Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Programa de Pós–Graduação em Matemática
Mestrado em Matemática

Aplicações da Geometria Riemanniana em Estatística Matemática

Felipe Fernando Ângelo Barreto

João Pessoa – PB
Agosto de 2013

Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Programa de Pós–Graduação em Matemática
Mestrado em Matemática

Aplicações da Geometria Riemanniana em Estatística Matemática

por

Felipe Fernando Ângelo Barreto

sob orientação do

Prof. Dr. Alexandre de Bustamante Simas

João Pessoa – PB
Agosto de 2013

B273a Barreto, Felipe Fernando Ângelo.
Aplicações da geometria riemanniana em estatística matemática / Felipe Fernando Ângelo Barreto.- João Pessoa, 2013.
59f.: il.
Orientador: Alexandre de Bustamante Simas
Dissertação (Mestrado) - UFPB/CCEN
1. Estatística Matemática. 2. Medida de Influência. 3. Variedade de Perturbação. 4. Tensor Métrico 5. Curvatura. 6. Modelo paramétrico. 7. Conexão afim.

UFPB/BC CDU: 519.2(043)

Aplicações da geometria riemanniana em estatística

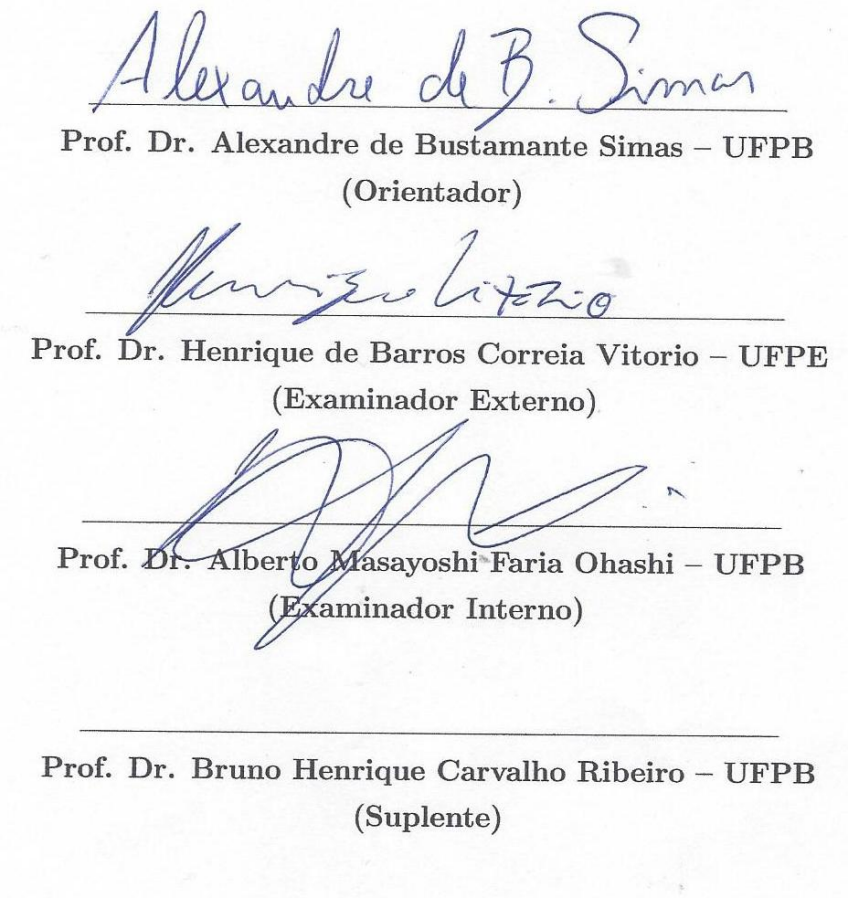
por

Felipe Fernando Ângelo Barreto

Dissertação apresentada ao Corpo Docente do Programa de Pós–Graduação em Matemática da Universidade Federal da Paraíba como requisito parcial para obtenção do título de Mestre em Matemática.

Aprovado em 25 de Setembro de 2013.

Banca Examinadora:



Prof. Dr. Alexandre de Bustamante Simas – UFPB
(Orientador)

Prof. Dr. Henrique de Barros Correia Vitorio – UFPE
(Examinador Externo)

Prof. Dr. Alberto Masayoshi Faria Ohashi – UFPB
(Examinador Interno)

Prof. Dr. Bruno Henrique Carvalho Ribeiro – UFPB
(Suplente)

Aos meus Pais

Agradecimentos

Primeiramente, agradeço a Deus por sempre me conceder sabedoria nas escolhas dos melhores caminhos, coragem para acreditar, força para não desistir e proteção para me amparar.

Agradeço a minha família, sempre tão presente em minha vida, me apoiando, dando força e compreendendo algumas ausências. Sem o grande esforço dos meus pais Carlos e Maristela eu não estaria aqui, escrevendo os agradecimentos de uma dissertação de mestrado. Obrigado!

Ao meu irmão João Carlos por ser mais que um irmão. Por todos esses anos, só tenho a agradecer pela amizade, brincadeiras, etc.

Agradeço especialmente a minha namorada Olívia por estar ao meu lado em todos os momentos bons e difícies desses últimos anos. Obrigado pelo incentivo, apoio, carinho, amor e compreensão. Você foi e será uma pessoa fundamental pra mim.

