V \subset \mathbb{R}^m \rightarrow \mathcal{M}_2$ em $f(p) \in \mathcal{M}_2$ existe uma parametrização $\varphi_1: U \subset \mathbb{R}^n \rightarrow \mathcal{M}_1$ de $p \in \mathcal{M}_1$ tal que $f(\varphi_1(U)) \subset \varphi_2(V)$ e a aplicação CARMO (1976a)*

$$\varphi_2^{-1} \circ f \circ \varphi_1: U \subset \mathbb{R}^n \rightarrow \mathbb{R}^m.$$

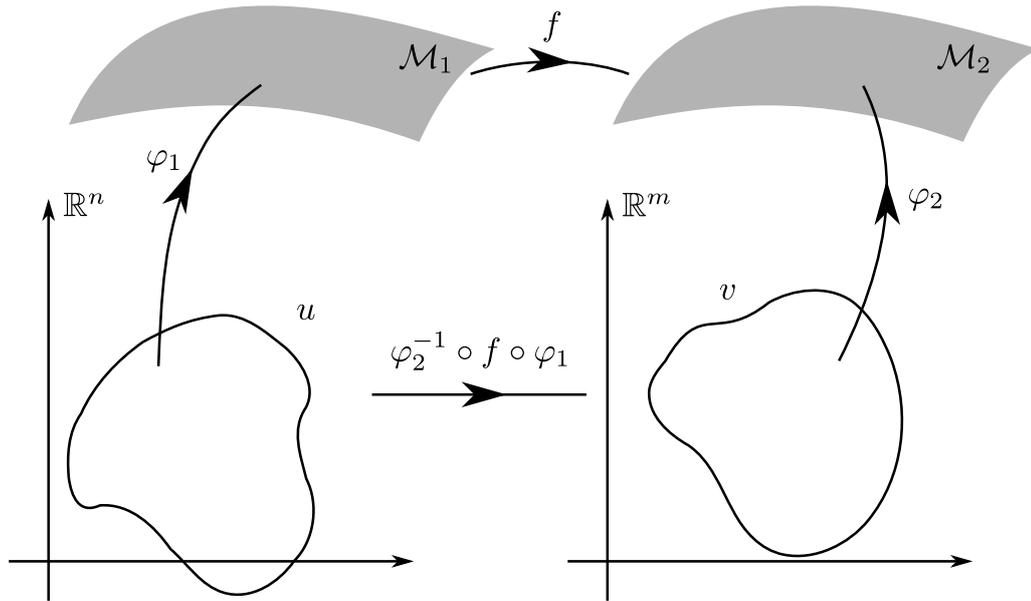


Figura 7: Função diferenciável f .

Observe que a definição anterior faz uso fortemente da Definição 6.1 para definir a diferenciabilidade de uma função definida em uma variável para a definição de diferenciabilidade dada para espaços euclidianos LIMA (1976).

Um outro conceito utilizado ao longo dessa tese remete à ideia de vetor tangente. Podemos estender a ideia que é aplicada em superfícies do \mathbb{R}^3 ABBENA, SALAMON, and GRAY (2006) para as variedades diferenciáveis. Com efeito, seja uma curva diferenciável $\alpha: (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}^n$, com $\alpha(0) = p$. Representando essa curva por suas coordenadas

$$\alpha(t) = (\alpha_1(t), \dots, \alpha_n(t)) \in \mathbb{R}^n, \quad (218)$$

o vetor tangente de α no ponto p é dado por

$$\alpha'(t) = (\alpha'_1(t), \dots, \alpha'_n(t)) \in \mathbb{R}^n. \quad (219)$$

Dada uma função diferenciável f definida em uma variedade, podemos restringi-la a uma curva α que satisfaz as condições (218) e (219). Dessa forma, a *derivada direcional de f na direção de $\alpha'(0)$* , em que $\alpha(0) = p$, é vetor no \mathbb{R}^n dado por

$$\frac{d(f \circ \alpha)}{dt} \Big|_{t=0} = \sum_{i=1}^n \frac{\partial f}{\partial \alpha_i} \Big|_{t=0} \frac{d\alpha_i}{dt} \Big|_{t=0} = \left(\sum_{i=1}^n \alpha'_i(0) \frac{\partial}{\partial \alpha_i} \right) f.$$

Portando, pode-se definir o vetor tangente em uma variedade diferenciável como abaixo.

