Universidade de Brasília

A Métrica de Fisher-Rao: Abordagem Geométrica em Probabilidade e Estatística

Saulo Henrique Furtado Leite

Orientador: Dr. Ary Vasconcelos Medino

Departamento de Matemática
Universidade de Brasília

Dissertação apresentada como requisito parcial para obtenção do grau de
Mestre em Matemática

Brasília, 06 de Outubro de 2023

Em memória ao meu tio Carlos Roberto Pimenta.

Agradecimentos

Primeiramente agradeço a Deus por tudo, principalmente pela saúde e por colocar pessoas incríveis na minha vida durante o mestrado, colegas, professores e porteiros do Mat-UnB.

Aos meus pais, Clemente Leite Vieira Júnior e Rosana Patrícia Furtado Leite e minha irmã Sara Furtado Leite, obrigado pelo amor e carinho. Agradeço também a Vitor Hugo por ter me recebido pela primeira vez que cheguei a Brasília e a Maria Fernanda pela paciência, amor e parceria durante essa caminhada.

Ao meu tio, Carlos Roberto Pimenta, que sempre me incentivou e me apoiou nos estudos durante toda nossa convivência no Xerox e minha tia Maria Inês Vieira pelo companheirismo. Aos meus avós pelo carinho. A minha prima Danielle Cristina leite por ter me ajudado com aulas de reforço de matemática no meu 9º ano do ensino fundamental, a qual, a partir daí, comecei a gostar e desfrutar dessa bela área. A toda minha família, primos e primas kelezeiros(a), tios e tias que se forem citar nomes, gera outra dissertação.

Aos amigos, Jadde Thaine e Márcio Henrique pela convivência e ajuda durante a graduação e mestrado. A Manoel, Henrylla, Daniel, Millena, Jônatas e Talita pela companhia na sala de mestrado. A Genilson Soares e a Katianny Freitas por me ajudarem com dúvidas em geometria Riemanniana. Aos meus amigos de graduação Wemenson que sempre esteve comigo em tudo durante o curso e João Marcos que conheci no final do curso.

A todos os professores e colegas da Unimontes pelo incentivo e por todo carinho, sou grato a todos vocês.

Aos professores do departamento de Matemática da UnB, em especial ao que compõem o quadro de probabilidade, em destaque o meu orientador Ary Vasconcelos Medino o qual tenho bastante admiração e carinho com as belíssimas reuniões que enriqueceram bastante o presente trabalho, agradeço muito pela paciência e muito obrigado pelos ensinamentos. A Chang Chung Yu Dorea por aceitar inicialmente a orientação e afins por questões burocráticas. Também agradeço aos professores Cátia Regina, Leandro Martins, Eduardo Antônio, Paulo Henrique e Daniele Baratela pelas contribuições em seminários.

Agradeço aos membros da banca da defesa, professora Daniele Baratela, professor Roberto Imbuzeiro e o professor Tarcísio Silva por aceitarem o convite de avaliar a minha dissertação e pelas sugestões, correções, as quais enriqueceram meu trabalho. Ao professor

Roberto Imbuzeiro que tive a honra de conhecê-lo inicialmente no congresso CLAPEM 2023 que ocorreu na USP, onde me ajudou por meio de uma fórmula a mostrar o item v) de regularidade do modelo estatístico no caso normal multivariado, essa ajuda foi fundamental para concluir o raciocínio da validade do item.

Por fim, agradeço a FAPDF e CNPq pelo apoio financeiro durante a elaboração deste trabalho.

Resumo

Nesta dissertação, veremos como a matriz de informação de Fisher dá origem a uma métrica riemanniana em modelos estatísticos paramétricos regulares e como daí se obtém o conceito de variedade estatística riemanniana. Veremos que essa métrica fornece uma medida de dissimilaridade entre distribuições de probabilidade, conhecida como distância de Fisher-Rao. Mostraremos que a família paramétrica das densidades gaussianas multivariadas é uma variedade estatística riemanniana. Apresentaremos uma relação entre a distância de Fisher-Rao e a divergência Kullback-Leibler. Por fim, ilustraremos através de exemplos como ferramentas da Geometria Riemanniana podem ser usadas em questões relacionadas à Inferência Estatística.

Palavras-chave: Matriz de Informação de Fisher, Métrica Riemanniana, Variedade Estatística Riemanniana, Distância de Fisher-Rao, Divergência Kullback-Leibler, Inferência Estatística.

Abstract

In this dissertation, we will see how the Fisher information matrix gives rise to a Riemannian metric in regular parametric statistical models and how the concept of Riemannian statistical manifold is derived from this. We will see that this metric provides a measure of dissimilarity between probability distributions, known as the Fisher-Rao distance. We will show that the parametric family of multivariate Gaussian densities is a Riemannian statistical manifold. We will present a relationship between the Fisher-Rao distance and the Kullback-Leibler divergence. Finally, we will illustrate through examples how tools from Riemannian Geometry can be used in questions related to Statistical Inference.
Keywords: Fisher Information Matrix, Riemannian Metric, Riemannian Statistical Manifold, Fisher-Rao Distance, Kullback-Leibler Divergence, Statistical Inference.

Conteúdo

Notações

$\mathbb{F}$	Corpo $\mathbb{R}$ dos reais ou $\mathbb{C}$ dos complexos (p. 6).
$\mathbb{F}^n$	Espaço vetorial dos vetores de dimensão n com entradas em $\mathbb{F}$ (p. 6).
$M_n(\mathbb{F})$	Espaço vetorial das matrizes quadradas de ordem n com entradas em $\mathbb{F}$ (p. 6).
A^T	Transposta da matriz A (p. 6).
$x \bullet y$	Produto interno canônico entre x e y em $\mathbb{F}^n$ (p. 6).
A^*	Transposta conjugada da matriz A (p. 6).
$[A]_j^c$	Coluna j da matriz A (p. 6).
$[A]_i^\ell$	Linha i da matriz A (p. 6).
V	Matriz de covariância (p. 10).
Λ	Inversa da matriz de Covariância (p. 10).
S	Modelo estatístico (p. 11).
Θ	Espaço paramétrico (p. 11).
$S_n(\mathbb{R})$	O subespaço das matrizes simétricas de $M_n(\mathbb{R})$ (p. 48).
$GL_n(\mathbb{R})$	O grupo das matrizes não-singulares sobre $M_n(\mathbb{R})$ (P. 31).
$P_n(\mathbb{R})$	O subconjunto das matrizes simétricas positivas definidas em $GL_n(\mathbb{R})$ (p. 31).
$Exp(\theta)$	Densidade da exponencial com parâmetro $\theta > 0$ (p. 12).
$\mathcal{N}(\mu, \sigma^2)$	Densidade da normal univariada (p. 12).
$\mathcal{N}(\mu, \mathrm{V})$	Densidade da normal multivariada (p. 10).
$G(\theta)$	Matriz de informação de Fisher (p. 42).
$\|\cdot\|_G$	Norma gerada pela matriz de informação de Fisher (p. 50).
$d_F(\cdot,\cdot)$	Distância de Fisher- Rao (p. 49).
$d(\cdot,\cdot)$	Distância euclidiana (p. 57).
$\mathcal{H}^2$	Plano de Poincaré (p. 26).

Lista de Figuras

Introdução

Nesta dissertação, forneço detalhamento para lacunas identificadas ao longo dos meus estudos de trabalhos relacionados à área de Geometria da Informação. Adicionalmente, chamo a atenção das comunidades matemática e estatística brasileiras para o estudo dessa área. Esse tema vem se consolidando como ramo de pesquisa em Matemática e pode ser visto como um ramo de confluência entre Probabilidade, Estatística, Geometria Diferencial e Riemanniana e Teoria da Informação.

Geometria da Informação estuda relações entre estruturas da Geometria Diferencial e famílias paramétricas de distribuições de probabilidade. O pioneirismo nesse assunto é atribuído a C. R. Rao em 1945 [30] que mostrou como gerar uma métrica riemanniana no espaço paramétrico de uma dada família de distribuições de probabilidade. Para isso, ele usou a matriz de informação de R. A. Fisher 1922 [13] e tal métrica é conhecida atualmente como métrica de Fisher. Considerando como distância entre duas distribuições de probabilidade a distância geodésica entre os respectivos parâmetros induzida pela métrica de Fisher, Rao calculou a distância entre distribuições para vários modelos estatísticos. Tal distância é comumente chamada de distância de Fisher-Rao.

A distância de Fisher–Rao é muito especial para modelos estatísticos de distribuições de probabilidade. Entre outras propriedades, ela é invariante por reparametrizações do espaço amostral e é covariante por reparametrizações do espaço dos parâmetros (ver, por exemplo, O. Calin and C. Udrişte [10] Teoremas 1.6.4 e 1.6.5)

C. Atkinson and A. F. Michell em 1981 [5] e J. Burbea em 1984 [9] calcularam a expressão da distância de Fisher-Rao para algumas famílias de distribuições de probabilidade. Por exemplo, uma forma explicita para a distância de Fisher-Rao no espaço das densidades gaussianas univariadas é obtida através de uma associação com o modelo clássico do plano hiperbólico (ver, por exemplo, C. Atkinson and A. F. Michell 1981 [5], J. Burbea 1984 [9], S. I. Costa, S. A. Santos and J.E. Strapasson 2015 [11] e C. R. Rao 1945 [30]).

A distância de Fisher-Rao é considerada uma medida adequada para a dissimilaridade entre distribuições de probabilidade, questão abordada em muitos problemas e aplicações. Muitos autores vêm estudando esse tópico sob a perspectiva das aplicações em assuntos tais

como inferência estatística, processos estocásticos, teoria da informação, processamento de imagens, aprendizado de máquina, classificação morfológica, análise de dados, entre outras, ver por exemplo S.-i. Amari 2016 [2].

Em J. Pinele, J. E. Strapasson and S. I. Costa 2020 [27], a distancia de Fisher-Rao é estudada no contexto do espaço das densidades das normais multivariadas. Por exemplo, estuda-se uma aplicação em "clustering" para segmentação de imagens, bem como em simplificação de misturas gaussianas usando o algoritmo de agrupamento hierárquico. Para mais detalhes, ver [27] e suas referências.

Em L. T. Skovgaard 1984 [34] é discutido como usar a distância de Fisher-Rao em modelos estatísticos regulares para fins de inferência estatística. Para tal, é usado o conceito de estimador de distância mínima bem como estatística de teste correspondente. Exemplos são usados para ilustrar a abordagem.

Alguns autores têm desenvolvido um tratamento unificado à teoria da Geometria da Informação, organizando e introduzindo novos conceitos, ferramentas e técnicas relativas a modelos estatísticos. Destacam-se os seguintes trabalhos: S.-i. Amari em 1985 [1], S.-i. Amari and H. Nagaoka em 2000 [3], O. Calin and C. Udrişte em 2014 [10] e S.-i. Amari em 2016 [2].

Em W. Santiago 2017 [31], é discutido o conceito de variedade estatística (ver Definição 1.5). Entre outros pontos, destaca-se o Teorema 2.1 em [31], onde se garante que modelos estatísticos paramétricos regulares são variedades diferenciáveis de Hausdorff e com base enumerável, Em tal contexto, por variedade diferenciável, entende-se de classe C^{∞}.

Para a família paramétrica das gaussianas multivariadas, não se conhece uma fórmula analítica para a distância de Fisher-Rao entre duas distribuições no caso geral. Alguns trabalhos consideram casos de subvariedades dessa família e calculam a distância entre duas distribuições na mesma. Por exemplo, em C. Atkinson and A. F. Michell 1981 [5], J. Burbea 1984 [9] e M. Moakher and M. Zerai 2011 [25] é apresentada uma expressão para a distância de Fisher-Rao restrita à subvariedade S_{μ} das densidades gaussianas multivariadas com vetor de média constante. Para a subvariedade S_V das gaussianas com matriz de covariância constante, uma tal fórmula é apresentada em C. Atkinson and A. F Michell 1981 [5]. Esse problema é abordado em J. Pinele, J. E. Strapasson e S. I. Costa 2020 [27], no caso da subvariedade S_D das gaussianas com matriz de covariância diagonal, bem como da subvariedade $S_{D\mu}$ das gaussianas com matriz de covariância V diagonal e vetor de médias μ um autovetor de V.

Alguns objetivos desse trabalho são: Provar que a família paramétrica das densidades gaussianas multivariadas é identificável; mostrar que o modelo gaussiano multivariado é uma variedade estatística; apresentar o cálculo da matriz de informação de Fisher e da métrica

de Fisher para esse modelo; estudar modelos estatísticos regulares como espaços métricos munidos da distância de Fisher-Rao; por fim, chamar a atenção para a teoria sobre inferência geométrica.

O presente trabalho está dividido em quatro capítulos. O Capítulo 1 está destinado a apresentar alguns resultados preliminares que serão uteis ao longo de todo o trabalho. Especificamente, iremos fazer uma breve revisão de alguns conceitos básicos de Álgebra Linear, Probabilidade, Estatística Matemática e Geometria Riemanniana. Primeiro, revisaremos algumas caracterizações de matrizes positivas definidas e alguns resultados, em seguida definimos matriz de covariância, densidade gaussiana multivariada, família paramétrica de distribuições de probabilidade e modelos estatísticos regulares com alguns exemplos e finalizamos incluindo as definições de variedade diferenciável, subvariedade, métrica riemanniana, geodésicas e alguns resultados.

O Capítulo 2 é dedicado a enxergar de que maneira modelos estatísticos regulares se apresentam como variedade estatística riemanniana. Mostramos que o modelo estatístico composto por densidades normais multivariadas é uma variedade estatística. Em seguida, definiremos a métrica de Fisher em modelos estatísticos regulares fazendo desdes uma variedade riemanniana, calcularemos a matriz de informação de Fisher na variedade normal multivariada e apresentaremos a métrica de Fisher nesse espaço e em seguida definiremos a distância de Fisher-Rao entre duas distribuições de probabilidade, verificaremos que com essa distância, modelos estatísticos regulares se comportam como espaços métricos e calcularemos a distância entre algumas distribuições de probabilidade já conhecidas e em subvariedades da normal multivariada.

No Capítulo 3, veremos o conceito de divergência de Kullback-Leibler e como ela se relaciona com a métrica de Fisher e a distância de Fisher-Rao.

Por fim, baseado em L. T. Skovgaard 1984 [34] e L. T. Skovgaard 1981 [33], o Capítulo 4 foi destinado a usar a distância de Fisher-Rao no modelo estatístico regular para fins de inferência estatística através de estimadores de distância mínima e estatística de teste correspondente. A teoria é ilustrada por alguns exemplos.

Capítulo 1

Preliminares

O objetivo deste capítulo é fazer uma breve revisão sobre alguns conceitos básicos de Álgebra Linear, Probabilidade, Estatística Matemática e Geometria Riemanniana. Por essa razão, as demonstrações de alguns resultados foram omitidas e apenas indicamos alguma referência onde tais provas podem ser encontradas. Este capítulo é baseado nas referências N. Johnston 2021 [16] e N. Loehr 2014 [23] para Álgebra Linear; A. F. Karr 1993 [18], B. James 2002 [15], R. W. Keener 2010 [20], J. Shao 2003 [32] e P. J. Bickel and K. A. Doksum 2001 [6] para Probabilidade e Estatística Matemática; M. P. do Carmo 2015 [12], J. Jost 2008 [17] e R. Biezuner 2017 [7] para Geometria Riemanniana. O capítulo foi dividido em três seções. Na Seção 1.1, revisaremos algumas caracterizações de matrizes positivas definidas e alguns resultados que usaremos no decorrer do trabalho. Na Seção 1.2, revisaremos alguns conceitos básicos de Probabilidade e Estatística Matemática, incluindo as definições de matriz de covariância, densidade gaussiana multivariada, família paramétrica de distribuições de probabilidade e modelos estatísticos regulares. Na Seção 1.3, são revisados alguns elementos básicos de Geometria Riemanniana, incluindo as definições de variedade diferenciável, subvariedade, métrica riemanniana e geodésicas. Assumem-se já conhecidos outros conceitos de teoria da probabilidade como função mensurável, variável aleatória, esperança, variância, covariância,vetores aleatórias, entre outros. Já para Geometria Riemanniana, assume-se que o leitor possui conhecimento mínimo de Geometria Diferencial.

1.1 Elementos de Álgebra Linear

Nesta seção, revisaremos alguns elementos de Álgebra Linear essenciais para o entendimento de certos tópicos que abordaremos nesta dissertação. Adicionalmente, aproveitamos para fixar parte da terminologia e notação a serem usadas ao longo do trabalho. Para tais, adotamos como fontes as referências N. Johnston 2021 [16] e N. Loehr 2014 [23].

No que se segue, $\mathbb{F}$ representa o corpo dos números reais $\mathbb{R}$ ou dos complexos $\mathbb{C}$. Denotamos por $\mathbb{F}^n$ o espaço dos vetores de dimensão n com entradas em $\mathbb{F}$, e por $M_n(\mathbb{F})$ o espaço das matrizes quadradas de ordem n com entradas em $\mathbb{F}$. De acordo com a conveniência, cada vetor $x \in \mathbb{F}^n$ poderá ser representado por $x = (x_1, \cdots, x_n)$ ou $x = [x_1 \cdots x_n]^T$, isto é, na forma de uma ênupla ou de uma matriz coluna $n \times 1$. Para qualquer matriz A, A^T indica a sua transposta.

Se $x = (x_1, \ldots, x_n)$ e $y = (y_1, \ldots, y_n)$ pertencem a $\mathbb{F}^n$, o produto interno canônico entre x e y é definido como

$$\langle x, y \rangle = \sum_{i=1}^{n} \overline{x}_i y_i,$$

em que $\overline{x}_i$ indica o conjugado do complexo x_i. Tal produto interno entre x e y será denotado por $\langle x, y \rangle = x \bullet y$. Se A é uma matriz qualquer com entradas em $\mathbb{F}$, denotaremos por A^* a transposta conjugada de A. Assim, identificando-se $M_1(\mathbb{F})$ com $\mathbb{F}$, o produto interno canônico pode então ser definido usando-se a operação de multiplicação de matrizes

$$x \bullet y = x^* y = [\overline{x}_1 \cdots \overline{x}_n] \, [y_1 \cdots y_n]^T .$$

Se $A \in M_n(\mathbb{F})$, denotaremos por $[A]_j^c$ a sua coluna j e por $[A]_i^\ell$ a sua linha i. Lembremos que A é inversível se, e somente se, $\{[A]_1^c, \cdots, [A]_n^c\}$ é uma base de $\mathbb{F}^n$. A seguinte situação será útil para o Teorema 1.1: Uma matriz $U \in M_n(\mathbb{F})$ é chamada de *unitária* se suas colunas formam uma base ortonormal de $\mathbb{F}^n$ em relação ao produto interno canônico (N. Johnston 2021 [16, p. 96]).

1.1.1 Matriz Positiva Definida

Uma matriz $A \in M_n(\mathbb{F})$ é dita Hermitiana ou autoadjunta se for igual à sua transposta conjugada, isto é, se $A = A^*$. No caso $\mathbb{F} = \mathbb{R}$, a matriz A é Hermitiana se, e somente se, é simétrica, ou seja $A = A^T$.

Por definição, uma matriz $A \in M_n(\mathbb{F})$ é dita positiva semidefinida se ela é Hermitiana e vale $x^* A x \geq 0$ para todo $x \in \mathbb{F}^n$. Se $x^* A x > 0$ para todo $x \in \mathbb{F}^n$ não nulo, diremos que A é positiva definida. O teorema a seguir, caracteriza matrizes positivas definidas de maneiras equivalentes, algumas das quais são mais esclarecedoras e mais fáceis de trabalhar.

Teorema 1.1 (Theorem 2.2.2, [16], p. 191)**.** *Suponha que $A \in M_n(\mathbb{F})$ é Hermitiana. As seguintes afirmações são equivalentes:*

a) A é positiva definida;

b) Todos os autovalores de A são estritamente positivos;

c) Existe uma matriz inversível $B \in M_n(\mathbb{F})$ tal que $A = B^*B$;

d) Existem uma matriz diagonal $D \in M_n(\mathbb{F})$ com entradas diagonais estritamente positivas e uma matriz unitária $U \in M_n(\mathbb{F})$ tal que $A = UDU^*$.

O Corolário 1.1 a seguir, será usado para provar a identificabilidade da parametrização da família das densidades normais multivariadas na Proposição 2.1. Especificamente, precisaremos desse corolário para garantir que a inversa da matriz de covariância da densidade normal multivariada é positiva definida.

Corolário 1.1. *Seja $A \in M_n(\mathbb{F})$ uma matriz positiva definida. Então A é inversível e sua inversa é positiva definida.*

Demonstração. Pelo item c) do Teorema 1.1, existe uma matriz inversível $B \in M_n(\mathbb{F})$ tal que $A = B^*B$. Logo, A é inversível, pois é produto de matrizes inversíveis. Além disso, vale

$$A^{-1} = B^{-1}(B^*)^{-1} = B^{-1}(B^{-1})^* = ((B^{-1})^*)^*(B^{-1})^* = C^*C,$$

sendo $C = (B^{-1})^* \in M_n(\mathbb{F})$ é inversível. Portanto, pelo Teorema 1.1, A^{-1} é positiva definida. Note que para qualquer matriz $A \in M_n(\mathbb{F})$ inversível, A^* é inversível e vale $(A^*)^{-1} = (A^{-1})^*$ (ver N. Loehr 2014 [23, p. 170]). □

O próximo resultado caracteriza produtos internos em $\mathbb{F}^n$ através de matrizes positivas definidas.

Teorema 1.2 (Theorem 2.2.5, [16], p. 193)**.** *Uma função $\langle \cdot, \cdot \rangle : \mathbb{F}^n \times \mathbb{F}^n \to \mathbb{F}$ é um produto interno se, e somente se, existe uma matriz positiva definida $A \in M_n(\mathbb{F})$ tal que*

$$\langle x, y \rangle = x^*Ay, \quad \forall x, y \in \mathbb{F}^n. \tag{1.1}$$

Observação 1.1. Fixadas duas bases para $\mathbb{F}^n$, a matriz A dada pelo Teorema 1.2 é única com respeito a essas bases. Assumiremos neste trabalho a base canônica de $\mathbb{F}^n$ fixa e o produto interno dado por (1.1) será chamado de produto interno gerado pela matriz positiva definida A e denotado por $\langle x, Ay \rangle$. Veja que para $A = I$ em (1.1), obtém-se o produto interno canônico, em que I é a matriz identidade.

Os resultados a seguir servirão de apoio para este trabalho. O Lema 1.1 será usado para verificar a validade da troca da ordem de integração e derivação em relação a cada parâmetro da densidade normal multivariada que se encontra na Proposição 2.4, para provar a independência linear dado no Lema 2.5 e também usaremos na prova do Teorema 2.4 que se encontra no Apêndice A.

Lema 1.1. *Dados $n \geq 1$ e $A \in M_n(\mathbb{F})$, então temos*

$$(Av) \bullet w = v \bullet (A^* w), \quad \forall w, v \in \mathbb{F}^n.$$

Demonstração. Segue diretamente, visto que

$$(Av) \bullet w = (Av)^* w = (v^* A^*) w = v^* (A^* w) = v \bullet (A^* w), \quad \forall w, v \in \mathbb{F}^n.$$

□

O resultado a seguir será utilizado na Subseção 2.3.3 para encontrarmos uma fórmula fechada para a distância de Fisher-Rao entre duas densidades na subvariedade S_V composta por densidades normais multivariadas com matriz de covariância fixada.

Lema 1.2 (Cholesky Factorization, [23], p. 234)**.** *Para toda matriz positiva semidefinida $A \in M_n(\mathbb{F})$, existe uma matriz triangular inferior P de ordem n com entradas diagonais não negativas tais que $A = PP^*$. Se A é positiva definida, então P é única.*

1.2 Elementos de Probabilidade e Estatística Matemática

Nesta seção, abordaremos alguns elementos, conceitos e resultados da Estatística Matemática necessário para o entendimento e desenvolvimento deste trabalho. Definiremos matriz de covariância, densidade de probabilidade normal multivariada, família paramétrica e modelo estatístico paramétrico regular e forneceremos alguns exemplos.Também estabeleceremos algumas notações para o desenvolvimento deste trabalho.

Para mais detalhes e aprofundamento sobre os elementos abordados nesta seção, indicamos as referências A. F. Karr 1993 [18], B. James 2002 [15], R. W. Keener 2010 [20], J. Shao 2003 [32] e P. J. Bickel and K. A. Doksum 2001 [6].

1.2.1 Matriz de Covariância

Denotaremos por $\mathbf{X} = (X_1, \cdots, X_n) = [X_1 \cdots X_n]^T$ um vetor aleatório cujas entradas são variáveis com segundo momento finito. Representaremos por

$$\mu = \mathbb{E}(\mathbf{X}) = (\mu_1, \cdots, \mu_n) = [\mu_1 \cdots \mu_n]^T$$

o vetor de médias de $\mathbf{X}$. Para cada índices i e j, designamos por

$$\sigma_{ij} = \mathrm{Cov}\left(X_i, X_j\right) = \mathbb{E}\left[\left(X_i - \mu_i\right)\left(X_j - \mu_j\right)\right] = \mathbb{E}\left(X_i X_j\right) - \mu_i \mu_j$$

a covariância entre X_i e X_j, em que $\mu_k = \mathbb{E}(X_k)$ indica a esperança de X_k. Recordemos que para $i = j$

$$\sigma_{ii} = \text{Cov}(X_i, X_i) = Var(X_i) = \sigma_i^2$$

indica a variância de X_i e $\sigma_i = \sqrt{Var(X_i)}$ é o desvio padrão de X_i. Para variáveis aleatórias com variância finita e positiva, o coeficiente de correlação entre X_i e X_j é denotado e definido por

$$\rho_{ij} = \text{Corr}\left(X_i, X_j\right) = \frac{\text{Cov}\left(X_i, X_j\right)}{\sqrt{Var(X_i)}\sqrt{Var(X_j)}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}.$$

As variáveis X_i e X_j são ditas não correlacionadas se $\text{Corr}\left(X_i, X_j\right) = 0$. Em particular, sob as condições acima, se X_i e X_j são independentes, então elas são não correlacionadas. Observe que em qualquer caso, vale a relação

$$\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j.$$

A seguir, veremos a definição de matriz de Covariância que é usada como parâmetro da densidade da normal multivariada.

Definição 1.1 (Matriz de Covariância, [20])**.** *Dado um vetor aleatório* $\boldsymbol{X} = (X_1, ..., X_n)$*, a matriz quadrada de ordem n cujas entradas são as covariâncias* $\text{Cov}\left(X_i, X_j\right)$ *será denotada por* $\text{Cov}(\mathbf{X})$ *e é chamada de* matriz de covariância *do vetor* $\boldsymbol{X}$*. Ou seja*

$$\text{Cov}(\mathbf{X}) = \left[\text{Cov}\left(X_i, X_j\right)\right]_n = \left[\sigma_{ij}\right]_n = \left[\rho_{ij}\sigma_i\sigma_j\right]_n.$$

Se $\mathbf{X} = (X_1, ..., X_n)$ é um vetor aleatório cujas entradas são variáveis não correlacionadas, ou em particular independentes, então

$$\text{Cov}(\mathbf{X}) = \text{diag}\left(\text{Var}(X_1), ..., \text{Var}(X_n)\right),$$

em que $\text{diag}(d_{11}, ..., d_{nn})$ indica a matriz diagonal com entradas $d_{11}, ..., d_{nn}$ na diagonal. É importante ressaltar que a matriz de covariância de um vetor aleatório é simétrica e positiva semidefinida (ver A. F. Karr 1993 [18, p. 126]).

1.2.2 Vetores Gaussianos

Nesta subseção, veremos a definição da densidade de probabilidade normal multivariada, que no próximo capítulo usaremos para mostrar que a família paramétrica composta por essas densidades satisfaz a propriedade de ser regular. Além disso, essa família de densidades é usada em todo este trabalho.

Uma distribuição de probabilidade ou densidade de probabilidade sobre um conjunto $\mathcal{X}$ é uma função $p : \mathcal{X} \to \mathbb{R}$ tal que

(i) $p(x) \geq 0$, para todo $x \in \mathcal{X}$;

(ii) $\int_{\mathcal{X}} p(x)\,dx = 1$, quando $\mathcal{X}$ é Lebesgue mensurável de $\mathbb{R}^n$.

Representaremos por $\mathbf{X} \sim \mathcal{N}(\mu, V)$ o vetor aleatório $\mathbf{X}$ tem densidade normal multivariada com vetor de médias μ e matriz de covariância V, como definimos a seguir.

Definição 1.2 (Densidade Normal Multivariada, [8])**.** *Um vetor aleatório $\boldsymbol{X} \in \mathbb{R}^n$ tem* densidade *normal multivariada com média $\mu = (\mu_1, ..., \mu_n) \in \mathbb{R}^n$ e matriz de covariância* V *simétrica, positiva definida, quando sua densidade de probabilidade é dada por*

$$p(x;\theta) = (2\pi)^{-n/2}[\det(\mathrm{V})]^{-1/2} \exp\left\{-\frac{1}{2}\langle x-\mu, \Lambda(x-\mu)\rangle\right\}, \quad x \in \mathbb{R}^n, \tag{1.2}$$

em que Λ denota a inversa da matriz de covariância, a qual é positiva definida pelo Corolário 1.1.

Usaremos o Lema 1.3 a seguir, também para verificar a validade da troca da ordem de integração e derivação em relação a cada parâmetro da densidade normal multivariada em que se encontra na Proposição 2.4.

Lema 1.3 ([10], p. 49)**.** *Seja $X \sim \mathcal{N}(\mu, \mathrm{V})$ a densidade gaussiana multivariada com média $\mu \in \mathbb{R}^n$ e matriz de covariância $\mathrm{V} = [\sigma_{ij}] \in P_n(\mathbb{R})$ sendo que $\Lambda = [\lambda_{ij}]_n$ é a sua inversa. Então, vale*

$$\frac{\partial\,(\det \mathrm{V})}{\partial \sigma_{ij}} = \frac{\lambda_{ij}}{\det(\Lambda)}.$$

Ou seja,

$$\frac{\partial(\det(\mathrm{V}))}{\partial \sigma_{ij}} = \det(\mathrm{V})\,\mathrm{Tr}\left(\Lambda \frac{\partial \mathrm{V}}{\sigma_{ij}}\right).$$

Para mais detalhes, consulte M. Gianquinta and G. Modica 2009 [14, p. 23].

O vetor aleatório $X \sim \mathcal{N}(\mu, \mathrm{V})$ tem densidade normal bivariada quando $n = 2$ na Definição 1.2. Veja a Figura 1.1, que representa o gráfico da densidade normal bivariada, sendo $\mu_1 = \mu_2 = 0$ e matriz de covariância igual à matriz identidade de ordem 2.

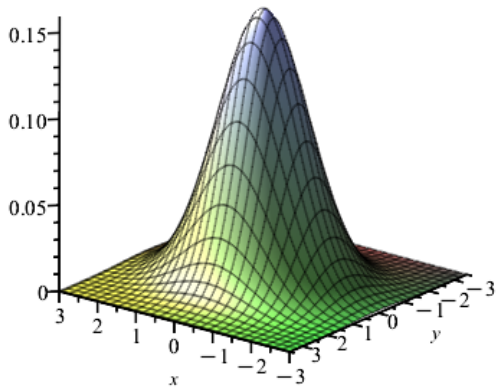

Figura 1.1 Densidade da normal bivariada com $\mu = 0$ e $\mathrm{V} = I_2$.
Produzida com programa Maple PPGMAT.

1.2.3 Modelos Estatísticos

Estamos interessados em saber de que forma família paramétrica de distribuições de probabilidade com a propriedade de ser regular é uma variedade diferenciável, a qual veremos futuramente na Seção 2.1. Para tanto, nesta subseção definiremos o conceito de família paramétrica de distribuições de probabilidade, modelo estatístico e modelo estatístico regular. Neste trabalho, a função log denota logaritmo na base natural.

A seguir, definiremos família paramétrica de distribuições de probabilidade.

Definição 1.3 (Família Paramétrica, [32])**.** *Um conjunto de medidas de probabilidade P_θ em $(\mathcal{X},\mathcal{B})$ indexadas por um parâmetro $\theta \in \Theta$ diz-se que é uma* família paramétrica *se, e somente se, $\Theta \subset \mathbb{R}^n$ para algum inteiro positivo fixo n e cada P_θ é de um tipo conhecido, ou forma conhecida, quando se conhece o θ.*

O conjunto Θ é chamado de espaço de parâmetros e n é a sua dimensão. Relembre que uma família paramétrica $\{P_\theta; \theta \in \Theta\}$ é dita ser identificável se, e somente se, para quaisquer $\theta_1, \theta_2 \in \Theta$ com $\theta_1 \neq \theta_2$ implica $P_{\theta_1} \neq P_{\theta_2}$, isto é, $\varphi : \Theta \to \{P_\theta; \theta \in \Theta\}$ é uma aplicação injetiva.

Damos o nome de modelo estatístico paramétrico às famílias paramétricas identificáveis. Veja a Definição 1.4.

Definição 1.4 (Modelo Estatístico Paramétrico, [32])**.** *Seja S uma família de densidades de probabilidade sobre $\mathcal{X}$. Suponha que cada elemento de S seja parametrizado por n valores reais $\theta = (\theta_1, \theta_2, ..., \theta_n)$ num subconjunto $\Theta \subset \mathbb{R}^n$, isto é,*

$$S = \{p_\theta = p(x;\theta); \theta = (\theta_1, ..., \theta_n) \in \Theta \subset \mathbb{R}^n\},$$

e além disso a aplicação $\varphi : \Theta \to S$ é injetiva, então dizemos que S é um modelo estatístico paramétrico *sobre $\mathcal{X}$ de dimensão n.*

Uma família de medida de probabilidade ou distribuição de probabilidade que não satisfaz a Definição 1.3 é chamada família não paramétrica. Um modelo não paramétrico refere-se a suposição de que a população P está em uma família não paramétrica.