Aos colegas de mestrado, pela colaboração e companheirismo, que de certa forma contribuíram com este trabalho (José Everton, José Ginaldo, Carlos, Renato, Mônica, Eudes, etc).

Agradeço aos meus dois amigos Eberson e Rafael, que desde a graduação sempre me ajudaram e contribuiram positivamente para essa conquista.

A todos os professores da graduação e do mestrado, em especial aos professores Maité, Jorge Hinojosa, Adriano, Rodrigo, Napoleón, Daniel e Jacqueline.

Agradeço ao meu orientador professor Doutor Alexandre de Bustamante Simas, pela paciência, confiança e orientações prestadas para a realização deste trabalho.

E para finalizar, não podia esquecer de agradecer à capes, pela concessão da bolsa de estudo.

"Para Tales... a questão primordial não era o que sabemos, mas como sabemos."

Aristóteles

Resumo:

A Abordagem de influência local de Cook [2] com base em curvatura normal é uma importante ferramenta de diagnóstico para avaliar a influência local de pequenas perturbações de um modelo estatístico. No entanto, tem sido desenvolvida nenhuma abordagem rigorosa para abordar duas questões fundamentais: a escolha de uma perturbação apropriada e o desenvolvimento de medidas de influência para funções objetos em um ponto com a primeira derivada diferente de zero. O objetivo deste trabalho é desenvolver uma estrutura diferencial-geométrica de um modelo de perturbação (chamado de variedade de perturbação) e utilizar o tensor métrico associado e as curvaturas afins para resolver esses problemas. Vamos mostrar que o tensor métrico da variedade de perturbação fornece informações importantes sobre a seleção de uma perturbação apropriada de um modelo.

Palavras-chave: Medida de Influência, Variedade de Perturbação, tensor métrico, curvatura, modelo paramétrico, conexão afim.

Abstract

Cook's local infuence approach based on normal curvature is an important diagnostic tool for assessing local infuence of minor perturbations to a statistical model. However, no rigorous approach has been developed to address two fundamental issues: the selection of an appropriate perturbation and the development of infuence measures for objective functions at a point with a nonzero first derivative. The aim of this paper is to develop a diferential-geometrical framework of a perturbation model (called the perturbation manifold) and utilize associated metric tensor and affine curvatures to resolve these issues. We will show that the metric tensor of the perturbation manifold provides important information about selecting an appropriate perturbation of a model.

Keywords: Infuence Measure, perturbation manifold, metric tensor, curvature, parametric model, affine connection.

Sumário

Introdução

Capítulo 1

Preliminares

1.1 Tópicos em Teoria da Probabilidade

Nosso objetivo nesta seção é citar alguns resultados importantes da teoria de probabilidade. Evidentemente, não nos prenderemos a detalhes e demonstrações, pois o interesse é apenas tornar o texto mais auto-suficiente. Para um aprofundamento maior, indicaremos as referências onde tais demonstrações podem ser encontradas. Começaremos definindo variável aleatória, supondo conhecidas as noções de teoria da medida.

1.1.1 Variáveis Aleatórias

Definição 1.1. Sejam $(\Omega, \mathcal{F})$ e $(S, \mathcal{S})$ espaços mensuráveis. Uma função $X : \Omega \to S$ é dita *mensurável* de $(\Omega, \mathcal{F})$ em $(S, \mathcal{S})$ se

$$X^{-1}(B) = \{w : X(w) \in B\} \in \mathcal{F} \qquad \forall\ B \in \mathcal{S}.$$

Se $(S, \mathcal{S}) = (\mathbf{R^d}, \mathcal{R}^d)$ e $d > 1$, então X é chamado um *vetor aleatório*. Naturalmente, se $d = 1$, X é chamada uma *variável aleatória*.

Vejamos alguns exemplos.

Exemplo 1.1. Se Ω é um espaço de probabilidade discreto, então qualquer função $X : \Omega \to \mathbf{R}$ é uma variável aleatória. Um outro exemplo trivial, porém útil, de variável aleatória é a **função indicadora** de um conjunto $A \in \mathcal{F}$:

$$1_A(w) = \begin{cases} 1, & \text{se} \quad w \in A \\ 0, & \text{se} \quad w \notin A \end{cases} .$$

Se X é uma variável aleatória, então ela induz uma medida de probabilidade sobre $\mathbf{R}$ chamada

distribuição de X dada por

$$\mu(A) = P(X \in A),$$

onde A são conjuntos bolerianos. O lado direito da equação acima pode ser escrito como $P(X^{-1}(A))$. A distribuição de uma variável aleatória X é geralmente descrita indicando a sua função distribuição, $F(x) = P(X \leq x)$.

O resultado a seguir é útil para provar mensurabilidade de funções.