Definição 6.3 *Dados uma variedade diferenciável \mathcal{M} de dimensão n e uma curva α , o*

vetor tangente à curva α no ponto $p \in \mathcal{M}$ aplicado à função diferenciável f é a função

$$\alpha'(0)(f) = \left. \frac{d(f \circ \alpha)}{dt} \right|_{t=0}, \quad (220)$$

em que $\alpha(0) = p$.

depende da escolha da curva α . Além disso, o conjunto de todos os vetores tangentes dados como em (220) é denominado espaço tangente de p em \mathcal{M} e é denotado por $T_p\mathcal{M}$. De uma forma geral, todos os vetores tangentes de um ponto identifica-se com espaço tangente em $p \in \mathcal{M}$ de uma variedade diferenciável. O espaço tangente tem a importância de dar uma noção de linearidade local na variedade. Além disso, $T_p\mathcal{M}$ é um espaço vetorial e faz sentido falar em uma base. Dessa forma, seja (θ^i) um sistema de coordenadas para \mathcal{M} e considere e_i o vetor tangente ao ponto p que é paralelo ao i -ésimo eixo. Então, o plano tangente pode ser gerado pelos vetores tangentes $\{e_1, \dots, e_n\}$. Por outro lado, dada uma parametrização $\varphi: U \rightarrow \mathcal{M}$ em $p = \varphi(0)$, podemos denotar a curva α e uma função f na parametrização por

$$(f \circ \varphi)(q) = f(x_1, \dots, x_n), \quad q = (x_1, \dots, x_n) \in U$$

e

$$(\varphi^{-1} \circ \alpha)(t) = (x_1(t), \dots, x_n(t)),$$

respectivamente. Dessa forma, considerando uma função diferenciável f definida em \mathcal{M} , podemos restringir f a α e obtemos

$$\begin{aligned} \alpha'(0) &= \left. \frac{d}{dt}(f \circ \alpha) \right|_{t=0} \\ &= \left. \frac{d}{dt} f(x_1(t), \dots, x_n(t)) \right|_{t=0} \\ &= \sum_{i=1}^n x'_i(0) \frac{\partial f}{\partial x_i} \\ &= \left(\sum_{i=1}^n x'_i(0) \left(\frac{\partial}{\partial x_i} \right)_0 \right) f. \end{aligned} \quad (221)$$

Portanto, comparando as expressões (220) e (221), podemos concluir que o vetor $\alpha'(0)$ por ser expresso em termos dos parâmetros por

$$\alpha'(0) = \sum_{i=1}^n x'_i(0) \left(\frac{\partial}{\partial x_i} \right)_0.$$

Portanto, concluímos que o vetor tangente é uma função $\alpha'(0)$ que atua em f , esta pertencente ao espaço das funções diferenciáveis definidas na variedade. Ademais, o conjunto $\{(\frac{\partial}{\partial x_i})_0, \dots, (\frac{\partial}{\partial x_n})_0\}$ é uma base do plano tangente $T_p\mathcal{M}$ associada à parametrização.

Além disso, vamos introduzir o conceito de aplicações suaves em variedades e um

importante resultado que relaciona espaços tangentes de duas variedades diferenciáveis \mathcal{M} e \mathcal{N} relacionadas por uma função suave.

Definição 6.4 *Uma aplicação $F : \mathcal{M} \rightarrow \mathbb{R}^n$ é suave se para cada ponto $p \in \mathcal{M}$, \mathcal{M} uma variedade diferenciável, existe uma vizinhança coordenada (U, φ) de p tal que $F \circ \varphi^{-1}$ é suave.*

Definição 6.5 *Uma aplicação $F : \mathcal{M} \rightarrow \mathcal{N}$, em que \mathcal{M} e \mathcal{N} são variedades diferenciáveis, é uma aplicação suave para $p \in \mathcal{M}$ existe uma vizinhança coordenada $(U, \varphi_{\mathcal{M}})$ de p e $(V, \varphi_{\mathcal{N}})$ de $F(p)$ tal que $F(U) \subseteq V$ e*

$$\varphi_{\mathcal{N}} \circ F \circ \varphi_{\mathcal{M}}^{-1} : \varphi_{\mathcal{M}}(U) \rightarrow \mathbb{R}^n$$

é suave. Além disso, a aplicação $\hat{F} = \varphi_{\mathcal{N}} \circ F \circ \varphi_{\mathcal{M}}^{-1}$ é chamada uma representação coordenada de F .