Veja alguns exemplos de Modelos estatísticos paramétricos.

Exemplo 1.1 (Modelo Exponencial)**.** *Seja $X \sim Exp(\theta)$ e considere a família $S = \{p_\theta; \theta \in \Theta\}$ formada por essas densidades, isto é, as densidades de probabilidade são dadas por*

$$p_\theta = p(x,\theta) = \theta e^{-\theta x},$$

com $\theta \in (0,\infty) = \Theta \subset \mathbb{R}$ e $x \in (0,\infty) = \mathcal{X}$. Pode-se ver que $\int_0^\infty p(x;\theta)\,dx = 1$ para todo $\theta \in \Theta$ e que $p(x;\theta) > 0$ para todo $x \in \mathcal{X}$. Agora mostremos que esta família é identificável. Para quaisquer θ_1 e θ_2 em Θ e $x \in \mathcal{X}$, segue

$$\begin{aligned} p(x;\theta_1) = p(x;\theta_2) \quad &\implies \quad \ln p(x;\theta_1) = \ln p(x;\theta_2) \implies \ln\theta_1 - \theta_1 x = \ln\theta_2 - \theta_2 x \\ &\implies \quad \ln\left(\frac{\theta_1}{\theta_2}\right) = (\theta_1 - \theta_2)x \implies \ln\left(\frac{\theta_1}{\theta_2}\right) - (\theta_1 - \theta_2)x = 0. \end{aligned}$$

Observe que pela ultima implicação, temos uma função afim de x que é identicamente nula. Logo, por definição

$$\theta_1 - \theta_2 = 0 \; e \; \ln\left(\frac{\theta_1}{\theta_2}\right) = 0 \implies \theta_1 = \theta_2.$$

Portanto, pela Definição 1.4 $S = \{p_\theta\}$é uma modelo estatístico paramétrico sobre $(0,\infty)$ de dimensão 1.

Exemplo 1.2 (Modelo Normal Univariado)**.** *Seja $X \sim N(\mu,\sigma^2)$, considere a família $S = \{p_\theta; \theta \in \Theta\}$ formada por densidades normais univariadas, isto é, as densidades de probabilidade são definidas por*

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

sendo $\theta = (\mu,\sigma) \in \Theta = \mathbb{R}\times(0,\infty) \subset \mathbb{R}^2$ e $\mathcal{X} = \mathbb{R}$. Pode-se ver que $\int_{-\infty}^{+\infty} p(x;\mu,\sigma)\,dx = 1$ para todo $(\mu,\sigma) \in \mathbb{R}\times(0,\infty)$ e que $p(x;\mu,\sigma) > 0$ para todo $x \in \mathcal{X}$. Esta família é identificável, o que provamos para o caso geral em que S é a família formada por densidades normais multivariadas na Proposição 2.1.

Neste trabalho vamos nos referir ao modelo estatístico paramétrico simplesmente por modelo estatístico e denotaremos S como $S = \{p_\theta; \theta \in \Theta\}$ e por $p_\theta = p$ um ponto de S quando não houver confusão. Fornecemos a seguir, condições de regularidade para um modelo estatístico S.

Definição 1.5 (Modelo Estatístico Regular, [10])**.** *Seja S um modelo estatístico sobre $\mathcal{X}$ de dimensão n. Diremos que S é um* modelo estatístico regular *se satisfaz os itens abaixo:*

i) $\Theta \subset \mathbb{R}^n$ *é aberto;*

ii) Fixado $x \in \mathcal{X}$, as funções $\theta \in \Theta \mapsto p(x,\theta)$ são suaves, ou seja, admitem derivadas parciais em relação a θ em todas as ordens e são contínuas;

iii) Podemos trocar livremente a ordem de integração e derivação, isto é,

$$\int_{\mathcal{X}} \frac{\partial p(x,\theta)}{\partial \theta_i} dx = \frac{\partial}{\partial \theta_i} \int_{\mathcal{X}} p(x,\theta) dx = 0,$$

para topo $p \in S$ e $i = 1,2,...,n$;

iv) O conjunto $Z_+ := \{x \in \mathcal{X} \mid p(x;\theta) > 0\}$ independe de θ, para todo $p \in S$;

v) Para cada $p \in S$ as funções $\partial p(x,\theta)/\partial \theta_i$, com $i = 1,2,...,n$, são linearmente independentes (LI) como funções de x, isto é, para escalares $\alpha_i \in \mathbb{R}$ com $i = 1,...,n$, vale

$$\sum_{i=1}^{n} \alpha_i \partial p(x,\theta)/\partial \theta_i = 0 \quad \forall x \in \mathcal{X} \iff \alpha_i = 0 \quad \forall i = 1,...,n.$$

Segue, da Definição 1.5 que a aplicação $\theta \in \Theta \mapsto p(x,\theta)$ define um sistema de coordenadas em S, tornando S uma variedade globalmente parametrizada. O referencial

$$\partial_i = \frac{\partial}{\partial \theta_i}\bigg|_p$$

denota os campos coordenados desta parametrização. Veremos no Teorema 2.1 do próximo capítulo.

Agora, veremos um resultado que se refere a uma equivalência ao item *v*) da Definição 1.5.

Proposição 1.1 ([10])**.** *A condição de regularidade v) do modelo estatístico S vale se, e somente se, para qualquer $\theta \in \Theta$ o conjunto*

$$\left\{ \frac{\partial}{\partial \theta_1} \ln(p(x;\theta)), ..., \frac{\partial}{\partial \theta_n} \ln(p(x;\theta)) \right\}$$

é um sistema de n funções linearmente independentes como funções de x.

Demonstração. Sabemos que

$$\frac{\partial}{\partial\theta_i}\ln(p(x;\theta)) = \frac{1}{p(x;\theta)}\frac{\partial}{\partial\theta_i}p(x;\theta)$$

Como pela condição *iv*) da Definição 1.5, $p(x;\theta) > 0$ para todo $x \in \mathcal{X}$, temos

$$\begin{aligned}
& \text{o conjunto } \left\{\frac{\partial}{\partial\theta_i}\ln(p(x;\theta))\right\} \text{ é L.I.} \\
\iff \quad & \sum_{i=1}^{n}\alpha_i\partial\ln p(x,\theta)/\partial\theta_i = 0 \quad \forall x\in\mathcal{X} \Longrightarrow \alpha_i = 0 \quad \forall i = 1,...,n. \\
\iff \quad & \sum_{i=1}^{n}\alpha_i p(x,\theta)^{-1}\partial p(x,\theta)/\partial\theta_i = 0 \quad \forall x\in\mathcal{X} \Longrightarrow \alpha_i = 0 \quad \forall i = 1,...,n. \\
\iff \quad & \sum_{i=1}^{n}\alpha_i\partial p(x,\theta)/\partial\theta_i = 0 \quad \forall x\in\mathcal{X} \Longrightarrow \alpha_i = 0 \quad \forall i = 1,...,n. \\
\iff \quad & \text{o conjunto } \left\{\frac{\partial}{\partial\theta_i}p(x;\theta)\right\} \text{ é L.I.}
\end{aligned}$$

□

A seguir, vamos fornecer dois exemplos para a Definição 1.5. Na proposição a seguir, mostraremos que o modelo exponencial é regular.

Proposição 1.2. *Seja S o modelo estatístico formado por densidades exponenciais sobre* $(0,\infty)$ *com parâmetro* $\theta > 0$ *de dimensão* 1*, como no Exemplo 1.1, então S é regular.*

Demonstração. Verificaremos os itens da Definição 1.5:

i) Claramente $\Theta = (0,\infty) \subset \mathbb{R}$ é aberto;

ii) Fixado $x \in \mathbb{R}_+^*$, as funções $\theta \mapsto p(x;\theta) = \theta e^{-\theta x}$ são suaves por se tratar de composição de funções suaves;

iii) Veja que

$$\frac{\partial p(x;\theta)}{\partial\theta} = e^{-\theta x} - \theta x e^{-\theta x}.$$

Assim,

$$\begin{aligned}
\int_0^\infty \frac{\partial p(x;\theta)}{\partial\theta}dx &= \int_0^\infty e^{-\theta x}dx - \int_0^\infty \theta x e^{-\theta x}dx = \int_0^\infty e^{-\theta x}dx - \mathbb{E}(X) = -\frac{1}{\theta}e^{-\theta x}\Big|_0^\infty - \frac{1}{\theta} \\
&= \frac{1}{\theta} - \frac{1}{\theta} = 0.
\end{aligned}$$

iv) Vale $Z_+ = \{x; p(x;\theta) > 0\} = (0,+\infty)$, logo, independe de θ, para todo $p \in S$.

v) Queremos mostrar que o conjunto $\{\partial p(x;\theta)/\partial\theta;$ como função de $x\}$ é L.I.. Usando a Proposição 1.1, basta provar que o conjunto $\{\partial \ln p(x;\theta)/\partial\theta;$ como função de $x\}$ é L.I.. Mas como temos apenas uma função neste conjunto, basta observar que a função não é identicamente nula. Temos

$$\ln p(x;\theta) = \ln\theta - \theta x \implies \frac{\partial \ln p(x;\theta)}{\partial\theta} = \frac{1}{\theta} - x, \tag{1.3}$$

que claramente se anula apenas no ponto $x = 1/\theta$.

Portanto, por i), ii), iii), iv) e v) $S = \{p(x;\theta); \theta \in (0,\infty)\}$ é um modelo estatístico regular. □

Agora, mostraremos que o modelo normal univariado é regular.

Proposição 1.3. *Seja S o modelo estatístico formado por densidades normais univariadas sobre* $\mathbb{R}$*, de dimensão* 2*, (como no Exemplo 1.2), então S é regular.*

Demonstração. Verificaremos os itens da Definição 1.5:

i) Como $\sigma > 0, \Theta = \mathbb{R}\times(0,\infty) \subset \mathbb{R}^2$ é o semiplano superior de $\mathbb{R}^2$, o qual é aberto;

ii) Fixado $x \in \mathcal{X}$, as funções $\theta \in \Theta \mapsto p(x;\mu,\sigma) = \dfrac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\dfrac{(x-\mu)^2}{2\sigma^2}\right\}$ são suaves por se tratar de composições de funções suaves;

iii) Veja que

$$\begin{aligned}
\frac{\partial p(x;\mu,\sigma)}{\partial\mu} &= \frac{(x-\mu)}{\sqrt{2\pi}\sigma^3}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},\\
\frac{\partial p(x;\mu,\sigma)}{\partial\sigma} &= -\frac{1}{\sqrt{2\pi}\sigma^2}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} + \frac{(x-\mu)^2}{\sqrt{2\pi}\sigma^4}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}
\end{aligned}$$

Assim,

$$\int_{\mathbb{R}} \frac{\partial p(x;\mu,\sigma)}{\partial\mu}dx = \frac{1}{\sigma^2}\mathbb{E}(X-\mu) = \frac{1}{\sigma^2}(\mathbb{E}(X)-\mu) = \frac{1}{\sigma^2}(\mu-\mu) = 0$$

e

$$\int_{\mathbb{R}} \frac{\partial p(x;\mu,\sigma)}{\partial\sigma}dx = -\frac{1}{\sigma}\int_{\mathbb{R}} p(x;\mu,\sigma)dx + \frac{1}{\sigma^3}\mathrm{Var}(X) = -\frac{1}{\sigma} + \frac{1}{\sigma} = 0;$$

iv) O conjunto $Z_+ = \{x; p(x;\mu,\sigma) = \left(1/\sqrt{2\pi}\sigma\right)\exp\left\{-\dfrac{(x-\mu)^2}{2\sigma^2}\right\} > 0\} = \mathbb{R}$, logo independe de (μ,σ), para todo $p \in S$;

v) Mostraremos que as funções $\dfrac{\partial}{\partial\mu}\ln p(x;\mu,\sigma)$ e $\dfrac{\partial}{\partial\sigma}\ln p(x;\mu,\sigma)$ são LI. Note que,

$$\ln p(x;\mu,\sigma) = -\ln(\sqrt{2\pi}) - \ln\sigma - \frac{(x-\mu)^2}{2\sigma^2}. \text{ Daí}$$

$$\frac{\partial}{\partial\mu}\ln p(x;\mu,\sigma) = \frac{(x-\mu)}{\sigma^2} \quad \text{e} \quad \frac{\partial}{\partial\sigma}\ln p(x;\mu,\sigma) = (x-\mu)^2\sigma^3 - \frac{1}{\sigma}.$$

Para quaisquer escalares α e β reais,

$$\frac{(x-\mu)}{\sigma^2}\alpha + \left[\frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}\right]\beta = 0 \implies \frac{y}{\sigma^2}\alpha + \left[\frac{y^2}{\sigma^3} - \frac{1}{\sigma}\right]\beta = 0,$$

sendo $y = (x-\mu)$ é uma variável em $\mathbb{R}$, assim temos um polinômio como função de y de grau 2 identicamente nulo. Daí, seus coeficientes são nulos, isto é, $\alpha/\sigma^2 = 0$, $\beta/\sigma^3 = 0$ e $-\beta/\sigma = 0$, o que implica $\alpha = \beta = 0$.

Portanto, por i), ii), iii), iv) e v) $S = \{p(x,\mu,\sigma);(\mu,\sigma) \in \mathbb{R}\times(0,\infty)\}$ composto por densidades normais univariadas é um modelo estatístico regular. □

Vale observar que apesar da definição de modelo estatístico regular ter bastantes restrições, os modelos estatísticos tradicionais na maioria das vezes são modelos regulares. No entanto, segue abaixo um exemplo de modelo estatístico que não é regular, a saber, o modelo de densidades uniformes.

Exemplo 1.3. *Sendo* $\mathcal{X} = \mathbb{R}, \theta = (a,b), \Theta = \left\{(a,b) \in \mathbb{R}^2 \mid b > a\right\}$ *e* $p_\theta = p(x;a,b)$ *dado por*

$$p_\theta = \begin{cases} \dfrac{1}{b-a}, & \text{se } a < x < b \\ 0, & \text{caso contrário.} \end{cases} \tag{1.4}$$

Considere que $p(x;a_1,b_1) = p(x;a_2,b_2)$ *para todo* $x \in \mathcal{X}$. *Suponha por contradição que* $(a_1,b_1) \neq (a_2,b_2)$, *sem perda de generalidade podemos supor que* $a_1 < a_2$. *Seja* x_0 *entre* a_1 *e* a_2 *tal que* $x_0 \in (a_1,b_1)$. *Aplicando* x_0 *na última igualdade tem-se que*

$$p(x_0;a_1,b_1) = \frac{1}{b_1-a_1} = 0 \implies \frac{1}{b_1-a_1} = 0.$$

Absurdo! Logo $(a_1,b_1) = (a_2,b_2)$, *a função* $\varphi : \Theta \to S$ *é injetiva e* $S = \{p_\theta\}$, *sendo que* p_θ *é definido na equação* (1.4) *é um modelo estatístico sobre* $\mathbb{R}$, *de dimensão 2. Mas* S *não é modelo regular pois as derivadas parcias de* $p(x;a,b)$ *com relação à a e b não são LI, como*

veremos a seguir. Por cálculos simples, temos

$$\frac{\partial p(x;a,b)}{\partial a} = \frac{1}{(b-a)^2} \quad e \quad \frac{\partial p(x;a,b)}{\partial b} = -\frac{1}{(b-a)^2}.$$

Logo, o conjunto formado pelas derivadas parciais de $p(x;a,b)$ em relação à a e b é linearmente dependente pois podemos escrever

$$\frac{\partial p(x;a,b)}{\partial a} = -\frac{\partial p(x;a,b)}{\partial b}$$

e a condição v) da Definição 1.5 não é satisfeita.

Neste trabalho, estamos interessados em fazer cálculos como comprimento de curva e distância entre densidades de probabilidade em um modelo estatístico regular. Para tanto, na seção seguinte revisamos alguns conceitos básicos de Geometria Riemanniana.

1.3 Elementos de Geometria Riemanniana

Nesta seção, revisaremos brevemente algumas definições tais como, variedade diferenciável, espaço tangente, subvariedade, métricas riemannianas e geodésicas que serão empregadas ao longo deste trabalho no contexto da Probabilidade e Estatística, isto é, para enxergamos de que forma um modelo estatístico regular é uma variedade para podermos estrar nas condições de fazer cálculos de comprimento de curva e distância entre densidades. Não demostraremos todos os resultados aqui apresentados, para tanto, sugerimos ao leitor algumas referências, como M. P. do Carmo 2015 [12], J. Jost 2008 [17] e R. Biezuner 2017 [7].

1.3.1 Variedades Diferenciáveis

A noção de variedade diferenciável estende os métodos do cálculo diferencial à espaços mais gerais que o espaço $\mathbb{R}^n$. Aqui, usaremos o termo diferenciável para as aplicação de classe C^∞. A seguir, veremos a definição de variedade diferenciável e na Seção 2.1, mostraremos que modelos estatísticos regulares satisfazem esta definição.

Definição 1.6 (Variedade Diferenciável, [12])**.** *Uma* variedade diferenciável *de dimensão n ou simplesmente, uma variedade, é um conjunto $M \neq \emptyset$ quando existe uma família de aplicações bijetivas $\{U_\lambda, \varphi_\lambda\}_{\lambda\in\mathcal{A}}$, de conjuntos abertos $U_\lambda \subset \mathbb{R}^n$ em M, $\varphi_\lambda : U_\lambda \to M$, satisfazendo as seguintes condições:*

i) $\bigcup_{\lambda\in\mathcal{A}} \varphi_\lambda(U_\lambda) = M;$

ii) Para cada par de índices $\lambda, \alpha \in \mathcal{A}$ *com* $\varphi_\lambda(U_\lambda) \bigcap \varphi_\alpha(U_\alpha) = W_{\lambda\alpha} \neq \emptyset$*, temos que* $\varphi_\lambda^{-1}(W_{\lambda\alpha})$ *e* $\varphi_\alpha^{-1}(W_{\lambda\alpha})$ *são conjuntos abertos em* $\mathbb{R}^n$ *e as funções transição (Figura 1.2)*

$$\varphi_\alpha^{-1} \circ \varphi_\lambda : \varphi_\lambda^{-1}(W_{\lambda\alpha}) \to \varphi_\alpha^{-1}(W_{\lambda\alpha}),$$
$$\varphi_\lambda^{-1} \circ \varphi_\alpha : \varphi_\alpha^{-1}(W_{\lambda\alpha}) \to \varphi_\lambda^{-1}(W_{\lambda\alpha}),$$

são aplicações diferenciáveis, onde, no nosso contexto, entende-se aplicações de classe C^∞*;*

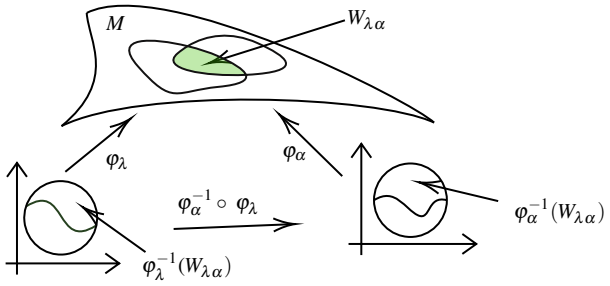

Figura 1.2 Variedade diferenciável.
Produzida online com Mathcha.

iii) A família $\{(U_\lambda, \varphi_\lambda)\}_{\lambda\in\mathcal{A}}$ *é máxima relativamente às condições i) e ii).*

Cada aplicação $\varphi_\lambda, \lambda \in \mathcal{A}$, é chamada uma carta ou uma parametrização ou um sistema de coordenadas locais para uma vizinhança de M, denotada $(\varphi_\lambda, U_\lambda)$, e $\varphi_\lambda(U_\lambda)$ é chamada uma vizinhança coordenada. Uma família $\{(U_\lambda, \varphi_\lambda)\}_{\lambda\in\mathcal{A}}$ satisfazendo *i)* e *ii)* é dita uma estrutura diferenciável em M.

A condição *iii)* apresenta-se por razões puramente técnicas. Em verdade, dada uma estrutura diferenciável em M, podemos facilmente completá-la em uma máxima, agregando-a todas as parametrizações de modo que a condição *ii)*, para esta nova família, continue satisfeita. Portanto, com um certo abuso de linguagem, podemos dizer que uma variedade diferenciável de dimensão n é um conjunto munido de uma estrutura diferenciável. Em geral, a extensão à estrutura máxima será admitida sem maiores comentários.

Observação 1.2. Uma estrutura diferenciável em um conjunto M induz de uma maneira natural uma topologia em M. Para tanto, basta definir que $A \subset M$ é um aberto de M se $\varphi_\lambda^{-1}(A \cap \varphi_\lambda(U_\lambda))$ é um aberto de $\mathbb{R}^n$ para todo $\lambda \in \mathcal{A}$. É imediato verificar que M e o vazio são abertos, que a união de abertos é um aberto e que a intersecção finita de abertos é um aberto. Observe que esta topologia é definida de tal modo que os conjuntos $\varphi_\lambda(U_\lambda)$ são abertos e as aplicações φ_λ são contínuas.

O espaço euclidiano $\mathbb{R}^n$, com a estrutura diferenciável dada pela identidade e as superfícies regulares são exemplos triviais de variedade diferenciável. Neste trabalho, apresentaremos principalmente variedades cujos pontos são densidades de probabilidade. Uma simplificação que se mostrará bastante útil é que, para esses espaços consideraremos apenas parametrizações globais.

Convém explorar um pouco mais as consequências da Definição 1.6, a qual não será nosso foco e não abordaremos em todos os detalhes. Primeiro, estenderemos a noção de diferenciabilidade às aplicações entre variedades.

Definição 1.7 (Aplicação diferenciável, [12])**.** *Sejam M_1 e M_2 variedades diferenciáveis de dimensões n e m, respectivamente. Uma aplicação $\varphi : M_1 \to M_2$ é diferenciável em $p \in M_1$ quando dada uma parametrização $\phi : V \subset \mathbb{R}^m \to M_2$ em torno de $f(p)$, existe uma parametrização $\psi : U \subset \mathbb{R}^n \to M_1$ em torno de p tal que $\varphi(\psi(U)) \subset \phi(V)$ e a aplicação*

$$\phi^{-1} \circ \varphi \circ \psi : U \subset \mathbb{R}^n \to \mathbb{R}^m$$

é diferenciável em $\psi^{-1}(p)$. A aplicação φ é diferenciável em um aberto de M_1 se é diferenciável em todos os pontos deste aberto. Veja Figura 1.3.

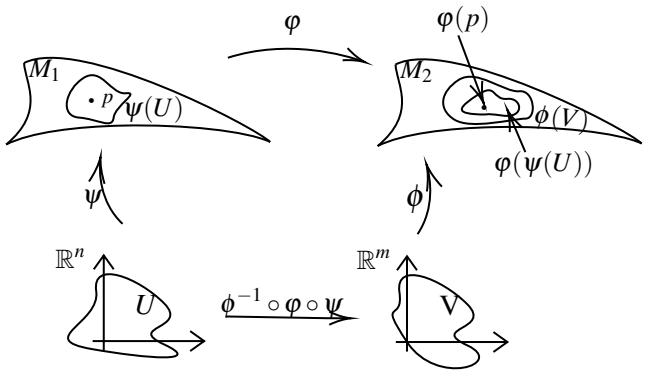

Figura 1.3 Aplicação Diferenciável.
Produzida online com Mathcha.

Observamos que se $\phi^{-1} \circ \varphi \circ \psi$ é diferenciável para as cartas $(\psi, U), (\phi, V)$, então para quaisquer cartas$(\widetilde{\psi}, \widetilde{U})$ de uma vizinhança de p e $(\widetilde{\phi}, \widetilde{V})$ de uma vizinhança de $\varphi(p)$ tais que $\varphi(\widetilde{\psi}(\widetilde{U})) \subset \widetilde{\phi}(\widetilde{V})$ temos que $\widetilde{\phi}^{-1} \circ \varphi \circ \widetilde{\psi}$ é diferenciável, pois

$$\widetilde{\phi}^{-1} \circ \varphi \circ \widetilde{\psi} = \left(\widetilde{\phi}^{-1} \circ \phi\right) \circ \phi^{-1} \circ \varphi \circ \psi \circ \left(\psi^{-1} \circ \widetilde{\psi}\right),$$

e $\widetilde{\phi}^{-1} \circ \phi, \psi^{-1} \circ \widetilde{\psi}$ são difeomorfismos. A aplicação $\phi^{-1} \circ \varphi \circ \psi$ é uma representação de φ em coordenadas. Ressaltamos novamente que, neste trabalho, entenderemos aplicação diferenciável por uma aplicação suave.

Definição 1.8 ([7]). *Dizemos que uma aplicação diferenciável $\varphi : M \longrightarrow N$ é um difeomorfismo se φ é diferenciável e φ^{-1} também é diferenciável.*

Se existir um difeomorfismo entre duas variedades diferenciáveis M e N, dizemos que elas são difeomorfas. Se duas variedades diferenciáveis são difeomorfas, em particular possuem a mesma dimensão.

Uma das aplicações diferenciáveis mais importantes entre variedades são as curvas diferenciáveis

Definição 1.9 ([7]). *Uma curva diferenciável em uma variedade diferenciável M é uma aplicação diferenciável $\alpha : I \longrightarrow M$ em que $I \subset \mathbb{R}$ é um intervalo.*

A seguir, definiremos espaço tangente a uma variedade em um ponto. Mas antes é aconselhado rever a definição de vetor tangente, a qual não abordaremos aqui. Para tanto, consulte as referências de Geometria Riemanniana mencionadas anteriormente.

Definição 1.10 (Espaço tangente, [12]). O *conjunto de todos os vetores tangentes a M em um ponto p é chamado de* espaço tangente *a M em p e é denotado por T_pM.*

Observamos, que o vetor tangente à uma curva α em p depende apenas das derivadas de α em um sistema de coordenadas. O conjunto T_pM, com as operações usuais de funções, forma um espaço vetorial de dimensão n e a escolha de uma parametrização $\varphi : U \to M$ determina uma base associada

$$\left\{ \left.\frac{\partial}{\partial x_1}\right|_p, \left.\frac{\partial}{\partial x_2}\right|_p, ..., \left.\frac{\partial}{\partial x_n}\right|_p \right\} \text{ em } T_pM.$$

Com a noção de espaço tangente podemos estender às variedades diferenciáveis a noção de diferencial de uma aplicação diferenciável.

Proposição 1.4 ([12]). *Sejam M_1 e M_2 variedades diferenciáveis de dimensão n e m respectivamente e seja $\varphi : M_1 \to M_2$ uma aplicação diferenciável. Para cada $p \in M_1$ e cada $v \in T_pM_1$, escolha uma curva diferenciável $\alpha : (-\varepsilon, \varepsilon) \to M_1$ com $\alpha(0) = p, \alpha'(0) = v$. Faça $\beta = \varphi \circ \alpha$. A aplicação $d\varphi_p : T_pM_1 \to T_{\varphi(p)}M_2$ dada por $d\varphi_p(v) = \beta'(0)$ é uma aplicação linear que não depende da escolha de α.*

Demonstração. A demonstração pode ser vista em M. P. do Carmo 2015 [12, p. 9]. □

A aplicação linear $d\varphi_p$ dada pela Proposição 1.4 é chamada diferencial de φ em p. A seguir introduzimos o conceito de subvariedade, a qual o interesse é verificar se alguns modelos estatísticos regulares são subvariedades estatísticas. Primeiro definiremos imersão e mergulho.

Definição 1.11 (Imersão e Mergulho, [12])**.** *Sejam M e N variedades diferenciáveis de dimensões m e n, respectivamente. Uma aplicação diferenciável $\varphi : M \to N$ é uma* imersão *quando $d\varphi_p : T_pM \to T_{\varphi(p)}N$ é injetiva para todo $p \in M$. Se, além disso, φ é um homeomorfismo de M sobre o subespaço $\varphi(M) \subset N$ dizemos que φ é um* mergulho*.*

Definição 1.12 (Subvariedade, [12])**.** *Seja N uma variedade de dimensão n. Se $M \subset N$ e a inclusão $i : M \hookrightarrow N$ é um mergulho, dizemos que M é uma* subvariedade *de N.*

A próxima subseção tem como foco a definição de métrica riemanniana em uma variedade diferenciável. Neste trabalho, veremos que a matriz de informação de Fisher gera uma métrica riemanniana em um modelo estatístico regular.

1.3.2 Métricas Riemannianas

Nesta subseção, veremos a definição de métrica riemanniana, que pelo Teorema 2.3 do próximo capítulo mostraremos que a matriz de informação de Fisher gera uma métrica riemanniana no modelo estatístico regular.

Definição 1.13 (Métrica Riemanniana, [7])**.** *Uma métrica riemanniana em uma variedade M de dimensão n é uma aplicação que associa a cada ponto $p \in M$ um produto interno. Isto é, uma forma bilinear simétrica, positiva definida*

$$g_p = \langle \cdot, \cdot \rangle_p$$

no espaço tangente T_pM que varia diferencialmente com p no sentido de que se a aplicação $\varphi : U \subset \mathbb{R}^n \to V \subset M$ é uma carta para uma vizinhança coordenada V de M e $\mathcal{B}_p = \left\{ \left.\frac{\partial}{\partial x_1}\right|_p, ..., \left.\frac{\partial}{\partial x_n}\right|_p \right\}$ é a base coordenada de T_pM associada a esta carta para cada $p \in V$, então as funções então as funções $g_{ij} : V \longrightarrow \mathbb{R}$

$$g_{ij}(p) = \left\langle \left.\frac{\partial}{\partial x_i}\right|_p, \left.\frac{\partial}{\partial x_j}\right|_p \right\rangle_p \text{ são diferenciáveis.}$$

Podemos representar a métrica riemanniana g, pela matriz $G_p = [g_{ij}(p)]_n$. Assim, dados $u = (u_1, ..., u_n), v = (v_1, ..., v_n) \in T_pM$, o produto interno associado a essa matriz é dado por

$$g_p = \langle u, v \rangle_p = u^T G_p v = \sum_{ij} g_{ij} u_i v_j.$$

Além disso, o elemento de comprimento ds na métrica G_p satisfaz

$$ds^2 = \sum_{i,j=1}^{n} g_{ij}(p) dx_i dx_j.$$

Uma variedade riemanniana é um par (M, g), onde M é uma variedade diferenciável e g a métrica riemanniana, convém relembrar que uma mesma variedade diferenciável pode admitir diferentes métricas riemannianas.

Veja a seguir um exemplo de métrica riemanniana.

Exemplo 1.4 ([12]). *A métrica euclidiana em $\mathbb{R}^n$ é uma métrica riemanniana, dado $p = (x_1, ..., x_n) \in \mathbb{R}^n$, sendo*

$$g_{ij}(p) = \langle e_i, e_j \rangle$$
$$G(p) = I_n,$$

em que e_i e e_j são vetores da base canônica do $\mathbb{R}^n$ e I_n é a matriz identidade de ordem n. Assim, dados $u = (u_1, ..., u_n), v = (v_1, ..., v_n) \in \mathbb{R}^n$,

$$\langle u, v \rangle = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n.$$

Definição 1.14 ([12]). *Um campo vetorial $\vec{V}$ ao longo de uma curva $\alpha : I \to M$ é uma aplicação que a cada $t \in I$ associa um vetor tangente $\vec{V}(t) \in T_{\alpha(t)}M$. O campo vetorial $\vec{V}$ é diferenciável em $t_0 \in I$ quando, para alguma parametrização $\varphi : U \subset \mathbb{R}^n \to M$ em $\alpha(t_0)$, as componentes $v_i : I \to \mathbb{R}$, $i = 1, 2, \cdots, n$ de*

$$\vec{V}(t) = \sum_{i=1}^{n} v_i(t) \frac{\partial}{\partial x_i},$$

na base $\left\{ \left.\frac{\partial}{\partial x_1}\right|_{\alpha(t)}, \cdots, \left.\frac{\partial}{\partial x_n}\right|_{\alpha(t)} \right\}$ são funções diferenciáveis de t em t_0. O campo de vetores $\vec{V}$ é diferenciável em I quando é diferenciável para todo $t \in I$.

Veremos como uma métrica riemanniana pode ser usada para calcular comprimentos de curvas.

Se uma curva α está restrita a um intervalo fechado $[a, b] \in I$, seu comprimento é dado por

$$\ell(\alpha) = \int_a^b \left\langle \frac{d\alpha}{dt}, \frac{d\alpha}{dt} \right\rangle_{\alpha(t)}^{\frac{1}{2}} dt. \tag{1.5}$$

Veremos agora uma proposição de existência para métricas riemannianas. Mas antes, veremos a definição de um espaço topológico de Hausdorff e espaço topológico com base enumerável. Para mais detalhes ver J. Munkres 2000 [26].

Definição 1.15 ([26])**.** *Um espaço topológico X é chamado Espaço de Hausdorff se para quaisquer pontos distintos x e y, existirem vizinhanças U de x e V de y com $U \cap V = \emptyset$.*

Definição 1.16 ([26])**.** *Dizemos que o espaço topológico (X, τ) possui uma base enumerável se existe uma base enumerável de τ. Neste caso, dizemos que (X, τ) satisfaz o segundo axioma de enumerabilidade.*

Note que o espaço $\mathbb{R}^n$ (com topologia usual) possui uma base enumerável, basta considerar bolas centradas em $\mathbb{Q}^n$.