Teorema 1.1. *Se $\{w : X(w) \in A\} \in \mathcal{F}$ para todo $A \in \mathcal{A}$ e $\mathcal{A}$ gera $\mathcal{S}$ (isto é, $\mathcal{S}$ é a menor σ-álgebra que contém $\mathcal{A}$), então X é mensurável.*

Demonstração. Considere $\{X \in B\}$ uma abreviação para $\{w : X(w) \in B\}$, temos que

$$\{X \in \cup_i B_i\} = \cup_i \{X \in B_i\},$$

e

$$\{X \in B^c\} = \{X \in B\}^c.$$

Assim, a classe de conjuntos $\mathcal{B} = \{B : \{X \in B\} \in \mathcal{F}\}$ é uma σ-álgebra. Uma vez que, $\mathcal{B} \supset \mathcal{A}$ e $\mathcal{A}$ gera $\mathcal{S}$, $\mathcal{B} \supset \mathcal{S}$. □

Teorema 1.2. *Se $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ e $f : (S, \mathcal{S}) \to (T, \mathcal{T})$ são funções mensuráveis, então $f(X)$ é uma função mensurável de $(\Omega, \mathcal{F})$ em $(T, \mathcal{T})$.*

Demonstração. Seja $B \in \mathcal{T}$. Temos que $\{\omega : f(X(\omega)) \in B\} = \{\omega : X(\omega) \in f^{-1}(B)\} \in \mathcal{F}$, uma vez que, por hipótese, $f^{-1}(B) \in \mathcal{S}$. □

Teorema 1.3. *Se $X_1, ..., X_n$ são variáveis aleatórias e $f : (R^n, \mathcal{R}^n) \to (R, \mathcal{R})$ é mensurável, então $f(X_1, ..., X_n)$ é uma variável aleatória.*

Demonstração. Tendo em vista o teorema (1.2), basta mostrar que $(X_1, ..., X_n)$ é um vetor aleatório. Para fazer isso, observe que se $A_1, ..., An$ são conjuntos borelianos então

$$\{(X_1, ..., X_n) \in A_1 \times \cdots \times A_n\} = \cap_i \{X_i \in A_i\} \in \mathcal{F}.$$

Uma vez que, os conjuntos da forma $A_1 \times \cdots \times A_n$ geram $\mathcal{R}^n$, o teorema (1.3) segue do teorema (1.1). □

Corolário 1.1. *Se $X_1, ..., X_n$ são variáveis aleatórias, então $X_1 + \cdots + X_n$ é uma variável aleatória.*

1.1.2 Valor Esperado

Definição 1.2. Se $X \geq 0$ é uma variável aleatória sobre $(\Omega, \mathcal{F}, P)$, então definimos o seu *Valor Esperado* (ou *Esperança*) como sendo $EX = \int X dP$.

Uma vez que EX é definido por meio da integral de X, consideraremos todas as propriedades de integração como verdadeiras. Vejamos agora dois teoremas.

Teorema 1.4. *Suponha $X, Y \geq 0$ e $E|X|, E|Y| < \infty$.*

1. *$E(X+Y) = EX + EY$.*
2. *$E(aX+b) = aE(X) + b$ para quaiquer $a, b \in \mathcal{R}$.*
3. *Se $X \geq Y$ então $EX \geq EY$.*

Teorema 1.5 (**Mudança de variável**). *Seja X um elemento aleatório de $(S, \mathcal{S})$ com distribuição μ, i.e., $\mu(A) = P(X \in A)$. Se $f : (S, \mathcal{S}) \to (\mathbf{R}, \mathcal{R})$ é uma função mensurável tal que $f \geq 0$ ou $E|f(X)| < \infty$, então*

$$Ef(X) = \int_S f(y)\mu(dy)$$

Observação. *Para explicar o nome, escreva h e $P \circ h^{-1}$ no lugar de X e μ, respectivamente, para obter*

$$\int_\Omega f(h(\omega))dP = \int_S f(y)d(P \circ h^{-1})$$

Demonstração. Ver referência [7] □

Uma consequência do teorema (1.5) é que podemos calcular valores esperados de funções de variáveis aleatórias efetuando integrais sobre a reta real. Se **k** é um inteiro positivo então EX^k é chamado o **k-ésimo momento** de X. O primeiro momento EX é geralmente chamado de média e denotado por μ. Se $EX^2 < \infty$ então a **variância** de X é definida por $var(x) = E(X-\mu)^2$. Para calcular a variância, a seguinte fórmula é útil:

$$var(X) = E(X-\mu)^2 = EX^2 - 2\mu EX + \mu^2 = EX^2 - \mu^2 \tag{1.1}$$

A partir de (1.1), é imediato que

$$var(X) \leq EX^2 \tag{1.2}$$

Uma vez que $E(aX+b) = aEX + b$, segue facilmente da definição que

$$var(aX+b) = E(aX+b-E(aX+b))^2 = a^2E(X-EX)^2 = a^2var(X) \tag{1.3}$$

1.1.3 Independência

Seja $(\Omega, \mathcal{F}, P)$ um espaço de probabilidade. Dois eventos A e B são *independentes* se

$$P(A \cap B) = P(A)P(B).$$

Duas variáveis aleatórias X e Y são *independentes* se $\forall\ C, D \in \mathcal{R}$,

$$P(X \in C, Y \in D) = P(X \in C)P(Y \in D)$$

isto é, os eventos $A = \{X \in C\}$ e $B = \{Y \in D\}$ são independentes.