Exemplo 17 *Consideremos $\mathcal{M} = \mathbb{S}^3$ e $\mathcal{N} = \mathbb{R}^4$ e a aplicação inclusão entre as variedades \mathcal{M} e \mathcal{N} definida por*

$$(x_1, x_2, x_3) \mapsto (x_1, x_2, x_3, \pm \sqrt{1 - (\sqrt{x_1^2 + \dots + x_n^2})})$$

é uma função suave desde que as coordenadas sejam funções suaves, o que é o caso.

Teorema 6.1 *Seja $F : \mathcal{M} \rightarrow \mathcal{N}$ uma aplicação suave entre duas variedades diferenciáveis \mathcal{M} e \mathcal{N} . Fixado um ponto $p \in \mathcal{M}$ e $v \in T_p\mathcal{M}$, escolha uma curva $\gamma : (-\varepsilon, \varepsilon) \rightarrow \mathcal{M}$ com $\gamma(0) = p$ e $\gamma'(0) = v$. Seja uma outra curva β em \mathcal{N} definida como a composição $\beta = F \circ \gamma$. A aplicação $F_* : T_p\mathcal{M} \rightarrow T_{F(p)}\mathcal{N}$ dada por $F_*(v) = \beta'(0)$ é uma aplicação linear que não depende da escolha de γ . A aplicação linear F_* é chamada de pushforward de F em p .*

Costumeiramente, a aplicação pushforward é definida como a diferencial de f em p . Algumas referências denotam por $df_p, df_p : T_p\mathcal{M} \rightarrow T_{F(p)}\mathcal{N}$.

Observação 1 *Ao longo dos capítulos da tese, identificamos o espaço paramétrico por (θ^i) . Embora difira do que abordamos aqui no apêndice, mas é a notação que a maior parte das referências de Geometria da Informação usam em seus trabalhos aqui citados.*

Um outro conceito utilizado foi o de subvariedade. Para tal, necessita-se definir as aplicações diferenciáveis denominadas por Imersão e Mergulho.

Definição 6.6 *Sejam \mathcal{M} e \mathcal{N} variedades diferenciáveis de dimensões m e n , respectivamente. Uma aplicação $f : \mathcal{M} \rightarrow \mathcal{N}$ é dita uma imersão quando a aplicação diferencial df_p definida acima seja injetiva para qualquer ponto $p \in \mathcal{M}$. Ademais, se f também é um homeomorfismo LEE (2003) de \mathcal{M} sobre sua imagem $f(\mathcal{M}) \subset \mathcal{N}$, denominamos f por mergulho.*

Com esse conceito citado acima, podemos definir o que é uma subvariedade.

Definição 6.7 *Seja \mathcal{M} uma variedade de dimensão n . Dizemos que \mathcal{N} é uma subvariedade de \mathcal{M} , quando $\mathcal{N} \subset \mathcal{M}$ e a função inclusão $i : \mathcal{N} \rightarrow \mathcal{M}$ é um mergulho.*

6.2 Fibrado Tangente e Campos Vetoriais

Um dos termos bem utilizados ao longo dessa tese foi o de *campo vetorial*.

Informalmente, um campo vetorial é uma função que toma um ponto na variedade e associa com um vetor tangente. Podemos formalizar melhor a ideia de campo vetorial apresentando antes o fibrado tangente SASAKI (1958). Portanto, seja \mathcal{M} uma variedade diferenciável. Chamamos de *fibrado tangente* ao conjunto

$$T\mathcal{M} = \bigcup_{p \in \mathcal{M}} \{p\} \times T_p\mathcal{M}. \quad (222)$$

Portanto, cada elemento de $T\mathcal{M}$ é um par ordenado da forma (p, v) , em que $p \in \mathcal{M}$ e $v \in T_p\mathcal{M}$ e assim, podemos definir o que seja um campo vetorial.

