Geometria Riemann-Finsler com Aplicações em Geometria da Informação

Tiago Moreira Vargas[1]

IME-Instituto de Matemática e Estatística, USP-Universidade de São Paulo, Brasil

Abstract

Information Geometry is a new branch in mathematics, originated from the applications of Diferential Geometry to Statistics. This area emerged from investigating the geometrical struture of a family of probability distribuitions, and has been applied to various areas including statistical inference and information theory. This dissertation is an introduction to the geometry of information for a more general point of view using Riemann-Finsler Geometry and Spray Geometry.

Palavras-Chave: Geometria Riemann-Finsler, Geometria da Informação.

1 Introdução

A Geometria da Informação surgiu para investigar a estrutura geométrica de famílias paramétricas de distribuições de probabilidade. Um exemplo seria, se considerarmos o conjunto S das distribuições normais com média μ e variância σ^2 :

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

S pode ser vista como uma variedade de dimensão 2 onde (μ, σ) é um sistema de coordenadas.

No entanto, este não é um espaço euclidiano, mas sim um espaço com uma métrica Riemanniana que, naturalmente, decorre das propriedades das distribuições de probabilidade subjacentes. Em particular, S além de uma família de distribuições normais, é um espaço de curvatura constante negativa.

No nosso trabalho, demos um enfoque mais geral à geometria da informação, utilizando para isto, a Geometria de Riemann-Finsler.

1.1 Métricas de Finsler

Uma métrica de Finsler em M é uma função $L : TM \longrightarrow [0, \infty)$ que satisfaz às seguintes condições:

(i) (Regularidade) $L \in C^\infty$ em $\widetilde{TM}$;

(ii) (Homogeneidade positiva) $L(x, \lambda y) = \lambda^2 L(x, y)$, $\forall \lambda > 0, (x, y) \in TM$;

(iii) (Convexidade forte) $g = (g_{ij}(x, y)) = \left(\frac{1}{2}[L(x, y)]_{y_i y_j}\right)$ é positiva definida em cada ponto de $\widetilde{TM}$.

Um espaço (M, L) é dito uma variedade de Finsler.

Função métrica L de Finsler será *Riemanniana* se $L(x, y) = g_{ij}(x)y^i y^j$. Nesse caso, $g_{ij}(x, y) = \frac{1}{2}[L]_{y^i y^j} = g_{ij}(x)$ independe de y, para cada i, j. Temos ainda que esta métrica será *Euclidiana* se os g_{ij}'s independerem de x também.

[1]E-mail: tiagomv@ime.usp.br

[1]Trabalho de Mestrado em Matemática, realizado na Universidade Federal de Goiás, no período de 2007 a 2009

1.2 Comprimento em espaços Finsler

Seja (M, L) uma variedade de Finsler, onde L é positivamente homogênea de grau 2. Seja $c : [0,1] \longrightarrow M$ uma curva diferenciável por partes com velocidade $\frac{dc}{dt} = \frac{dc^i}{dt}\frac{\partial}{\partial x^i} \in T_{c(t)}M$. O comprimento integral $L(C)$ é definido como

$$L(C) = \int_0^1 \sqrt{L\left(c(t), \frac{dc}{dt}\right)}dt. \tag{1}$$

Utilizando essa estrutura de comprimento , podemos definir uma função $d = d(p,q)$ em $M \times M$ por

$$d(p,q) = \inf(L(C)),$$

onde o ínfimo é calculado sobre todas as curvas que ligam p a q , com $c(0) = p$ e $c(1) = q$. A função distância d satisfaz

(i) $d(p,q) \geq 0$ e a igualdade acontece se, e somente se, $p = q$;

(ii) $d(p,q) \leq d(p,r) + d(r,q)$.

A função d é chamada de *função distância* de L.

1.3 H-função

Uma H-função em uma variedade M é uma função escalar $H = H(x,y)$ em TM com as seguintes propriedades:

(i) $H(x, \lambda y) = \lambda^3 H(x,y)$, $\lambda > 0$.

(ii) $H(x,y)$ é de classe C^∞ em $\widetilde{TM}$.