Proposição 1.1. *Se A e B são independentes, então A e B^c também são independentes (e também A^c e B, e ainda A^c e B^c).*

Demonstração. Suponha que A e B são independentes. Então

$$P(A \cap B^c) = P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(B^c).$$

□

Definição 1.3. σ-Álgebras $\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_n$ são *independentes* se sempre que $A_i \in \mathcal{F}_i$ com $i = 1, ..., n$, nós tivermos

$$P(\cap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i).$$

Variáveis aleatórias $X_1, ..., X_n$ são *independentes* se sempre que $B_i \in \mathcal{R}$ com $i = 1, ..., n$, nós tivermos

$$P(\cap_{i=1}^n \{X_i \in B_i\}) = \prod_{i=1}^n P(X_i \in B_i).$$

Conjuntos $A_1, ..., A_n$ *são independentes* se sempre que $I \subset \{1, ..., n\}$, nós tivermos

$$P(\cap_{i \in I} A_i) = \prod_{i \in I} P(A_i).$$

Uma das primeiras coisas a compreender sobre a definição de eventos independentes é que dada uma sequência de eventos $A_1, ..., A_n$ que satisfaz a propriedade $P(A_i \cap A_j) = P(A_i)P(A_j),\ \forall\ i \neq j$, não nos garante que $A_1, ..., A_n$ são independentes. Para verificarmos isso, vejamos um exemplo.

Exemplo 1.2. Sejam X_1, X_2, X_3 variáveis aleatórias independentes com

$$P(X_i = 0) = P(X_i = 1) = 1/2.$$

Considere $A_1 = \{X_2 = X_3\}$, $A_2 = \{X_3 = X_1\}$ e $A_3 = \{X_1 = X_2\}$. Note que, a propriedade

$$P(A_i \cap A_j) = P(X_1 = X_2 = X3) = 1/4 = P(A_i)P(A_j)$$

é satisfeita, mas eles não são independentes, uma vez que

$$P(A_1 \cap A_2 \cap A_3) = 1/4 \neq 1/8 = P(A_1)P(A_2)P(A_3).$$

Definição 1.4. *Dizemos que as variáveis aleatórias X e Y são independentes se $\sigma(X)$ e $\sigma(Y)$ são σ-álgebras independentes, onde $\sigma(X)$ denota a menor σ-álgebra tal que X é mensurável.*

Em virtude da definição acima, a fim de mostrar que as variáveis aleatórias X e Y são independentes, temos de verificar que $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ para todos os conjuntos borelianos A e B. Uma vez que, há um grande número de conjuntos borelianos, iremos ver agora algumas condições suficientes para independência.

Estamos interessados no resultado do teorema (1.6). Para isso, vejamos uma definição que generaliza todas as nossas definições anteriores sobre independência.

Definição 1.5. As coleções de conjuntos $\mathcal{A}_1, ..., \mathcal{A}_n \subset \mathcal{F}$ são ditas *independentes* se sempre que $A_i \in \mathcal{A}_i$ e $I \subset \{1, ..., n\}$, nós tivermos

$$P(\cap_{i=1} A_i) = \prod_{i \in I} P(A_i).$$

Se cada coleção é um conjunto unitário, isto é, $\mathcal{A}_i = \{A_i\}$, esta definição reduz ao utilizado para conjuntos. Se cada $\mathcal{A}_i$ contém Ω, por exemplo, $\mathcal{A}_i$ é uma σ-álgebra, a condição é equivalente a $P(\cap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i)$ sempre que $A_i \in \mathcal{A}_i$, uma vez que podemos definir $A_i = \Omega$ para $i \notin I$. Reciprocamente, se $\mathcal{A}_1, ..., \mathcal{A}_n$ são independentes e $\overline{\mathcal{A}}_i = \mathcal{A}_i \cup \{\Omega\}$, então $\overline{\mathcal{A}}_1, ..., \overline{\mathcal{A}}_n$ são independentes, por isso não há perda de generalidade em supor $\Omega \in \mathcal{A}_i$.

Antes de enuciarmos o teorema (1.6), precisamos da seguinte definição.

Definição 1.6. Dizemos que $\mathcal{A}$ é um sistema π se dados $A, B \in \mathcal{A}$ então $A \cap B \in \mathcal{A}$.

Teorema 1.6. *Suponha $\mathcal{A}_1, ..., \mathcal{A}_n$ independentes e cada $\mathcal{A}_i$ é um sistema π. Então $\sigma(\mathcal{A}_1), \sigma(\mathcal{A}_2), ..., \sigma(\mathcal{A}_n)$ são independentes.*

Demonstração. Ver a referência [7] □

Corolário 1.2. *A fim de que $X_1, ..., X_n$ sejam independentes, é suficiente mostrar que $\forall\ x_1, ..., x_n \in (-\infty, \infty]$*

$$P(X_1 \leq x_1, ..., X_n \leq x_n) = \prod_{i=1}^{n} P(X_i \leq x_i).$$

1.1.4 Leis dos grandes números e o teorema central do limite

Quando $X_1, X_2, ...$ são variáveis aleatórias independentes onde todas elas possuem a mesma distribuição, dizemos que são *independentes e identicamente distribuídas* (i.i.d.).

Teorema 1.7 (**Lei fraca dos grandes números**). *Se $X_1, X_2, ...$ são variáveis aleatórias i.i.d. com média $EX_i = \mu$, então para todo $\epsilon > 0$*

$$P(|S_n/n - \mu| > \epsilon) \to 0 \qquad \textit{quando} \quad n \to \infty$$

onde $S_n = X_1 + \cdots + X_n$.