Definição 6.8 *Um campo vetorial é uma função $X : \mathcal{M} \rightarrow T\mathcal{M}$ tal que*

$$p \mapsto X(p) \in T_p\mathcal{M}, \forall p \in \mathcal{M}.$$

Naturalmente, podemos associar uma função de projeção $\pi : T\mathcal{M} \rightarrow \mathcal{M}$ tal que

$$\pi(X(p)) \mapsto p, \forall p \in \mathcal{M}. \quad (223)$$

Por esse motivo, ao tomarmos a inversa da projeção $\pi^{-1} : \mathcal{M} \rightarrow T_p\mathcal{M}$, chamamos a imagem da inversa de *fibra*, o que faz jus ao termo fibrado tangente. Então, pela própria expressão da função projeção dada em (223), uma função X é um campo vetorial se, e somente se,

$$\pi \circ X \equiv id_{\mathcal{M}}. \quad (224)$$

Aplicando π^{-1} em (224), obtemos que $X = \pi^{-1}(p) \in T_p\mathcal{M}$.

Uma outra forma de se interpretar um campo vetorial é seguindo a ideia sugerida na expressão (221), em que o campo X é interpretado como uma função que atua em uma função diferenciável f como segue abaixo

$$(Xf)(p) = \sum_{i=1}^n X_i(p) \frac{\partial f}{\partial x_i}(p). \quad (225)$$

Dessa forma, observamos que a expressão (225) nos sugere que o campo vetorial atua como um operador derivação da função f no ponto p . Além disso, vejamos que se X e Y são campos diferenciáveis definidos em \mathcal{M} e $f: \mathcal{M} \rightarrow \mathbb{R}$ é uma função diferenciável, as derivadas de segunda ordem $X(Yf)$ não necessariamente ainda são campos vetoriais. Com essa motivação, podemos introduzir uma técnica de combinar dois campos de vetores X e Y dados pelo lema que se segue e pode ser encontrado em CARMO (1976a).

Lema 6.1 *Sejam X e Y dois campos diferenciáveis de vetores em uma variedade diferenciável \mathcal{M} . Então, existe um único campo vetorial Z tal que, para toda função diferenciável f definida em \mathcal{M} , $Zf = (XY - YX)f$.*

Esse campo vetorial Z é chamado o colchete $[X, Y] = XY - YX$ e este pode ser interpretado como uma derivação de Y ao longo de X . Observe que quando temos os campos vetoriais provenientes de um sistema de coordenadas (x_1, \dots, x_n) , o colchete de Lie satisfaz

$$\left[\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j} \right] = 0, \quad \text{para todo } i, j.$$

6.3 Variedades Riemannianas

Vamos introduzir a noção de produto interno em uma variedade \mathcal{M} , com a qual se é possível ter uma ideia de comprimento de vetores tangentes, ângulos entre vetores, comprimento de uma curva e também definir geodésicas. Dessa forma, apresentamos a definição abaixo.

Definição 6.9 *Uma variedade riemanniana \mathcal{M} de dimensão finita n é uma variedade diferenciável que associa a cada ponto $p \in \mathcal{M}$ uma forma bilinear simétrica definida positiva $\langle \cdot, \cdot \rangle_p$ no espaço tangente $T_p\mathcal{M}$. A função $\langle \cdot, \cdot \rangle_p$ é denominada de produto interno.*

Dada uma função de parametrização $\varphi: U \subset \mathbb{R}^n \rightarrow \mathcal{M}$ numa vizinhança do ponto p ,

para $\varphi(x_1, \dots, x_n) = p$, as funções

$$g_{ij}(x_1, \dots, x_n) = \left\langle \frac{\partial}{\partial x_i}(p), \frac{\partial}{\partial x_j}(p) \right\rangle_p, \quad i, j = 1, \dots, n, \quad (226)$$

são diferenciáveis em U , em que a derivada parcial $\frac{\partial}{\partial x_i}(p) = d\varphi_p \cdot (0, \dots, 1, \dots, 0)$ e o número 1 se encontra na i -ésima posição do vetor de coordenadas. Essa métrica proveniente do produto interno (226) é chamada de *métrica riemanniana*. No mesmo sentido, ao par (\mathcal{M}, g) , em que \mathcal{M} é uma variedade diferenciável e g uma métrica riemanniana em \mathcal{M} , é denominada de *variedade riemanniana*.

Dessa forma, dados dois campos de vetores $u = (u_1, \dots, u_n)$, $v = (v_1, \dots, v_n) \in T_p\mathcal{M}$ de uma variedade riemanniana \mathcal{M} , podemos o produto interno $\langle u, v \rangle_p$ é dado por

$$\langle u, v \rangle_p = u^T G_p v,$$

em que denotamos por G_p a matriz da métrica que identifica-se pelos seus termos g_{ij} .