Proposição 1.5 ([12])**.** *Uma variedade diferenciável M de Hausdorff e com base enumerável possui uma métrica riemanniana.*

Demonstração. A prova desta proposição pode ser vista em M. P. do Carmo 2015 [12, p. 47]. □

1.3.3 Geodésicas

Nesta subsecção, veremos o conceito de geodésicas e algumas de suas propriedades. Ao longo deste trabalho, não estamos muito interessados na definição analítica de uma curva geodésica, e sim que é a curva com menor comprimento que liga dois pontos na variedade riemanniana. Vamos generalizar o conceito de geodésica para variedades riemannianas. Para isso é necessária a noção do que é uma derivada covariante para variedades riemannianas. Essa noção é mais complexa do que aquela apresentada no contexto das superfícies regulares e requer um estudo mais aprofundado em Geometria Riemanniana. A seguir, apresentaremos sua definição. Para mais detalhes sugerimos M. P. do Carmo 2015 [12], J. Jost 2008 [17] e R. Biezuner 2017 [7].

Sejam $\alpha : I \to M$ uma curva diferenciável e $\varphi : U \subset \mathbb{R}^n \to M$ um sistema de coordenadas para M com $\alpha(I) \cap \varphi(U) \neq \emptyset$. Dado $t \in I$, a expressão local de $\alpha(t)$ é dada por $\alpha(t) = (x_1(t), \cdots, x_n(t))$. Seja $\vec{V}$ um campo de vetores ao longo da curva α. Podemos expressar o campo $\vec{V}$ localmente como

$$\vec{V} = \sum_j v_j X j$$

$j = 1, \cdots, n$, em que $v_j = v_j(t)$ e $X_j = \dfrac{\partial}{\partial x_j}(\alpha(t))$. A derivada covariante, $\dfrac{D\vec{V}}{dt}$, de $\vec{V}$ ao longo de α no sistema de coordenadas (U, φ) é dada por

$$\frac{D\vec{V}}{dt} = \sum_k \left\{ \frac{dv_k}{dt} + \sum_{i,j} \Gamma_{ij}^k v_j \frac{dx_i}{dt} \right\} X_k,$$

em que os coeficientes Γ_{ij}^k são funções diferenciáveis definidas em U conhecidos como símbolos de Christoffel de M na parametrização φ. Sendo $g_{ij} = \langle X_i, X_j \rangle$, as entradas de uma matriz e escrevermos a sua inversa como $[g^{ij}]$, os símbolos de Christoffel são definidos por

$$\Gamma_{ij}^k = \frac{1}{2} \sum_l \left\{ \frac{\partial}{\partial x_j} g_{il} + \frac{\partial}{\partial x_i} g_{jl} - \frac{\partial}{\partial x_l} g_{ij} \right\} g^{kl}.$$

Definição 1.17 (Geodésica, [12])**.** *Uma curva parametrizada $\gamma : I \to M$ é uma* geodésica *em $t_0 \in I$ quando*

$$\frac{D}{dt}\left(\frac{d\gamma}{dt}\right)\bigg|_{t_0} = 0.$$

Quando γ é geodésica em t, para todo $t \in I$, dizemos que γ é uma geodésica.

Vamos agora determinar as equações locais satisfeitas por uma geodésica γ em um sistemas de coordenadas (U, φ) em torno de $\gamma(t_0)$. Em U,

$$\gamma(t) = (x_1(t), \cdots, x_n(t)),$$

γ será uma geodésica se, e somente se,

$$0 = \frac{D}{dt}\left(\frac{d\gamma}{dt}\right) = \sum_k \left(\frac{d^2 x_k}{dt^2} + \sum_{i,j} \Gamma_{ij}^k \frac{dx_i}{dt} \frac{dx_j}{dt} \right) \frac{\partial}{\partial x_k}.$$

Logo o sistema de equações diferenciais de 2ª ordem

$$\frac{d^2 x_k}{dt^2} + \sum_{i,j} \Gamma_{ij}^k \frac{dx_i}{dt} \frac{dx_j}{dt} = 0, \quad k = 1, \cdots, n, \tag{1.6}$$

fornece as equações procuradas.

A seguir, veremos a Definição 1.18, a qual usaremos na prova do Teorema 3.3.

Definição 1.18 ([7])**.** *Uma geodésica $\gamma : I \longrightarrow M$ é normalizada (ou unitária) se*

$$\left\| \gamma'(t) \right\| \equiv 1$$

Toda geodésica que não é um ponto (ou seja, $\|\gamma'(t)\| \neq 0$) pode ser normalizada através de uma parametrização por comprimento de arco. Se $\gamma : I \longrightarrow M$ é uma parametrização qualquer para uma geodésica, ela pode ser reparametrizada para se tornar uma geodésica normalizada escolhendo-se um ponto $t_0 \in I$ e definindo o parâmetro comprimento de arco

$$s(t) = \int_{t_0}^{T} \|\gamma'(t)\| \, dt,$$

pela regra da cadeia

$$\|\gamma'(s)\| = \|\gamma'(t)\| \, |t'(s)| = \|\gamma'(t)\| \, \frac{1}{|s'(t)|} = \|\gamma'(t)\| \, \frac{1}{\|\gamma'(t)\|} = 1.$$

A seguir, veremos um teorema de existência e unicidade de geodésicas.

Teorema 1.3 (Teorema de Existência e Unicidade de Geodésicas, [7])**.** *Seja M uma variedade riemanniana. Então para todos $p \in M$ e $v \in T_pM$, e para cada $t_0 \in \mathbb{R}$, existe um intervalo aberto $I \subset \mathbb{R}$ contendo t_0 e uma única geodésica $\gamma : I \longrightarrow M$ tal que $\gamma(t_0) = p$ e $\gamma'(t_0) = v$.*

Demonstração. Ver demostração em R. Biezuner 2017 [7, 4.4 Teorema]. □

Definição 1.19. *Se $\gamma : I \longrightarrow M$ é uma geodésica e $[a,b] \subset I$, a restrição $\gamma|_{[a,b]}$ é chamada o segmento de geodésica ligando $\gamma(a)$ a $\gamma(b)$.*

Denote o comprimento de uma curva α em M ligando os pontos p e q por $\ell(\alpha)$ dado em (1.5). Dizemos que o segmento de geodésica $\gamma : [a,b] \longrightarrow M$ é minimizante se $\ell(\gamma) \leqslant \ell(\alpha)$ para toda curva α ligando $\gamma(a)$ a $\gamma(b)$.

Proposição 1.6 (Geodésicas minimizam distâncias localmente, [7])**.** *Sejam M uma variedade riemanniana, $p \in M$ e $B(p)$ uma bola normal centrada em p. Seja $\gamma : [0,1] \longrightarrow B(p)$ um segmento de geodésica com $\gamma(0) = p$ e denote $q = \gamma(1)$. Se $\alpha : [0,1] \longrightarrow M$ é qualquer curva suave por partes ligando p a q, então*

$$\ell(\gamma) \leqslant \ell(\alpha).$$

Se $\ell(\alpha) = \ell(\gamma)$, então necessariamente $\alpha([0,1]) = \gamma([0,1])$.

Demonstração. Ver demonstração em R. Biezuner 2017 [7, 4.28 Proposição]. □

Agora, veja a definição de distância em uma variedade riemanniana.

Definição 1.20 ([7]). *Seja M uma variedade riemanniana conexa. Dados $p,q \in M$, a distância entre p e q é definida por*

$$dist(p,q) = \inf\{\ell(\gamma) : \gamma \text{ é uma curva suave por partes ligando } p \text{ e } q\}.$$

A seguir, veremos que as geodésicas minimizam distâncias, localmente. Usaremos para definir geodésica de estimativas no Capítulo 4.

Proposição 1.7 ([7]). *Se existe uma geodésica minimizante γ ligando p e q, então $dist(p,q) = \ell(\gamma)$.*

Demonstração. Ver R. Biezuner 2017 [7, 4.37 Proposição]. □

Definição 1.21 (Variedade Totalmente Geodésica, [12]). *Uma subvariedade M de uma variedade riemanniana N é dita* totalmente geodésica *quando toda geodésica de M é geodésica de N.*

A seguir, veremos a definição de espaço hiperbólico em que usaremos como apoio o plano de Poincaré para determinar uma fórmula fechada para distância de Fisher-Rao entre duas densidades de probabilidade normais univariadas na Subseção 2.3.2 do Capítulo 2.

Definição 1.22 (Espaço hiperbólico, [4]). *O* espaço hiperbólico *de dimensão n é o semiespaço de $\mathbb{R}^n$ dado por*

$$\mathcal{H}^n = \{(x_1,...,x_n) \in \mathbb{R}^n; x_n > 0\}$$

com métrica riemanniana $ds^2 = \dfrac{dx_1^2 + \cdots + dx_n^2}{x_n^2}$.

Consideremos o plano hiperbólico $\mathcal{H}^2 = \{(x,y) \in \mathbb{R}^2; y > 0\}$ também conhecido como plano de Poincaré. A métrica riemanniana nesse plano é gerada por

$$G(x,y) = \text{diag}(1/y^2, 1/y^2),$$

e a expressão da métrica é dada por $ds_1^2 = \dfrac{dx^2 + dy^2}{y^2}$. A escolha da notação ds_1^2 será conveniente para este trabalho.

Dados dois pontos $x = (x_1, x_2)$ e $y = (y_1, y_2)$ em $\mathcal{H}^2$ uma expressão para distância hiperbólica é

$$d_{\mathcal{H}^2}(x,y) = \log\left(\frac{\|x - \bar{y}\| + \|x + y\|}{\|x - \bar{y}\| - \|x + y\|}\right), \tag{1.7}$$

em que $\bar{y} = (y_1, -y_2)$ e $\|z\| = \sqrt{z_1^2 + z_2^2}$, para $z = (z_1, z_2)$.

As geodésicas de $\mathcal{H}^2$ são as semirretas verticais $\gamma_1 : (0, \infty) \to \mathcal{H}^2$ e as semicircunferências $\gamma_2 : (0, \pi) \to \mathcal{H}^2$ de centro $(c, 0)$ e raio ρ dadas por (ver Figura 1.4)

$$\gamma_1(t) = (x_0, t) \text{ e } \gamma_2(t) = (\rho \cos(t) + c, \rho \operatorname{sen}(t)),$$

ver exemplos de geodésicas em R. Biezuner 2017 [7, p. 81].

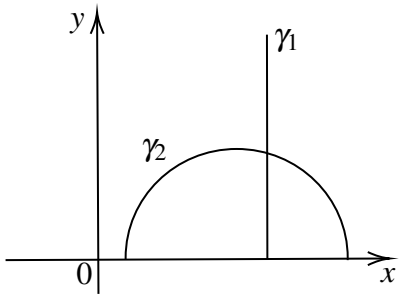

Figura 1.4 Geodésicas de $\mathcal{H}^2$.
Produzida online com Mathcha.

Relembre que $\cosh^{-1}(x) = \log\left(x + \sqrt{x^2 - 1}\right)$. Logo, podemos representar a distância hiperbólica como

$$d_{\mathcal{H}^2}(x, y) = \cosh^{-1}\left(1 + \frac{(y_1 - x_1)^2 + (x_2 - y_2)^2}{2x_2 y_2}\right). \tag{1.8}$$

Capítulo 2

Estrutura Geométrica de Modelos Estatísticos

Neste capítulo, veremos de que maneira modelos estatísticos regulares se apresentam como variedade diferenciável. Definiremos a métrica de Fisher em modelos estatísticos regulares e a distância de Fisher-Rao. Nosso objeto de trabalho são famílias paramétricas de densidades de probabilidade com a propriedade de ser regular. Na Seção 2.1, apresentaremos um teorema que garante que estas famílias são variedades diferenciáveis, de Hausdorff e com base enumerável, a qual denotaremos por variedade estatística, também mostraremos que a familia composta por densidade normal multivariada é regular. Na Seção 2.2, definiremos uma métrica na variedade estatística, a métrica de Fisher, a qual gera uma métrica riemanniana nesta variedade, fazendo desta variedade uma variedade estatística riemanniana. Na Seção 2.3 definiremos a distância de Fisher-Rao e com alguns exemplos em que calcularemos a distância entre algumas densidades de probabilidade já conhecida e em subvariedades da normal multivariada.

2.1 Variedade Estatística

Nesta seção, veremos através do Teorema 2.1, como um modelo estatístico regular é uma variedade estatística. O referido teorema foi apresentado em W. Santiago 2017 [31, p. 33]. Apresentaremos e forneceremos uma prova mais detalhada do mesmo.

Teorema 2.1. *Seja S um modelo estatístico regular sobre $\mathcal{X}$ de dimensão n, então S é uma variedade diferenciável de dimensão n, de Hausdorff e com base enumerável.*

Demonstração. Da hipótese, segue a existência de uma função bijetiva $\varphi : \Theta \to S$ dada por $\varphi(\theta) = p_\theta$, sendo $\Theta \subset \mathbb{R}^n$ aberto. Pelo item *ii*) da Definição 1.5

$$\Theta \ni \theta \mapsto \varphi(\theta) = p_\theta \text{ é suave,}$$

então φ é uma parametrização global de classe C^∞ de S. Seja o conjunto

$$M = \{f_\alpha : \Theta \to U_\alpha \subset \mathbb{R}^n; U_\alpha \text{ é aberto e } f_\alpha \text{ é difeomorfismo de classe } C^\infty\}.$$

Veja que este conjunto é não vazio, pois basta tomar $f_\alpha = I : \Theta \to \Theta$ a identidade.

Defina $y_\alpha := \varphi \circ f_\alpha^{-1} : U_\alpha \to S$ e note que y_α é bijetiva por se tratar de composição de funções bijetivas. Veja a Figura 2.1

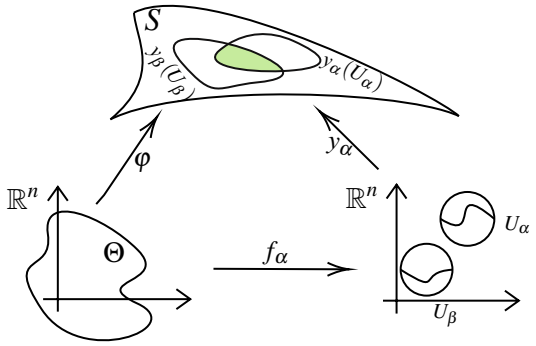

Figura 2.1 Representação geométrica da demostração do Teorema 2.1.
Produzida online com Mathcha.

Mostremos que a família $\mathcal{A} = \{(U_\alpha, y_\alpha)\}$ é uma estrutura diferenciável em S. Com efeito,

i) Devemos mostrar que $\bigcup_{\alpha\in\mathcal{I}} y_\alpha(U_\alpha) = S$. Dado $p \in \bigcup_{\alpha\in\mathcal{I}} y_\alpha(U_\alpha)$ temos que, $p \in y_\alpha(U_\alpha)$ para algum $\alpha \in \mathcal{I}$, em que $\mathcal{I}$ é um conjunto de índices, segue que $p \in S$.
Dado $p \in S$, existe $\theta \in \Theta$ tal que $p = \varphi(\theta)$. Veja que $\theta = f^{-1}(U_\alpha)$ para algum $\alpha \in \mathcal{I}$. Logo,

$$p = \varphi(f^{-1}(U_\alpha)) \implies p \in \bigcup_{\alpha\in\mathcal{I}} \varphi(f^{-1}(U_\alpha) \implies p \in \bigcup_{\alpha\in\mathcal{I}} y_\alpha(U_\alpha).$$

ii) Para cada par α, β com $y_\alpha(U_\alpha) \bigcap y_\beta(U_\beta) = W_{\alpha\beta} \neq \emptyset$, temos que $y_\alpha^{-1}(W_{\alpha\beta})$ e $y_\beta^{-1}(W_{\alpha\beta})$ são conjuntos abertos em $\mathbb{R}^n$. Além disso,

$$y_\beta^{-1} \circ y_\alpha = f_\beta \circ \varphi^{-1} \circ \varphi \circ f_\alpha^{-1} = f_\beta \circ f_\alpha^{-1},$$

e

$$y_\alpha^{-1} \circ y_\beta = f_\alpha \circ \varphi^{-1} \circ \varphi \circ f_\beta^{-1} = f_\alpha \circ f_\beta^{-1},$$

são aplicações diferenciáveis. Por *i*) e *ii*), a família $\mathcal{A} = \{(U_\alpha, y_\alpha)\}$ é uma estrutura diferenciável em S. Portanto pela Definição 1.6, S é uma variedade diferenciável de dimensão n. A topologia de S é dada pela estrutura diferencial, ver Observação 1.2.

A afirmação ser de Hausdorff é preservada sob homeomorfismo, pois Θ é um espaço topológico de Hausdorff e φ é um homeomorfismo diferenciável, assim segue que S é um espaço topológico de Hausdorff. Como Θ é aberto e φ leva aberto em aberto, S possui base enumerável, pois em $\Theta \subset \mathbb{R}^n$ a topologia natural é a do subespaço e uma base enumerável pode ser as bolas abertas centradas em pontos com coordenadas racionais interceptadas com Θ, assim a base de S é a imagem dessas intercessões pela φ. □

Na subseção a seguir, mostraremos que o modelo formado por densidades normais multivariadas é variedade estatística, ou seja, satisfaz as propriedades de regularidade do modelo.

2.1.1 Normais Multivariadas

O objetivo desta subseção é mostrar que a família paramétrica composta por densidades normais multivariadas é uma variedade estatística, para isso, mostraremos que é identificável e que satisfaz a propriedade de ser regular. Sabemos que para mostrar a regularidade do modelo tem-se que verificar se satisfaz todos os itens da Definição 1.5. Para tanto, apresentaremos proposições com base em cada item da definição referida para que a demonstração não fique longa.

O espaço paramétrico das densidades normais multivariadas é dada por

$$\Theta = \{(\mu, \mathrm{V}); \mu \in \mathbb{R}^n \text{ e } \mathrm{V} \in P_n(\mathbb{R})\} = \mathbb{R}^n \times P_n(\mathbb{R}).$$

Relembre que $P_n(\mathbb{R})$ é o subconjunto das matrizes simétricas positivas definidas em $GL_n(\mathbb{R})$. Além disso, sabemos que todo espaço vetorial n-dimensional sobre $\mathbb{R}$ é isomorfo a $\mathbb{R}^n$. Daí, como $P_n(\mathbb{R})$ é um subespaço de vetorial de $M_n(\mathbb{R}) \cong \mathbb{R}^{n^2}$ e $\dim(P_n(\mathbb{R})) = n(n+1)/2$, segue que $P_n(\mathbb{R}) \cong \mathbb{R}^{n(n+1)/2}$. Assim, temos

$$\Theta = \mathbb{R}^n \times P_n(\mathbb{R}) \cong \mathbb{R}^n \times \mathbb{R}^{n(n+1)/2} \subset \mathbb{R}^n \times \mathbb{R}^{n^2}.$$

Pela Definição 1.3, o conjunto $\{p(x;\mu,\mathrm{V});(\mu,\mathrm{V})\in\mathbb{R}^n\times P_n(\mathbb{R})\}$ formado pelas densidades de probabilidade dada em (1.2) é uma família paramétrica sobre $\mathcal{X}=\mathbb{R}^n$ de dimensão $n+n(n+1)/2$.

A seguir, veremos que a família composta por densidades normais multivariada é identificável.

Proposição 2.1. *Seja*

$$S=\{p(x;\theta);x\in\mathbb{R}^n \text{ e } \theta=(\mu,\mathrm{V})\in\mathbb{R}^n\times P_n(\mathbb{R})\}$$

a família paramétrica das densidades normais multivariadas de dimensão n. Então S é identificável.

Demonstração. Dados $\theta_1=(\mu_1,\mathrm{V}_1)$ e $\theta_2=(\mu_2,\mathrm{V}_2)$ tais que

$$p(x;\theta_1)=p(x;\theta_2)\quad\forall x\in\mathbb{R}^n,$$

devemos provar que $\theta_1=\theta_2$. Temos para todo $x\in\mathbb{R}^n$

$$\frac{(2\pi)^{-n/2}}{\sqrt{\det(\mathrm{V}_1)}}\exp\left(-\frac{1}{2}\langle x-\mu_1,\Lambda_1(x-\mu_1)\rangle\right)=\frac{(2\pi)^{-n/2}}{\sqrt{\det(\mathrm{V}_2)}}\exp\left(-\frac{1}{2}\langle x-\mu_2,\Lambda_2(x-\mu_2)\rangle\right).$$

Assim

$$\exp\left\{-\frac{1}{2}\left[\langle x-\mu_1,\Lambda_1(x-\mu_1)\rangle-\langle x-\mu_2,\Lambda_2(x-\mu_2)\rangle\right]\right\}=\frac{\sqrt{\det(\mathrm{V}_1)}}{\sqrt{\det(\mathrm{V}_2)}}\quad\forall x\in\mathbb{R}^n.$$

Segue que

$$\langle x-\mu_1,\Lambda_1(x-\mu_1)\rangle-\langle x-\mu_2,\Lambda_2(x-\mu_2)\rangle=k, \tag{2.1}$$

sendo k contante em relação ao argumento $x\in\mathbb{R}^n$. Vamos provar que $k=0$. Pelo Corolário 1.1, temos que Λ_1 e Λ_2 são positivas definidas, logo para todo $x\in\mathbb{R}^n$,

$$\langle x-\mu_1,\Lambda_1(x-\mu_1)\rangle\geq 0\quad\text{ e }\quad\langle x-\mu_2,\Lambda_2(x-\mu_2)\rangle\geq 0.$$

Suponha que $k<0$. Então, fazendo $x=\mu_2$ em (2.1), chegaremos à contradição

$$k=\langle\mu_2-\mu_1,\Lambda_1(\mu_2-\mu_1)\rangle-\langle\mu_2-\mu_2,\Lambda_2(\mu_2-\mu_2)\rangle=\langle\mu_2-\mu_1,\Lambda_1(\mu_2-\mu_1)\rangle\geq 0,$$

pois Λ_1 é positiva definida. Por outro lado, se $k > 0$, fazendo $x = \mu_1$ em (2.1), obteremos

$$k = \langle \mu_1 - \mu_1, \Lambda_1(\mu_1 - \mu_1)\rangle - \langle \mu_1 - \mu_2, \Lambda_2(\mu_1 - \mu_2)\rangle = -\langle \mu_1 - \mu_2, \Lambda_2(\mu_1 - \mu_2)\rangle \leq 0,$$

pois Λ_1 é positiva definida, e teremos novamente uma contradição. Portanto $k = 0$.

Agora, fazendo $x = \mu_2$ em (2.1), temos

$$\langle \mu_2 - \mu_1, \Lambda_1(\mu_2 - \mu_1)\rangle = 0$$

o que implica $\mu_2 = \mu_1$, pois Λ_1 é positiva definida. Assim, de (2.1)

$$\langle x - \mu_1, \Lambda_1(x - \mu_1)\rangle = \langle x - \mu_1, \Lambda_2(x - \mu_1)\rangle \quad x \in \mathbb{R}^n. \tag{2.2}$$

Resta mostrar que $V_1 = V_2$. Para todo $u \in \mathbb{R}^n$, fazendo $x = u + \mu_1$ em (2.2), teremos

$$\langle u, \Lambda_1 u\rangle = \langle u, \Lambda_2 u\rangle, \quad \forall u \in \mathbb{R}^n. \tag{2.3}$$

Sendo, para todo $u, v \in \mathbb{R}^n$

$$\langle u, v\rangle_1 = \langle u, \Lambda_1 v\rangle \text{ e } \langle u, v\rangle_2 = \langle u, \Lambda_2 v\rangle$$

o produto interno em $\mathbb{R}^n$ gerado pelas matrizes Λ_1 e Λ_2 respectivamente, temos as correspondentes normas em $\mathbb{R}^n$

$$\|u\|_1^2 = \langle u, u\rangle_1 \text{ e } \|u\|_2^2 = \langle u, u\rangle_2, \quad \forall u \in \mathbb{R}^n$$

e por (2.3) vale

$$\|u\|_1^2 = \|u\|_2^2, \quad \forall u \in \mathbb{R}^n.$$

Pela identidade de polarização, temos para todo $u, v \in \mathbb{R}^n$

$$\langle u, v\rangle_1 = \frac{1}{4}\left[\|u + v\|_1^2 - \|u - v\|_1^2\right] = \frac{1}{4}\left[\|u + v\|_2^2 - \|u - v\|_2^2\right] = \langle u, v\rangle_2.$$

Ou seja,

$$\langle u, \Lambda_1 v\rangle = \langle u, \Lambda_2 v\rangle \quad \forall u, v \in \mathbb{R}^n. \tag{2.4}$$

Logo, pela Observação 1.1, $\Lambda_1 = \Lambda_2$ e portanto, $V_1 = V_2$. □

Observação 2.1. Uma prova simples e direta da Proposição 2.1 seria apenas observar que

$$\mu_1 = \int_{\mathbb{R}^n} p(x;\theta_1)x\,dx = \int_{\mathbb{R}^n} p(x;\theta_2)x\,dx = \mu_2,$$

$$\mathrm{V}_1 = \int_{\mathbb{R}^s} (x-\mu_1)(x-\mu_1)^T p(x;\theta_1)\,dx = \int_{\mathbb{R}^s} (x-\mu_2)(x-\mu_2)^T p(x;\theta_2)\,dx = \mathrm{V}_2,$$

onde $s = n(n+1)/2$.

Corolário 2.1. *Sob as condições da Proposição 2.1, S é um modelo estatístico de dimensão* $k = n+n(n+1)/2$.

A seguir, estamos interessados em mostrar que o modelo S dado no Corolário 2.1 é regular. Para isto, apresentaremos o Lema 2.2 que servirá de auxílio para demostração da Proposição 2.2. Mas antes, vamos destacar o Lema 2.1.

Lema 2.1 ([22], p. 174)**.** *A função det* : $P_n(\mathbb{R}) \to \mathbb{R}$ *que associa a cada matriz* $n \times n$*, o seu determinante, é suave.*

Lema 2.2. *O conjunto das matrizes inversíveis* $n \times n$ *é aberto em* $\mathbb{R}^{n^2}$.

Demonstração. Considere a função

$$\begin{aligned} \det : GL_n(\mathbb{R}) &\to \mathbb{R} - \{0\} \\ A &\mapsto \det(A) \end{aligned}$$

Veja que essa função está bem definida. Vamos mostrar que é sobrejetiva para podermos escrever $GL_n(\mathbb{R}) = \det^{-1}(\mathbb{R} - \{0\})$. Dado $x \in \mathbb{R}^+_*$, tome a matriz

$$A = \mathrm{diag}\left(\sqrt[n]{x}, \sqrt[n]{x}, ..., \sqrt[n]{x}\right) \in GL_n(\mathbb{R})$$

Assim, $\det(A) = \left(\sqrt[n]{x}\right)^n = x$. Agora, se $x < 0$, temos que considerar dois casos:

1. Se n for impar, tome

$$A = \mathrm{diag}\left(-\sqrt[n]{|x|}, -\sqrt[n]{|x|}, ..., -\sqrt[n]{|x|}\right) \in GL_n(\mathbb{R})$$

   Assim, $\det(A) = \left(-\sqrt[n]{|x|}\right)^n = -|x| = x$;

2. Caso n for par, tome

$$A = \mathrm{diag}\left(-\sqrt[n-1]{|x|}, -\sqrt[n-1]{|x|}, ..., -\sqrt[n-1]{|x|}, 1\right) \in GL_n(\mathbb{R})$$

Assim, $\det(A) = \left(-\sqrt[n-1]{|x|}\right)^{n-1} = -|x| = x$.

Portanto, para qualquer $x \in \mathbb{R} - \{0\}$ existe pelo menos uma matriz inversível A em $GL_n(\mathbb{R})$ tal que $\det(A) = x$. Logo, temos a sobrejetividade da função det, isto é, $\mathrm{Im}(\det) = \mathbb{R} - \{0\} = (-\infty, 0) \cup (0, +\infty)$.

Sabemos que a função determinante é contínua pelo Lema 2.1 e que $GL_n(\mathbb{R}) \cong \mathbb{R}^{n(n+1)/2}$. Logo, $\det^{-1}((-\infty, 0) \cup (0, +\infty)) = GL_n(\mathbb{R})$ é aberto. □

A Proposição 2.2 a seguir é dedicada ao item *i*) da Definição 1.5, a qual veremos que o espaço de parâmetros do modelo S é aberto de $\mathbb{R}^n$.

Proposição 2.2. *Seja $S = \{p(x; \mu, \mathrm{V}); x \in \mathbb{R}^n \text{ e } (\mu, \mathrm{V}) \in \mathbb{R}^n \times P_n(\mathbb{R})\}$ o modelo estatístico composto por densidades normais multivariadas sobre $\mathbb{R}^n$ de dimensão $k = n + n(n+1)/2$, então o espaço paramétrico $\mathbb{R}^n \times P_n(\mathbb{R})$ é um subconjunto aberto em $\mathbb{R}^{n+n^2}$.*

Demonstração. Pelo Lema 2.2, segue que $P_n(\mathbb{R}) \cong \mathbb{R}^{n(n+1)/2}$ é um aberto em $\mathbb{R}^{n^2}$. Logo $\mathbb{R}^n \times P_n(\mathbb{R})$ é um subconjunto aberto em $\mathbb{R}^n \times \mathbb{R}^{n^2} \cong \mathbb{R}^{n+n^2}$, pois o produto cartesiano finito de abertos é aberto, ver E. L. Lima 2014 [22]. □

Adiante, veremos dois lemas sem nos preocupara com a demostração que usaremos para demonstrar a Proposição 2.3.

Lema 2.3 ([22], p. 138)**.** *A função $f = \langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ tal que $f((x,y)) = \langle x, y \rangle$ para quaisquer $x, y \in \mathbb{R}^n$ é suave.*

Lema 2.4 ([22], p. 258)**.** *A função $\psi : GL_n(\mathbb{R}) \to GL_n(\mathbb{R})$ tal que $\psi(A) = A^{-1}$, é suave.*

A proposição a seguir, é dedicada ao item *ii*) da Definição 1.5, a qual veremos que as densidades são suaves para $x \in \mathbb{R}^n$ fixo.

Proposição 2.3. *Seja $S = \{p(x; \mu, \mathrm{V}); x \in \mathbb{R}^n \text{ e } (\mu, \mathrm{V}) \in \mathbb{R}^n \times P_n(\mathbb{R})\}$ o modelo estatístico sobre $\mathbb{R}^n$ composto por densidades normais multivariadas. Fixado $x \in \mathbb{R}^n$, as funções $\theta \in \Theta \mapsto p(x, \theta)$ são suaves, ou seja, admitem derivadas parciais em relação a θ em todas as ordens e são contínuas.*

Demonstração. Fixado $x \in \mathbb{R}^n$, para qualquer $p \in S$, sabemos que

$$p(x; \mu, \mathrm{V}) = \frac{(2\pi)^{-n/2}}{\sqrt{\det(\mathrm{V})}} \exp\left(-\frac{1}{2}\langle x - \mu, \Lambda(x - \mu)\rangle\right).$$

Para mostrar a suavidade da densidade de probabilidade $p(x; \mu, \mathrm{V})$, olharemos como uma composição de funções suaves. Veja que:

1. A função $\exp : \mathbb{R} \to \mathbb{R}$ que associa a cada número real x, o número e^x é suave, pois a função exp é analítica;

2. A função $\phi : (0,\infty) \to \mathbb{R}_+^*$ tal que $\phi(x) = (2\pi)^{-n/2}/\sqrt{x}$, é uma suave por se tratar de composições de funções suaves;

3. A função $\xi : \mathbb{R}^n \times P_n(\mathbb{R}) \to \mathbb{R}$ tal que $\xi((\mu, \mathrm{V})) = -\langle x-\mu, \Lambda(x-\mu)\rangle/2$ é suave por se tratar de composição de funções suaves, ver Lema 2.3 e Lema 2.4.

Como o produto e composição de funções suaves são suaves [22] e pelo Lema 2.1, segue que:

$$p(x;\mu,\mathrm{V}) = \phi(\det(\mathrm{V})) \cdot \exp(\xi(\mu,\mathrm{V}))$$

é uma função suave. □

Agora, vamos nos dedicar ao item *iii*) da Definição 1.5, a qual vale a troca da ordem de integração e derivação em relação a cada parâmetro da densidade normal multivariada. Destaca-se a Observação 2.2 como apoio para demonstração da Proposição 2.4 e do Lema 2.5.