Teorema 1.8 (**Lei forte dos grandes números**). *Sejam* $X_1, X_2, \ldots$ *variáveis aleatórias i.i.d. com* $E|X_i| < \infty$. *Considere* $EX_i = \mu$ *e* $S_n = X_1 + \cdots + X_n$. *Então* $S_n/n \to \mu$ *quase certamente quando* $n \to \infty$.

Teorema 1.9 (**Teorema central do limite**). *Se* $X_1, X_2, \ldots$ *são variáveis aleatórias i.i.d. com média* $EX_i = \mu$ *e variância* $\sigma^2 = E(X_i - \mu)^2$, *então para qualquer* y

$$P\left(\frac{S_n - n\mu}{\sigma n^{1/2}} \leq y\right) \to \mathcal{N}(y)$$

onde $\mathcal{N}(y) = \int_{-\infty}^{y} (2\pi)^{-1/2} \exp^{-x^2/2}\, dx$ *é a distribuição normal padrão.*

Demonstração. Ver a referência [7] □

Definição 1.7. Dados um espaço de probabilidade $(\Omega, \mathcal{F}_0, \mathbb{P})$, uma σ-álgebra $\mathcal{F} \subset \mathcal{F}_0$, e uma variável aleatória $X \in \mathcal{F}_0$ com $E|X| < \infty$. Definimos a *esperança condicional de* X *dado* $\mathcal{F}$, $E(X|\mathcal{F})$, sendo qualquer variável aleatória Y que

(i) $Y \in \mathcal{F}$, isto é, é $\mathcal{F}$ mensurável;

(ii) Para todo $A \in \mathcal{F}$, $\int_A X dP = \int_A Y dP$

Qualquer Y satisfazendo (i) e (ii) é chamado uma *versão de* $E(X|\mathcal{F})$.

Definição 1.8. Sejam $(\Omega, \mathcal{F}, P)$ um espaço de probabilidade, $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ uma função mensurável, e $\mathcal{G}$ uma σ-álgebra $\subset \mathcal{F}$. Dizemos que $\mu : \Omega \times \mathcal{S} \to [0, 1]$ é uma *distribuição condicional regular* para X dado $\mathcal{G}$, se

(i) Para cada A, $\omega \to \mu(\omega, A)$ é uma versão de $P(X \in A|\mathcal{G})$.

(ii) Para quase todo ω, $A \to \mu(\omega, A)$ é uma medida de probabilidade sobre $(S, \mathcal{S})$.

Quando $S = \Omega$ e X é a função identidade, μ é chamada uma *probabilidade condicional regular.*

Vamos agora apresentar um teorema que mostra a existência da probabilidade condicional regular.

Teorema 1.10. *Seja* $(M, \mathcal{M}, P)$ *um espaço de probabilidade, onde* M *é um espaço métrico completo e separável, e* $\mathcal{M}$ *é a* σ*-álgebra de Borel de* M. *Então, dados vetores aleatórios* X *e* T, *existe uma função* $P_{X|T} : \sigma(X) \times \mathbb{R}^d \to [0, 1]$ *tal que, para cada* $A \in \sigma(X)$, $P(A \mid \cdot)$ *é uma função mensurável, e para quase todo* t, $P(\cdot \mid t)$ *é uma medida de probabilidade. Além disso, para toda função* f *mensurável, tal que* $f(X)$ *é integrável, temos*

$$E(f(X) \mid T = t) = \int f(x) P_{X|T}(dx, t),$$

para todo t num conjunto de medida total. Finalmente, se h é uma função mensurável tal que $h(T)$ é integrável, temos que

$$E(f(X)h(T) \mid T = t) = h(t)E(f(X) \mid T = t).$$

Observe que como

$$E(E(f(X) \mid T)) = E(f(X)),$$

segue que

$$E(f(X)) = E(E(f(X) \mid T)) = \int E(f(X) \mid T = t)dP^T(dt) = \int\int f(x)P_{X|T}(dx, t)dP^T(dt),$$

onde P^T é a distribuição de T.

Outra observação relevante é o caso em que o vetor aleatório X se "decompõe" como $X = (Y, T)$. Então, temos que

$$E(f(X)) = E(f(Y, T)) = \int_{\mathbb{R}^d}\int f(y, t)P^{Y|T}(dy, t)P^T(dt).$$

Corolário 1.3. *Seja Y um vetor aleatório p-dimensional e T um vetor aleatório d-dimensional. Seja $P^{Y,T}$ uma medida de probabilidade no espaço produto $\mathbb{R}^p \times \mathbb{R}^d$, na σ-álgebra boreliana produto. Seja $Q^{Y,T}$ outra medida de probabilidade no mesmo espaço produto tal que*

$$dQ(y, t) = a(y)b(t)dP(y, t),$$

onde $a(y) > 0$ para todo y. Então, a distribuição marginal de T sobre Q pode ser dada em termos da distribuição marginal de T sobre P da forma:

$$dQ^T(t) = b(t)\left[\int a(y)P^{Y|T}(dy, t)\right]dP^T(t).$$

Além disso, a distribuição condicional de Y dado T sobre Q pode ser dada em termos da distribuição condicional sobre P como:

$$dQ^{Y|T}(y, t) = \frac{a(y)P^{Y|T}(dy, t)}{\int a(\tilde{y})P^{Y|T}(d\tilde{y}, t)},$$

onde esta fração está bem definida já que $a(y) > 0$.