2 Sprays

2.1 Sprays de Métricas de Finsler

Um spray sobre uma variedade Finsler M munida de uma métrica L é um campo de vetores definidos sobre TM o qual é expresso num sistema padrão de coordenadas locais (x^i, y^i) por

$$\mathcal{G} = y - \mathcal{G}^i(x,y)\frac{\partial}{\partial y^i},$$

onde $y = y^i\frac{\partial}{\partial x^i}$ e $\mathcal{G}^i(x,y)$ são funções de classe C^∞, definidas localmente sobre $TM \setminus \{0\}$ satisfazendo

$$\mathcal{G}^i(x, \lambda y) = \lambda^2 \mathcal{G}^i(x,y),$$

para $\lambda > 0$ com

$$\mathcal{G}^i(x,y) = \frac{1}{4} g^{il}(x,y)\{L_{x^k y^l}(x,y)y^k - L_{x^l}(x,y)\}. \tag{2}$$

A equação

$$\ddot{c^i}(t) + 2\mathcal{G}^i(c, \dot{c}(t)) = 0. \tag{3}$$

é a equação das geodésicas do spray da métrica de Finsler L, e sua origem está no fato de se utilizar um método variacional para se achar a equação das geodésicas de uma Métrica de Finsler no funcional (1).

2.2 Sprays em Variedades

O conceito de sprays induzidos por Métricas de Finsler pode ser generalizado. Seja M uma variedade. Um spray G em M é um campo de vetores no fibrado tangente TM tal que em algum sistema de coordenadas locais usual (x^i, y^i) em TM, é expressado da seguinte forma:

$$G = y^i\frac{\partial}{\partial x^i} - 2G^i\frac{\partial}{\partial y^i}, \tag{4}$$

onde $G^i \in C^\infty$ com $y \neq 0$ e

$$G^i(x, \lambda y) = \lambda^2 G^i(x,y), \qquad \lambda > 0.$$

Uma curva $\sigma = \sigma(t)$ é chamada de *geodésica* de G em uma variedade M se satisfaz o seguinte sistema de equações:

$$\ddot{x}^i(t) + 2G^i(x, \dot{x}(t)) = 0,$$

onde $x(t) = (x^i(t))$ denota as coordenadas de $\sigma(t)$. Geodésicas também são chamadas de *caminhos*. A coleção de todos os caminhos de um spray é chamada *estrutura de caminhos*. Um spray G é denominado *afim* se em algum sistema de coordenadas locais,

$$G^i(x, y) = \frac{\Gamma^i_{jk}(x)}{2} y^i y^j, \qquad \Gamma^i_{jk}(x) = \Gamma^i_{kj}(x). \tag{5}$$

Um spray G em uma variedade é chamado **flat** se para todo ponto , existe um sistema de coordenadas locais (x^i, y^i) em TM tal que $G = y^i \frac{\partial}{\partial x^i}$, isto é, $G^i = 0$. Nesse caso (x^i, y^i) é chamado de sistema de coordenadas adaptado. Os sprays do tipo *flat* são casos muito especiais de sprays afins. Se G é flat, então em um sistema de coordenadas locais, as geodésicas são linhas retas, isto é, as coordenadas $(x^i(t))$ de toda geodésica $\sigma(t)$ assumem a forma

$$x^i(t) = a^i t + b^i, \qquad \forall t.$$

3 Teorema de Euler para funções homogêneas

Seja A um subconjunto aberto do $\mathbb{R}^m$, tal que, se $y \in A$, então $ty \in A$ e seja $L : A \to \mathbb{R}$ uma função diferenciável. Então para $k \in \mathbb{R}$, L é homogênea de grau k, isto é, $L(ty) = t^k L(y)$ se, e somente se,

$$y^i \frac{\partial L}{\partial y^i}(y) = kL(y), \forall y \in A. \tag{6}$$

4 Divergências e Estruturas de Informação

4.1 Divergência : Definição

Considere $\mathcal{F}$ um conjunto de objetos que podem ser imagens $2D/3D$, distribuições de probabilidade, etc... Podemos calcular a diferença entre dois objetos de $\mathcal{F}$ definindo uma função chamada de **divergência** no espaço $\mathcal{F} \times \mathcal{F}$ dada por:

$$\mathcal{D} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$$
$$(p, q) \to \mathcal{D}(p, q)$$

com as seguintes propriedades:

- $\mathcal{D}(p, q) \geq 0$
- $\mathcal{D}(p, q) = 0 \Leftrightarrow p = q$.