Definição 6.10 *Um campo vetorial V ao longo de uma curva $\gamma : I \rightarrow \mathcal{M}$ é uma aplicação que associa a cada $t \in I$ um vetor tangente $V(t)$ em $T_{\gamma(t)}\mathcal{M}$. Mais formalmente, dada uma parametrização $\varphi : U \subset \mathbb{R}^n \rightarrow \mathcal{M}$, temos*

$$V(t) = \sum_{i=1}^n v_i(t) \frac{\partial}{\partial x_i},$$

em que as componentes coordenadas v_i de V são diferenciáveis e na base

$$\left\{ \left(\frac{\partial}{\partial x_1} \right)_{\gamma(0)}, \dots, \left(\frac{\partial}{\partial x_n} \right)_{\gamma(0)} \right\}$$

.

Nas superfícies regulares CARMO (1976b), mostra-se uma forma de calcular comprimento de curvas. Podemos estender essa medida para uma variedade riemanniana. Portanto, se a curva γ está definida em um intervalo fechado $[a, b]$, o seu comprimento $l(\gamma)$ é dado por

$$l(\gamma) = \int_a^b \left\langle \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \right\rangle^{\frac{1}{2}} dt.$$

6.4 Conexões

Definição 6.11 *Uma conexão D em uma variedade suave \mathcal{M} é uma aplicação $D : \mathcal{TM} \times \mathcal{TM} \rightarrow \mathcal{TM}$, a qual associa um par de campos de vetores X e Y com um outro*

campo de vetor $D_X Y$, satisfazendo as seguintes propriedades:

- (D1) $D_{fX+gY} Z = fD_X Z + gD_Y Z$,
- (D2) $D_X(Y + Z) = D_X Y + D_X Z$,
- (D3) $D_X(fY) = fD_X Y + X(f)Y$,

para cada campos de vetores $X, Y, Z \in \mathcal{TM}$ e quaisquer funções $f, g \in C^\infty(\mathcal{M})$.

Sejam X e Y dois campos de vetores suave, os quais podem ser expressos em termos de coordenadas locais como

$$X = \sum_i a^i \frac{\partial}{\partial x_i} \quad \text{e} \quad Y = \sum_j b^j \frac{\partial}{\partial x_j}.$$

Pelas propriedades da conexão, podemos escrever

$$\begin{aligned} D_X Y &= \sum_i a^i D_{\partial/\partial x_i} \left(\sum_j b^j \frac{\partial}{\partial x_j} \right) \\ &= \sum_i \sum_j a^i \frac{\partial b^j}{\partial x_i} \frac{\partial}{\partial x_j} + \sum_i \sum_j a^i b^j D_{\partial/\partial x_i} \left(\frac{\partial}{\partial x_j} \right). \end{aligned} \quad (227)$$

Introduzimos os *símbolos de Christoffel* Γ_{ij}^k por $D_{\partial/\partial x_i} \partial/\partial x_j = \sum_k \Gamma_{ij}^k \partial/\partial x^k$, a equação (227) resulta em

$$D_X Y = \sum_k \left(\sum_i a^i \frac{\partial b^k}{\partial x_i} + \sum_{i,j} a^i b^j \Gamma_{ij}^k \right) \frac{\partial}{\partial x_k}, \quad (228)$$

a qual é a forma da expressão local da derivada covariante $D_X Y$ em algum sistema de coordenadas (x_i) .

Definição 6.12 *Uma conexão D em uma variedade riemanniana (\mathcal{M}, g) é dita ser compatível com a métrica se a expressão*

$$Xg(Y, Z) = g(D_X Y, Z) + g(Y, D_X Z) \quad (229)$$

é satisfeita para quaisquer $X, Y, Z \in \mathcal{TM}$.

A equação (229) tem um significado especial. Dada uma curva γ sobre a variedade

riemanniana \mathcal{M} , temos que o produto interno $\langle X, Y \rangle$ entre os campos de vetores X e Y é constante. Em outras palavras, o ângulo formado entre os campos de vetores X e Y permanecem constante ao longo de γ . Além do mais, uma conexão D que satisfaz a equação (229) é denominada *compatível* com a métrica. Uma outra definição que se segue é o de simetria.