Observação 2.2. Note que

$$\Lambda \mathrm{V} = I \implies \frac{\partial \Lambda}{\partial \sigma_{ij}} = -\Lambda \frac{\partial \mathrm{V}}{\partial \sigma_{ij}} \Lambda.$$

Para $i = j$, podemos escrever

$$\frac{\partial \mathrm{V}}{\partial \sigma_{ii}} = e_i e_i^T \quad \text{e para } i \neq j \quad \frac{\partial \mathrm{V}}{\partial \sigma_{ij}} = e_i e_j^T + e_j e_i^T,$$

em que e_i e e_j denotam vetores da base canônica do $\mathbb{R}^n$. Assim, segue que

$$\frac{\partial \Lambda}{\partial \sigma_{ii}} = -\Lambda \frac{\partial \mathrm{V}}{\partial \sigma_{ii}} \Lambda = -\Lambda e_i e_i^T \Lambda, \quad i = 1, ..., n$$

e

$$\frac{\partial \Lambda}{\partial \sigma_{ij}} = -\Lambda \frac{\partial \mathrm{V}}{\partial \sigma_{ij}} \Lambda = -\Lambda (e_i e_j^T + e_j e_i^T) \Lambda, \quad i, j = 1, ..., n, \, i < j.$$

Proposição 2.4. *Seja* $S = \{p(x;\mu,\mathrm{V}); x \in \mathbb{R}^n \text{ e } (\mu,\mathrm{V}) \in \mathbb{R}^n \times P_n(\mathbb{R})\}$ *um modelo estatístico composto por densidades normais multivariadas sobre* $\mathbb{R}^n$ *de dimensão* $k = n + n(n+1)/2$*, então para todo* $p \in S$ *e* $i = 1, 2, ..., k$ *temos*

$$\int_{\mathbb{R}^n} \frac{\partial p(x,\theta)}{\partial \theta_i} dx = 0.$$

Demonstração. Sabemos que $\theta = (\mu, \mathrm{V})$, sendo que $\mu = (\mu_1, ..., \mu_n) \in \mathbb{R}^n$ e $\mathrm{V} = [\sigma_{ij}]_{n\times n}$, em que $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$ para $i, j = 1, ..., n$. Temos que calcular as seguintes derivadas

$$\frac{\partial p(x;\mu,\mathrm{V})}{\partial \mu_k}, \quad \forall k = 1, ..., n \quad \text{e} \quad \frac{\partial p(x;\mu,\mathrm{V})}{\partial \sigma_{ij}} \quad 1 \leq i \leq j \leq n.$$

Assim, para todo $k = 1, ..., n$

$$\begin{aligned}
\frac{\partial p(x;\mu,\mathrm{V})}{\partial \mu_k} &= \frac{(2\pi)^{-n/2}}{\sqrt{\det(\mathrm{V})}} \frac{\partial}{\partial \mu_k} \left[\exp\left(-\frac{1}{2} \langle x-\mu, \Lambda(x-\mu) \rangle \right) \right] \\
&= -\frac{1}{2} \frac{(2\pi)^{-n/2}}{\sqrt{\det(\mathrm{V})}} \exp\left(-\frac{1}{2} \langle x-\mu, \Lambda(x-\mu) \rangle \right) \frac{\partial}{\partial \mu_k} \langle x-\mu, \Lambda(x-\mu) \rangle,
\end{aligned}$$

ou seja, pelo Lema 1.1,

$$\begin{aligned}
\frac{\partial p(x;\mu,\mathrm{V})}{\partial \mu_k} &= -\frac{1}{2} p(x;\mu,\mathrm{V}) \left(\left\langle \frac{\partial (x-\mu)}{\partial \mu_k}, \Lambda(x-\mu) \right\rangle + \left\langle x-\mu, \Lambda \frac{\partial (x-\mu)}{\partial \mu_k} \right\rangle \right) \\
&= -\frac{1}{2} p(x;\mu,\mathrm{V}) \left(\langle -e_k, \Lambda(x-\mu) \rangle + \langle x-\mu, -\Lambda e_k \rangle \right) \\
&= \frac{1}{2} p(x;\mu,\mathrm{V}) \left(\langle e_k, \Lambda(x-\mu) \rangle + \langle x-\mu, \Lambda e_k \rangle \right) = p(x;\mu,\mathrm{V}) \langle e_k, \Lambda(x-\mu) \rangle \\
&= p(x;\mu,\mathrm{V}) e_k^T \Lambda(x-\mu) = p(x;\mu,\mathrm{V}) \sum_{j=1}^{n} \lambda_{kj}(x_j - \mu_j). \qquad (2.5)
\end{aligned}$$

Daí,

$$\begin{aligned}
\int_{\mathbb{R}^n} \frac{\partial p(x;\mu,\mathrm{V})}{\partial \mu_k} dx &= \int_{\mathbb{R}^n} p(x;\mu,\mathrm{V}) \mathbf{e}_k^T \Lambda(x-\mu)\, dx = \int_{\mathbb{R}^n} \mathbf{e}_k^T \Lambda(x-\mu) p(x;\mu,\mathrm{V})\, dx \\
&= \mathbf{e}_k^T \Lambda \mathbb{E}(\mathbf{X}-\mu) = 0,
\end{aligned}$$

sendo $\mathbf{X} \sim \mathcal{N}(\mu, \mathrm{V})$.

Agora, para $1 \leq i \leq j \leq n$, temos

$$\begin{aligned}
\frac{\partial p(x;\mu,\mathrm{V})}{\partial \sigma_{ij}} &= -(2\pi)^{-n/2} \cdot \frac{\partial\left(\sqrt{\det(\mathrm{V})}\right)}{\partial \sigma_{ij}} \frac{1}{\det(\mathrm{V})} \exp\left(-\frac{1}{2}\langle x-\mu, \Lambda(x-\mu)\rangle\right) \\
&\quad + \frac{(2\pi)^{-n/2}}{\sqrt{\det(\mathrm{V})}} \frac{\partial}{\partial \sigma_{ij}} \left[\exp\left(-\frac{1}{2}\langle x-\mu, \Lambda(x-\mu)\rangle\right)\right] \\
&= -(2\pi)^{-n/2} \cdot \frac{1}{2} \frac{\det(\mathrm{V})^{-1/2}}{\det(\mathrm{V})} \frac{\partial\left(\det(\mathrm{V})\right)}{\partial \sigma_{ij}} \exp\left(-\frac{1}{2}\langle x-\mu, \Lambda(x-\mu)\rangle\right) \\
&\quad - \frac{1}{2} \frac{(2\pi)^{-n/2}}{\sqrt{\det(\mathrm{V})}} \exp\left(-\frac{1}{2}\langle x-\mu, \Lambda(x-\mu)\rangle\right) \frac{\partial}{\partial \sigma_{ij}} \langle x-\mu, \Lambda(x-\mu)\rangle,
\end{aligned}$$

ou seja, pelo Lema 1.3

$$\begin{aligned}
\frac{\partial p(x;\mu,\mathrm{V})}{\partial \sigma_{ij}} &= -(2\pi)^{-n/2} \cdot \frac{1}{2} \det(\mathrm{V})^{-1/2} \mathrm{tr}\left(\Lambda \frac{\partial \mathrm{V}}{\partial \sigma_{ij}}\right) \exp\left(-\frac{1}{2}\langle x-\mu, \Lambda(x-\mu)\rangle\right) \\
&\quad - \frac{1}{2} p(x;\mu,\mathrm{V}) \left\langle x-\mu, \frac{\partial \Lambda}{\partial \sigma_{ij}}(x-\mu)\right\rangle \\
&= -\frac{1}{2}\mathrm{tr}\left(\Lambda \frac{\partial \mathrm{V}}{\partial \sigma_{ij}}\right) p(x;\mu,\mathrm{V}) - \frac{1}{2} p(x;\mu,\mathrm{V}) \left\langle x-\mu, \frac{\partial \Lambda}{\partial \sigma_{ij}}(x-\mu)\right\rangle \\
&= -\frac{1}{2}\mathrm{tr}\left(\Lambda \frac{\partial \mathrm{V}}{\partial \sigma_{ij}}\right) p(x;\mu,\mathrm{V}) - \frac{1}{2} p(x;\mu,\mathrm{V}) \left\langle x-\mu, B^{ij}(x-\mu)\right\rangle, \qquad (2.6)
\end{aligned}$$

sendo pela Observação 2.2 $B^{ij} = \dfrac{\partial \Lambda}{\partial \sigma_{ij}} = -\Lambda \dfrac{\partial \mathrm{V}}{\partial \sigma_{ij}} \Lambda$. Daí, temos

$$\begin{aligned}
\int_{\mathbb{R}^n} \frac{\partial p(x;\mu,\mathrm{V})}{\partial \sigma_{ij}} dx &= -\int_{\mathbb{R}^n} \frac{1}{2}\mathrm{tr}\left(\Lambda \frac{\partial \mathrm{V}}{\partial \sigma_{ij}}\right) p(x;\mu,\mathrm{V})\, dx \\
&\quad - \int_{\mathbb{R}^n} \frac{1}{2} p(x;\mu,\mathrm{V}) \sum_{k,l=1}^{n} B^{ij}_{kl}(x_k-\mu_k)(x_l-\mu_l)\, dx \\
&= -\frac{1}{2}\mathrm{tr}\left(\Lambda \frac{\partial \mathrm{V}}{\partial \sigma_{ij}}\right) \int_{\mathbb{R}^n} p(x;\mu,\mathrm{V})\, dx \\
&\quad - \frac{1}{2} \sum_{k,l=1}^{n} B^{ij}_{kl} \int_{\mathbb{R}^n} p(x;\mu,\mathrm{V})(x_k-\mu_k)(x_l-\mu_l)\, dx \\
&= -\frac{1}{2}\mathrm{tr}\left(\Lambda \frac{\partial \mathrm{V}}{\partial \sigma_{ij}}\right) - \frac{1}{2} \sum_{k,l=1}^{n} B^{ij}_{kl} \mathbb{E}[(X_k-\mu_k)(X_l-\mu_l)],
\end{aligned}$$

isto é,

$$\begin{aligned}\int_{\mathbb{R}^n} \frac{\partial p(x;\mu,\mathrm{V})}{\partial \sigma_{ij}}\,dx &= -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial \sigma_{ij}}\right) - \frac{1}{2}\sum_{k,l=1}^{n} B_{kl}^{ij}\sigma_{kl} \\ &= -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial \sigma_{ij}}\right) - \frac{1}{2}\sum_{k,l=1}^{n} B_{kl}^{ij}\sigma_{lk} \\ &= -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial \sigma_{ij}}\right) - \frac{1}{2}\mathrm{tr}\left(B^{ij}\,\mathrm{V}\right) \\ &= -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial \sigma_{ij}}\right) + \frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial \sigma_{ij}}\Lambda\,\mathrm{V}\right) \\ &= -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial \sigma_{ij}}\right) + \frac{1}{2}\mathrm{tr}\left(\mathrm{V}\,\frac{\partial \mathrm{V}}{\partial \sigma_{ij}}\right) = 0.\end{aligned}$$

Portanto,

$$\int_{\mathbb{R}^n} \frac{\partial p(x,\theta)}{\partial \theta_i} dx = 0$$

Para todo $p \in S$ e $i = 1,2,...,k$. □

A seguir, veremos a última condição de regularidade para o modelo S, a independência linear.

Lema 2.5. *Seja $X \in \mathbb{R}^n$ um vetor aleatório com $X \sim \mathcal{N}(\mu,\mathrm{V})$, então $\partial \ln(p(x;\theta))/\partial\theta_i$ para $i = 1,...,k$ são linearmente independente como funções de x, sendo que $k = n + n(n+1)/2$.*

Demonstração. Por hipótese, a densidade de probabilidade do vetor aleatório X é dada por

$$p(x;\theta) = \frac{(2\pi)^{-n/2}}{\sqrt{\det(\mathrm{V})}}\exp\left(-\frac{1}{2}\langle x-\mu, \Lambda(x-\mu)\rangle\right), \quad x \in \mathbb{R}^n,$$

em que $\mu = (\mu_1,...,\mu_n)$, $\mathrm{V} = [\sigma_{ij}]_{n\times n}$ e $\Lambda = [\lambda_{ij}]_{n\times n}$, ver Definição 1.2.

Faremos uma combinação linear das funções $\partial \ln(p(x;\theta))/\partial\theta_i$ para $i = 1,...,k$ dando zero e olharemos essa combinação como um polinômio sobre o corpo do reais de grau 2 identicamente nulo. Para tanto, por (2.5) e (2.6), temos que:

i) $\forall k = 1,2,...,n$

$$\frac{\partial}{\partial \mu_k}\ln p(x;\theta) = \sum_{j=1}^{n}\lambda_{kj}(x_j-\mu_j) = \langle [\Lambda]_k^{\ell}, u\rangle,$$

sendo $u = (x-\mu) = (x_1-\mu_1,...,x_n-\mu_n) \in \mathbb{R}^n$ e $[\Lambda]_k^{\ell} = [\lambda_{k1}\cdots\lambda_{kn}]$ indicando a linha k da matriz Λ;

ii) $\forall i,j=1,...,n \quad i\leq j$

$$\begin{aligned}\frac{\partial}{\partial\sigma_{ij}}\ln p(x;\theta) &= -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\sigma_{ij}}\right)-\frac{1}{2}(x-\mu)^T\frac{\partial\Lambda}{\partial\sigma_{ij}}(x-\mu)\\ &= \frac{1}{2}\left(-\lambda_{ij}-(x-\mu)^T\frac{\partial\Lambda}{\partial\sigma_{ij}}(x-\mu)\right)\\ &= \frac{1}{2}\left(-\lambda_{ij}+(x-\mu)^T\Lambda\frac{\partial \mathrm{V}}{\partial\sigma_{ij}}\Lambda(x-\mu)\right).\end{aligned}$$

Pela Observação 2.2, a qual descrevemos uma forma alternativa de escrever a $\partial \mathrm{V}/\partial\sigma_{ij}$ a fim de facilitar os cálculos. Mostremos que $\alpha_i\in\mathbb{R} \quad i=1,...,n$ e $\beta^{rs}\in\mathbb{R}, \quad r\leq s$, tais que

$$\begin{aligned}&\alpha_1\langle\Lambda_1^\ell,u\rangle+\alpha_2\langle\Lambda_2^\ell,u\rangle+\cdots+\alpha_n\langle\Lambda_n^\ell,u\rangle\\ &+\beta^{11}\left[-\lambda_{11}+u^T\Lambda e_1e_1^T\Lambda u\right]+\beta^{12}\left[-\lambda_{12}+u^T\Lambda(e_1e_2^T+e_2e_1^T)\Lambda u\right]+\cdots\\ &+\beta^{1n}\left[-\lambda_{1n}+u^T\Lambda(e_1e_n^T+e_ne_1^T)\Lambda u\right]\\ &+\beta^{22}\left[-\lambda_{22}+u^T\Lambda e_2e_2^T\Lambda u\right]+\cdots+\beta^{2n}\left[-\lambda_{2n}+u^T\Lambda(e_2e_n^T+e_ne_2^T)\Lambda u\right]+\cdots\\ &+\beta^{nn}\left[-\lambda_{nn}+u^T\Lambda e_ne_n^T\Lambda u\right]=0\end{aligned}\tag{2.7}$$

se, e somente se, $\alpha_i=\beta^{rs}=0$ para todo $i=1,...,n$ e $r,s=1,...,n$, $r\leq s$.

Note que, pelo Lema 1.1 para $i=1,...,n$, temos

$$u^T\Lambda e_ie_i^T\Lambda u=(u^T\Lambda e_i)(e_i^T\Lambda u)=\langle[\Lambda]_i^\ell,u\rangle^2.\tag{2.8}$$

e para $i\neq j$

$$\begin{aligned}u^T\Lambda\left(e_ie_j^T+e_je_i^T\right)\Lambda u &= u^T\Lambda e_ie_j^T\Lambda u+u^T\Lambda e_je_i^T\Lambda u\\ &= 2(u^T\Lambda e_i)(e_j^T\Lambda u)\\ &= 2\langle[\Lambda]_i^\ell,u\rangle\langle[\Lambda]_j^\ell,u\rangle\end{aligned}\tag{2.9}$$

Pelas equações (2.8) e (2.9), substituindo em (2.7), temos

$$\begin{aligned}&\alpha_1\langle\Lambda_1^\ell,u\rangle+\alpha_2\langle\Lambda_2^\ell,u\rangle+\cdots+\alpha_n\langle\Lambda_n^\ell,u\rangle\\ &+\beta^{11}\left[-\lambda_{11}+\langle\Lambda_1^\ell,u\rangle^2\right]+\beta^{12}\left[-\lambda_{12}+2\langle\Lambda_1^\ell,u\rangle\langle\Lambda_2^\ell,u\rangle\right]+\cdots\\ &+\beta^{1n}\left[-\lambda_{1n}+2\langle\Lambda_1^\ell,u\rangle\langle\Lambda_n^\ell,u\rangle\right]\\ &+\beta^{22}\left[-\lambda_{22}+\langle\Lambda_2^\ell,u\rangle^2\right]+\cdots+\beta^{2n}\left[-\lambda_{2n}+2\langle\Lambda_2^\ell,u\rangle\langle\Lambda_n^\ell,u\rangle\right]+\cdots\\ &+\beta^{nn}\left[-\lambda_{nn}+\langle\Lambda_n^\ell,u\rangle^2\right]=0 \quad \forall x,\end{aligned}$$

equivalentemente

$$\begin{aligned}
&\beta^{11}\langle\Lambda_1^\ell,u\rangle^2+\beta^{22}\langle\Lambda_2^\ell,u\rangle^2+\cdots+\beta^{nn}\langle\Lambda_n^\ell,u\rangle^2+2\beta^{12}\langle\Lambda_1^\ell,u\rangle\langle\Lambda_2^\ell,u\rangle+\cdots\\
&+2\beta^{1n}\langle\Lambda_1^\ell,u\rangle\langle\Lambda_n^\ell,u\rangle+2\beta^{23}\langle\Lambda_2^\ell,u\rangle\langle\Lambda_3^\ell,u\rangle+\cdots+2\beta^{2n}\langle\Lambda_2^\ell,u\rangle\langle\Lambda_n^\ell,u\rangle+\cdots\\
&+2\beta^{(n-1)n}\langle\Lambda_{n-1}^\ell,u\rangle\langle\Lambda_n^\ell,u\rangle+\alpha_1\langle\Lambda_1^\ell,u\rangle+\alpha_2\langle\Lambda_2^\ell,u\rangle+\cdots+\alpha_n\langle\Lambda_n^\ell,u\rangle-\beta^{11}\lambda_{11}\\
&-\beta^{12}\lambda_{12}-\cdots-\beta^{1n}\lambda_{1n}-\beta^{22}\lambda_{22}-\cdots-\beta^{2n}\lambda_{2n}-\cdots-\beta^{rr}\lambda_{rr}-\cdots\beta^{rn}\lambda_{rn}-\cdots\\
&-\beta^{nn}\lambda_{nn}=0\quad\forall x.
\end{aligned}$$

Ou seja,

$$\sum_{1\le i\le j\le n}\beta^{ij}\langle\Lambda_i^\ell,u\rangle\langle\Lambda_j^\ell,u\rangle+\sum_{i=1}^{n}\alpha_i\langle\Lambda_i^\ell,u\rangle-\sum_{1\le i\le j\le n}\beta^{ij}\lambda_{ij}=0\quad\forall x.$$

Como $u=(x-\mu)\in\mathbb{R}^n$, temos um polinômio multivariado de grau dois sobre o corpo dos reais identicamente nulo.

Vamos observar que dado $y\in\mathbb{R}^n$, existe x tal que $\langle\Lambda_k^\ell,(x-\mu)\rangle=y_k$, para todo $k=1,...,n$. Basta tomar $x=\mu+\mathrm{V}\,y$ e usar o fato de que

$$\left(\Lambda_k^\ell\right)^T\mathrm{V}=k-\text{ésima linha de }\Lambda\mathrm{V}=e_k^T.$$

Logo, temos

$$p(y)=\sum_{1\le i\le j\le n}\beta^{ij}y_iy_j+\sum_{i=1}^{n}\alpha_iy_i-\sum_{1\le i\le j\le n}\beta^{ij}\lambda_{ij}=0\quad\forall y\in\mathbb{R}^n.$$

Assim, segue diretamente que

$$\alpha_1=\cdots=\alpha_i=\cdots=\alpha_n=\beta^{11}=\cdots=\beta^{rs}=\cdots=\beta^{nn}=0.$$

Portanto, $\partial\ln(p(x;\theta))/\partial\theta_i$ são linearmente independente para $i=1,...,k$. □

Adiante, apresentaremos o teorema principal. Neste, colocamos em ordem os resultados estudados anteriormente.

Teorema 2.2. *Seja $S=\{p(x;\mu,\mathrm{V});x\in\mathbb{R}^n$ e $(\mu,\mathrm{V})\in\mathbb{R}^n\times P_n(\mathbb{R})\}$ um modelo estatístico composto por densidades normais multivariadas sobre $\mathbb{R}^n$ de dimensão $k=n+n(n+1)/2$. Então S é uma variedade diferenciável, de Hausdorff e com base enumerável.*

Demonstração. Pelo Teorema 2.1 basta mostrar que o modelo S é regular, isto é, que satisfaz os itens da Definição 1.5. Os itens $i),ii)$ e $iii)$ são verificados pelas Proposições 2.2, 2.3, 2.4

respectivamente e o item v) é satisfeito pelo Lema 2.5. Agora, note que o conjunto

$$Z_+ = \left\{ x; p(x;\mu,\mathrm{V}) = \frac{(2\pi)^{-n/2}}{\sqrt{\det(\mathrm{V})}} \exp\left(-\frac{1}{2}\langle x-\mu, \Lambda(x-\mu)\rangle \right) > 0 \quad \forall x \in \mathbb{R}^n \right\} = \mathbb{R}^n,$$

logo, é independente de (μ, V), para todo $p \in S$. Assim, garantimos o item iv). Portanto, S é regular, isto é, S é uma variedade diferenciável, de Hausdorff e com base enumerável. □

A partir daqui, todos os modelos estatísticos regulares serão chamados de variedades estatísticas.

2.2 Métrica de Fisher

Em 1945, C. R. Rao, ver referência [30], propôs um método para calcular a distância entre distribuições de probabilidade introduzindo uma métrica riemanniana em termos da chamada matriz de informação de Fisher em uma variedade estatística. Nesta seção, veremos a definição da matriz de informação de Fisher de um modelo estatístico regular dado por R. A. Fisher em 1922 [13], com alguns exemplos de modelos aqui apresentados. Essa matriz gera uma métrica riemanniana na variedade estatística e fornecemos alguns resultados de maneiras diferentes de escrever as entradas dessa matriz para fins de facilitar os cálculos quando for conveniente. Os resultados apresentados aqui se baseiam na referência O. Calin and C. Udrişte 2014 [10] e P. J. Bickel and K. A. Doksum 2001 [6].

Definição 2.1 (Matriz de Informação de Fisher, [10])**.** *Dada uma variedade estatística* $S = \{p_\theta = p(x;\theta); \theta = (\theta_1, ..., \theta_n) \in \Theta \subset \mathbb{R}^n\}$. *Considere um ponto* $\theta \in \Theta$*, a* matriz de informação de Fisher *de S em* θ *é a matriz* $G(\theta) = [g_{ij}(\theta)]$ *de ordem n, com*

$$\begin{aligned} g_{ij}(\theta) &= \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta_i}\log p(x;\theta) \frac{\partial}{\partial\theta_j}\log p(x;\theta) \right) \\ &= \int_{\mathcal{X}} \frac{\partial}{\partial\theta_i}\log p(x;\theta) \frac{\partial}{\partial\theta_j}\log p(x;\theta) p(x;\theta)\, dx, \end{aligned} \tag{2.10}$$

caso essa integral exista.

Note que $\mathbb{E}_\theta$ é a esperança com respeito a densidade p_θ. Para $n = 1$, $G(\theta)$ é denominada de informação de Fisher. Para mais detalhes ver referências [10, section 1.6, p. 21] e [6, section 3.4, p. 176].

Embora algumas vezes a integral dada na equação (2.10) seja divergente, neste trabalho, vamos considerar apenas modelos estatísticos nos quais $g_{ij}(\theta)$ é finita para todo θ e todo i, j e que $g_{ij} : \Theta \to \mathbb{R}$ é suave.

A seguir, apresentaremos um exemplo em que calculamos a informação de Fisher de uma variedade estatística formada por densidades exponenciais.

Exemplo 2.1. *Seja $S = \{p_\theta; \theta > 0\}$ a variedade estatística de densidades exponenciais, ou seja, se $p_\theta \in S$, então p_θ é uma população de $X \sim Exp(\theta)$ e é dada por*

$$p(x;\theta) = \theta e^{-\theta x}, \quad x > 0.$$

Usando a equação (1.3) *apresentada na Proposição 1.2, temos que a informação de Fisher de S em θ é dada por*

$$\begin{aligned} G(\theta) &= \mathbb{E}_\theta\left(\left(\frac{1}{\theta} - x\right)^2\right) = \int_0^\infty \left(\frac{1}{\theta} - x\right)^2 p(x;\theta)dx \\ &= \frac{1}{\theta^2} - \frac{2}{\theta}\int_0^\infty x p(x;\theta)dx + \int_0^\infty x^2 p(x;\theta)dx = \frac{1}{\theta^2} - \frac{2}{\theta}\mathbb{E}(X) + \mathbb{E}(X^2) \\ &= \frac{1}{\theta^2} - \frac{2}{\theta^2} + \frac{2}{\theta^2} = \frac{1}{\theta^2}. \end{aligned}$$

Agora, apresentaremos um exemplo em que calculamos a matriz de informação de Fisher de uma variedade estatística formada por densidades normais univariadas.

Exemplo 2.2. *Seja $S = \{p(x;\mu,\sigma); (\mu,\sigma^2) \in \mathbb{R}^2 \times (0,\infty)\}$ a variedade estatística de densidades normais univariadas, ou seja, se $p_\theta \in S$, então p_θ é uma população de $X \sim N(\mu,\sigma^2)$ e é dada por*

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

Neste caso o parâmetro θ é dado por duas variáveis reais $\theta = (\mu,\sigma) \in \Theta = \mathbb{R} \times (0,+\infty)$. Note que, $\ln p(x;\mu,\sigma) = -\ln(\sqrt{2\pi}) - \ln\sigma - \dfrac{(x-\mu)^2}{2\sigma^2}$. *Daí,*

$$\frac{\partial}{\partial\mu}\ln p(x;\mu,\sigma) = \frac{(x-\mu)}{\sigma^2} \quad e \quad \frac{\partial}{\partial\sigma}\ln p(x;\mu,\sigma) = \frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}.$$

Logo, os coeficientes da matriz de informação de Fisher são dados por

$$g_{11}(\theta) = \int_{-\infty}^{\infty}\frac{(x-\mu)^2}{\sigma^4}p(x;\mu,\sigma)dx = \frac{1}{\sigma^4}\int_{-\infty}^{\infty}(x-\mu)^2 p(x;\mu,\sigma)dx = \frac{1}{\sigma^4}Var(X) = \frac{1}{\sigma^2};$$

$$\begin{aligned} g_{12}(\theta) &= \int_{-\infty}^{\infty}\frac{(x-\mu)}{\sigma^2}\left(\frac{-1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}\right)p(x;\mu,\sigma)dx = \frac{-1}{\sigma^3}\mathbb{E}[X-\mu] + \frac{1}{\sigma^5}\mathbb{E}\left[(X-\mu)^3\right] \\ &= 0 + 0 = 0 = g_{21}(\theta); \end{aligned}$$

$$\begin{aligned} g_{22}(\theta) &= \int_{-\infty}^{\infty} \left(-\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3} \right)^2 p(x;\mu.\sigma)dx \\ &= \frac{1}{\sigma^2} \int p(x;\mu,\sigma) + -\frac{2}{\sigma^4} Var[X] + \frac{1}{\sigma^6} \mathbb{E}\left[(X-\mu)^4\right] = \frac{1}{\sigma^2} - \frac{2}{\sigma^2} + \frac{3\sigma^4}{\sigma^6} = \frac{2}{\sigma^2}. \end{aligned}$$

Portanto, a matriz de informação de Fisher de S em $\theta = (\mu, \sigma)$ é dada por

$$G(\theta) = \text{diag}(1/\sigma^2, 2/\sigma^2). \tag{2.11}$$

A seguir, veremos uma proposição que mostra que podemos escrever as entradas da matriz de informação de Fisher em (2.10) em termos da raiz quadrada das densidades de probabilidade e usaremos este resultado como apoio para demostração do Teorema 2.3. Ver O. Calin and C. Udrişte 2014 [10, p.22].

Proposição 2.5 ([10])**.** *A matriz de informação de Fisher pode ser representada em termos da raiz quadrada das densidades de probabilidade como*

$$g_{ij}(\theta) = 4 \int_{\mathcal{X}} \frac{\partial \sqrt{p_\theta(x)}}{\partial \theta_i} \cdot \frac{\partial \sqrt{p_\theta(x)}}{\partial \theta_j} dx.$$

Demonstração.

$$\begin{aligned} g_{ij}(\theta) &= \int_{\mathcal{X}} \frac{\partial \ln p_\theta(x)}{\partial \theta_i} \cdot \frac{\partial \ln p_\theta(x)}{\partial \theta_j} p_\theta(x)\, dx \\ &= \int_{\mathcal{X}} \frac{1}{p_\theta(x)} \frac{\partial p_\theta(x)}{\partial \theta_i} \cdot \frac{1}{p_\theta(x)} \frac{\partial p_\theta(x)}{\partial \theta_j} p_\theta(x)\, dx \\ &= 4 \int_{\mathcal{X}} \frac{1}{2 \cdot p_\theta(x)} \frac{\partial p_\theta(x)}{\partial \theta_i} \cdot \frac{1}{2} \frac{\partial p_\theta(x)}{\partial \theta_j} dx, \end{aligned}$$

ou seja,

$$\begin{aligned} g_{ij}(\theta) &= 4 \int_{\mathcal{X}} \frac{1}{2\sqrt{p_\theta(x)}} \frac{\partial p_\theta(x)}{\partial \theta_i} \cdot \frac{1}{2\sqrt{p_\theta(x)}} \frac{\partial p_\theta(x)}{\partial \theta_j} dx \\ &= 4 \int_{\mathcal{X}} \frac{\partial \sqrt{p_\theta(x)}}{\partial \theta_i} \cdot \frac{\partial \sqrt{p_\theta(x)}}{\partial \theta_j} dx. \end{aligned}$$

□

O Teorema 2.3 estabelece que a matriz de informação de Fisher gera uma métrica riemanniana em uma variedade estatística *S*, fazendo de *S*, uma variedade riemanniana. A ideia é mostrar que a referida matriz é simétrica e positiva definida. Ver O. Calin and C. Udrişte 2014 [10, Proposition 1.6.2].

Teorema 2.3 ([10])**.** *Seja S uma variedade estatística. Então a matriz de informação de Fisher define uma métrica riemanniana sobre S.*

Demonstração. Por hipótese, S é uma família paramétrica de densidades de probabilidade com a propriedade de ser regular. Fixemos um sistema de coordenadas em S, ou seja, $S = \{p_\theta\}$, com θ definida em num espaço de parâmetros $\Theta \subset \mathbb{R}^n$ aberto. Primeiro, vamos provar que $[g_{ij}(\theta)]_{n\times n}$ é uma matriz positiva definida. Note que a matriz de informação de Fiher é positiva semidefinida, pois para todo θ e $v \in T_pS$, $v \neq 0$ e pela Proposição 2.5, temos

$$\begin{aligned}
g_p(v,v) &= \langle v,v\rangle_p = v^T[g_{ij}(\theta)]v = \sum_{i,j=1}^{n} g_{ij}v_iv_j \\
&= \sum_{i,j=1}^{n}\left(4\int_{\mathcal{X}} \frac{\partial\sqrt{p_\theta(x)}}{\partial\theta_i}\cdot\frac{\partial\sqrt{p_\theta(x)}}{\partial\theta_j}\right)v_iv_j\,dx \\
&= 4\sum_{i,j=1}^{n}\left(\int_{\mathcal{X}} v_i\frac{\partial\sqrt{p_\theta(x)}}{\partial\theta_i}\cdot v_j\frac{\partial\sqrt{p_\theta(x)}}{\partial\theta_j}\right)dx \\
&= 4\int_{\mathcal{X}}\left(\sum_{i,j=1}^{n} v_i\frac{\partial\sqrt{p_\theta(x)}}{\partial\theta_i}\cdot v_j\frac{\partial\sqrt{p_\theta(x)}}{\partial\theta_j}\right)dx \\
&= 4\int_{\mathcal{X}}\left(\sum_{i=1}^{n} v_i\frac{\partial\sqrt{p_\theta(x)}}{\partial\theta_i}\right)\left(\sum_{j=1}^{n} v_j\frac{\partial\sqrt{p_\theta(x)}}{\partial\theta_j}\right)dx \\
&= 4\int_{\mathcal{X}}\left(\sum_{i=1}^{n} v_i\frac{\partial\sqrt{p_\theta(x)}}{\partial\theta_i}\right)^2 dx \geq 0.
\end{aligned}$$

Mostremos que g é não-degenerado. Pela hipótese, $p_\theta(x) > 0$ para todo x, $p_\theta(x)$ é suave e $\partial p_\theta(x)/\partial\theta_i\ \forall i = 1,...,n$ são linearmente independentes como funções de x. Daí

$$\begin{aligned}
g_p(v,v) = 0 \iff & \int_{\mathcal{X}}\left(\sum_{i=1}^{n} v_i\frac{\partial\sqrt{p_\theta(x)}}{\partial\theta_i}\right)^2 dx = 0 \iff \left(\sum_{i=1}^{n} v_i\frac{\partial\sqrt{p_\theta(x)}}{\partial\theta_i}\right)^2 = 0,\ \forall x\in\mathbb{R}^n \\
\iff & \sum_{i=1}^{n} v_i\frac{\partial\sqrt{p_\theta(x)}}{\partial\theta_i} = 0 \iff \sum_{i=1}^{n} v_i\frac{1}{2\sqrt{p_\theta(x)}}\frac{\partial p_\theta(x)}{\partial\theta_i} = 0 \\
\iff & \sum_{i=1}^{n} v_i\frac{\partial p_\theta(x)}{\partial\theta_i} = 0 \iff v_i = 0, \forall i = 1,...,n,
\end{aligned}$$

Como $[g_{ij}(\theta)]_{n\times n}$ é não-degenerado, garantimos que

$$4\int_{\mathcal{X}}\left(\sum_{i=1}^{n} v_i\frac{\partial\sqrt{p_\theta(x)}}{\partial\theta_i}\right)^2 dx > 0 \quad \text{quando} v \neq 0.$$

e portanto, $[g_{ij}(\theta)]_{n\times n}$ é positiva definida.