O número $\mathcal{D}(p, q)$ mede a divergência de p para q. O par $(\mathcal{F}, \mathcal{D})$ é chamado de **espaço de divergência**. Para o espaço de divergência $(\mathcal{F}, \mathcal{D})$, $\mathcal{F}$ nem sempre tem dimensão finita. Na prática, consideramos uma família de objetos em $\mathcal{F}$ parametrizados em um domínio de $\mathbb{R}^n$. Essa família é chamada de **modelo** em $(\mathcal{F}, \mathcal{D})$. Sendo mais preciso, um modelo de um espaço de divergência $(\mathcal{F}, \mathcal{D})$ é uma variedade M n-dimensional de classe C^∞ mergulhada em um subconjunto de $\mathcal{F}$ com divergência induzida $D = \mathcal{D}_{|M}$. Esse modelo (M, D) também é um espaço de divergência. A divergência $\mathcal{D}$ nem sempre é reversível, ou seja, em geral, $\mathcal{D}(p, q) \neq \mathcal{D}(q, p)$. Seja $d = d(p, q)$ a função distância de uma métrica de Finsler L em M. Seja

$$D(p, q) := \frac{1}{2} d(p, q)^2, \qquad p, q \in M.$$

D é a divergência da métrica de Finsler em M. Em geral, a divergência D não é C^∞ ao longo da diagonal $\Delta = \{(p, p) \in M \times M\}$ a não ser que L seja Riemanniana.

Lema 1. *Se D é a divergência de uma métrica de Finsler L em uma variedade M, então para todo ponto p de M, existe um sistema de coordenadas locais (U, Φ) em M tal que*

$$2D(\Phi^{-1}(x), \Phi^{-1}(x + y)) = L(x, y) + \frac{1}{2} L_{x^k}(x, y) y^k + o(|y|^3). \tag{7}$$

4.2 Divergências Regulares

Seja M uma variedade. Uma função divergência D em M é chamada **regular** se em qualquer sistema de coordenadas locais (U, Φ) e todo ponto em M, restringindo a um pequeno domínio se necessário,

$$2D(\Phi^{-1}(x), \Phi^{-1}(x+y)) = L(x,y) + P(x,y) + o(|y|^3), \tag{8}$$

onde $L = L(x,y)$ é uma métrica de Finsler em U e $P = P(x,y) \in C^\infty$ é uma função em $TU - \{0\}$ com

$$P(x, \lambda y) = \lambda^3 P(x,y), \lambda > 0.$$

Lema 2. *Seja D uma divergência regular em M. Seja L e P as funções locais definidas na equação* (13) *em um sistema de coordenadas locais (U, Φ). Então*

$$H := P(x,y) - \frac{1}{2} L_{x^k}(x,y) y^k \tag{9}$$

é uma H-função bem definida em M.

Proposição 1. *Se D é uma divergência de uma métrica de Finsler em uma variedade M, então ela é regular com $H = 0$.*

5 Estruturas de Informação

Uma **Estrutura de informação** em uma variedade M é o par $\{L, H\}$, onde $L = L(x,y)$ é uma métrica de Finsler em M e $H = H(x,y)$ é uma H-função.

5.1 α-sprays de uma estrutura de informação

Seja (L, H) uma estrutura de informação em uma variedade M. Seja

$$\mathcal{G} = y^i \frac{\partial}{\partial x^i} - 2\mathcal{G}^i \frac{\partial}{\partial y^i} \tag{10}$$

um spray de L, onde

$$\mathcal{G}^i(x, \lambda y) = \lambda^2 \mathcal{G}^i(x,y),$$

para $\lambda > 0$ com

$$\mathcal{G}^i(x,y) = \frac{1}{4} g^{il}(x,y) \{L_{x^k y^l}(x,y) y^k - L_{x^l}(x,y)\}.$$

Utilizando H, vamos definir uma família de sprays

$$G_\alpha = y^i \frac{\partial}{\partial x^i} - 2G^i_\alpha \frac{\partial}{\partial y^i}, \tag{11}$$

onde

$$G^i_\alpha(x,y) = \mathcal{G}^i(x,y) + \frac{\alpha}{2} g^{ij}(x,y) H_{y^j}(x,y), \tag{12}$$

ou seja, vamos pertubar o spray de L . G_α é chamado de α-spray de (L, H).

Lema 3. *Seja D uma divergência regular em uma variedade M, (L, H) a estrutura de informação induzida por D e G_α o α-spray de (L, H). Seja $\sigma = \sigma(t)$ uma geodésica. Então para toda geodésica σ do spray G_α,*

$$2D(\sigma(t_0), \sigma(t_0 + \epsilon)) = L(x,y)\epsilon^2 + (1 - 3\alpha) H(x,y) \epsilon^3 + o(\epsilon^3), \tag{13}$$

e

$$d(\sigma(t_0), \sigma(t_0 + \epsilon))^2 = L(x,y)\epsilon^2 - 3\alpha H(x,y)\epsilon^3 + o(\epsilon^3), \tag{14}$$

onde $\epsilon > 0$, $x = \sigma(t_0)$, $y = \dot{\sigma}(t_0)$.