Definição 6.13 *Seja \mathcal{M} uma variedade diferenciável com uma conexão D . O tensor torção da conexão D é uma aplicação*

$$T : \mathcal{T}\mathcal{M} \times \mathcal{T}\mathcal{M} \rightarrow \mathcal{T}\mathcal{M}$$

definida por

$$T(X, Y) = D_X Y - D_Y X.$$

Observação 2 *Quando o tensor torção é identicamente nulo dizemos que a conexão D é simétrica. Ao considerarmos os campos base $X = \partial/\partial x_i$ e $Y = \partial/\partial x_j$, percebe-se que $T \equiv 0$, pois*

$$D_{\partial/\partial x_i} \left(\frac{\partial}{\partial x_j} \right) - D_{\partial/\partial x_j} \left(\frac{\partial}{\partial x_i} \right) = \left[\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j} \right] = 0.$$

Como resultado, os símbolos de Christoffel satisfazem $\Gamma_{ij}^k = \Gamma_{ji}^k$. Isto justifica escrever $[X, Y] = D_X Y - D_Y X$ para as conexões.

Teorema 6.2 (Levi-Civita). *Dada uma variedade riemanniana (\mathcal{M}, g) , existe uma única conexão simétrica D em \mathcal{M} que é compatível com a métrica riemanniana.*

A conexão dada segundo o Teorema de Levi-Civita é chamada de *conexão de Levi-Civita* em \mathcal{M} . Costumeiramente usa-se ∇ para representar a conexão de Levi-Civita. Os símbolos de Christoffel Γ_{ij}^k da conexão de Levi-Civita podem ser expressados em termos de g_{ij} . Devido à conexão ser compatível com a métrica,

$$\begin{aligned} \frac{\partial g_{jk}}{\partial x_i} &= g \left(\nabla_{\frac{\partial}{\partial x_i}} \frac{\partial}{\partial x_j}, \frac{\partial}{\partial x_k} \right) + g \left(\frac{\partial}{\partial x_j}, \nabla_{\frac{\partial}{\partial x_i}} \frac{\partial}{\partial x_k} \right), \\ \frac{\partial g_{ki}}{\partial x_j} &= g \left(\nabla_{\frac{\partial}{\partial x_j}} \frac{\partial}{\partial x_k}, \frac{\partial}{\partial x_i} \right) + g \left(\frac{\partial}{\partial x_k}, \nabla_{\frac{\partial}{\partial x_j}} \frac{\partial}{\partial x_i} \right), \\ \frac{\partial g_{ij}}{\partial x_k} &= g \left(\nabla_{\frac{\partial}{\partial x_k}} \frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j} \right) + g \left(\frac{\partial}{\partial x_i}, \nabla_{\frac{\partial}{\partial x_k}} \frac{\partial}{\partial x_j} \right). \end{aligned}$$

Fazendo uso da simetria de ∇ , estas expressões implicam em

$$\frac{\partial g_{jk}}{\partial x_i} + \frac{\partial g_{ki}}{\partial x_j} - \frac{\partial g_{ij}}{\partial x_k} = 2g \left(\frac{\partial}{\partial x_k}, \nabla_{\frac{\partial}{\partial x_j}} \frac{\partial}{\partial x_i} \right).$$

Da definição de símbolo de Christoffel, segue-se que

$$\Gamma_{ijk} := \sum_m \Gamma_{ij}^m g_{mk} = \frac{1}{2} \left(\frac{\partial g_{jk}}{\partial x_i} + \frac{\partial g_{ki}}{\partial x_j} - \frac{\partial g_{ij}}{\partial x_k} \right),$$

Denotando por (g^{km}) a inversa de (g_{km}) , obtemos

$$\Gamma_{ij}^m = \frac{1}{2} \sum_k \left(\frac{\partial g_{jk}}{\partial x_i} + \frac{\partial g_{ki}}{\partial x_j} - \frac{\partial g_{ij}}{\partial x_k} \right) g^{km}. \quad (230)$$

A expressão (230) é a expressão clássica para os símbolos de Christoffel da conexão de Levi-Civita em termos de g_{ij} .

Um campo vetorial ao longo de uma curva $\gamma : I \rightarrow \mathcal{M}$ é suave se, para cada função f em \mathcal{M} , a função $t \mapsto Vf(t)$ é suave em I . O campo vetorial $\gamma_*(d/dt)$, denotado por $d\gamma/dt$ é chamado de campo vetorial tangente da curva γ .