A simetria da matriz de informação de Fisher segue do fato de

$$\begin{aligned} g_{ij}(\theta) &= E_\theta\left(\frac{\partial}{\partial\theta_i}\log p(x;\theta)\frac{\partial}{\partial\theta_j}\log p(x;\theta)\right) = E_\theta\left(\frac{\partial}{\partial\theta_j}\log p(x;\theta)\frac{\partial}{\partial\theta_i}\log p(x;\theta)\right) \\ &= g_{ji}(\theta), \quad \forall i,j=1,...,n. \end{aligned}$$

Agora, dado $p = p_\theta$, considere vetores tangentes $u, v \in T_pS$ e escreva u e v em termos da base coordenada, $u = \sum_{i=1}^{n} u_i \frac{\partial}{\partial\theta_i}$ e $v = \sum_{j=1}^{n} v_j \frac{\partial}{\partial\theta_j}$. Usando que $\frac{\partial \ln p_\theta(x)}{\partial\theta_i} = \frac{1}{p_\theta(x)}\frac{\partial p_\theta(x)}{\partial\theta_i}$, temos o seguinte:

$$\begin{aligned} g_p(u,v) &= \sum_{i,j=1}^{n} g_{ij}(\theta)u_iv_j = \int_{\mathcal{X}}\left(\sum_{i,j=1}^{n} u_iv_j\frac{\partial \ln p_\theta(x)}{\partial\theta_i}\cdot\frac{\partial \ln p_\theta(x)}{\partial\theta_j}p_\theta(x)\right)dx \\ &= \int_{\mathcal{X}}\left(\sum_{i=1}^{n} u_i\frac{1}{p_\theta(x)}\frac{\partial p_\theta(x)}{\partial\theta_i}\right)\left(\sum_{j=1}^{n} v_j\frac{1}{p_\theta(x)}\frac{\partial p_\theta(x)}{\partial\theta_j}\right)p_\theta(x)\,dx \\ &= \int_{\mathcal{X}}\frac{u(p_\theta(x))}{p_\theta(x)}\cdot\frac{v(p_\theta(x))}{p_\theta(x)}p_\theta(x)\,dx. \end{aligned}$$

Da última igualdade, segue que a matriz de informação de Fisher determina o seguinte 2-tensor sobre S

$$g_p(u,v) = \int_{\mathcal{X}}\frac{u(p_\theta(x))}{p_\theta(x)}\cdot\frac{v(p_\theta(x))}{p_\theta(x)}p_\theta(x)\,dx,$$

que depende apenas do ponto p e dos vetores tangentes $u, v \in T_pS$. Como $[g_{ij}(\theta)]_n$ é uma matriz positiva definida e simétrica que varia suavemente com o parâmetro θ, segue que $g_p(u,v) = \sum_{i,j=1}^{n} g_{ij}(\theta)u_iv_j$ define uma métrica riemanniana em S. □

Essa métrica é chamada de métrica de Fisher-Rao, métrica de informação de Fisher ou simplesmente, métrica de Fisher. Pelo Teorema 2.3, segue que uma variedade estatística S sobre $\mathcal{X}$, de dimensão n, munida da métrica de Fisher, é uma variedade riemanniana, a qual denotaremos por variedade estatística riemanniana.

Portanto, a matriz de informação de Fisher fornece os coeficientes de uma métrica riemanniana na variedade S. Isso nos permite medir distâncias, ângulos e definir conexões em modelos estatísticos regulares.

A próxima fórmula é útil em aplicações práticas. Veremos uma outra forma de escrever as entradas da matriz de informação de Fisher dada em (2.10) em termos da esperança negativa das entradas da matriz Hessiana da função log-verossimilhança. Usaremos futuramente na

Proposição 3.3 para relacionar a divergência de Kullback-Leibler com a matriz de informação de Fisher.

Proposição 2.6 ([10], Proposition 1.6.3)**.** *A matriz de informação de Fisher pode ser escrita como a esperança negativa do Hessiano da função de log-verossimilhança*

$$g_{ij} = -\mathbb{E}_\theta \left[\frac{\partial^2 \ln(p_\theta)}{\partial \theta_i \partial \theta_j} \right].$$

Demonstração. Sabe-se que $\int_{\mathcal{X}} p(x;\theta)\, dx = 1$, então

$$\frac{\partial}{\partial \theta_i} \int_{\mathcal{X}} p(x;\theta)\, dx = 0,$$

que pelo item *iii*) da Definição 1.5 pode ser escrita como

$$\mathbb{E}_\theta \left[\frac{\partial}{\partial \theta_i} \ln p_\theta \right] = \int_{\mathcal{X}} \frac{\partial}{\partial \theta_i} \ln p(x;\theta) p(x;\theta)\, dx = 0.$$

Diferenciando em relação a $\dfrac{\partial}{\partial \theta_j}$, obtemos

$$\int_{\mathcal{X}} \frac{\partial}{\partial \theta_j} \frac{\partial}{\partial \theta_i} \ln p(x;\theta) p(x;\theta)\, dx + \int_{\mathcal{X}} \frac{\partial}{\partial \theta_i} \ln p(x;\theta) \frac{\partial}{\partial \theta_j} p(x;\theta)\, dx = 0$$

se, e somente se,

$$\mathbb{E}_\theta \left[\frac{\partial}{\partial \theta_j} \frac{\partial}{\partial \theta_i} \ln p_\theta \right] + \int_{\mathcal{X}} \frac{\partial}{\partial \theta_i} \ln p(x;\theta) \frac{\partial}{\partial \theta_j} \ln p(x;\theta) p(x;\theta)\, dx = 0$$

assim,

$$\mathbb{E}_\theta \left[\frac{\partial}{\partial \theta_j} \frac{\partial}{\partial \theta_i} \ln p_\theta \right] + g_{ij}(\theta) = 0.$$

□

A seguir, estudaremos a métrica de Fisher na variedade estatística normal multivariada.

2.2.1 Normal Multivariada

Nesta subseção, S será a variedade estatística formada por densidades normais multivariadas. Estudaremos o teorema a seguir, em que forneceremos a matriz de informação de Fisher de S apresentada na referencia por B. Porat and B. Friedlander em 1986 [28], sendo que aqui apresentamos uma prova mais detalhada e também veremos a métrica de Fisher em S fazendo de S uma variedade estatística riemanniana. Deixamos a prova do Teorema 2.4 no apêndice para que a leitura do texto fique mais fluida pois a demonstração é muito técnica.

Teorema 2.4 (Porat and Benjamin, [28]). *Seja a variedade estatística* $S = \{p(x;\mu,\mathrm{V}); x \in \mathbb{R}^n$ *e* $(\mu,\mathrm{V}) \in \mathbb{R}^n \times P_n(\mathbb{R})\}$*, então a matriz de informação de Fisher de S é*

$$g_{ij}(\theta) = \left(\frac{\partial \mu}{\partial \theta_i}\right)^T \Lambda \frac{\partial \mu}{\partial \theta_j} + \frac{1}{2} tr\left(\Lambda \frac{\partial \mathrm{V}}{\partial \theta_i} \Lambda \frac{\partial \mathrm{V}}{\partial \theta_j}\right). \tag{2.12}$$

Demonstração. Apresentamos uma prova mais detalhada no Apêndice A do que foi apresentada em B. Porat and B. Friedlander em 1986 [28, p. 128].

□

Dado $\theta \in \Theta$, o espaço tangente a Θ em θ é o conjunto $T_\theta\Theta = \{(x,A); x \in \mathbb{R}^n$ e $A \in S_n(\mathbb{R})\}$, em que $S_n(\mathbb{R})$ é o espaço das matrizes simétricas de ordem n com entradas reais. Sejam $V = (x,A)$ e $W = (y,B)$ vetores pertencentes a $T_\theta\Theta$, o produto interno no ponto $\theta = (\mu,\mathrm{V})$ associado a matriz de informação de Fisher $G(\theta)$, dada em (2.12) conforme J. P. S. Porto 2017 [29] é

$$\langle V,W\rangle_\theta = x^T \Lambda y + \frac{1}{2}\mathrm{tr}(\Lambda A \Lambda B). \tag{2.13}$$

A seguir, forneceremos o elemento de comprimento infinitesimal da métrica de Fisher em S usando (2.13).

Proposição 2.7. *Seja a variedade estatística* $S = \{p_\theta; \theta = (\mu,\mathrm{V}) \in \Theta = \mathbb{R}^n \times P_n(\mathbb{R})\}$*, então a métrica de Fisher em S é*

$$ds^2 = (d\mu)^T \Lambda d\mu + \frac{1}{2} tr[(\Lambda d\,\mathrm{V})^2].$$

Demonstração. Sejam p_{θ_1} e $p_{\theta_2} \in S$ parametrizadas por $\theta_1 = (\mu_1,\mathrm{V}_1)$ e $\theta_2 = (\mu_2,\mathrm{V}_2)$. Considere uma curva suave por partesγ em Θ conectando esses dois parâmetros, ou seja, $\gamma : [a,b] \to \Theta$ tal que $\gamma(t) = (\mu(t),\mathrm{V}(t))$, $t \in [a,b]$, sendo que $\gamma(a) = (\mu_1,\mathrm{V}_1)$ e $\gamma(b) = (\mu_2,\mathrm{V}_2)$. Assim, por (2.13), o elemento infinitesimal da métrica de Fisher é expresso da seguinte maneira

$$ds^2 = \langle \gamma'(t),\gamma'(t)\rangle_G = \langle (d\mu,d\,\mathrm{V}),(d\mu,d\,\mathrm{V})\rangle_G = (d\mu)^T \Lambda d\mu + \frac{1}{2}\mathrm{tr}[(\Lambda d\,\mathrm{V})^2].$$

Portanto, a equação da métrica da informação de Fisher de S é

$$ds^2 = (d\mu)^T \Lambda d\mu + \frac{1}{2}\mathrm{tr}[(\Lambda d\,\mathrm{V})^2].$$

sendo $d\mu = (d\mu_1,...,d\mu_n) \in \mathbb{R}^n$ e $d\,\mathrm{V} = [d\sigma_{ij}]_{n\times n} \in P_n(\mathbb{R})$ é a matriz cujas entradas são as derivadas das entradas correspondentes da matriz V.

□

Desse modo, temos que S é uma variedade estatística riemanniana com a métrica dada pela Proposição 2.7.

Seja γ uma curva suave por partes em Θ, definida no intervalo $[a,b] \subset \mathbb{R}$, dada por $\gamma(t) = (\mu(t), \mathrm{V}(t))$. A curva γ é geodésica em Θ se suas funções coordenadas satisfazem as seguintes equações,

$$\begin{cases} \dfrac{d^2\mu}{dt^2} - \left(\dfrac{d\mathrm{V}}{dt}\right)\Lambda\left(\dfrac{d\mu}{dt}\right) = 0 \\ \dfrac{d^2\mathrm{V}}{dt^2} + \left(\dfrac{d\mu}{dt}\right)\left(\dfrac{d\mu}{dt}\right)^T - \left(\dfrac{d\mathrm{V}}{dt}\right)\Lambda\left(\dfrac{d\mathrm{V}}{dt}\right) = 0. \end{cases} \tag{2.14}$$

Essas equações são obtidas calculando os símbolos de Cristoffel de S na parametrização Θ e substituindo na equação (1.6). Para mais detalhes consulte L. T. Skovgaard 1981 [33, Theorem 6.1] e L. T. Skovgaard 1984 [34, section 3].

2.3 Distância de Fisher-Rao

A distância de Fisher-Rao é uma medida de dissimilaridade entre duas densidades de probabilidade. Assim como outras medidas de divergência, isto é, formas de medir diferenças entre densidades de probabilidade, está relacionada à entropia e está no centro da área de pesquisa chamada Geometria da Informação.

Para o que se segue, denotaremos por

$$S = \{p_\theta = p(x;\theta); x \in \mathcal{X} \text{ e } \theta \in \Theta \subset \mathbb{R}^n\}$$

uma variedade estatística riemanniana munida da métrica de Fisher de dimensão n. Dadas as densidades p_{θ_1} e p_{θ_2} em S, identificadas pelos parâmetros θ_1 e θ_2, defina

$$d_F(p_{\theta_1}, p_{\theta_2}) \equiv d_F(\theta_1, \theta_2) = \inf\left\{\int_a^b \|\gamma'(t)\|_G \, dt, \ \gamma \in \Gamma_a^b\right\} \tag{2.15}$$

sendo que

$$\Gamma_a^b = \{\gamma; \ \gamma : [a,b] \to \Theta \text{ é curva suave por partes, } \gamma(a) = \theta_1, \ \gamma(b) = \theta_2\}$$

e

$$\|\gamma'(t)\|_G = \sqrt{\langle \gamma'(t), \gamma'(t)\rangle_G} = \sqrt{\gamma'(t)^T G(\theta)\gamma'(t)},$$

sendo $\gamma'(t)^T$ vetor transposto e $G(\theta)$ é a matriz de informação de Fisher dada em (2.10).

Note que $\gamma(t) = (\gamma_1(t), ..., \gamma_n(t)) = \theta(t)$ e relembre que pela métrica de Fisher dada pela equação

$$ds^2 = \sum_{i,j=1}^{n} g_{ij}(\theta) d\theta_i d\theta_j.$$

Temos que o comprimento de arco de uma curva γ entre θ_1 e θ_2 é dada por

$$\ell(\gamma) \;=\; \int_a^b \|\gamma'(t)\|_G dt = \int_a^b \left(\langle \gamma'(t), \gamma'(t)\rangle_G\right)^{1/2} dt = \int_a^b \left(\gamma'(t)^T G(\theta)\gamma'(t)\right)^{1/2} dt.$$

A curva que minimiza esse comprimento é chamado de curva geodésica. Admitiremos que o espaço paramétrico é um aberto conexo do $\mathbb{R}^n$. A seguir veremos que a função d_F está bem definida.

Lema 2.6. *A função $d_F : S \times S \to \mathbb{R}$ dada por* (2.15) *está bem definida.*

Demonstração. Sabemos que a aplicação $\varphi : \Theta \to S$ é uma parametrização global C^∞, a qual é um homeomorfismo diferenciável. O domínio Θ é um aberto conexo, logo por φ ser um homomorfismo, S é uma variedade estatística conexa e por consequência é conexa por caminhos. Assim, dadas quaisquer duas densidades em S, essas podem ser conectados por uma curva suave por partes e pela identificabilidade de S, existe a respectiva curva suave por partes conectando os parâmetros em Θ, logo d_F está bem definida. □

Agora veremos através do Lema 2.7 que d_F define uma métrica em S que será importante para a teoria do Capítulo 4.

Lema 2.7. *A função $d_F : S \times S \to \mathbb{R}$ dada por* (2.15) *define uma métrica na variedade estatística riemanniana S, isto é, satisfaz*

i) $d_F(p_{\theta_1}, p_{\theta_2}) \geq 0$ *e* $d_F(p_{\theta_1}, p_{\theta_2}) = 0$ *se, e somente se,* $p_{\theta_1} = p_{\theta_2}$*;*

ii) $d_F(p_{\theta_1}, p_{\theta_2}) = d_F(p_{\theta_2}, p_{\theta_1})$*;*

iii) $d_F(p_{\theta_1}, p_{\theta_3}) \leq d_F(p_{\theta_1}, p_{\theta_2}) + d_F(p_{\theta_2}, p_{\theta_3})$.

O resultado acima é clássico na Geometria Riemanniana. Sua demonstração detalhada requer outros conceitos e resultados dessa área, cujos pormenores fogem aos objetivos desta dissertação. Para mais informações, ver: M. P. do Carmo 2015 [12, 2.5 Proposição p. 161] ou J. Jost 2008 [17, Lemma 1.4.1. p. 16].

O Lema anterior assegura que o par (S, d_F) é um espaço métrico, em que S é uma variedade estatística riemanniana munida da métrica de Fisher.

A seguir, definiremos a distância de Fisher-Rao dada por (2.15) que inicialmente foi introduzida por C. R. Rao em [30] como uma medida adequada para o cálculo da distância entre duas populações.

Definição 2.2 (Distância de Fisher-Rao, [27])**.** *Dadas duas densidades de probabilidade p_{θ_1} e p_{θ_2} na variedade estatística riemanniana S munida da métrica de Fisher, a* distância de Fisher-Rao *entre elas é dada por* (2.15)*. Através da identificabilidade entre cada densidade e o respectivo parâmetro, por simplicidade, escreveremos*

$$d_F(p_{\theta_1}, p_{\theta_2}) = d_F(\theta_1, \theta_2).$$

Na prática é muito difícil o cálculo da distância de Fisher-Rao para grande parte das densidades de probabilidade, uma vez que envolve a solução de equações diferenciais de segunda ordem. Em alguns casos, podemos simplificar o cálculo dessa distância relacionando a métrica do espaço com a métrica de espaços já conhecidos (por exemplo, o espaço Euclidiano, hiperbólico ou esférico). C. Atkinson and A. F. Mitchell em 1981 [5] e J. Burbea em 1984 [9] calcularam a distância de Fisher-Rao entre algumas densidades de probabilidade já conhecidas: densidade de Poisson, Multinomial, Gamma, normal, entre outras. Na Subseção 2.3.1 calcularemos a distância entre duas densidades exponenciais e na Subsecção 2.3.2, descreveremos a distância de Fisher-Rao no espaço das densidades normais univariadas e veremos que a métrica de Fisher nesse espaço está relacionada com a métrica do espaço hiperbólico. No caso do espaço formado por densidades normais multivariadas, ainda não se tem uma fórmula fechada para a distância de Fisher-Rao no caso geral.

2.3.1 Densidade da Exponencial

Nesta subseção, veremos um exemplo onde calcularemos a distância de Fisher-Rao entre duas densidades de probabilidade exponenciais, as quais sabemos que a variedade estatística riemanniana composta por essas densidades tem dimensão 1.

Para tanto, seja $S = \{p_\theta; \theta > 0\}$ a variedade estatística riemanniana formada por densidades exponenciais, ou seja, se $p_\theta \in S$, então p_θ é uma população de $X \sim \text{Exp}(\theta)$ e é dada por

$$p(x;\theta) = \theta e^{-\theta x}, \quad x \geq 0.$$

Dadas duas densidades p_{θ_1} e p_{θ_2} em S identificadas por θ_1 e θ_2 respectivamente, pelo Exemplo 2.1 a informação de Fisher de S em θ é

$$G(\theta) = \frac{1}{\theta^2}.$$

Assim,

$$\begin{aligned}\int_a^b \|\gamma'\|_G\, dt &= \int_a^b \sqrt{\gamma'(t)\frac{1}{\theta^2}\gamma'(t)}\, dt = \int_a^b \left|\frac{\gamma'(t)}{\theta(t)}\right| dt = \int_{\theta_1}^{\theta_2} \frac{\theta'(t)}{u}\frac{du}{\theta'(t)} = \int_{\theta_1}^{\theta_2} \frac{1}{u}\, du \\ &= \ln\left|\frac{\theta_2}{\theta_1}\right|.\end{aligned}$$

Portanto,

$$d_F(p_{\theta_1}, p_{\theta_2}) \equiv d_F(\theta_1,\theta_2) = \inf\left\{\int_a^b \|\gamma'(t)\|_G\, dt,\ \gamma \in \Gamma_a^b\right\} = \ln\left|\frac{\theta_2}{\theta_1}\right|.$$

2.3.2 Normal Univariada

Nesta subseção, apresentaremos uma fórmula fechada para distância de Fisher-Rao entre densidades normais univariadas com o apoio do plano hiperbólico de Poincaré, relacionando métrica do modelo formado por densidades normais univariadas com a métrica do espaço hiperbólico. Baseamos nas referências S. I. Costa, S. A. Santos and J. E. Strapasson 2015 [11] e C. Atkinson and A. F. Mitchell 1981 [5]. Aqui, apresentaremos o que foi feito nesses referidos artigos. No Capítulo 4, Seção 4.2, usaremos este estudo da distância entre normais univariadas.

Seja S a variedade estatística riemanniana formada por densidades normais univariadas com média μ e variância σ^2, a qual sabemos que as densidade de probabilidade são dadas por

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

Neste caso, o parâmetro θ é dado por duas variáveis reais $\theta = (\mu,\sigma) \in \Theta = \mathbb{R}\times(0,\infty)$.

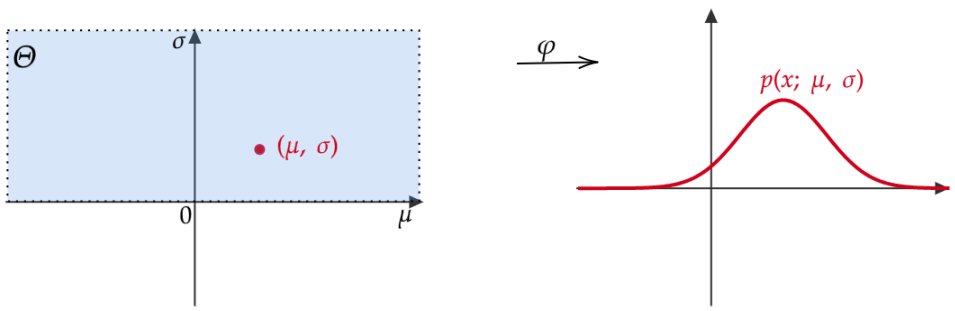

Figura 2.2 Representação da densidade da normal univariada.
Produzida online com Mathcha.

Já vimos no Exemplo 2.2 que a matriz de informação de Fisher de S em $\theta = (\mu, \sigma)$ é dada por

$$G(\theta) = \text{diag}(1/\sigma^2, 2/\sigma^2),$$

a expressão da métrica de Fisher é dada por

$$ds^2 = \sum_{i,j=1}^{2} g_{ij}(\theta) d\theta_i d\theta_j = g_{11}(\theta) d\theta_1 d\theta_1 + g_{22}(\theta) d\theta_2 d\theta_2 = \frac{(d\mu)^2 + 2(d\sigma)^2}{\sigma^2}.$$

Essa métrica é muito similar a métrica dada no plano hiperbólico, ver Definição 1.22 . Sabemos que a matriz da métrica no modelo do plano hiperbólico de Poincaré, $\mathcal{H}^2 = \{(x,y) \in \mathbb{R}^2; y > 0\}$, obtida por

$$G_p(x,y) = \text{diag}(1/y^2, 1/y^2),$$

a expressão da métrica é

$$ds_1^2 = \frac{(dx)^2 + (dy)^2}{y^2}.$$

No espaço S, uma fórmula fechada para a distância de Fisher-Rao é conhecida via uma associação com o modelo do plano hiperbólico. A métrica de Fisher

$$ds^2 = \frac{(d\mu)^2 + 2(d\sigma)^2}{\sigma^2},$$

é redutível à métrica do plano hiperbólico. O método de redução a uma métrica de uma geometria hiperbólica procede da seguinte forma

$$x = \frac{\mu}{\sqrt{2}}, y = \sigma \text{ e } ds_1^2 = \frac{ds^2}{2} \implies ds_1^2 = \frac{(dx)^2 + (dy)^2}{y^2}.$$

Portanto, a métrica de Fisher da densidade normal univariada é essencialmente duas vezes a métrica do plano da metade superior de Poincaré com a seguinte mudança de variáveis

$$x = \frac{\mu}{\sqrt{2}}, y = \sigma.$$

Assim,

$$\begin{aligned} ds^2 &= \frac{(d\mu)^2 + 2(d\sigma)^2}{\sigma^2} = \frac{2}{\sigma^2}\left(\frac{(d\mu)^2}{2} + (d\sigma)^2\right) = \frac{2}{\sigma^2}\left(\frac{(d(x\sqrt{2}))^2}{2} + (dy)^2\right) \\ &= \frac{2}{\sigma^2}\left(\frac{\left(\sqrt{2}\right)^2 (dx)^2}{2} + (dy)^2\right) = \frac{2}{\sigma^2}\left((dx)^2 + (dy)^2\right) = 2ds_1^2. \end{aligned}$$

Logo, $ds^2/2 = ds_1^2$. Portanto, segue que a métrica em Θ está relacionada com a métrica de $\mathcal{H}^2$ através da transformação

$$\begin{aligned} f : \Theta &\to \mathcal{H}^2 \\ (\mu, \sigma) &\mapsto \left(\frac{\mu}{\sqrt{2}}, \sigma\right) \end{aligned}$$

A distância de Fisher-Rao entre os pontos $\theta_1 = (\mu_1, \sigma_1)$ e $\theta_2 = (\mu_2, \sigma_2)$ pertencentes a Θ, pode ser expressa em termos da distância hiperbólica de Poincaré, $d_{\mathcal{H}^2}$. Como

$$\begin{aligned} d_F(\theta_1, \theta_2) &= \inf\left\{\int_a^b \left(ds^2\right)^{1/2} dt, \gamma \in \Gamma_a^b\right\} = \inf\left\{\int_a^b \left(2ds_1^2\right)^{1/2} dt, \bar{\gamma} \in \bar{\Gamma}_a^b\right\} \\ &= \sqrt{2}\inf\left\{\int_a^b \left(ds_1^2\right)^{1/2} dt, \bar{\gamma} \in \bar{\Gamma}_a^b\right\} = \sqrt{2}\, d_{\mathcal{H}^2}\left(\left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right), \left(\frac{\mu_2}{\sqrt{2}}, \sigma_2\right)\right), \end{aligned}$$

em que

$$\bar{\Gamma}_a^b = \left\{\bar{\gamma}; \bar{\gamma} \text{ é uma curva suave por partes tal que } \bar{\gamma}(a) = \left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) \text{ e } \bar{\gamma}(b) = \left(\frac{\mu_2}{\sqrt{2}}, \sigma_2\right)\right\}.$$

Pela Equação (1.7), temos uma expressão analítica para d_F pode ser adquirida por

$$\begin{aligned} & d_F(p_{\theta_1}, p_{\theta_2}) = d_F((\mu_1, \sigma_1), (\mu_2, \sigma_2)) \\ = \ & \sqrt{2}\log\left(\frac{\left\|\left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}}, -\sigma_2\right)\right\| + \left\|\left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}}, \sigma_2\right)\right\|}{\left\|\left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}}, -\sigma_2\right)\right\| - \left\|\left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}}, \sigma_2\right)\right\|}\right) \end{aligned}$$

ou pela equação (1.8) também pode ser representada por

$$\begin{aligned} d_F(p_{\theta_1}, p_{\theta_2}) &= d_F((\mu_1, \sigma_1), (\mu_2, \sigma_2)) \\ &= \sqrt{2}\cosh^{-1}\left(\frac{(\mu_2 - \mu_1)^2 + 2(\sigma_1^2 + \sigma_2^2)}{4\sigma_1\sigma_2}\right). \end{aligned} \tag{2.16}$$

As geodésicas de Θ são as imagens inversas, por meio da transformação f, das geodésicas de $\mathcal{H}^2$. Essas geodésicas são as semirretas verticais positivas $\gamma_1 : (0, \infty) \to \Theta$ e as semielipses $\gamma_2 : (0, \pi) \to \Theta$ centradas em $\sigma = 0$ com excentricidade $\dfrac{1}{\sqrt{2}}$ dadas por

$$\gamma_1(t) = (\sqrt{2}t_0, t) \text{ e } \gamma_2(t) = (\sqrt{2}(\rho\cos(t) + c), \rho\sin(t)). \tag{2.17}$$

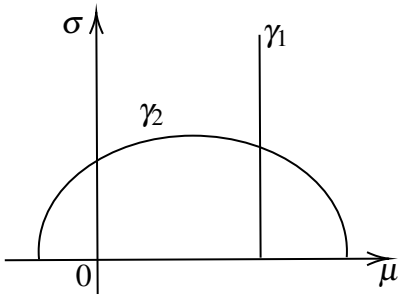

Figura 2.3 Geodésicas em Θ com a métrica de Fisher.
Produzida online com Mathcha.

Subvariedade $\widehat{S}$ de S

Considerando a subvariedade $\widehat{S}$ de S, a qual provaremos para o caso geral multivariado no Lema 2.8, formada pelas densidades normais com média constante, $\widehat{S} = \{p(x;\mu,\sigma); \mu = \mu_0$ constante, $\sigma \in (0,\infty)\}$, temos que a distância de Fisher-Rao entre duas densidades dessa subvariedade parametrizadas por (μ_0,σ_1) é (μ_0,σ_2) é

$$d_{F_\mu}((\mu_0,\sigma_1),(\mu_0,\sigma_2)) = \sqrt{2}\ln\left|\frac{\sigma_2}{\sigma_1}\right|, \tag{2.18}$$

pois a métrica $ds^2 = \dfrac{(d\mu)^2+2(d\sigma)^2}{\sigma^2}$ se reduz para $ds^2 = \dfrac{2(d\sigma)^2}{\sigma^2}$. Daí,

$$\begin{aligned} d_{F_\mu}((\mu_0,\sigma_1),(\mu_0,\sigma_2)) &= \inf\left\{\int_a^b \left(ds^2\right)^{1/2} dt, \gamma \in \Gamma_a^b\right\} \\ &= \inf\left\{\int_a^b \left(\frac{2d\sigma^2}{\sigma^2}\right)^{1/2} dt, \gamma \in \Gamma_a^b\right\} \\ &= \sqrt{2}\inf\left\{\int_a^b \left(\frac{d\sigma}{\sigma}\right) dt, \gamma \in \Gamma_a^b\right\} = \sqrt{2}\inf\int_{\sigma(a)}^{\sigma(b)} \frac{1}{u}\,du \\ &= \sqrt{2}\left(\ln|u||_{\sigma_1}^{\sigma_2}\right) = \sqrt{2}\left(\ln|\sigma_2| - \ln|\sigma_1|\right) = \sqrt{2}\ln\left|\frac{\sigma_2}{\sigma_1}\right|. \end{aligned}$$

As curvas geodésicas em Θ_μ são as semirretas verticais positivas e, portanto, $\widehat{S}$ é uma subvariedade totalmente geodésica. Em outras palavras, a distância em Fisher-Rao restrita à subvariedade $\widehat{S}$ é igual à distância na variedade S, $d_F = d_{F_\mu}$.

Subvariedade $\overline{S}$ de S

Considerando agora a subvariedade $\overline{S}$ de S formada pelas densidades normais com desvio padrão constante, $\overline{S} = \{p(x;\mu,\sigma); \sigma = \sigma_0 \text{ constante}, \mu \in \mathbb{R}\}$, temos que a distância de Fiher-Rao entre duas densidades dessa subvariedade parametrizadas por (μ_1,σ_0) e (μ_2,σ_0) é

$$d_{F_\sigma}((\mu_1,\sigma_0),(\mu_2,\sigma_0)) = \frac{|\mu_1-\mu_2|}{\sigma_0},$$

já que a métrica $ds^2 = \dfrac{(d\mu)^2+2(d\sigma)^2}{\sigma^2}$ reduz para $ds^2 = \dfrac{(d\mu)^2}{\sigma_0^2}$. Daí,

$$\begin{aligned} d_{F_\sigma}((\mu_1,\sigma_0),(\mu_2,\sigma_0)) &= \inf\left\{\int_a^b \left(ds^2\right)^{1/2} dt, \gamma \in \Gamma_a^b\right\} \\ &= \inf\left\{\int_a^b \left(\frac{(d\mu)^2}{\sigma_0^2}\right)^{1/2} dt, \gamma \in \Gamma_a^b\right\} \\ &= \inf\left\{\int_a^b \left(\frac{d\mu}{\sigma_0}\right) dt, \gamma \in \Gamma_a^b\right\}, \end{aligned}$$

ou seja,

$$\begin{aligned} d_{F_\sigma}((\mu_1,\sigma_0),(\mu_2,\sigma_0)) &= \frac{1}{\sigma_0}\inf\left\{\int_a^b (d\mu)\, dt, \gamma \in \Gamma_a^b\right\} \\ &= \frac{1}{\sigma_0}(\mu(b)-\mu(a)) = \frac{1}{\sigma_0}(\mu_2-\mu_1). \end{aligned}$$

Considere sem perda de generalidade que

$$d_{F_\sigma}((\mu_1,\sigma_0),(\mu_2,\sigma_0)) = \frac{|\mu_2-\mu_1|}{\sigma_0}.$$

Diferente de $\widehat{S}$, a subvariedade $\overline{S}$ não é totalmente geodésica. O espaço Θ_σ é formado por retas horizontais paralelas ao eixo μ, as quais não são geodésicas em Θ com a métrica de Fisher.

Dados dois pontos $\theta_1 = (\mu_1,\sigma)$ e $\theta_2 = (\mu_2,\sigma)$ em Θ_σ, por (2.16) temos que

$$\begin{aligned} d_F((\mu_1,\sigma),(\mu_2,\sigma)) &= \sqrt{2}\cosh^{-1}\left(\frac{(\mu_2-\mu_1)^2+4\sigma^2}{4\sigma^2}\right). \\ &\leq \frac{|\mu_1-\mu_2|}{\sigma} = d_{F_\sigma}((\mu_1,\sigma),(\mu_2,\sigma)). \end{aligned}$$

A seguir, veremos um exemplo onde calculamos a distância de Fisher-Rao entre duas densidades normais univariadas parametrizadas por $\theta_1 = (1.5, 0.75)$ e $\theta_2 = (3.5, 0.75)$ e

compararemos com a distância euclidiana. Esses parâmetros específicos foram retirados do artigo S. I. Costa, S. A. Santos and J. E. Strapasson em 2015 [11].

Exemplo 2.3. *A Figura 2.4, ilustra a geodésica em* $\Theta = \mathbb{R} \times (0, \infty)$ *com a métrica de Fisher conectando* θ_1 *e* θ_2 *e as respectivas densidades em S, em que* θ *é um parâmetro qualquer entre* θ_1 *e* θ_2 *na geodésica às conectando. Essa geodésica é uma semielipse dada por* (2.17) *sendo* $c = 1.76776$ *e* $\rho = 1.03077$.

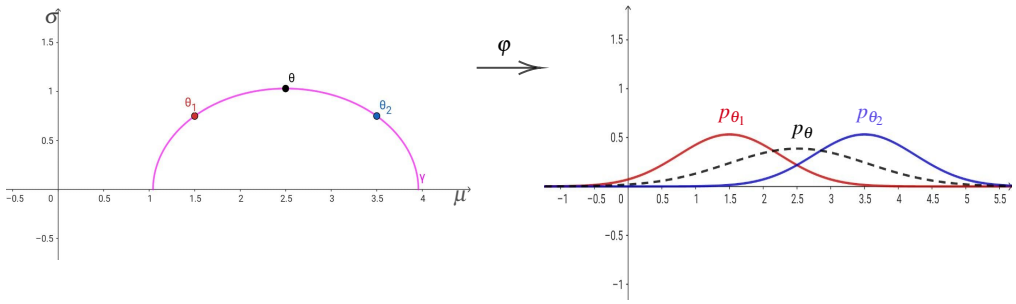

Figura 2.4 Semielipse em Θ conectando θ_1 e θ_2 e as respectivas densidades em S.
Produzida com GeoGebra.