Definição 1. *Uma estrutura de informação (L, H) em uma variedade M é chamada de α-**flat** para algum α se o α-spray G_α de (L, H) é flat. (L, H) é chamada de **flat** se é 1-flat.*

Lema 4. *Seja (L, H) uma estrutura de informação em uma variedade M. Para algum $\alpha \neq 0$, (L, H) é α-flat se, e somente se, para todo ponto existe um sistema de coordenadas locais (x^i) tal que*

$$L_{x^k y^l} y^k = 2L_{x^l}, \tag{15}$$

$$\alpha H = -\frac{1}{6} L_{x^k} y^k. \tag{16}$$

5.2 Divergências Afins e Estruturas de Informação afins

Definição 2. *Uma divergência regular D em uma variedade M é chamada* **afim** *se D é uma função C^∞ em uma vizinhança da diagonal $M \times M$.*

Lema 5. *Seja D uma divergência regular afim em uma variedade M. Então a informação induzida (L, H) satisfaz*

(i) $L = g_{ij}(x)y^i y^j$ *é Riemanniana.*

(ii) $H = H_{ijk}(x)y^i y^j y^k$.

Definição 3. *Uma estrutura de informação (L, H) em uma variedade M é chamada de afim se L é Riemanniana e $H = H_{ijk}y^i y^j y^k$ é um polinômio homogêneo.*

6 Aplicações aos Modelos Estatísticos

6.1 f-divergências em espaços de probabilidade

Seja $\mathcal{X} = (\mathcal{X}, \mathcal{B}, \mathcal{V})$ um espaço de medida onde $\mathcal{X}$ é um conjunto não vazio, $\mathcal{B}$ é uma classe aditiva completa que consiste em $\mathcal{X}$ e seus subconjuntos, e $\mathcal{V}$ uma medida σ-finita em $(\mathcal{X}, \mathcal{B})$. Seja $\mathcal{P} = \mathcal{P}(\mathcal{X})$ o espaço das distribuições de probabilidade em $\mathcal{X}$,ou seja,

$$\mathcal{P}(\mathcal{X}) = \{p : \mathcal{X} \to [0, \infty)| \int_{\mathcal{X}} p(r)dr = 1\}.$$

Temos que $\mathcal{P}$ é um conjunto convexo .

Seja f uma função convexa $f : (0, \infty) \to \mathbb{R}$ com $f(1) = 0$ e $f"(1) = 1$. Defina

$$D_f : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$$

$$D_f(p, q) = \int_{\mathcal{X}} p(r) f\left(\frac{q(r)}{p(r)}\right) dr \tag{17}$$

$p, q \in \mathcal{P}$. Temos que D_f é uma divergência em $\mathcal{P}$.

Para $f(t) = -ln(t)$, temos

$$D_{-1}(p, q) = \int_{\mathcal{X}} p(r) ln\left(\frac{p(r)}{q(r)}\right) dr,$$

que é chamada de *divergência de Kullback-Leiber*.

6.2 Os α-sprays em Modelos Estatísticos

Seja $\mathcal{P}$ um espaço de distribuições de probabilidade em um espaço mensurável $\mathcal{X}$ e $\mathcal{D}$ uma divergência em $\mathcal{P}$. Um modelo estatístico em $(\mathcal{P}, \mathcal{D})$ é um par (M, D) , onde M é uma variedade C^∞ mergulhada em $\mathcal{P}$ e $D = \mathcal{D}_{|M}$. Se f é uma função $f : (0, \infty) \to \mathbb{R}$ convexa, tal que $f(1) = 0$ e $f"(1) = 1$, então podemos definir a f-divergência D_f em $\mathcal{P}$ como em (3.2). Agora, vamos provar que para toda variedade $M \subset \mathcal{P}$, a divergência induzida $D_f = \mathcal{D}_{f|M}$ é afim , a saber , a métrica induzida $L = g_{ij}(x)y^i y^j$ é Riemanniana e a H-função induzida é polinomial.