Demonstração. Note que

$$\begin{aligned} Q^T(A) &= Q(T \in A) = E_Q[1_B(T)] = E_P[1_B(T)b(T)a(Y)] \\ &= E_P[1_B(T)b(T)E[a(Y) \mid T]] = \int_B b(t) \int a(y)P^{Y|T}(dy,t)dP^T(t), \end{aligned}$$

donde segue a primeira afirmação. Vamos para a segunda afirmação. Temos que mostrar que para quaisquer A e B mensuráveis, temos que

$$E_Q[1_A(T)\frac{\int 1_B(y)a(y)P^{Y|T}(dy,T)}{\int a(\tilde{y})dP^{Y|T}(\tilde{y},T)}] = E_Q[1_A(T)1_B(Y)].$$

Note que

$$\begin{aligned} E_Q[1_A(T)\frac{\int 1_B(y)a(y)P^{Y|T}(dy,T)}{\int a(\tilde{y})P^{Y|T}(d\tilde{y},T)}] &= E_P[a(Y)b(T)1_A(T)\frac{\int 1_B(y)a(y)P^{Y|T}(dy,T)}{\int a(\tilde{y})P^{Y|T}(d\tilde{y},T)}] \\ &= E_P[a(Y)b(T)1_A(T)\frac{E_P(1_B(Y)a(Y) \mid T)}{E_P(a(Y) \mid T)}] \\ &= E_P[a(Y)E_P\left(\frac{b(T)1_A(T)1_B(Y)a(Y)}{E_P(a(Y) \mid T)} \mid T\right)] \\ &= E_P[E_P(a(Y) \mid T)E_P\left(\frac{b(T)1_A(T)1_B(Y)a(Y)}{E_P(a(Y) \mid T)} \mid T\right)] \\ &= E_P[E_P(b(T)1_A(T)1_B(Y)a(Y) \mid T)] \\ &= E_P[b(T)a(Y)1_A(T)1_B(Y)] \\ &= E_Q[1_A(T)1_B(Y)], \end{aligned}$$

o que prova o resultado. □

Capítulo 2

Tópicos em Fundamentos de Estatística

2.1 População, Amostra e Modelos

Usualmente, é impraticável observar toda uma população, seja pelo custo alto seja por dificuldades operacionais. Examina-se então uma amostra, de preferência bastante representativa, para que os resultados obtidos possam ser generalizados para toda a população. Um experimento pode ter por finalidade a determinação da estimativa de um parâmetro de uma função. Toda conclusão tirada por amostragem, quando generalizada para a população, apresentará um grau de incerteza. Ao conjunto de técnicas e procedimentos que permitem dar ao pesquisador um grau de confiabilidade nas afirmações que faz para a população, baseadas nos resultados das amostras, damos o nome de Inferência Estatística.

Na Inferência Estatística, o conjunto de dados é visto como uma realização ou observação de um elemento aleatório definido em um espaço de probabilidade $(\Omega, \mathcal{F}, P)$ relacionado com o experimento aleatório. A medida de probabilidade P é chamada de *população*. O conjunto de dados ou elemento aleatório que gera os dados é chamado de *amostra* de P e a cardinalidade do conjunto de dados é chamado de *dimensão da amostra*. A população P é *conhecida* se, e somente se, $P(A)$ é um valor conhecido para cada evento $A \in \mathcal{F}$.

Exemplo 2.1. Para medir uma quantidade desconhecida θ (por exemplo, uma distância, peso, ou temperatura), n medições, $x_1, ..., x_n$, são tomados em um experimento de medição θ. Se θ pode ser medido sem erros, então $x_i = \theta$ para todo i, caso contrário, cada x_i tem um possível erro de medição. Na análise descritiva dos dados, certas medidas podem ser calculadas, como por exemplo, a média da amostra

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.1}$$

e a variância da amostra

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{2.2}$$

No entanto, qual é a relação entre $\overline{x}$ e θ? Eles estão próximos (se não iguais) em algum caso? A variância da amostra s^2 é claramente uma média de desvios ao quadrados dos x_i's com a sua média. Mas que tipo de informação o s^2 nos fornece? Finalmente, não é suficiente apenas olhar para $\overline{x}$ e s^2 para poder medir θ? Essas questões não podem ser respondidas em análise descritiva de dados.

Na Inferencia estatística e teoria da decisão, o conjunto de dados, $(x_1, ..., x_n)$, é visto como um resultado do experimento cujo espaço amostral é $\Omega = \mathcal{R}^n$. Nós geralmente assumimos que as n medições são obtidas em n testes independentes do experimento. Por isso, nós podemos definir um n-vetor aleatório $X = (X_1, ..., X_n)$ sobre $\Pi_{i=1}^n(\mathcal{R}, \mathcal{B}, P)$ cuja realização é $(x_1, ..., x_n)$. A população neste problema é P (note que a medida produto obtida é de probabilidade, sendo determinada por P) e é pelo menos parcialmente desconhecida. O vetor aleatório X é uma amostra e n é a dimensão da amostra. Defina

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.3}$$

e

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{2.4}$$

Então $\overline{X}$ e S^2 são variáveis aleatórias que produzem $\overline{x}$ e s^2, respectivamente.