A distância de Fisher-Rao entre $p(x; 1.5, 0.75)$ *e* $p(x; 3.5, 0.75)$ *usando* (2.16) *é dado por*

$$\begin{aligned} d_F(p_{\theta_1}, p_{\theta_2}) &= \sqrt{2}\cosh^{-1}\left(\frac{(3.5-1.5)^2 + 2(0.75^2 + 0.75^2)}{2.25}\right) \\ &= \sqrt{2}\cosh^{-1}\left(\frac{4+2.25}{2.25}\right) = \sqrt{2}\cosh^{-1}(2.77777778) \\ &= 2.37687. \end{aligned}$$

Note que a distância euclidiana é calcular o comprimento de um segmento de reta que liga θ_1 *e* θ_2, *assim*

$$d(\theta_1, \theta_2) = 2.$$

Daí, temos

$$d(\theta_1, \theta_2) < d_F(\theta_1, \theta_2).$$

A seguir, estudaremos a distância de Fisher-Rao e as geodésicas de algumas subvariedades da variedade composta por densidades normais multivariadas, a qual usaremos na Seção 4.2 do Capítulo 4.

2.3.3 Em Subvariedades da Normal Multivariada

Nesta subseção, veremos algumas subvariedades da normal multivariada e a distância de Fisher-Rao estudadas em J. Pinele, J. E. Strapasson and S. I. Costa 2020 [27, p. 5] e em J. P. S. Porto 2017 [29]. Consideramos subvariedades $S_* \subset S$ com a distância induzida pela métrica de Fisher em S. É importante observar que, em geral, dadas duas densidades p_{θ_1} e p_{θ_2} em S_*, a distância entre p_{θ_1} e p_{θ_2}, quando restrita a uma subvariedade S_* é maior que a distância entre p_{θ_1} e p_{θ_2} em S, ou seja, $d_{S_*}(p_{\theta_1}, p_{\theta_2}) \geqslant d_S(p_{\theta_1}, p_{\theta_2})$. Isso se deve ao fato de que para obter d_{S_*}, consideramos o comprimento mínimo das curvas restritas, que são aquelas contidas na subvariedade S_*. Conforme a Definição 1.21, dizemos que S_* é totalmente geodésico se, e somente se, $d_{S_*}(p_{\theta_1}, p_{\theta_2}) = d_S(p_{\theta_1}, p_{\theta_2})$, para qualquer $p_{\theta_1}, p_{\theta_2} \in S_*$, o que significa que a geodésica em S, conectando θ_1 e θ_2, está contida em S_*. Na variedade n-dimensional composta por densidades normais multivariadas com vetor de médias comum μ_0, denotada por

$$S_\mu = \{p_\theta; \theta = (\mu, \mathrm{V}) \in \Theta, \mu = \mu_0 \in \mathbb{R}^n \text{ constante}\},$$

é uma subvariedade de S. Veja o Lema 2.8.

Observação 2.3. S_μ é um modelo estatístico regular e a demonstração é análoga ao caso geral S.

Lema 2.8. *A variedade estatística S_μ é uma subvariedade de S.*

Demonstração. De maneira análoga à prova de que S é variedade, prova-se que S_μ é variedade de dimensão $n(n+1)/2$. Claramente $S_\mu \subset S$, então considere a função inclusão

$$\begin{aligned} i : S_\mu &\hookrightarrow S \\ p_\theta &\mapsto p_\theta \end{aligned}$$

sendo que $\theta = (\mu_0, \mathrm{V}(t))$. Mostraremos que i é um mergulho.
Para cada $p_\theta \in S_\mu$ e $v \in T_{p_{\theta_\mu}} S_\mu$, considere uma curva suave por partes α em S_μ dada da seguinte forma

$$\begin{aligned} \alpha : [-\varepsilon, \varepsilon] &\to S_\mu \\ t &\mapsto \alpha(t) = p_\theta. \end{aligned}$$

Com $\alpha(0) = p_{\theta_0}, \theta_0 = (\mu_0, \mathrm{V}(0))$ e $\alpha'(0) = v$. Veja a Figura 2.5.

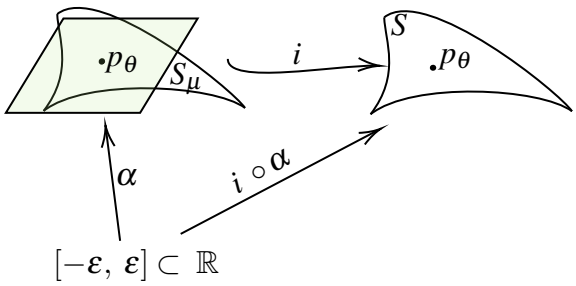

Figura 2.5 Representação da aplicação inclusão.
Produzida online com Mathcha.

Daí, pela Proposição 1.4, temos que

$$di_{p_\theta}(v) = \frac{d}{dt}(i(\alpha(t))\Big|_{t=0} = \frac{d}{dt}\alpha(t)\Big|_{t=0} = v,$$

é injetiva para todo $p_\theta \in S_\mu$. Além disso, como S_μ é identificável, segue que i é injetora e é sobrejetora sobre sua imagem. Pelo fato de S_μ ser um modelo regular, segue que $p_\theta \mapsto i(p_\theta) = p_\theta$ é contínua e sua inversa também é contínua. Logo, i é um homeomorfismo de S_μ sobre o subespaço $i(S_\mu) = S_\mu \subset S$. Portanto, i é um mergulho. □

As demais variedades que aparecerão nas próximas subseções são subvariedades de S e a demonstração é análoga ao caso S_μ.

A subvariedade S_μ em que μ é constante

Seja a subvariedade de S dada por $S_\mu = \{p_\theta; \theta = (\mu, \mathrm{V}), \mu = \mu_0 \in \mathbb{R}^n \text{constante}\}$ de dimensão $n(n+1)/2$ composta por densidades que possuem o mesmo vetor de médias μ_0. Apresentaremos aqui o teorema dado por S. T. Jensen em 1976, no qual ele determina a distância na subvariedade S_μ.

Teorema 2.5 ([5])**.** *Considere a família de densidades normais multivariadas S_μ com o vetor da média comum μ_0, mas com diferentes matrizes de covariância* V*. Dados dois elementos dessa família, parametrizados por $\theta_1 = (\mu_0, \mathrm{V}_1)$ e $\theta_2 = (\mu_0, \mathrm{V}_2)$, a distância entre dois elementos dessa família é obtida por*

$$d_\mu(\theta_1, \theta_2) = \sqrt{\frac{1}{2}\sum_{i=1}^{n}[\log(\lambda_i)]^2},$$

em que $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ são os autovalores de $\Lambda_1 \mathrm{V}_2$.

Demonstração. A demonstração desse teorema pode ser encontrada em C. Atkinson and A. F. Mitchell 1981 [5, Appendix 1]. □

Observe que as equações que determinam as geodésicas de Θ, determinadas em (2.14), quando restritas à $\Theta_\mu = \{(\mu_0, \mathrm{V}); \mu_0 \in \mathbb{R}^n \text{constante}\}$ se reduzem a

$$\frac{d^2\mathrm{V}}{dt^2} - \left(\frac{d\mathrm{V}}{dt}\right)\mathrm{V}^{-1}\left(\frac{d\mathrm{V}}{dt}\right) = 0.$$

A curva $\gamma_\mu(t) = (\mu(t), \mathrm{V}(t))$ que satisfaz a equação acima ligando dois pontos $\theta_1 = (\mu_0, \mathrm{V}_1)$ e $\theta_2 = (\mu_0, \mathrm{V}_2)$ em Θ_μ, com $\gamma_\mu(0) = \theta_1$ e $\gamma_\mu(1) = \theta_2$, é dada por

$$\gamma_\mu(t) = (\mu_0, \mathrm{V}_1^{1/2}\exp(t\log(\mathrm{V}_1^{-1/2}\mathrm{V}_2\mathrm{V}_1^{-1/2}))\mathrm{V}_1^{1/2}).$$

para todo $t \in [0,1]$, ver J. Pinele, J. E. Strapasson and S. I. Costa 2020 [27].

A subvariedade S_μ é uma subvariedade totalmente geodésica, ver L. T. Skovgaard 1984 [34, p. 214]. Logo $d_\mu(\theta_1, \theta_2) = d_F(\theta_1, \theta_2)$ para todo $\theta_1, \theta_2 \in S_\mu$.

A subvariedade S_V em que V é constante

Seja $S_\mathrm{V} = \{p_\theta; \theta = (\mu, \mathrm{V}) \in \Theta, \mathrm{V} = \mathrm{V}_0 \in P_n(\mathbb{R}) \text{constante}\}$, uma subvariedade de dimensão n composta por densidades normais multivariadas com matriz de covariância V comum. Nesse espaço podemos relacionar a métrica de Fisher com a métrica de um espaço Euclidiano. A métrica de Fisher em S_V é

$$ds^2 = (d\mu)^T \Lambda d\mu.$$

Como $\mathrm{V} = \mathrm{V}_0$ é uma matriz simétrica definida positiva, pela Fatoração de Cholesky dada no Lema 1.2, temos que existe, e é única, uma matriz triangular inferior P de ordem n tal que

$$\mathrm{V}_0 = PP^T \implies P^T\Lambda_0 P = I_n. \tag{2.19}$$

em que I_n é a matriz identidade de ordem n. Seja agora um vetor $\nu = (\nu_1, \nu_2, ..., \nu_n) \in \mathbb{R}^n$ tal que

$$\mu = P\nu \implies d\mu = Pd\nu.$$

Sendo assim,

$$\begin{aligned} ds^2 &= (d\mu)^T\Lambda d\mu = (Pd\nu)^T\Lambda Pd\nu = (d\nu)^T P^T\Lambda Pd\nu = (d\nu)^T d\nu \\ &= (d\nu_1, ..., d\nu_n)^T(d\nu_1, ..., d\nu_n) = \sum_{i=1}^{n}(d\nu_i)^2, \end{aligned}$$

ou seja, a métrica de Fisher em S_{V} coincide com a métrica do espaço Euclidiano sob a transformação $\mu \mapsto P^{-1}\mu$.

Assim, para ν_1 e $\nu_2 \in \mathbb{R}^n$,

$$d(\nu_1,\nu_2) = \sqrt{(\nu_1-\nu_2)^T(\nu_1-\nu_2)}.$$

Portanto, segue que a distância Fisher–Rao entre duas densidades normais multivariadas parametrizadas por $\theta_1 = (\mu_1, \mathrm{V}_0)$ e $\theta_2 = (\mu_2, \mathrm{V}_0)$, utilizando (2.19) é

$$\begin{aligned} d_{\mathrm{V}}(\theta_1,\theta_2) &= d(\nu_1,\nu_2) = \sqrt{\left(P^{-1}(\mu_1-\mu_2)\right)^T\left(P^{-1}(\mu_1-\mu_2)\right)} \\ &= \sqrt{(\mu_1-\mu_2)^T P^{-t}P^{-1}(\mu_1-\mu_2)} \\ &= \sqrt{(\mu_1-\mu_2)^T\Lambda_0(\mu_1-\mu_2)}. \end{aligned}$$

A distância dada pela equação acima é igual à distância de Mahalanobis, ver referência P. C. Mahalanobis 1936 [24]. P. C. Mahalanobis foi um dos pioneiros no estudo sobre distâncias entre distribuições de probabilidade. Essa distância foi uma das primeiras medidas de dissimilaridade entre conjuntos de dados com alguma correlação entre variáveis.

Esta subvariedade não é totalmente geodésica, ver J. Pinele, J. E. Strapasson and S. I. Costa 2020 [27, p. 5]. Assim,

$$d_{\mathrm{V}}(\theta_1,\theta_2) \geq d_F(\theta_1,\theta_2).$$

Uma curva geodésica $\gamma_{\mathrm{V}}(t)$ tal que $\gamma_{\mathrm{V}}(0) = \theta_1$ e $\gamma_{\mathrm{V}}(1) = \theta_2$ em Θ_{V} pode ser fornecida por, ver J. Pinele, J. E. Strapasson and S. I. Costa 2020 [27]:

$$\gamma_{\mathrm{V}}(t) = ((1-t)\mu_1 + t\mu_2, \mathrm{V}_0).$$

A subvariedade S_D em que V é diagonal

Seja $S_D = \{p_\theta; \theta = (\mu,\mathrm{V}) \in \Theta, \mathrm{V} = \mathrm{diag}(\sigma_1^2,\sigma_2^2,...,\sigma_n^2), \sigma_i > 0, i = 1,...n\}$ uma subvariedade de S formada pelas densidades cuja matriz de covariância é uma matriz diagonal. A métrica de Fisher de S_D é

$$\begin{aligned} ds^2 &= (d\mu)^T\Lambda d\mu + \frac{1}{2}\mathrm{tr}[(\Lambda d\mathrm{V})^2] = \sum_{i=1}^n \frac{1}{\sigma_i^2}(d\mu_i)^2 + \frac{1}{2}\mathrm{tr}[\Lambda^2(d\mathrm{V})^2] \\ &= \sum_{i=1}^n \frac{1}{\sigma_i^2}(d\mu_i)^2 + \frac{1}{2}\sum_{i=1}^n \frac{1}{\sigma_i^4}(2\sigma_i d\sigma_i)^2 = \sum_{i=1}^n \frac{(d\mu_i)^2}{\sigma_i^2} + \frac{1}{2}\sum_{i=1}^n \frac{4\sigma_i^2(d\sigma_i)^2}{\sigma_i^4} \\ &= \sum_{i=1}^n \frac{(d\mu_i)^2}{\sigma_i^2} + 2\sum_{i=1}^n \frac{(d\sigma_i)^2}{\sigma_i^2} = \sum_{i=1}^n \frac{(d\mu_i)^2 + 2(d\sigma_i)^2}{\sigma_i^2}. \end{aligned}$$

A subvariedade S_D é parametrizada por uma interseção de espaços metade do $\mathbb{R}^{2n}$ com $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2, ..., \mu_n, \sigma_n)$, a matriz de informação de Fisher dada pelo Teorema 2.4 em S_D é

$$\text{diag}\left(1/\sigma_1^2, 2/\sigma_1^2, 1/\sigma_2^2, 2/\sigma_2^2, ..., 1/\sigma_n^2, 2/\sigma_n^2\right).$$

Observe que $\Theta_D = \{(\mu, \mathrm{V}); \mathrm{V} = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2), \sigma_i > 0, i = 1, ...n\}$ é um espaço de dimensão $2n$. Como a métrica em $S = \{p(x; \mu, \sigma); (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+^*\}$, o modelo estatístico formado por densidades normais univariadas, está relacionada com a métrica do modelo do plano superior de Poincaré $\mathcal{H}^2$, a métrica em Θ_D está relacionada com a métrica produto no espaço produto $(\mathcal{H}^2)^n = \mathcal{H}^2 \times \cdots \times \mathcal{H}^2$, n vezes.

Essa relação é dada pela transformação

$$\begin{gathered} \phi : \Theta_D \to \mathcal{H}^{2n} \\ (\mu_1, \sigma_1, \mu_2, \sigma_2, ..., \mu_n, \sigma_n) \mapsto \left(\frac{\mu_1}{\sqrt{2}}, \sigma_1, \frac{\mu_2}{\sqrt{2}}, \sigma_2, ..., \frac{\mu_n}{\sqrt{2}}, \sigma_n\right) \end{gathered}$$

Dados $\theta_1 = (\mu_{11}, \sigma_{11}, \mu_{12}, \sigma_{12}, ..., \mu_{1n}, \sigma_{1n})$ e $\theta_2 = (\mu_{21}, \sigma_{21}, \mu_{22}, \sigma_{22}, ..., \mu_{2n}, \sigma_{2n})$, a distância entre duas densidades de probabilidade com esses parâmetros é dada por

$$\begin{aligned} d_D(\theta_1, \theta_2) &= d_D((\mu_{11}, \sigma_{11}, \mu_{12}, \sigma_{12}, ..., \mu_{1n}, \sigma_{1n}), (\mu_{21}, \sigma_{21}, \mu_{22}, \sigma_{22}, ..., \mu_{2n}, \sigma_{2n})) \\ &= \sqrt{2} d_{\mathcal{H}^{2n}}\left(\left(\frac{\mu_{11}}{\sqrt{2}}, \sigma_{11}, ..., \frac{\mu_{1n}}{\sqrt{2}}, \sigma_{1n}\right), \left(\frac{\mu_{21}}{\sqrt{2}}, \sigma_{21}, ..., \frac{\mu_{2n}}{\sqrt{2}}, \sigma_{2n})\right)\right), \end{aligned}$$

ou seja,

$$\begin{aligned} d_D(\theta_1, \theta_2) &= \sqrt{2 \sum_{i=1}^{n} \left(d_{\mathcal{H}^2}\left(\left(\frac{\mu_{1i}}{\sqrt{2}}, \sigma_{1i}\right), \left(\frac{\mu_{2i}}{\sqrt{2}}, \sigma_{2i}\right)\right)\right)^2} \\ &= \sqrt{2 \sum_{i=1}^{n} \cosh^{-2}\left(1 + \frac{(\mu_{2i} - \mu_{1i})^2 + 2(\sigma_{1i} - \sigma_{2i})^2}{4\sigma_{1i}\sigma_{2i}}\right)} \\ &= \sqrt{2 \sum_{i=1}^{n} \cosh^{-2}\left(\frac{(\mu_{2i} - \mu_{1i})^2 + 2(\sigma_{1i}^2 + \sigma_{2i}^2)}{4\sigma_{1i}\sigma_{2i}}\right)} \end{aligned} \tag{2.20}$$

Neste espaço, uma curva $\gamma_D(t) = (\gamma_1(t), \gamma_2(t), ..., \gamma_n(t))$ é uma geodésica se, e somente se, $\gamma_i(t)$ é uma curva geodésica curva no caso univariado, para todo $i = 1, ..., n$. As curvas geodésicas no espaço paramétrico (meio plano superior $\mathbb{R} \times \mathbb{R}^+$) das densidades normais univariadas são semirretas verticais e semielipses centradas em $\sigma = 0$, com excentricidade $1/\sqrt{2}$. É importante notar que $S_D \subset S$ não é totalmente geodésico. A subvariedade de S_D composta apenas por densidades normais com matrizes de covariância que são múltiplos da identidade é totalmente geodésica, ver [11].

Capítulo 3

Divergência de Kullback-Leibler

Outra medida de dissimilaridade (com relação a distância de Fisher-Rao) entre duas densidades de probabilidade que estudaremos é a divergência de Kullback-Leibler, que é usada na Teoria da Informação e comumente referida como a entropia relativa de probabilidade entre duas densidades. Neste capítulo, veremos o conceito de divergência de Kullback-Leibler e como ela se relaciona com a métrica de Fisher e a distância de Fisher-Rao. Este capítulo é baseado em O. Calin and C. Udrişte 2014 [10, Chapter 4].

Dadas duas densidades p e q na mesma variedade estatística, a divergência de kullback-Leibler, D_{KL}, é dada por

$$D_{KL}(p\|q) = \int_{\mathcal{X}} p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx$$

no caso contínuo, e

$$D_{KL}(p\|q) = \sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{q_i}\right)$$

no caso discreto, e é referido p_i simplesmente por $p(x_i)$ em que $x_i \in \mathcal{X}$ um conjunto discreto, para todo $i = 1, ..., n$. Vamos nos concentrar neste trabalho ao caso contínuo.

Na teoria da informação, a densidade p é considerada a verdadeira densidade determinada a partir de observações, enquanto q é a densidade teórica do modelo. A entropia relativa de Kullback-Leibler pode ser usada para encontrar uma qualidade de ajuste dessas duas densidades dada pelo valor esperado de uma informação extra necessária para codificar usando q em vez de usar p. Às vezes, a entropia relativa de Kullback-Leibler é considerada como uma medida de ineficiência de assumir dados distribuídos de acordo com q, quando na verdade é distribuído como p. Os resultados e demostrações apresentados neste capítulo podem ser encontrados em O. Calin and C. Udrişte 2014 [10].

Observação 3.1. Em alguns trabalhos $D_{KL}(p\|q)$ é chamada de distância. No entanto, formalmente ela não define uma métrica sobre as densidades, pois não é simétrica, nem satisfaz a desigualdade triangular, mas é não-negativa e não-degenerada.

A seguir, veremos um lema e um teorema que servirá de base para mostrar que a divergência de Kullback-Leibler é positiva e não-degenerada no item *i*) da Proposição 3.1.

Lema 3.1 (Lemma 4.1.1, [10])**.** *Se p e q são duas densidades contínuas estritamente positivas em $\mathcal{X}$, então*

$$\int_{\mathcal{X}} p(x)\ln p(x)dx \geq \int_{\mathcal{X}} p(x)\ln q(x)dx.$$

A desigualdade anterior torna-se igualdade quando as densidades são iguais.

Demonstração. Usando que $x > 0$ e a seguinte desigualdade $\ln x \leq x-1$, temos a seguinte igualdade

$$\begin{aligned}\int_{\mathcal{X}} p(x)\ln q(x)dx - \int_{\mathcal{X}} p(x)\ln p(x)dx &= \int_{\mathcal{X}} p(x)\ln\left(\frac{q(x)}{p(x)}\right)dx \leq \int_{\mathcal{X}} p(x)\left(\frac{q(x)}{p(x)}-1\right)dx \\ &= \int_{\mathcal{X}} q(x)\,dx - \int_{\mathcal{X}} p(x)\,dx = 0.\end{aligned}$$

A igualdade é alcançada para $q(x) = p(x),$ para todo $x \in \mathcal{X}$, isto é, caso de densidades iguais. □

Utilizaremos o Teorema 3.1 a seguir, para verificar que a divergência de Kullback-Leibler é não-degenerada na Proposição 3.1.

Teorema 3.1 (Desigualdade de Jensen, [21])**.** *Se φ é uma função convexa definida sobre um intervalo I, e X é uma variável aleatória com $\mathbb{P}(X \in I) = 1$ e esperança finita, então*

$$\varphi[\mathbb{E}(X)] \leq \mathbb{E}[\varphi(X)].$$

Se φ é estritamente convexo, a desigualdade é estrita, a menos que X seja uma constante com probabilidade 1.

Demonstração. Ver E. L. Lehmann and G. Casella 1998 [21, Theorem 7.5, p. 46]. □

Agora, fornecemos algumas condições em relação as densidades de probabilidade para que a divergência de Kullback-Leibler seja distância.

Proposição 3.1 ([10])**.** *Seja S uma variedade estatística. Então,*

i) A divergência de Kullback-Leibler é positiva e não-degenerada: $D_{KL}(p\|q) \geq 0$, para todo $p,q \in S$, com $D_{KL}(p\|q) = 0$ se, e somente se, $p = q$.

ii) A divergência de Kullback-Leibler é simétrica, ou seja, $D_{KL}(p\|q) = D_{KL}(q\|p)$ se, e somente se,

$$\int_{\mathcal{X}} [p(x) + q(x)] \ln\left(\frac{p(x)}{q(x)}\right) dx = 0.$$

iii) A divergência de Kullback-Leibler satisfaz a desigualdade triangular $D_{KL}(p\|q) + D_{KL}(q\|r) \geq D_{KL}(p\|r)$ se, e somente se,

$$\int_{\mathcal{X}} [p(x) - q(x)] \ln\left(\frac{q(x)}{r(x)}\right) dx \leq 0.$$

Demonstração. i) Primeiramente mostraremos que a divergência de Kulback-Leibler é positiva. Para tanto, aplicando o Lema 3.1, obtemos

$$\begin{aligned} D_{KL}(p\|q) &= \int_{\mathcal{X}} p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx = \int_{\mathcal{X}} p(x) \left(\ln p(x) - \ln q(x)\right) dx \\ &= \int_{\mathcal{X}} p(x) \ln p(x)\, dx - \int_{\mathcal{X}} p(x) \ln q(x)\, dx \geq 0. \end{aligned} \tag{3.1}$$

Agora, mostraremos que é não-degenerada. Com efeito, se $p = q$, então $p(x) = q(x)$, para todo $x \in \mathcal{X}$. Daí, $\ln(p(x)/q(x)) = 0$. Logo, $D_{KL}(p\|q) = 0$. Por outro lado, isto é, se $D_{KL}(p\|q) = 0$, mostraremos que $p = q$. Iniciamos escrevendo a divergência de Kulback-Laibler em termos de esperança com respeito a densidade p da seguinte forma

$$D_{KL}(p\|q) = \mathbb{E}_p\left[\ln(p(X)/q(X))\right]$$

e sabemos que por (3.1)

$$\mathbb{E}_p\left[\ln(p(X)/q(X))\right] \geq 0.$$

Relembrando que a função $-\ln$ é estritamente convexa e da desigualdade de Jensen segue $\mathbb{E}[\ln(X)] < \ln[\mathbb{E}(X)]$. Pelo Teorema 3.1 a igualdade ocorre se a variável aleatória $Y = (p(X)/q(X))$ é constante e existe $c \in \mathbb{R}$ com $\mathbb{P}(Y = c) = 1$. Logo,

$$\mathbb{E}_p\left[\ln(p(X)/q(X))\right] = 0 \implies \ln(p(X)/q(X)) = 0 \implies p(X)/q(X) = 1 \implies p(X) = q(X).$$

Portando, $p = q$.

ii) Se $D_{KL}(p\|q) = D_{KL}(q\|p)$, então

$$\int_{\mathcal{X}} p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx = \int_{\mathcal{X}} q(x) \ln\left(\frac{q(x)}{p(x)}\right) dx.$$

Daí,

$$\int_{\mathcal{X}} p(x)\ln\left(\frac{p(x)}{q(x)}\right)dx - \int_{\mathcal{X}} q(x)\ln\left(\frac{q(x)}{p(x)}\right)dx = 0,$$

assim,

$$\int_{\mathcal{X}} \left[p(x)\ln\left(\frac{p(x)}{q(x)}\right) - q(x)\ln\left(\frac{q(x)}{p(x)}\right)\right]dx = 0.$$

Logo,

$$\int_{\mathcal{X}} \left[p(x)\ln\left(\frac{p(x)}{q(x)}\right) + q(x)\ln\left(\frac{p(x)}{q(x)}\right)\right]dx = 0.$$

Agora, caso

$$\int_{\mathcal{X}} [p(x)+q(x)]\ln\left(\frac{p(x)}{q(x)}\right)dx = 0,$$

temos

$$\int_{\mathcal{X}} p(x)\ln\left(\frac{p(x)}{q(x)}\right)dx + \int_{\mathcal{X}} q(x)\ln\left(\frac{p(x)}{q(x)}\right)dx = 0,$$

daí,

$$D_{KL}(p\|q) = -\int_{\mathcal{X}} q(x)\ln\left(\frac{p(x)}{q(x)}\right)dx = +\int_{\mathcal{X}} q(x)\ln\left(\frac{q(x)}{p(x)}\right)dx = D_{KL}(q\|p).$$

iii) Se $D_{KL}(p\|q) + D_{KL}(q\|r) \geq D_{KL}(p\|r)$, então

$$\int_{\mathcal{X}} p(x)\ln\left(\frac{p(x)}{q(x)}\right)dx + \int_{\mathcal{X}} q(x)\ln\left(\frac{q(x)}{r(x)}\right)dx \geq \int_{\mathcal{X}} p(x)\ln\left(\frac{p(x)}{r(x)}\right)dx$$

assim,

$$\int_{\mathcal{X}} p(x)\ln\left(\frac{p(x)}{r(x)}\right)dx - \int_{\mathcal{X}} p(x)\ln\left(\frac{p(x)}{q(x)}\right)dx - \int_{\mathcal{X}} q(x)\ln\left(\frac{q(x)}{r(x)}\right)dx \leq 0$$

consequentemente,

$$\int_{\mathcal{X}} p(x)\left[\ln\left(\frac{p(x)}{r(x)}\right) + \ln\left(\frac{q(x)}{p(x)}\right)\right]dx - \int_{\mathcal{X}} q(x)\ln\left(\frac{q(x)}{r(x)}\right)dx \leq 0$$

logo,

$$\int_{\mathcal{X}} p(x)\ln\left(\frac{q(x)}{r(x)}\right)dx - \int_{\mathcal{X}} q(x)\ln\left(\frac{q(x)}{r(x)}\right)dx \leq 0$$

portanto,

$$\int_{\mathcal{X}} (p(x)-q(x))\ln\left(\frac{q(x)}{r(x)}\right)dx \leq 0.$$

Agora, se

$$\int_{\mathcal{X}} [p(x) - q(x)] \ln\left(\frac{q(x)}{r(x)}\right) dx \leq 0,$$

temos,

$$\int_{\mathcal{X}} p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx + \int_{\mathcal{X}} p(x) \ln\left(\frac{q(x)}{r(x)}\right) dx - \int_{\mathcal{X}} q(x) \ln\left(\frac{q(x)}{r(x)}\right) dx \leq \int_{\mathcal{X}} p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx$$

com isso,

$$\int_{\mathcal{X}} p(x) \ln\left(\frac{p(x)}{r(x)}\right) dx \leq \int_{\mathcal{X}} q(x) \ln\left(\frac{q(x)}{r(x)}\right) dx + \int_{\mathcal{X}} p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx$$

logo,

$$D_{KL}(p\|r) \leq D_{KL}(q\|r) + D_{KL}(p\|q).$$

□

A Proposição 3.1 mostra que a divergência de Kullback-Leibler não satisfaz todos os axiomas de uma métrica na variedade estatística S, salvo sob algumas restrições. Vale a pena notar que a não-simetria pode ser removida definindo uma versão simétrica da entropia relativa, $\mathcal{D}(p,q) = \left(D_{KL}(p\|q) + D_{KL}(q\|p)\right)/2$ chamado quase métrica.

A seguir, calcularemos a divergência de Kullback-Leibler entre duas densidades exponenciais e veremos que não vale a propriedade de simetria e desigualdade triangular.

Exemplo 3.1 ([10])**.** *Seja uma variável aleatória $X \sim Exp(\theta)$. Calcularemos a entropia relativa de Kullback-Leibler para pares de densidades na mesma classe de densidades exponenciais.*
Considere

$$p_1(x) = \theta_1 e^{-\theta_1 x} \text{ e } p_2(x) = \theta_2 e^{-\theta_2 x}$$

duas densidades exponenciais. Note que,

$$\begin{aligned} \ln\left(\frac{p_1(x)}{p_2(x)}\right) &= \ln\left(\frac{\theta_1 e^{-\theta_1 x}}{\theta_2 e^{-\theta_2 x}}\right) = \ln\left(\frac{\theta_1}{\theta_2}\right) + \ln\left(\frac{e^{-\theta_1 x}}{e^{-\theta_2 x}}\right) = \ln\left(\frac{\theta_1}{\theta_2}\right) + (-\theta_1 x + \theta_2 x) \\ &= \ln\left(\frac{\theta_1}{\theta_2}\right) + (-\theta_1 + \theta_2)x. \end{aligned}$$

Assim,

$$\begin{aligned} D_{KL}(p_1\|p_2) &= \int_0^\infty p_1(x)\ln\left(\frac{p_1(x)}{p_2(x)}\right)dx = \int_0^\infty p_1(x)\left(\ln\left(\frac{\theta_1}{\theta_2}\right)+(\theta_2-\theta_1)x\right)dx \\ &= \int_0^\infty p_1(x)\ln\left(\frac{\theta_1}{\theta_2}\right)dx + \int_0^\infty (\theta_2-\theta_1)x\,p_1(x)\,dx \\ &= \ln\left(\frac{\theta_1}{\theta_2}\right)+(\theta_2-\theta_1)\int_0^\infty x\theta_1 e^{-\theta_1 x}dx = \ln\left(\frac{\theta_1}{\theta_2}\right)+(\theta_2-\theta_1)\mathbb{E}(X) \\ &= \ln\left(\frac{\theta_1}{\theta_2}\right)+(\theta_2-\theta_1)\frac{1}{\theta_1} = \ln\left(\frac{\theta_1}{\theta_2}\right)+\frac{\theta_2}{\theta_1}-1 = \frac{\theta_2}{\theta_1}-\ln\left(\frac{\theta_2}{\theta_1}\right)-1. \end{aligned}$$

Daí,

$$D_{KL}(p_1\|p_2) = f\left(\frac{\theta_2}{\theta_1}\right),$$

com $f(x) = x-\ln(x)-1 \geq 0$ *para todo* $x \in (0,\infty)$.

Logo $D_{KL}(p_1\|p_2) \geq 0$, *e* $D_{KL}(p_1\|p_2) = 0$ *se, e somente se,* $\theta_1 = \theta_2$, *ou seja,* $p_1 = p_2$. *Agora note que*

$$D_{KL}(p_1\|p_2) = f\left(\frac{\theta_2}{\theta_1}\right) \neq f\left(\frac{\theta_1}{\theta_2}\right) = D_{KL}(p_2\|p_1),$$

não é simétrico e veja que não satisfaz a condição de simetria da Proposição 3.1, isto é,

$$\int_0^\infty [p_1(x)+p_2(x)]\ln\left(\frac{p_1(x)}{p_2(x)}\right)dx = 2\ln\left(\frac{\theta_1}{\theta_2}\right)+\frac{\theta_2}{\theta_1}-\frac{\theta_1}{\theta_2} \neq 0, \quad \forall \theta_1 \neq \theta_2.$$

Também não vale a desigualdade triangular

$$f\left(\frac{\theta_2}{\theta_1}\right)+f\left(\frac{\theta_3}{\theta_2}\right) \not\geq f\left(\frac{\theta_3}{\theta_1}\right) \implies \frac{\theta_2}{\theta_1}+\frac{\theta_3}{\theta_2} \not\geq \frac{\theta_3}{\theta_1}.$$

Pode-se ver que essa relação não vale para qualquer $\theta_1,\theta_2,\theta_3 > 0$. *Vejamos também que não satisfaz a condição de desigualdade triangular da Proposição 3.1, isto pois*

$$\int_{\mathcal{X}} [p_1(x)-p_2(x)]\ln\left(\frac{p_2(x)}{p_3(x)}\right)dx = \frac{\theta_2}{\theta_1}-\frac{\theta_3}{\theta_2},$$

e essa relação não é menor ou igual a zero para qualquer $\theta_1,\theta_2,\theta_3 > 0$.