As conexões mostram sua importância quando, a partir delas, temos a ideia de derivada de campos ao longo de uma curva.

Proposição 6.1 *Seja \mathcal{M} uma variedade diferenciável com conexão D . existe uma única correspondência $\frac{D}{dt}$, a qual associa a um campo vetorial V ao longo de uma curva diferencial $\gamma : I \rightarrow \mathcal{M}$ um outro campo vetorial $\frac{DV}{dt}$ ao longo de γ , denominado de derivada covariante de V ao longo de γ , tal que*

- (a) $\frac{D}{dt}(V + W) = \frac{DV}{dt} + \frac{DW}{dt}$;
- (b) $\frac{D}{dt}(fV) = \frac{df}{dt}V + f\frac{DV}{dt}$, em que W é o campo de vetores ao longo de γ e f é uma função diferenciável em I ;
- (c) Se V é induzido por um campo de vetores $Y \in \mathcal{TM}$, isto é, $V(t) = Y(\gamma(t))$, então $\frac{DV}{dt} = \nabla_{d\gamma/dt} Y$.

Seja $(\gamma_1(t), \dots, \gamma_n(t))$ a expressão local da curva γ em termos de um sistema de coordenadas (x_i) . Portanto, nesse sistema de coordenadas, podemos expressar o campo vetorial $d\gamma/dt$ e V ao longo de γ como

$$\frac{d\gamma}{dt} = \sum_i \frac{d\gamma_i}{dt} \frac{\partial}{\partial x_i} \quad \text{e} \quad V = \sum_j v_j \frac{\partial}{\partial x_j}.$$

A expressão (228) e a propriedade (c) implica que

$$\frac{DV}{dt} = \sum_k \left(\frac{dv_k}{dt} + \sum_{i,j} \frac{d\gamma_i}{dt} v_j \Gamma_{ij}^k \right) \frac{\partial}{\partial x_k}. \quad (231)$$

A expressão (231) é uma ferramenta importante para desenvolvermos a ideia de paralelismo. Com efeito, vejamos a seguinte definição.

Definição 6.14 *Seja \mathcal{M} uma variedade diferenciável com uma conexão D . Um campo vetorial V ao longo de uma curva $\gamma : I \rightarrow \mathcal{M}$ é chamado de paralelo se $\frac{DV}{dt} = 0$, para todo $t \in I$.*

Teorema 6.3 *Seja uma variedade diferenciável \mathcal{M} com uma conexão D . Seja $\gamma : I \rightarrow \mathcal{M}$ uma curva suave em \mathcal{M} . Fixemos um vetor tangente V_0 no ponto $\gamma(t_0)$, para algum $t_0 \in I$. Então, existe um único campo paralelo de vetores V ao longo de γ tal que $V(t_0) = V_0$. O vetor $V(t)$ é chamado de transporte paralelo de V_0 ao longo de γ .*

O teorema acima é de grande importância, pois com ele, pode-se obter uma conexão a partir do transporte paralelo. Sejam X e Y campos de vetores em \mathcal{M} . Seja $p \in \mathcal{M}$ e $\gamma : I \rightarrow \mathcal{M}$ uma curva de X passando por p , isto é, $\gamma(t_0) = p$ e $\frac{d\gamma}{dt} = X(\gamma(t))$. Considere a aplicação

$$P_{\gamma, t_0, t} : T_{\gamma(t_0)}\mathcal{M} \rightarrow T_{\gamma(t)}\mathcal{M},$$

a qual é definida como o de um vetor ao longo de γ de t_0 para t . Então, temos

$$(D_X Y)(p) = \frac{d}{dt} P_{\gamma, t_0, t}^{-1} Y(\gamma(t))|_{t=t_0}.$$

Nesse contexto de transporte, a expressão (229) nos proporciona uma outra interpretação.

Teorema 6.4 *Seja (\mathcal{M}, g) uma variedade Riemanniana. Uma conexão D em \mathcal{M} é compatível com a métrica se, e somente se,*

$$\frac{d}{dt} g(V, W) = g\left(\frac{DV}{dt}, W\right) + g\left(V, \frac{DW}{dt}\right), \quad (232)$$

para cada campo de vetor V e W ao longo de uma curva γ .

6.5 Geodésicas

O conceito de geodésica estende a ideia de linha reta do espaço euclidiano para as variedades diferenciáveis e está fortemente ligado à noção de paralelismo.