2.1.1 Modelos Paramétricos, Famílias Exponenciais e Localização-escala

Definição 2.1. Um conjunto de medidas de probabilidade P_θ em $(\Omega, \mathcal{F})$, indexado por um parâmetro $\theta \in \Theta$, é dito ser uma *família paramétrica* se, e somente se, $\Theta \subset \mathcal{R}^d$ para algum inteiro fixo positivo d e cada P_θ é uma medida de probabilidade *conhecida* quando θ for conhecido. O conjunto Θ é chamado *espaço de parâmetro* e d é chamado de *dimensão.*

Um modelo paramétrico refere-se à suposição de que a população P pertence a uma família paramétrica. Uma família paramétrica $\{P_\theta : \theta \in \Theta\}$ é dita ser *identificável* se, e somente se, $\theta_1 \neq \theta_2$ e $\theta_i \in \Theta$ implica $P_{\theta_1} \neq P_{\theta_2}$.

Na maioria dos casos, uma família paramétrica identificável pode ser obtida por meio de uma reparametrização. Assumiremos a partir de agora que cada família paramétrica é identificável.

Sejam $\mathcal{P}$ uma família de populações e ν uma medida $\sigma-$finita sobre (Ω, F). Se $P \ll \nu$ para todo $P \in \mathcal{P}$ então dizemos que a família $\mathcal{P}$ é *dominada por* ν.

Exemplo 2.2. Para um inteiro positivo k fixado, considere a distribuição normal $N_k(\mu, \Sigma)$ dado por

$$f(x) = (2\pi)^{-k/2}[Det(\Sigma)]^{-1/2}e^{-(x-\mu)^\tau\Sigma^{-1}(x-u)/2}, \quad x \in \mathcal{R}^k \tag{2.5}$$

onde $\mu \in \mathcal{R}^k$, Σ é uma matriz positiva de ordem k, Det(Σ) é o determinante de Σ e cada k-vetor c é visto como uma matriz coluna $(k \times 1)$, onde c^τ denota a sua transposta. Uma importante

família paramétrica na estatística é a família de distribuição normal

$$\mathcal{P} = N_k(\mu, \Sigma) : \mu \in R^k, \Sigma \in M_k,$$

onde M_k é uma coleção de matrizes positivas simétricas de ordem k. Esta família é dominada pela medida de Lebesgue sobre $\mathcal{R}^k$.

No exemplo 2.1, $X_i's$ são frequentemente i.i.d. a partir da distribuição $N(\mu, \sigma^2)$. Por isso, nós podemos impor um modelo paramétrico sobre a população, isto é, $P \in \mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathcal{R}, \sigma^2 > 0\}$.

Uma família de medida de probabilidade que não satisfaz a definição 2.1 é chamada *família não paramétrica*. Um modelo *não paramétrico* refere-se a suposição de que a população P é uma família não paramétrica.

Definição 2.2. [Famílias Exponenciais] Uma família paramétrica $\{P_\theta : \theta \in \Theta\}$ dominada por uma medida σ-finita ν sobre $(\Omega, \mathcal{F})$ é chamada uma *família exponencial* se, e somente se,

$$\frac{dP_\theta}{d\nu}(w) = \exp\{[\eta(\theta)]^\tau T(w) - \xi(\theta)\}h(w), \quad w \in \Omega, \tag{2.6}$$

onde $exp\{x\} = e^x$, T é um p-vetor aleatório com p sendo um inteiro positivo fixado, η é uma função de Θ em $\mathcal{R}^p$, $\xi(\theta) = \log\{\int_\Omega \exp[\eta(\theta)]^\tau T(\omega)h(\omega)d\nu(\omega)\}$, e h é uma função Boreliana não negativa sobre $(\Omega, \mathcal{F})$.

Na definição 2.2, T e h são funções de ω, enquanto que η e ξ são funções de θ. Geralmente, Ω é $\mathcal{R}^k$. A representação (2.6) de uma família exponencial não é única. Na verdade, qualquer transformação $\tilde{\eta}(\theta) = D\eta(\theta)$ com uma matriz não singular D de ordem p dá uma outra representação (com T substituído por $\tilde{T} = (D^\tau)^{-1}T$). A alteração da medida que domina a família também muda a representação. Por exemplo, se definirmos $\lambda(A) = \int_A h d\nu$ para algum A$\in \mathcal{F}$, então nós obtemos uma família exponencial com densidade

$$\frac{dP_\theta}{d\lambda}(\omega) = exp\{[\eta(\theta)]^\tau T(\omega) - \xi(\theta)\}. \tag{2.7}$$

Numa família exponencial, considere a reparametrização $\eta = \eta(\theta)$ e

$$f_\eta(\omega) = exp\{\eta^\tau T(\omega) - \zeta(\eta)\}h(w), \quad \omega \in \Omega, \tag{2.8}$$

onde $\zeta(\eta) = log\{\int_\Omega exp\{\eta^\tau T(\omega)\}h(\omega)d\nu(\omega)\}$. Esta é a forma canônica para a família. O novo parâmetro η é chamado de *parâmetro natural*. O novo espaço de parâmetro $\Xi = \{\eta(\theta) : \theta \in \Theta\}$, um subconjunto de $\mathcal{R}^p$, é chamado de *espaço de parâmetro natural*. Uma família exponencial na forma canônica é chamada de *Família exponencial Natural*. Se existe um conjunto aberto

contido no espaço de parâmetros natural de uma família exponencial, então a família é dita ser de *posto completo*.