A seção a seguir é sobre a relação da divergência de Kullback-Leibler com a distância de Fisher-Rao.

3.1 Relação com a distância de Fisher-Rao

Começaremos com uma motivação da relação com a métrica de Fisher. A divergência de Kullback-Leibler de suas densidades exponenciais p_{θ_1} e p_{θ_2} é

$$D_{KL}(p_{\theta_1}\|p_{\theta_2}) = \frac{\theta_2}{\theta_1} - \ln\frac{\theta_2}{\theta_1} - 1, \quad \theta_1,\theta_2 > 0,$$

(ver Exemplo 3.1). As duas primeiras derivadas em relação a θ_2 são

$$\begin{aligned}\frac{\partial}{\partial\theta_2}D_{KL}(p_{\theta_1}\|p_{\theta_2}) &= \frac{\theta_1}{(\theta_1)^2} - \frac{\theta_1}{\theta_2}\frac{\theta_1}{\theta_1^2} = \frac{1}{\theta_1} - \frac{1}{\theta_2},\\ \frac{\partial^2}{\partial\theta_2\partial\theta_2}D_{KL}(p_{\theta_1}\|p_{\theta_2}) &= \frac{1}{(\theta_2)^2}.\end{aligned}$$

Notamos que as partes diagonais dessas derivadas parciais, obtidas para $\theta_2 = \theta_1$, são

$$\frac{\partial}{\partial\theta_2}D_{KL}(p_{\theta_1}\|p_{\theta_2})\bigg|_{\theta_2=\theta_1} = 0, \quad \frac{\partial^2}{\partial\theta_2\partial\theta_2}D_{KL}(p_{\theta_1}\|p_{\theta_2})\bigg|_{\theta_2=\theta_1} = \frac{1}{(\theta_1)^2} = g_{11}(\theta_1)$$

sendo g_{11} é a informação de Fisher conforme visto no Exemplo 2.1.

Essas duas relações vistas acima não são uma coincidência, como os dois resultados a seguir mostrarão.

Proposição 3.2 ([10])**.** *Sejam p_θ e p_α pertencentes a mesma variedade estatística parametrizadas por θ e α respectivamente e $D_{KL}(p_\theta\|p_\alpha)$ a divergência de Kullback-Leibler. Então vale*

$$\frac{\partial}{\partial\alpha_i}D_{KL}(p_\theta\|p_\alpha)\bigg|_{\alpha=\theta} = 0$$

Demonstração. A diferenciação na definição da entropia relativa de Kullback-Leibler, resulta em

$$\begin{aligned}\frac{\partial}{\partial\alpha_i}D_{KL}(p_\theta\|p_\alpha) &= \frac{\partial}{\partial\alpha_i}\int_{\mathcal{X}} p(x;\theta)\ln p(x;\theta)\,dx - \frac{\partial}{\partial\alpha_i}\int_{\mathcal{X}} p(x;\theta)\ln p(x;\alpha)\,dx\\ &= -\int_{\mathcal{X}} p(x;\theta)\frac{\partial}{\partial\alpha_i}\ln p(x;\alpha)\,dx.\end{aligned}$$

Portanto,

$$\frac{\partial}{\partial\alpha_i}D_{KL}(p_\theta\|p_\alpha)\bigg|_{\alpha=\theta} = -\int_{\mathcal{X}} p(x;\theta)\frac{\partial}{\partial\theta_i}\ln p(x;\theta)\,dx = -\frac{\partial}{\partial\theta_i}\int_{\mathcal{X}} p(x;\theta)\,dx = 0.$$

□

Proposição 3.3 ([10]). *A matriz Hessiana diagonal da divergência de Kullback-Leibler é a métrica de Fisher*

$$g_{ij}(\theta) = \left.\frac{\partial^2 D_{KL}(p_\theta \| p_\alpha)}{\partial \alpha_i \partial \alpha_j}\right|_{\alpha=\theta}$$

Demonstração. A diferenciação na definição da divergência de Kullback-Leibler implica

$$\begin{aligned}\frac{\partial^2 D_{KL}(p_\theta \| p_\alpha)}{\partial \alpha_i \partial \alpha_j} &= \frac{\partial^2}{\partial \alpha_i \partial \alpha_j}\int_{\mathcal{X}} p_\theta \ln(p_\theta)\, dx - \frac{\partial^2}{\partial \alpha_i \partial \alpha_j}\int_{\mathcal{X}} p_\theta \ln(p_\alpha)\, dx \\ &= -\int_{\mathcal{X}} p_\theta \frac{\partial^2 \ln(p_\alpha)}{\partial \alpha_i \partial \alpha_j}\, dx.\end{aligned}$$

Tomando o valor diagonal em $\alpha = \theta$ e usando a Proposição 2.6, temos

$$\left.\frac{\partial^2 D_{KL}(p_\theta \| p_\alpha)}{\partial \alpha_i \partial \alpha_j}\right|_{\alpha=\theta} = -\int_{\mathcal{X}} p_\theta \frac{\partial^2 \ln(p_\theta)}{\partial \theta_i \partial \theta_j}\, dx = -\mathbb{E}_\theta\left[\frac{\partial^2 \ln(p_\theta)}{\partial \theta_i \partial \theta_j}\right] = g_{ij}(\theta).$$

□

A Proposição 3.2 afirma que p_θ é um ponto crítico para o mapeamento $p_\alpha \to D_{KL}(p_\theta \| p_\alpha)$. Usando a Proposição 3.3 obtemos o seguinte resultado, que é específico para funções de distância.

Proposição 3.4 ([10]). *A densidade p_θ é um ponto de mínimo para o funcional*

$$p_\alpha \to D_{KL}(p_\theta \| p_\alpha).$$

Demonstração. Como g_{ij} é definida positiva, pela Proposição 3.3, o ponto p_θ é um mínimo local. Da não-negatividade da divergência de Kullback-Leibler, segue-se que p_θ é de fato um mínimo global. □

O próximo resultado trata da aproximação quadrática da divergência de Kullback-Leibler em termos da métrica de Fisher.

Teorema 3.2 ([10]). *Dada uma variedade estatística $S = \{p_\theta; \theta \in \Theta\}$ e seja $\Delta\alpha_i = \alpha_i - \theta_i$. Então*

$$D_{KL}(p_\theta \| p_\alpha) = \frac{1}{2}\sum_{i,j} g_{ij}(\theta)\Delta\alpha_i \Delta\alpha_j + R\left(\|\Delta\alpha\|^2\right), \tag{3.2}$$

em que $g_{ij}(\theta)$ são as entradas da matriz de informação de Fisher vista em (2.10).

Demonstração. Seja $f : \Theta \to \mathbb{R}$ um função definida por $f(\alpha) = D_{KL}(p_\theta \| p_\alpha)$, e considere a aproximação quadrática da f como sendo:

$$f(\alpha) = f(\theta) + \sum_i \frac{\partial f}{\partial \alpha_i}(\theta)\Delta\alpha_i + \frac{1}{2}\sum_{i,j}\frac{\partial^2 f}{\partial\alpha_i\partial\alpha_j}(\theta)\Delta\alpha_i\Delta\alpha_j + R\left(\|\Delta\alpha\|^2\right), \tag{3.3}$$

em que

$$\frac{f(\alpha)-\psi(\theta)}{\|\Delta\alpha\|^2} = \frac{R\left(\|\Delta\alpha\|^2\right)}{\|\Delta\alpha\|^2} \longrightarrow 0 \quad \text{quando } \|\Delta\alpha\| \to 0,$$

sendo $\psi(\theta) = f(\theta) + \sum_i \frac{\partial f}{\partial \alpha_i}(\theta)\Delta\alpha_i + \frac{1}{2}\sum_{i,j}\frac{\partial^2 f}{\partial\alpha_i\partial\alpha_j}(\theta)\Delta\alpha_i\Delta\alpha_j$.

Usando as Proposições 3.1, 3.2, e 3.3, temos

$$\begin{aligned} f(\theta) &= D_{KL}(p_\theta||p_\theta) = 0; \\ \frac{\partial f}{\partial\alpha_i}(\theta) &= \frac{\partial}{\partial\alpha_i}D_{KL}(p_\theta\|p_\alpha)\Big|_{\alpha=\theta} = 0; \\ \frac{\partial^2 f}{\partial\alpha_i\partial\alpha_j}(\theta) &= \frac{\partial^2 D_{KL}(p_\theta\|p_\alpha)}{\partial\alpha_i\partial\alpha_j}\Big|_{\alpha=\theta} = g_{ij}(\theta). \end{aligned}$$

Substituindo em (3.3) obtemos (3.2) e assim obtemos o resultado desejado. □

Sejam $p, q \in S$ duas densidades na variedade estatística S. A distância de Fisher-Rao representa a distância de informação entre as densidades p e q é definida como o comprimento da curva mais curta em S entre p e q, ou seja, o comprimento da curva geodésica que une essas populações. Em seguida, investigaremos a relação entre a entropia relativa de Kullback-Leibler $D_{KL}(p\|q)$ e a distância de Fisher-Rao $d_F(p,q)$. Começaremos com um exemplo. A partir da Subseção 2.3.1 e do Exemplo 3.1, a distância de Fisher-Rao e a entropia relativa de Kullback-Leibler entre duas densidades exponenciais são

$$d_F(p_{\theta_1}, p_{\theta_2}) = \ln\left(\frac{\theta_2}{\theta_1}\right) \quad \text{e} \quad D_{KL}\left(p_{\theta_1}\|p_{p_{\theta_2}}\right) = \frac{\theta_2}{\theta_1} - \ln\left(\frac{\theta_2}{\theta_1}\right) - 1$$

em que $0 < \theta_1 < \theta_2$. Assim, temos

$$\lim_{\theta_2 \searrow \theta_1} \frac{D_{KL}(p_{\theta_1} \| p_{\theta_2})}{\frac{1}{2} d_F(p_{\theta_1}, p_{\theta_2})^2} = \lim_{\theta_2 \searrow \theta_1} \frac{\frac{\theta_2}{\theta_1} - \ln\left(\frac{\theta_2}{\theta_1}\right) - 1}{\frac{1}{2}\left(\ln\left(\frac{\theta_2}{\theta_1}\right)\right)^2} = \lim_{x \searrow 1} \frac{x - \ln x - 1}{\frac{1}{2}(\ln x)^2}$$
$$= \lim_{u \searrow 0} \frac{e^u - u - 1}{\frac{1}{2}u^2} = 1,$$

pela regra de L'Hôspital. Assim, a assintótica de $D_{KL}(p_{\theta_1} \| p_{\theta_2})$ a medida que $\theta_2 \to \theta_1$ é $1/2 d_F(p_{\theta_1}, p_{\theta_2})^2$. Este resultado será válido num resultado mais geral. O próximo resultado é uma variante do Teorema 3.2, a qual relaciona a Kullback-Leibler com a distância de Fisher-Rao.

Teorema 3.3 ([10]). *Sejam $d_F(p,q)$ a distância de Fisher-Rao entre as densidades p e q e $D_{KL}(p\|q)$ a divergência de Kullback-Leibler. Então*

$$D_{KL}(p\|q) = \frac{1}{2} d_F^2(p,q) + R\left(d_F^2(p,q)\right),$$

em que

$$\frac{D_{KL}(p\|q) - 1/2 d_F^2(p,q)}{d_F^2(p,q)} = \frac{R\left(d_F^2(p,q)\right)}{d_F^2(p,q)} \longrightarrow 0 \quad \textit{quando } d_F(p,q) \to 0.$$

Demonstração. Considere uma geodésica $\xi : [0,t] \to S$ na variedade estatística riemanniana S conectando as densidades $p_\theta = p$ e $p_\alpha = q$ parametrizadas por θ e α, respectivamente. Esta curva satisfaz $\xi(s) = (\imath \circ \gamma)(s) = \imath(\gamma(s)) = \imath(\theta(s)) = p_{\theta(s)}$, com $\xi(0) = p_{\theta(0)} = p_\theta = p$ e $\xi(t) = p_{\alpha(t)} = p_\alpha = q$. A curva $\gamma(s) = \theta(s)$ pertence ao espaço de parâmetros cuja as extremidades são $\theta(0) = \theta$ e $\theta(t) = \alpha$. Veja Figura 3.1.

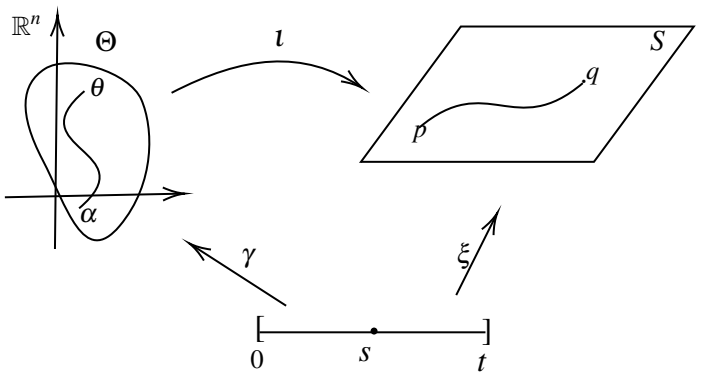

Figura 3.1 Curva Geodésica em S.
Produzida online com Mathcha.

Pela Definição 1.18, ξ pode ser normalizada através de uma parametrização por comprimento de arco e como o comprimento do arco ao longo da geodésica é a distância riemanniana, temos $t = d_F(p,q)$.

Considere a função $\varphi(s) = f(\theta(s))$, com $f(\alpha) = D_{KL}(p_\theta \| p_\alpha)$. Uma expansão de segunda ordem de φ em torno $t = 0$ resulta

$$\varphi(t) = \varphi(0) + t\varphi'(0) + \frac{t^2}{2}\varphi''(0) + R\left(t^2\right), \tag{3.4}$$

em que

$$\frac{\varphi(t) - \phi(0)}{t^2} = \frac{R\left(t^2\right)}{t^2} \longrightarrow 0 \quad \text{quando } t \to 0,$$

sendo $\phi(0) = \varphi(0) + t\varphi'(0) + \frac{t^2}{2}\varphi''(0)$. Usando as Proposições 3.1, 3.2 e 3.3, temos

$$\begin{aligned}
\varphi(0) &= f(\theta) = D_{KL}\left(p_\theta \| p_\theta\right) = 0; \\
\varphi'(0) &= \sum_i \frac{\partial f}{\partial \alpha_i}(\theta)\,\theta'(0) = 0; \\
\varphi''(0) &= \sum_{i,j} \frac{\partial^2 f}{\partial \alpha_i \partial \alpha_j}(\theta)\,\alpha_i'(0)\alpha_j'(0) + \sum_i \frac{\partial f}{\partial \alpha_i}(\theta)\,\alpha_i''(0) \\
&= \sum_{i,j} g_{ij}(\theta)\,\alpha_i'(0)\alpha_j'(0).
\end{aligned}$$

Substituindo em (3.4) tem-se

$$\varphi(t) = \frac{t^2}{2}\sum_{i,j} g_{ij}(\theta)\,\alpha_i'(0)\alpha_j'(0) + o\left(t^2\right) = \frac{t^2}{2}\|\alpha'(0)\|_G + o\left(t^2\right) = \frac{t^2}{2} + o\left(t^2\right) = \frac{1}{2}d^2 + o\left(t^2\right),$$

haja vista que as geodésicas parametrizada pelo comprimento do arco são curvas de velocidade unitária. Expressar o lado esquerdo como $\varphi(t) = f(\theta(t)) = D_{KL}(p\|q)$ leva ao resultado desejado. □

Capítulo 4

Inferência Geométrica

Neste capítulo, mostraremos como utilizar a estrutura geométrica de modelos estatísticos desenvolvida no Capítulo 2, para fins de inferência estatística. Basicamente, a ideia é tomar um modelo estatístico regular e usar a distância de Fisher-Rao (que estudamos no Lema 2.7) para definir estimadores de distância mínima e estatísticas de teste. Dividiremos o capítulo em duas seções. Na Seção 4.1 vamos definir o conceito de estimativa geodésica, geodésica de estimativas bem como estatística de teste geodésico e estudaremos alguns resultados que enriquecem a teoria. Concluiremos essa seção destacando alguns pontos cujas demonstrações não foram conclusivas e encontradas e que podem ser tomados como perspectiva de trabalhos futuros. Por fim, a Seção 4.2 é dedicada a exemplos de algumas das possíveis aplicações da Geometria. Este capítulo é baseado em L.T Skovgaard 1984 [34] e L.T Skovgaard 1981 [33].

4.1 Estimação de Distância Mínima

Dado um modelo estatístico regular S de dimensão n, suponha que estamos interessados em uma hipótese estatística sobre S dada por um subconjunto não vazio $N \subset S$. Admita que N é uma subvariedade de dimensão $0 < k \leq n$ munida da topologia herdada de S. A distância de Fisher-Rao de uma população $p \in S$ à hipótese N é definida por

$$d_F(p,N) = \inf\{d_F(p,r);\ r \in N\},$$

sendo $d_F(p,r)$ dado em (2.15). Observe que $d_F(p,N)$ está bem definida, pois N é não vazio e d_F é limitada inferiormente. A seguir, definiremos o conceito de estimativa geodésica para uma dada população na hipótese N e o conjunto dessas estimativas.

Definição 4.1 (Estimativa Geodésica, [33]). *Dada uma população p no modelo estatístico regular S, um ponto $r \in N$ é chamado de uma* estimativa geodésica *para p na hipótese N se*

$$d_F(p,r) = d_F(p,N),$$

caso um tal r exista. O conjunto das estimativas geodésicas para p na hipótese N é definido por

$$\Pi_N(p) = \{r \in N \mid d_F(p,r) = d_F(p,N)\} \subset N.$$

Note que $\Pi_N(p)$ pode ser vazio, conter apenas um ponto ou vários pontos. Admitimos que p em geral não está em N.

A distância geodésica entre dois pontos é igual ao comprimento do segmento geodésico entre eles, veja Proposição 1.7.

Definição 4.2 (Geodésica de Estimativas, [33]). *Se γ_p^r for um segmento geodésico conectando uma população $p \in S$ com $r \in \Pi_N(p)$, tal que*

$$d_F(p,r) = \ell(\gamma_p^r),$$

chamaremos γ_p^r de geodésica de estimativas.

Se a população q estiver no arco geodésico $\gamma_p^{r_0}$ conectando as populações p e a estimativa geodésica r_0, então por algumas condições teremos que r_0 pertence ao conjunto das estimativas geodésicas para q na hipótese N, como veremos no teorema a seguir.

Teorema 4.1. *Considere uma população p no modelo estatístico regular S, uma estimativa geodésica $r_0 \in \Pi_N(p)$ e um arco geodésico $\gamma_p^{r_0}$ conectando p a r_0, isto é, $\gamma : [a,b] \to S$ é tal que $\gamma(a) = p$, $\gamma(b) = r_0$ e*

$$d_F(p,r_0) = \ell(\gamma_p^{r_0}).$$

Se $q \in \gamma_p^{r_0}$, então $r_0 \in \Pi_N(q)$.

Demonstração. Suponha que $r_0 \notin \Pi_N(q) = \{r \in N \mid d_F(q,r) = d_F(q,N)\}$. Como por hipótese, $r_0 \in \Pi_N(p)$, temos por definição que $r_0 \in N$. Assim

$$d_F(q,r_0) < d_F(q,N) \tag{4.1}$$

ou

$$d_F(q,r_0) > d_F(q,N). \tag{4.2}$$

De $r_0 \in N$ e $d_F(q,N) = \inf\{d_F(q,r);\ r \in N\}$, segue que $d_F(q,N) \leq d_F(q,r_0)$, assim eliminamos a desigualdade (4.1) e ficamos apenas com a desigualdade (4.2). Então existe

$r_1 \in N$ tal que

$$d_F(q,r_1) < d_F(q,r_0). \tag{4.3}$$

Usando (4.3), a desigualdade triangular e a hipótese de que q é um ponto no arco geodésico $\gamma_p^{r_0}$, temos

$$d_F(p,r_1) \leq d_F(p,q) + d_F(q,r_1) < d_F(p,q) + d_F(q,r_0) = d_F(p,r_0),$$

Ou seja, existe um ponto $r_1 \in N$ tal que

$$d_F(p,r_1) < d_F(p,r_0),$$

o que é uma contradição, pois por hipótese $r_0 \in \Pi_N(p)$, o que significa que $d_F(p,r_0) \leq d_F(q,r),$ para todo $r \in N$. Assim, também não vale a desigualdade (4.2). Portanto,

$$d_F(q,r_0) = d_F(q,N), \text{ ou seja, } r_0 \in \Pi_N(q).$$

□

Com base em $X_1, X_2, ..., X_n$ uma amostra i.i.d. de tamanho n do modelo estatístico regular S, seja $\hat{p}_n$ uma estimativa inicial em S com base nas observações $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$, como por exemplo, a estimativa de máxima verossimilhança. Vimos que pela Definição 4.1, o ponto $r \in \Pi_N(\hat{p}_n)$ é chamado de estimativa geodésica de $\hat{p}_n$ na subvariedade N. A seguir, definiremos o conceito de estatística de teste geodésico da hipótese N.

Definição 4.3 (Estatística de teste Geodésico, [34])**.** *Baseado em $X_1, ..., X_n$ observações independentes, definimos a* estatística de teste geodésico *da hipótese N por*

$$T_N = n \cdot d_F(\hat{p}_n, N)^2,$$

sendo $\hat{p}_n$ uma estimativa inicial.

A seguir, veremos um lema como apoio para a demonstração do próximo teorema.

Lema 4.1. *Sejam $\hat{p}_n$ o estimador de máxima verossimilhança para a densidade do vetor aleatório $\boldsymbol{X} \sim \mathcal{N}(\mu, \mathrm{V})$, com base na amostra $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ de $\boldsymbol{X}$, isto é,*

$$\hat{p}_n = p(x; \hat{\mu}, \hat{\mathrm{V}}),$$

sendo

$$\hat{\mu} = \overline{\boldsymbol{X}}, \quad \hat{\mathrm{V}} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T$$

e p_0 a verdadeira densidade do vetor $\boldsymbol{X}$ com densidade normal multivariada com média μ e matriz de covariância V. *Então $\hat{p}_n$ é fortemente consistente.*

Demonstração. Devemos mostrar que

$$\hat{p}_n \longrightarrow p_0 \quad \text{q.c..}$$

Veja que pela lei forte dos grandes números

$$\overline{\mathbf{X}}_n \longrightarrow \mu, \quad \text{q.c.}$$

e que

$$\hat{\mathrm{V}}_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T = \left(\frac{n-1}{n}\right)\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T.$$

Como

$$\left(\frac{n-1}{n}\right) \longrightarrow 1$$

e

$$\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}}_n)(\mathbf{X}_i - \overline{\mathbf{X}}_n)^T \longrightarrow \mathrm{V}, \quad \text{q.c..}$$

ver referência L. T. Skovgaard 1981 [32, p. 373]. Então, por propriedade operacional

$$\hat{\mathrm{V}}_n \longrightarrow \mathrm{V} \quad \text{q.c..} \tag{4.4}$$

Portanto, por 4.4

$$\hat{\Lambda}_n \longrightarrow \Lambda \quad \text{q.c.}$$

e daí,

$$p(x;\hat{\mu},\hat{\mathrm{V}}) = \frac{(2\pi)^{-n/2}}{\sqrt{\det(\hat{\mathrm{V}}_n)}}\exp\left(-\frac{1}{2}\langle x - \overline{\mathbf{X}}_n, \hat{\Lambda}_n(x - \overline{\mathbf{X}}_n)\rangle\right) \longrightarrow p_0 \quad \text{q.c..}$$

□

No Teorema 4.2, veremos que pelo estimador de máxima verossimilhança $\hat{p}_n$ e algumas condições iniciais, existirá uma estimativa geodésica $\hat{r}_n$ $n \geq n_0$ para $\hat{p}_n$ na hipótese N e algumas propriedades quando tomamos qualquer sequência dessa estimativa geodésica.

Teorema 4.2 ([33])**.** *Seja $\hat{p}_n$ um estimador de máxima verossimilhança para a densidade do vetor aleatório $\boldsymbol{X} \sim \mathcal{N}(\mu, \mathrm{V})$ com base na amostra $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ de $\boldsymbol{X}$ de tamanho n, isto é,*

$$\hat{p}_n = p(\boldsymbol{x}; \hat{\mu}, \hat{\mathrm{V}}),$$

em que $\hat{\mu} = \overline{\boldsymbol{X}} = (X_1 + \cdots + X_n)/n \quad e \quad \hat{\mathrm{V}} = \sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T / n.$

Consideramos a hipótese $N \subset S$, subvariedade, sendo $S = \{p(\boldsymbol{x}; \mu, \mathrm{V}); (\mu, \mathrm{V}) \in \Theta = \mathbb{R}^n \times P_n(\mathbb{R})\}$ é o modelo estatístico regular formado por densidades normais multivariadas. Então, existirá uma estimativa geodésica $\hat{r}_n \in \Pi_N(\hat{p}_n)$ quase certamente, $n \geq n_0$.

Demonstração. Considere a hipótese $N \subset S$ uma subvariedade e $U(p_0)$ uma vizinhança aberta arbitrária do ponto $p_0 \in N$ em N. Como a topologia em N é herdada de S, existe um conjunto $O_1 \subset S$, tal que

$$U(p_0) = O_1 \cap N.$$

Denote

$$\delta = d_F(p_0, O_1^c)$$

e observe que $\delta > 0$. Defina a bola aberta

$$O_2 := B(p_0, \delta/2).$$

Desde que $\hat{p}_n$ é fortemente consistente pelo Lemma 4.1, isto é,

$$\hat{p}_n \to p_0 \quad q.c.$$

então q.c. existe $n_0 \geq 1$ tal que para todo $n \geq n_0,$ temos $\hat{p}_n \in O_2$. Daí,

$$d_F(\hat{p}_n, p_0) < \delta/2.$$

Assim, para $n \geq n_0$

$$d_F(p_0, O_1^c) \leq d_F(p_0, \hat{p}_n) + d_F(\hat{p}_n, O_1^c) \implies d_F(\hat{p}_n, O_1^c) > \delta/2,$$

de modo que a estimativa geodésica sob N deve ser procurada apenas em $U(p_0)$, pois caso contrário, se a estimativa geodésica $r_n \in (U(p_0))^c$, então $r_n \in N \cap O_1^c$.

Neste caso, olhar para os pontos sob N em O_1^c já não faz mais sentido pois $d_F(\hat{p}_n, O_1^c) > \delta/2$ e que $d_F(\hat{p}_n, p_0) < \delta/2$, ou seja, a verdadeira densidade p_0 está sendo um candidato

para a estimativa geodésica. No conjunto fechado $cl\,(U(p_0))$, o ínfimo

$$\inf\{d_F(\hat{p}_n;a);a\in N\}$$

será alcançado, isto é, existe $\hat{r}_n\in N$ tal que $d_F(\hat{p}_n,N)=d_F(\hat{p}_n,\hat{r}_n)$ o que implica $\hat{r}_n\in\Pi_N(\hat{p}_n)$. Das condições acima, mostram que de fato que $\hat{r}_n\in U(p_0)$, pois se

$$\hat{r}_n\in fr(U_{p_0})\implies \hat{r}_n\in O_1^c\implies d_F(p_0,\hat{r}_n)\geq\delta,$$

o que nos leva a uma contradição, pois pela desigualdade triangular

$$d_F(p_0,\hat{r}_n)\leq d_F(p_0,\hat{p}_n)+d_F(\hat{p}_n,\hat{r}_n)<\delta/2+\delta/2,$$

temos que $d_F(p_0,\hat{r}_n)<\delta$. □

Sob as hipóteses do Teorema 4.2, segue o resultado.

Corolário 4.1 ([33])**.** *A sequência de estimadores $(\hat{r}_n)_{n\geq n_0}$ dado no Teorema 4.2 é fortemente consistente para p_0, isto é,*

$$\hat{r}_n\to p_0\quad q.c.$$

em que $p_0\in N$ denota a verdadeira densidade.

Demonstração. Segue diretamente da demonstração do Teorema 4.2, pois vimos que dado uma vizinhança arbitrária $U(p_0)$ em N, então existe $n_0\geq 1$ tal que $\hat{r}_n\in U(p_0)$ para todo $n\geq n_0$ q.c.. □

Observação 4.1. Concluímos esta seção mencionando os itens abaixo, afirmados em Skovgaard [34] e [33]:

i) A sequência de estimadores $(\hat{r}_n)_{n\geq n_0}$ dada pelo Teorema 4.2 é assintoticamente normalmente distribuída e eficiente de primeira ordem;

ii) A estatística de teste geodésico

$$T_N=n\cdot d_F(\hat{p}_n,N)^2=n\cdot d_F(\hat{p}_n,\hat{r}_n)^2$$

tem distribuição limite qui-quadrado com $m-k$ graus de liberdade.

iii) O Teorema 4.2 também vale se usarmos como estimativa inicial de p_0, uma estimativa não-viesada ou alguma outra estimativa eficiente assintoticamente normal.

iv) O Teorema 4.2 pode ser estendido para famílias de distribuições que admitem um estimador eficiente assintoticamente normal, como por exemplo, a família exponencial.

Tais afirmações não são provadas em Skovgaard [34] e [33], que apenas citam R. E. Kass 1980 [19] para verificar a validade das afirmações i) e ii). Em nossas pesquisas, não encontramos tais demonstrações e nossas tentativas de fornecer tais provas não foram conclusivas. Indicamos como perspectiva de trabalho futuro o estudo da validade de cada afirmação acima.

4.2 Exemplos

Nesta seção, ilustraremos a teoria desenvolvida nos capítulos anteriores considerando alguns exemplos de uso da estrutura geométrica para fins de estimativa e teste. Uma vez que as geodésicas na geometria considerada aqui em geral diferem das retas no sistema de coordenadas naturais, a inferência derivada desta geometria também diferirá da inferência de verossimilhança. O objetivo é esboçar algumas das possíveis aplicações da geometria.

4.2.1 Modelo Normal Univariado com μ Desconhecido

Seja $X_1,...,X_n$ uma amostra de tamanho n de uma população $X \sim \mathcal{N}(\mu,\sigma^2)$ sendo a média μ desconhecida. Considere a hipótese

$$H_0 : \mu = \mu_0.$$

A subvariedade N é então dada como

$$N = \{p(x,\mu_0,\sigma^2); (\mu_0,\sigma^2) \in \Theta_0 \text{ sendo } \mu_0 \text{ fixado e } \sigma^2 \in (0,\infty)\}$$

formada por todas as densidades normais univariadas com média μ_0. Sabemos que o espaço paramétrico de N é uma semireta vertical paralela ao eixo σ em μ_0, veja Figura 4.1.

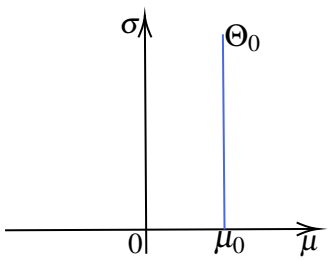

Figura 4.1 Espaço de parâmetros da Subvariedade N.
Produzida online com Mathcha.

As geodésicas de estimativas para a hipótese N são arcos de semielipses dada em (2.17). Sabemos que a estatística de teste geodésico é dado por

$$T_N = n \cdot d_F(\hat{p}_n, N)^2,$$

sendo $\hat{p}_n$ o estimador de máxima verossimilhança parametrizada por $(\bar{X}, s)$, em que

$$\bar{X} = (X_1 + \cdots + X_n)/n \quad \text{e} \quad s^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / n.$$

Pelo Teorema 4.2, existe $\hat{r}_n \in \Pi_N(\hat{p}_n)$, em que o parâmetro de $\hat{r}_n$ é (μ_0, s). Logo, por (2.16)

$$\begin{aligned} T_N &= n \cdot d_F((\bar{X}, s), (\mu_0, s))^2 = 2n \left[\cosh^{-1} \left(\frac{(\bar{X} - \mu_0)^2 + 4s^2}{4s^2} \right) \right]^2 \\ &= 2n \left[\cosh^{-1} \left(\frac{n(\bar{X} - \mu_0)^2}{4ns^2} + 1 \right) \right]^2 = 2n \left[\cosh^{-1} \left(\frac{t^2}{4(n-1)} + 1 \right) \right]^2, \end{aligned}$$

em que $t = \sqrt{n}(\bar{X} - \mu_0)/\bar{s}$ é a estatística t-student, sendo $\bar{s}^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / (n-1)$ a variância amostral não-enviesada.

4.2.2 Modelo Normal Univariado com σ^2 Desconhecida

Seja $X_1, ..., X_n$ uma amostra de tamanho n de uma população $X \sim \mathcal{N}(\mu, \sigma^2)$ em que a variância σ^2 é desconhecida. Considere a hipótese

$$H_0 : \sigma^2 = \sigma_0^2,$$

ou seja, a subvariedade

$$N = \{p(x, \mu, \sigma_0^2); (\mu, \sigma_0) \in \Theta_1 \text{ sendo } \mu \in \mathbb{R} \text{ e } \sigma_0 \text{ fixado}\}$$

formada por todas as densidades normais univariadas com variância σ_0^2. Sabemos que o espaço paramétrico de N é uma reta paralela ao eixo μ passando por σ_0, ver Figura 4.2.