Definição 6.15 *Seja \mathcal{M} uma variedade riemanniana. Uma curva $\gamma : I \rightarrow \mathcal{M}$ é uma geodésica se $\frac{D}{dt}\left(\frac{d\gamma}{dt}\right) = 0$ para todo $t \in I$.*

Em outras palavras, quando temos uma geodésica γ , isso significa que o campo velocidade não muda de direção. Observe também que se γ é uma geodésica sobre uma variedade riemanniana \mathcal{M} que possui uma conexão ∇ compatível com a métrica, a expressão (232) nos garante que o vetor velocidade de γ $\|\gamma'(t)\|_g$ tem comprimento constante c segundo a métrica g , desde que γ' seja um campo paralelo ao longo de γ . Quando isso ocorre, podemos definir a função comprimento de arco $s(t)$ dada por

$$s(t) = \int_{t_0}^t \|\gamma'(t)\|_g dt = c(t - t_0),$$

em que c é o comprimento do vetor γ' .

Expressando uma curva $\gamma = (\gamma_1(t), \dots, \gamma_n(t))$ em termos de suas coordenadas com respeito à um sistema de coordenadas (x_i) , a expressão (231) implica que

$$\frac{D}{dt}\left(\frac{d\gamma}{dt}\right) = \sum_k \underbrace{\left(\frac{d^2\gamma_k}{dt^2} + \sum_{i,j} \frac{d\gamma_i}{dt} \frac{d\gamma_j}{dt} \Gamma_{ij}^k\right)}_{\text{sistema de 2ª ordem}} \frac{\partial}{\partial x_k}.$$

Como os campos $\left(\frac{\partial}{\partial x_k}\right)_{\gamma(t)}$ formam um base para o espaço tangente em cada ponto $\gamma(t) \in \mathcal{M}$. Logo, se $\frac{D}{dt}\left(\frac{d\gamma}{dt}\right) = 0$, o sistema de 2ª ordem é identicamente nulo. Logo, uma curva γ que satisfaça ao sistema

$$\frac{d^2\gamma_k}{dt^2} + \sum_{i,j} \frac{d\gamma_i}{dt} \frac{d\gamma_j}{dt} \Gamma_{ij}^k = 0. \quad (233)$$

é uma geodésica. A equação (233) é chamada de *equação geodésica*.

Por fim, citamos o teorema que garante a existência e a unicidade de geodésicas.

Teorema 6.5 *Seja \mathcal{M} uma variedade riemanniana. Então, para qualquer ponto $p \in \mathcal{M}$ e $v \in T_p\mathcal{M}$, e para todo $t_0 \in \mathbb{R}$ contendo t_0 , existe uma única geodésica $\gamma : I \rightarrow \mathcal{M}$ em um intervalo $I \subset \mathbb{R}$ tal que $\gamma(t_0) = p$ e $\gamma'(t_0) = v$.*

6.6 Curvaturas

O tensor torção T de uma conexão D é definido por

$$T(X, Y) = D_X Y - D_Y X - [X, Y].$$

Os componentes T_{ij}^k do tensor torção, os quais são dados como coeficientes da expressão

$$T\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right) = \sum_k T_{ij}^k \frac{\partial}{\partial x_k},$$

satisfazem a relação

$$T_{ij}^k = \Gamma_{ij}^k - \Gamma_{ji}^k.$$

Portanto, uma conexão D é simétrica se, e somente se, o tensor torção é identicamente nulo.

O tensor curvatura R de uma conexão D é definida como

$$R(X, Y)Z = D_X D_Y Z - D_Y D_X Z - D_{[X, Y]}Z,$$

cujos componentes R_{ijk}^l são dados por

$$R\left(\frac{\partial}{\partial x_k}, \frac{\partial}{\partial x_l}\right) \frac{\partial}{\partial x_i} = \sum_j R_{ijk}^l \frac{\partial}{\partial x_j}.$$

Os componentes R_{ijk}^l são expressados em termos dos símbolos de Christoffel Γ_{ij}^k por

$$R_{ijk}^l = \frac{\partial \Gamma_{lj}^i}{\partial x_k} - \frac{\partial \Gamma_{kj}^i}{\partial x_l} + \sum_m (\Gamma_{lj}^m \Gamma_{km}^i - \Gamma_{kj}^m \Gamma_{lm}^i).$$