Teorema 1. *Seja $f = f(t)$ uma função $f(1) = 0$ e $f''(1) = 1$. Para todo modelo estatístico regular (M, D_f) de $(\mathcal{P}, \mathcal{D}_f)$, a estrutura de informação induzida em M é dada por $(L_f, H_f) = (L, \rho N)$, onde $\rho := 3 + f'''(1)$ e*

$$L = \int_{\mathcal{X}} \left[y^i \frac{\partial}{\partial x^i} lnp\right]^2 pdr = \mathbb{E}\left[\left(\frac{\partial lnp(r; x)}{\partial x^i} y^i\right)^2\right], \tag{18}$$

$$N = \frac{1}{6}\int_{\mathcal{X}} \left[y^i \frac{\partial}{\partial x^i} lnp\right]^3 pdr = \mathbb{E}\left[\frac{1}{6}\left(\frac{\partial lnp(r; x)}{\partial x^i} y^i\right)^3\right], \tag{19}$$

onde $\mathbb{E}$ denota a esperança matemática. O α-spray $G_{\alpha,\rho}$ de D_f é dado por

$$G^i_{\alpha,\rho} = \overline{G}^i + (\rho\alpha + 1)A^i,$$

onde

$$\overline{G}^i = \frac{g^{il}(x)}{2}\int_{\mathcal{X}}\left[y^r y^j \frac{\partial^2}{\partial x^r \partial x^j} lnp(r;x)\right]\frac{\partial}{\partial x^l}p(r;x)dr \tag{20}$$

e

$$A^i = \frac{g^{il}(x)}{4}\int_{\mathcal{X}}\left[y^i \frac{\partial}{\partial x^i} lnp(r;x)\right]^2\frac{\partial}{\partial x^l}p(r;x)dr. \tag{21}$$

6.3 Família exponencial de distribuições de probabilidade

Definição 4. *Uma variedade M em $\mathcal{P}$ é chamada de* ***família exponencial*** *ou* ***variedade exponencial*** *se é coberta por injeções*

$$\overline{\omega} : \Omega \subset \mathbb{R}^n \to M \subset \mathcal{P},$$

tal que, $p := \overline{\omega}(x) \in \mathcal{P}$ pode ser expressa por

$$p(r;x) = e^{x^i f_i(r) + k(r) - \psi(x)}. \tag{22}$$

Proposição 2. *Seja M uma família exponencial de distribuições de probabilidade como em* (3.44)*. A estrutura de informação induzida de D_f é dada por $(L_f, H_f) = (L, \rho N)$ com $\rho = 3 + 2f'''(1)$ e*

$$L = \frac{\partial \psi^2(x)}{\partial x^i \partial x^j} y^i y^j, \tag{23}$$

$$N = \frac{1}{6}\frac{\partial \psi^3(x)}{\partial x^i \partial x^j \partial x^k} y^i y^j y^k. \tag{24}$$

Exemplo 1. *Considere agora uma família M de distribuições normais gaussianas com média μ e variância σ^2 :*

$$p(r;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left[-\frac{1}{2}\frac{(r-\mu)^2}{2\sigma^2}\right]}.$$

Vamos reparametrizar M por

$$p(r;x) = e^{x^1 f_1(r) + x^2 f_2(r) + k(r) - \psi(x)},$$

onde

$x^1 = \frac{\mu}{\sigma^2}$, $x^2 = \frac{1}{2\sigma^2}$, $f_1(r) = r$, $f_2(r) = -r^2$ *e* $k(r) = 0$.

Teremos então

$$\psi(x) = \frac{\mu^2}{2\sigma^2} + log(\sqrt{2\pi}\sigma) = \frac{(x^1)^2}{4x^2} ln\sqrt{\frac{\pi}{x^2}}. \tag{25}$$

Assim M é uma variedade exponencial em $\mathcal{P}$ e

$$L = \frac{\partial^2 \psi(x)}{\partial x^i \partial x^j} y^i y^j,$$

onde

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

Para a variedade M, obtemos a matriz da métrica:

$$[g_{ij}(x)] = \begin{bmatrix} \frac{1}{2x^2} & -\frac{x^1}{2(x^2)^2} \\ -\frac{x^1}{2(x^2)^2} & \frac{(x^1)^2}{2(x^2)^3} + \frac{1}{2x^2} \end{bmatrix},$$

calculando a curvatura Gaussiana dessa métrica, obteremos $K = -\frac{1}{2}$.

References

[1] Amari, S-I. Methods in Information Geometry AMS, 20 (2000)

[2] Z.Shen Riemann-Finsler Geometry with applications to Information Geometry, to appear in *Chinn. Ann. Math 27B(1),2006,73-94.*