A geodésica de estimativas para hipótese N, são segmentos de retas verticais com

$$\mu = \text{const.},$$

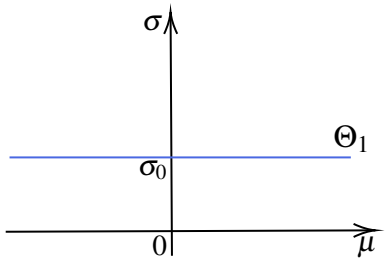

Figura 4.2 Espaço de parâmetros da Subvariedade N.
Produzida online com Mathcha.

de modo que sob H_0, a estimativa $\tilde{\mu}$ de μ, permanecerá a mesma que a estimativa inicial, por exemplo, a estimativa de máxima verossimilhança

$$\tilde{\mu} = \hat{\mu} = \bar{X}.$$

Dado $\hat{p}_n = p(x, \hat{\mu}, s^2)$ o estimador de máxima verossimilhança parametrizada por $(\bar{X}, s)$, pelo Teorema 4.2, existe $\hat{r}_n \in \Pi_N(\hat{p}_n)$, em que o parâmetro de $\hat{r}_n$ é $(\bar{X}, \sigma_0)$ para n suficientemente grande, $n \geq n_0$. Então, usando (2.18), a estatística de teste geodésico para H_0 é dada por

$$\begin{aligned} T_N &= n \cdot d_F(\hat{p}_n, N)^2 = n \cdot d_F(\hat{p}_n, \hat{r}_n)^2 = n\left[\sqrt{2}\ln\left(\frac{\sigma_0}{s}\right)\right]^2 = n\left[\frac{1}{\sqrt{2}}\ln\left(\frac{\sigma_0^2}{s^2}\right)\right]^2 \\ &= (n/2)\left[\ln\left(\frac{\sigma_0^2}{s^2}\right)\right]^2. \end{aligned}$$

se usarmos $\left(\bar{X}, s^2\right)$ como nossa estimativa inicial, isto é, a estimativa de máxima verossimilhança.

4.2.3 Modelo Normal Bivariado com V Diagonal

Passamos agora a considerar uma amostra $Z_1, ..., Z_n$ de tamanho n de uma população Z com densidade de probabilidade normal bivariada, $\mathcal{N}(\mu, \mathrm{V})$, em que V é diagonal. Considere a variedade estatística formada por essas densidades

$$S = \{p(x, \mu, \mathrm{V}); \mu \in \mathbb{R}^2 \text{ e } \mathrm{V} = \mathrm{diag}\left(\sigma_1^2, \sigma_2^2\right)\}.$$

Nesta variedade, consideramos a subvariedade

$$N = \{p(x, \mu, \sigma^2 I); \mu \in \mathbb{R}^2, \sigma^2 \in (0, \infty) \text{ e } I_2 \text{ é a matriz identidade de ordem } 2\}$$

correspondente à hipótese

$$H_0 : \sigma_1^2 = \sigma_2^2.$$

A partir da expressão (2.20) da distância geodésica em S, é facilmente visto que, se usarmos como estimativa inicial a estimativa imparcial, a estimativa geodésica do vetor de valor médio sob H_0 será dada por

$$\tilde{\mu} = \bar{Z}.$$

Dado $\hat{p}_n = p(x, \bar{Z}, \hat{\mathrm{V}})$ o estimador de máxima verossimilhança parametrizada por $(\bar{Z}, \hat{\mathrm{V}})$, em que $\hat{\mathrm{V}} = \mathrm{diag}(s_1^2, s_2^2)$. Pelo Teorema 4.2, existe $\hat{r}_n \in \Pi_N(\hat{p}_n)$, em que o parâmetro de $\hat{r}_n$ é $(\bar{Z}, \tilde{\sigma}^2 I_2)$ para n suficientemente grande, $n \geq n_0$. Então a estatística de teste geodésico para H_0 é dada por

$$\begin{aligned} T_N &= n \cdot d_F(\hat{p}_n, N)^2 = n \cdot d_F(\hat{p}_n, \hat{r}_n)^2 = 2n \sum_{i=1}^{2} \left(\cos^{-1} \left(\frac{(\tilde{\sigma}^2 + s_i^2)}{2\tilde{\sigma} s_i} \right) \right)^2 \\ &= 2n \sum_{i=1}^{2} \left(\ln \left(\frac{(\tilde{\sigma}^2 + s_i^2)}{2\tilde{\sigma} s_i} + \sqrt{\left(\frac{(\tilde{\sigma}^2 + s_i^2)}{2\tilde{\sigma} s_i} \right)^2 - 1} \right) \right)^2 = 2n \sum_{i=1}^{2} \left(\ln \left(\frac{\tilde{\sigma}}{s_i} \right) \right)^2 \\ &= \frac{n}{2} \sum_{i=1}^{2} \left(\ln \left(\frac{\tilde{\sigma}^2}{s_i^2} \right) \right)^2. \end{aligned}$$

Bibliografia

[1] Amari, S.-i. (1985). *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statistics 0028. Springer.

[2] Amari, S.-i. (2016). *Information Geometry and its Applications*, volume 194. Springer.

[3] Amari, S.-i. and Nagaoka, H. (2000). *Methods of information geometry*, volume 191. American Mathematical Soc.

[4] Angulo, J. and Velasco-Forero, S. (2014). Morphological Processing of Univariate Gaussian Distribution-Valued Images Based on Poincaré Upper-Half Plane Representation. *Geometric Theory of Information*, pages 331–366.

[5] Atkinson, C. and Mitchell, A. F. (1981). Rao's Distance Measure. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 345–365.

[6] Bickel, P. J. and Doksum, K. A. (2001). *Mathematical Statistics*, volume vol.1. PH, 2ed. edition.

[7] Biezuner, R. (2017). *Notas de Aula de Geometria Riemanniana*. Departamento de Matemáticas, UFMG.

[8] Bosq, D. and Limnios, N. (2012). *Mathematical Statistics and Stochastic Processes*. Wiley Online Library.

[9] Burbea, J. (1984). *Informative geometry of probability spaces*. Center for Multivariate Analysis, University of Pittsburgh.

[10] Calin, O. and Udrişte, C. (2014). *Geometric Modeling in Probability and Statistics*, volume 121. Springer.

[11] Costa, S. I., Santos, S. A., and Strapasson, J. E. (2015). Fisher Information Distance: A Geometrical Reading. *Discrete Applied Mathematics*, 197:59–69.

[12] do Carmo, M. P. (2015). *Geometria Riemanniana*. Instituto de Matemática Pura e Aplicada.

[13] Fisher, R. A. (1922). 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*, 222(594-604):309–368.

[14] Giaquinta, M. and Modica, G. (2009). *Mathematical Analysis: An Introduction to Functions of Several Variables*. Birkhäuser Basel, 1 edition.

[15] James, B. R. (2002). *Probabilidade: Um Curso em Nível Intermediário, 2a. edição*. IMPA, Rio de Janeiro.

[16] Johnston, N. (2021). *Advanced Linear and Matrix Algebra*. Springer.

[17] Jost, J. (2008). *Riemannian Geometry and Geometric Analysis*, volume 42005. Springer.

[18] Karr, A. F. (1993). *Probability*. Springer Texts in Statistics. Springer-Verlag New York, 1 edition.

[19] Kass, R. E. (1980). *The Riemannian Structure of Model Spaces: A Geometrical Approach to Inference.* PhD thesis, The University of Chicago.

[20] Keener, R. W. (2010). *Theoretical Statistics: Topics for a Core Course*. Springer Texts in Statistics. Springer-Verlag New York, 1 edition.

[21] Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer texts in statistics. Springer, 2nd ed edition.

[22] Lima, E. L. (2014). *Curso de Análise*, volume 2 of *Projeto Euclides*. IMPA.

[23] Loehr, N. (2014). *Advanced Linear Algebra*. CRC Press.

[24] Mahalanobis, P. C. (1936). On the Generalized Distance in Statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55.

[25] Moakher, M. and Zéraï, M. (2011). The Riemannian Geometry of the Space of Positive-Definite Matrices and its Application to the Regularization of Positive-Definite Matrix-Valued Data. *Journal of Mathematical Imaging and Vision*, 40(2):171–187.

[26] Munkres, J. (2000). *Topology*. Prentice Hall, Inc, 2nd ed edition.

[27] Pinele, J., Strapasson, J. E., and Costa, S. I. (2020). The Fisher–Rao Distance Between Multivariate Normal Distributions: Special Cases, Bounds and Applications. *Entropy*, 22(4):404.

[28] Porat, B. and Friedlander, B. (1986). Computation of the Exact Information Matrix of Gaussian Time Series With Stationary Random Components. *IEEE transactions on acoustics, speech, and signal processing*, 34(1):118–130.

[29] Porto, J. P. S. (2017). *Geometria do Modelo Estatístico das Distribuições Normais Multivariadas*. PhD thesis, [sn].

[30] Rao, C. R. (1945). Information and Accuracy Attainable in the Estimation of Statistical Parameters. *Bulletin of the Calcutta Mathematical Society*, pages 37:81–91.

[31] Santiago, W. (2017). *Geometria da Informação - O Teorema de Cramér-Rao*. Dissertação de Mestrado UFRJ.

[32] Shao, J. (2003). *Mathematical Statistics*. Springer Science & Business Media.

[33] Skovgaard, L. T. (1981). *A Riemannian Geometry of the Multivariate Normal Model*. Research Report 81/3. Statistical Research Unit, Danish Medical Research Council, Danish Social Science Research Council.

[34] Skovgaard, L. T. (1984). A Riemannian Geometry of the Multivariate Normal Model. *Scandinavian journal of statistics*, pages 211–223.

Apêndice A

Prova do Teorema 2.4

Demonstração. Sabemos que

$$p(x;\theta) = \frac{(2\pi)^{-n/2}}{\sqrt{\det(\mathrm{V})}} \exp\left(-\frac{1}{2}\langle x-\mu, \Lambda(x-\mu)\rangle\right).$$

sendo que $\mu = (\mu_1, ..., \mu_n) \in \mathbb{R}^n$, matriz de covariância $\mathrm{V} = [\sigma_{ij}] \in P_n(\mathbb{R})$ e $\Lambda = [\lambda_{ij}]$ é a sua inversa. Assim,

$$\begin{aligned}
\ln p(x;\theta) &= \ln\left[\frac{(2\pi)^{-n/2}}{\sqrt{\det(\mathrm{V})}} \exp\left(-\frac{1}{2}\langle x-\mu, \Lambda(x-\mu)\rangle\right)\right] \\
&= \ln\left[\frac{(2\pi)^{-n/2}}{\sqrt{(\mathrm{V})}}\right] + \ln\left[\exp\left(-\frac{1}{2}\langle x-\mu, \Lambda(x-\mu)\rangle\right)\right] \\
&= \ln\left[(2\pi)^{-n/2}\right] - \ln\left[\sqrt{\det(\mathrm{V})}\right] - \frac{1}{2}\langle x-\mu, \Lambda(x-\mu)\rangle \\
&= -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln\det(\mathrm{V}) - \frac{1}{2}\langle x-\mu, \Lambda(x-\mu)\rangle.
\end{aligned}$$

Derivando $\ln p(x;\theta)$ em relação a θ_i, pelo Lema 1.3 temos

$$\begin{aligned}
\frac{\partial}{\partial\theta_i}\ln p(x;\theta) &= \frac{\partial}{\partial\theta_i}\left(-\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln\det(\mathrm{V}) - \frac{1}{2}\langle x-\mu, \Lambda(x-\mu)\rangle\right) \\
&= \frac{\partial}{\partial\theta_i}\left(-\frac{n}{2}\ln 2\pi\right) + \frac{\partial}{\partial\theta_i}\left(-\frac{1}{2}\ln\det(\mathrm{V})\right) + \frac{\partial}{\partial\theta_i}\left(-\frac{1}{2}\langle x-\mu, \Lambda(x-\mu)\rangle\right) \\
&= -\frac{1}{2}\frac{\partial}{\partial\theta_i}\left(\ln\det(\mathrm{V})\right) - \frac{1}{2}\frac{\partial}{\partial\theta_i}\langle x-\mu, \Lambda(x-\mu)\rangle \\
&= -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right) - \frac{1}{2}\left\langle \frac{\partial}{\partial\theta_i}(x-\mu), \Lambda(x-\mu)\right\rangle \\
&\quad -\frac{1}{2}\left\langle x-\mu, \frac{\partial}{\partial\theta_i}\Lambda(x-\mu)\right\rangle,
\end{aligned}$$

isto é, pelo Lema 1.1

$$\begin{aligned}
\frac{\partial}{\partial\theta_i}\ln p(x;\theta) &= -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right) - \frac{1}{2}\left\langle \frac{\partial}{\partial\theta_i}(-\mu), \Lambda(x-\mu)\right\rangle \\
&\quad -\frac{1}{2}\left\langle x-\mu, \left(\frac{\partial\Lambda}{\partial\theta_i}\right)(x-\mu) + \Lambda\frac{\partial}{\partial\theta_i}(x-\mu)\right\rangle \\
&= -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right) + \frac{1}{2}\left\langle \frac{\partial\mu}{\partial\theta_i}, \Lambda(x-\mu)\right\rangle \\
&\quad -\frac{1}{2}\left\langle x-\mu, \left(\frac{\partial\Lambda}{\partial\theta_i}\right)(x-\mu)\right\rangle - \frac{1}{2}\left\langle x-\mu, \Lambda\frac{\partial}{\partial\theta_i}(x-\mu)\right\rangle \\
&= -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right) + \frac{1}{2}\left\langle \frac{\partial\mu}{\partial\theta_i}, \Lambda(x-\mu)\right\rangle \\
&\quad -\frac{1}{2}\left\langle x-\mu, \left(-\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\Lambda\right)(x-\mu)\right\rangle + \frac{1}{2}\left\langle x-\mu, \Lambda\frac{\partial\mu}{\partial\theta_i}\right\rangle \\
&= -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right) + \frac{1}{2}\left\langle \frac{\partial\mu}{\partial\theta_i}, \Lambda(x-\mu)\right\rangle \\
&\quad +\frac{1}{2}\left\langle x-\mu, \Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\Lambda(x-\mu)\right\rangle + \frac{1}{2}\left\langle x-\mu, \Lambda\frac{\partial\mu}{\partial\theta_i}\right\rangle \\
&= -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right) + \left\langle \frac{\partial\mu}{\partial\theta_i}, \Lambda(x-\mu)\right\rangle + \frac{1}{2}\left\langle x-\mu, \Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\Lambda(x-\mu)\right\rangle.
\end{aligned}$$

Logo, derivando $\ln p(x;\theta)$ em relação a θ_j, obteremos

$$\frac{\partial}{\partial\theta_j}\ln p(x;\theta) = -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right) + \left\langle \frac{\partial\mu}{\partial\theta_j}, \Lambda(x-\mu)\right\rangle + \frac{1}{2}\left\langle x-\mu, \Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\Lambda(x-\mu)\right\rangle.$$

Agora calculamos o produto de $\dfrac{\partial}{\partial\theta_i}\ln p(x;\theta)$ com $\dfrac{\partial}{\partial\theta_j}\ln p(x;\theta)$:

$$\begin{aligned}
&\frac{\partial}{\partial\theta_i}\ \ln p(x;\theta)\cdot\frac{\partial}{\partial\theta_j}\ln p(x;\theta) = \\
&= \left[-\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right) + \left\langle \frac{\partial\mu}{\partial\theta_i}, \Lambda(x-\mu)\right\rangle + \frac{1}{2}\left\langle x-\mu, \Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\Lambda(x-\mu)\right\rangle\right]\cdot \\
&\quad \left[-\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right) + \left\langle \frac{\partial\mu}{\partial\theta_j}, \Lambda(x-\mu)\right\rangle + \frac{1}{2}\left\langle x-\mu, \Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\Lambda(x-\mu)\right\rangle\right] =
\end{aligned}$$

$$
\begin{aligned}
= \quad & \frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right) - \frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\left\langle\frac{\partial\mu}{\partial\theta_j},\Lambda(x-\mu)\right\rangle \\
& -\frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\left\langle x-\mu,\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\Lambda(x-\mu)\right\rangle - \frac{1}{2}\left\langle\frac{\partial\mu}{\partial\theta_i},\Lambda(x-\mu)\right\rangle\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right) \\
& +\left\langle\frac{\partial\mu}{\partial\theta_i},\Lambda(x-\mu)\right\rangle\left\langle\frac{\partial\mu}{\partial\theta_j},\Lambda(x-\mu)\right\rangle \\
& +\frac{1}{2}\left\langle\frac{\partial\mu}{\partial\theta_i},\Lambda(x-\mu)\right\rangle\left\langle x-\mu,\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\Lambda(x-\mu)\right\rangle \\
& -\frac{1}{4}\left\langle x-\mu,\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\Lambda(x-\mu)\right\rangle\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right) \\
& +\frac{1}{2}\left\langle x-\mu,\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\Lambda(x-\mu)\right\rangle\left\langle\frac{\partial\mu}{\partial\theta_j},\Lambda(x-\mu)\right\rangle \\
& +\frac{1}{4}\left\langle x-\mu,\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\Lambda(x-\mu)\right\rangle\left\langle x-\mu,\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\Lambda(x-\mu)\right\rangle. \text{ Daí,}
\end{aligned}
$$

$$
\begin{aligned}
& \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta_i}\ln p(x;\theta)\cdot\frac{\partial}{\partial\theta_j}\ln p(x;\theta)\right) = \frac{1}{4}\mathbb{E}_\theta\left[\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right)\right] \\
& -\frac{1}{2}\mathbb{E}_\theta\left[\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\left\langle\frac{\partial\mu}{\partial\theta_j},\Lambda(x-\mu)\right\rangle\right] \\
& -\frac{1}{4}\mathbb{E}_\theta\left[\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\left\langle x-\mu,\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\Lambda(x-\mu)\right\rangle\right] \\
& -\frac{1}{2}\mathbb{E}_\theta\left[\left\langle\frac{\partial\mu}{\partial\theta_i},\Lambda(x-\mu)\right\rangle\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right)\right] \\
& +\mathbb{E}_\theta\left[\left\langle\frac{\partial\mu}{\partial\theta_i},\Lambda(x-\mu)\right\rangle\left\langle\frac{\partial\mu}{\partial\theta_j},\Lambda(x-\mu)\right\rangle\right] \\
& +\frac{1}{2}\mathbb{E}_\theta\left[\left\langle\frac{\partial\mu}{\partial\theta_i},\Lambda(x-\mu)\right\rangle\left\langle x-\mu,\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\Lambda(x-\mu)\right\rangle\right] \\
& -\frac{1}{4}\mathbb{E}_\theta\left[\left\langle x-\mu,\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\Lambda(x-\mu)\right\rangle\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right)\right] \\
& +\frac{1}{2}\mathbb{E}_\theta\left[\left\langle x-\mu,\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\Lambda(x-\mu)\right\rangle\left\langle\frac{\partial\mu}{\partial\theta_j},\Lambda(x-\mu)\right\rangle\right] \\
& +\frac{1}{4}\mathbb{E}_\theta\left[\left\langle x-\mu,\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\Lambda(x-\mu)\right\rangle\left\langle x-\mu,\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\Lambda(x-\mu)\right\rangle\right]. \qquad \text{(A.1)}
\end{aligned}
$$

Calcularemos parcela por parcela separadamente. Tomaremos $A = \Lambda\dfrac{\partial \mathrm{V}}{\partial\theta_i}\Lambda$ e $B = \Lambda\dfrac{\partial \mathrm{V}}{\partial\theta_j}\Lambda$.

i) $\frac{1}{4}\mathbb{E}_\theta\left[\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right)\right] = \frac{1}{4}\int_{\mathbb{R}^n}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right)p(x;\theta)\,dx$
$= \frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right)\int_{\mathbb{R}^n}p(x;\theta)\,dx = \frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right).$

ii) $-\frac{1}{2}\mathbb{E}_\theta\left[\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\left\langle\frac{\partial\mu}{\partial\theta_j},\Lambda(x-\mu)\right\rangle\right] = -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathbb{E}_\theta\left[\left\langle\frac{\partial\mu}{\partial\theta_j},\Lambda(x-\mu)\right\rangle\right]$
$= -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathbb{E}_\theta\left[\left(\frac{\partial\mu}{\partial\theta_j}\right)^T\Lambda(x-\mu)\right] = -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\left(\frac{\partial\mu}{\partial\theta_j}\right)^T\Lambda\mathbb{E}_\theta\left[(x-\mu)\right]$
$= 0.$

iii) $-\frac{1}{4}\mathbb{E}_\theta\left[\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\left\langle x-\mu,\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\Lambda(x-\mu)\right\rangle\right]$
$= -\frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathbb{E}_\theta\left[\left\langle x-\mu,\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\Lambda(x-\mu)\right\rangle\right]$
$= -\frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathbb{E}_\theta\left[(x-\mu)^T\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\Lambda(x-\mu)\right]$
$= -\frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathbb{E}_\theta\left[(x-\mu)^T B(x-\mu)\right]$
$= -\frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathbb{E}_\theta\left[\sum_{r,s=1}^{n}B_{rs}(x_r-\mu_r)(x_s-\mu_s)\right]$
$= -\frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\sum_{r,s=1}^{n}B_{rs}\mathbb{E}_\theta\left[(x_r-\mu_r)(x_s-\mu_s)\right] = -\frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\sum_{r,s=1}^{n}B_{rs}\sigma_{rs}$
$= -\frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\sum_{r,s=1}^{n}B_{rs}\sigma_{sr} = -\frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathrm{tr}\left(B\,\mathrm{V}\right)$
$= -\frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\Lambda\,\mathrm{V}\right) = -\frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right).$

iv) $-\frac{1}{2}\mathbb{E}_\theta\left[\left\langle\frac{\partial\mu}{\partial\theta_i},\Lambda(x-\mu)\right\rangle\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right)\right]$
$= -\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right)\mathbb{E}_\theta\left[\left\langle\frac{\partial\mu}{\partial\theta_i},\Lambda(x-\mu)\right\rangle\right] = 0.$

Pela Observação A.1 e por propriedades do traço de uma matriz, temos

v) $\mathbb{E}_\theta\left[\left\langle\frac{\partial\mu}{\partial\theta_i},\Lambda(x-\mu)\right\rangle\left\langle\frac{\partial\mu}{\partial\theta_j},\Lambda(x-\mu)\right\rangle\right]$
$= \mathbb{E}_\theta\left[\sum_{r,s=1}^{n}\lambda_{rs}\frac{\partial\mu_r}{\partial\theta_i}(x_s-\mu_s)\sum_{t,z=1}^{n}\lambda_{tz}\frac{\partial\mu_t}{\partial\theta_j}(x_z-\mu_z)\right]$
$= \mathbb{E}_\theta\left[\sum_{r,s,t,z=1}^{n}\lambda_{rs}\frac{\partial\mu_r}{\partial\theta_i}(x_s-\mu_s)\lambda_{tz}\frac{\partial\mu_t}{\partial\theta_j}(x_z-\mu_z)\right]$

$$= \sum_{r,s,t,z=1}^{n} \lambda_{rs} \frac{\partial \mu_r}{\partial \theta_i} \lambda_{tz} \frac{\partial \mu_t}{\partial \theta_j} \mathbb{E}_\theta[(x_s-\mu_s)(x_z-\mu_z)] = \sum_{r,s,t,z=1}^{n} \lambda_{rs} \frac{\partial \mu_r}{\partial \theta_i} \lambda_{tz} \frac{\partial \mu_t}{\partial \theta_j} \sigma_{sz}$$

$$= \sum_{r,s,t,z=1}^{n} \lambda_{rs}\sigma_{sz}\lambda_{zt} \frac{\partial \mu_t}{\partial \theta_j} \frac{\partial \mu_r}{\partial \theta_i} = \operatorname{tr}\left(\Lambda \mathrm{V} \Lambda \frac{\partial \mu}{\partial \theta_j} \left(\frac{\partial \mu}{\partial \theta_i}\right)^T\right) = \operatorname{tr}\left(\Lambda \frac{\partial \mu}{\partial \theta_j} \left(\frac{\partial \mu}{\partial \theta_i}\right)^T\right)$$

$$= \operatorname{tr}\left(\left(\frac{\partial \mu}{\partial \theta_i}\right)^T \Lambda \frac{\partial \mu}{\partial \theta_j}\right) = \left(\frac{\partial \mu}{\partial \theta_i}\right)^T \Lambda \frac{\partial \mu}{\partial \theta_j}.$$

Como o momento impar de um vetor aleatório com densidade normal multivariado é 0, temos

vi) $$\frac{1}{2}\mathbb{E}_\theta\left[\left\langle \frac{\partial \mu}{\partial \theta_i}, \Lambda(x-\mu)\right\rangle \left\langle x-\mu, \Lambda \frac{\partial \mathrm{V}}{\partial \theta_j} \Lambda(x-\mu)\right\rangle\right]$$

$$= \frac{1}{2}\mathbb{E}_\theta\left[\left\langle \frac{\partial \mu}{\partial \theta_i}, \Lambda(x-\mu)\right\rangle \langle x-\mu, B(x-\mu)\rangle\right]$$

$$= \frac{1}{2}\mathbb{E}_\theta\left[\sum_{r,s=1}^{n} \lambda_{rs} \frac{\partial \mu_r}{\partial \theta_i}(x_s-\mu_s) \sum_{t,z=1}^{n} B_{tz}(x_t-\mu_t)(x_z-\mu_z)\right]$$

$$= \frac{1}{2}\mathbb{E}_\theta\left[\sum_{r,s,t,z=1}^{n} \lambda_{rs} \frac{\partial \mu_r}{\partial \theta_i}(x_s-\mu_s) B_{tz}(x_t-\mu_t)(x_z-\mu_z)\right]$$

$$= \frac{1}{2}\sum_{r,s,t,z=1}^{n} \lambda_{rs} \frac{\partial \mu_r}{\partial \theta_i} B_{tz}\mathbb{E}_\theta\left[(x_s-\mu_s)(x_t-\mu_t)(x_z-\mu_z)\right] = 0.$$

Pelo item *iii*), temos

vii) $$-\frac{1}{4}\mathbb{E}_\theta\left[\left\langle x-\mu, \Lambda \frac{\partial \mathrm{V}}{\partial \theta_i} \Lambda(x-\mu)\right\rangle \operatorname{tr}\left(\Lambda \frac{\partial \mathrm{V}}{\partial \theta_j}\right)\right] = -\frac{1}{4}\operatorname{tr}\left(\Lambda \frac{\partial \mathrm{V}}{\partial \theta_i}\right)\operatorname{tr}\left(\Lambda \frac{\partial \mathrm{V}}{\partial \theta_j}\right).$$

Pelo item *vi*), temos

viii) $$\frac{1}{2}\mathbb{E}_\theta\left[\left\langle x-\mu, \Lambda \frac{\partial \mathrm{V}}{\partial \theta_i} \Lambda(x-\mu)\right\rangle \left\langle \frac{\partial \mu}{\partial \theta_j}, \Lambda(x-\mu)\right\rangle\right] = 0.$$

iv) $$\frac{1}{4}\mathbb{E}_\theta\left[\left\langle x-\mu, \Lambda \frac{\partial \mathrm{V}}{\partial \theta_i} \Lambda(x-\mu)\right\rangle \left\langle x-\mu, \Lambda \frac{\partial \mathrm{V}}{\partial \theta_j} \Lambda(x-\mu)\right\rangle\right]$$

$$= \frac{1}{4}\mathbb{E}_\theta\left[\langle x-\mu, A(x-\mu)\rangle \langle x-\mu, B(x-\mu)\rangle\right]$$

$$= \frac{1}{4}\mathbb{E}_\theta\left[(x-\mu)^T A(x-\mu)(x-\mu)^T B(x-\mu)\right]$$

$$= \frac{1}{4}\mathbb{E}_\theta\left[\sum_{r,s=1}^{n} A_{rs}(x_r-\mu_r)(x_s-\mu_s) \sum_{t,z=1}^{n} B_{tz}(x_t-\mu_t)(x_z-\mu_z)\right]$$

$$= \frac{1}{4}\mathbb{E}_\theta\left[\sum_{r,s,t,z=1}^{n} A_{rs}(x_r-\mu_r)(x_s-\mu_s) B_{tz}(x_t-\mu_t)(x_z-\mu_z)\right]$$

$$= \frac{1}{4}\sum_{r,s,t,z=1}^{n} A_{rs}B_{tz}\mathbb{E}_\theta\left[(x_r-\mu_r)(x_s-\mu_s)(x_t-\mu_t)(x_z-\mu_z)\right]$$

$$
\begin{aligned}
&= \frac{1}{4}\sum_{r,s,t,z=1}^{n} A_{rs}B_{tz}\left(\sigma_{rs}\sigma_{tz}+\sigma_{rt}\sigma_{sz}+\sigma_{rz}\sigma_{st}\right)\\
&= \frac{1}{4}\left(\sum_{r,s,t,z=1}^{n} A_{rs}B_{tz}\sigma_{rs}\sigma_{tz}+\sum_{r,s,t,z=1}^{n} A_{rs}B_{tz}\sigma_{rt}\sigma_{sz}+\sum_{r,s,t,z=1}^{n} A_{rs}B_{tz}\sigma_{rz}\sigma_{st}\right)\\
&= \frac{1}{4}\left(\sum_{r,s=1}^{n} A_{rs}\sigma_{rs}\sum_{t,z=1}^{n} B_{tz}\sigma_{tz}+\sum_{r,s,t,z=1}^{n} A_{rs}B_{tz}\sigma_{rt}\sigma_{sz}+\sum_{r,s,t,z=1}^{n} A_{rs}B_{tz}\sigma_{rz}\sigma_{st}\right)\\
&= \frac{1}{4}\left(\sum_{r,s=1}^{n} A_{rs}\sigma_{rs}\sum_{t,z=1}^{n} B_{tz}\sigma_{tz}+\sum_{r,s,t,z=1}^{n} A_{rs}\sigma_{sz}B_{zt}\sigma_{tr}+\sum_{r,s,t,z=1}^{n} A_{rs}\sigma_{st}B_{tz}\sigma_{zr}\right)\\
&= \frac{1}{4}\left[\mathrm{tr}\left(A\mathrm{V}\right)\mathrm{tr}\left(B\mathrm{V}\right)+2\mathrm{tr}\left(A\mathrm{V}B\mathrm{V}\right)\right]\\
&= \frac{1}{4}\left[\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\Lambda\mathrm{V}\right)\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\Lambda\mathrm{V}\right)+2\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\Lambda\mathrm{V}\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\Lambda\mathrm{V}\right)\right]\\
&= \frac{1}{4}\left[\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right)+2\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right)\right]\\
&= \frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right)+\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right).
\end{aligned}
$$

Portanto, substituindo os itens *i*) ao *iv*) em (A.1), obtemos

$$
\begin{aligned}
g_{ij}(\theta) &= \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta_i}\ln p(x;\theta)\cdot\frac{\partial}{\partial\theta_j}\ln p(x;\theta)\right)=\\
&= \frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right)-\frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right)\\
&\quad+\left(\frac{\partial\mu}{\partial\theta_i}\right)^T\Lambda\frac{\partial\mu}{\partial\theta_j}-\frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right)\\
&\quad+\frac{1}{4}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\right)\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right)+\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right)\\
&= \left(\frac{\partial\mu}{\partial\theta_i}\right)^T\Lambda\frac{\partial\mu}{\partial\theta_j}+\frac{1}{2}\mathrm{tr}\left(\Lambda\frac{\partial \mathrm{V}}{\partial\theta_i}\Lambda\frac{\partial \mathrm{V}}{\partial\theta_j}\right).
\end{aligned}
$$

Observação A.1. Sabemos que

$$
\frac{\partial\mu}{\partial\theta_j}=\left(\frac{\partial\mu_1}{\partial\theta_j},...,\frac{\partial\mu_t}{\partial\theta_j},...,\frac{\partial\mu_n}{\partial\theta_j}\right)\ \text{e}\ \frac{\partial\mu}{\partial\theta_i}=\left(\frac{\partial\mu_1}{\partial\theta_i},...,\frac{\partial\mu_r}{\partial\theta_i},...,\frac{\partial\mu_n}{\partial\theta_i}\right)
$$

O produto $\frac{\partial \mu_t}{\partial \theta_j}\frac{\partial \mu_r}{\partial \theta_i}$ é elemento da seguinte matriz $n \times n$

$$\begin{aligned}
\frac{\partial \mu}{\partial \theta_j}\left(\frac{\partial \mu}{\partial \theta_i}\right)^T &= \left(\frac{\partial \mu_1}{\partial \theta_j} \quad \cdots \quad \frac{\partial \mu_n}{\partial \theta_j}\right)^T \left(\frac{\partial \mu_1}{\partial \theta_i} \quad \cdots \quad \frac{\partial \mu_n}{\partial \theta_i}\right) \\
&= \begin{pmatrix}
\frac{\partial \mu_1}{\partial \theta_j}\frac{\partial \mu_1}{\partial \theta_i} & \cdots & \frac{\partial \mu_1}{\partial \theta_j}\frac{\partial \mu_r}{\partial \theta_i} & \cdots & \frac{\partial \mu_1}{\partial \theta_j}\frac{\partial \mu_n}{\partial \theta_i} \\
\vdots & \ddots & \vdots & \vdots & \vdots \\
\frac{\partial \mu_t}{\partial \theta_j}\frac{\partial \mu_1}{\partial \theta_i} & \cdots & \frac{\partial \mu_t}{\partial \theta_j}\frac{\partial \mu_r}{\partial \theta_i} & \cdots & \frac{\partial \mu_t}{\partial \theta_j}\frac{\partial \mu_n}{\partial \theta_i} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{\partial \mu_n}{\partial \theta_j}\frac{\partial \mu_1}{\partial \theta_i} & \cdots & \frac{\partial \mu_n}{\partial \theta_j}\frac{\partial \mu_r}{\partial \theta_i} & \cdots & \frac{\partial \mu_n}{\partial \theta_j}\frac{\partial \mu_n}{\partial \theta_i}
\end{pmatrix}.
\end{aligned}$$

□