Exemplo 2.3. Seja P_θ a distribuição binomial $Bin(\theta, n)$ com parâmetro θ, onde n é um inteiro fixo positivo. Então $\{P_\theta : \theta \in (0,1)\}$ é uma família exponencial, uma vez que a função densidade probabilidade de P_θ com relação à medida de contagem é

$$f_\theta(x) = \exp\{x \log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\}\binom{n}{x} I_{\{0,1,\ldots,n\}}(x)$$

$(T(x) = x, \eta(\theta) = \log\frac{\theta}{1-\theta}, \xi(\theta) = -n\log(1-\theta),$ e $h(x) = \binom{n}{x} I\{0,1,...,n\}(x))$.
Se deixarmos $\eta = \log\frac{\theta}{1-\theta}$, então $\Xi = \mathcal{R}$ e a família com densidades

$$f_\eta(x) = exp\{x\eta - n\log(1+e^\eta)\}\binom{n}{x} I\{0,1,...,n\}(x)$$

é uma família exponencial natural de posto completo.

Exemplo 2.4. A família normal $\{N(\mu,\sigma^2) : \mu \in \mathcal{R}, \sigma > 0\}$ é uma família exponencial, uma vez que a f.d.p. de Lebesgue de $N(\mu,\sigma^2)$ pode ser escrita como

$$\frac{1}{\sqrt{2\pi}} exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma\right\}.$$

Por isso, $T(x) = (x, -x^2), \eta(\theta) = (\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2}), \theta = (\mu,\sigma^2), \xi(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sigma$, e $h(x) = 1/\sqrt{2\pi}$. Seja $\eta = (\eta_1,\eta_2) = (\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2})$. Então $\Xi = \mathcal{R}\times(0,\infty)$ e nós podemos obter uma família exponencial natural de posto completo com $\zeta(n) = \eta_1^2/(4\eta_2) + \log(1/\sqrt{2\eta_2})$.

Uma subfamília da família normal citada acima, $\{N(\mu,\mu^2) : \mu \in \mathcal{R}, \mu \neq 0\}$, é também uma família exponencial com o parâmetreo natural $\mu = (\frac{1}{\mu}, \frac{1}{2\mu^2})$ e espaço de parâmetro natural $\Xi = \{(x,y) : y = 2x^2, x \in \mathcal{R}, y > 0\}$. Esta família exponencial não é de posto completo.

Para uma família exponencial, (2.7) implica que existe uma medida λ diferente de zero tal que

$$\frac{dP_\theta}{d\lambda}(\omega) > 0, \quad \forall\ \omega \quad \text{e} \quad \theta. \tag{2.9}$$

Nós vamos usar esse fato para mostrar que uma família de distribuição não é uma família exponencial. De fato,

Considere a família de distribuição uniforme, i.e., P_θ é $U(0,\theta)$ com um desconhecido $\theta \in (0,\infty)$. Se $\{P_\theta : \theta \in (0,\infty)\}$ é uma família exponencial, então pela discussão anterior existe uma medida $\lambda \neq 0$ tal que (2.9) é verdadeira. Para cada $t > 0$, existe um $\theta < t$ tal que $P_\theta([t,\infty)) = 0$ que por (2.9) implica $\lambda([t,\infty)) = 0$. Também, para cada $t \leq 0, P_\theta((-\infty,t]) = 0$ que por (2.9) implica que $\lambda((-\infty,t]) = 0$. Uma vez que t é arbitrário, $\lambda \equiv 0$. Esta contradição implica que $\{\mathcal{P}_\theta : \theta \in (0,\infty)\}$ não pode ser uma família exponencial.

Exemplo 2.5. [A família Multinomial]. Considere um experimento dividido em n ensaios independentes, no qual cada ensaio resulta em um número finito $k+1$ de valores possíveis com probabilidade $p_0, p_1, ..., p_k$ (de modo que $p_i \geq 0$, para $i = 0, ..., k$, $\Sigma_{i=1}^{k} p_i = 1$). Seja X_i a variável aleatória, que representa o número de vezes que o indice i foi observado nos n ensaios. Então a função densidade de probabilidade (com relação à medida de contagem) de $(X_0, X_1, ..., X_k)$ é

$$f_\theta(x_0, x_1, ..., x_k) = \frac{n!}{x_0! x_1! \cdots x_k!} p_0^{x_0} p_1^{x_1} \cdots p_k^{x_k} I_B(x_0, x_1, ..., x_k),$$

onde $B = \{(x_0, x_1, ..., x_k) : x_i's \text{ são inteiros} \geq 0, \Sigma_{i=0}^{k} x_i = n\}$ e $\theta = (p_0, p_1, ..., p_k)$. A distribuição de $(X_0, X_1, ..., X_k)$ é chamada de distribuição Multinomial, o qual é uma extensão da distribuição binomial. De fato, a distribuição marginal de cada X_i é a distribuição binomial $Bin(p_i, n)$. Seja $\Theta = \{\theta = (p_0, \ldots, p_k) \in \mathcal{R}^{k+1} : 0 < p_i < 1, \Sigma_{i=0}^{k} p_i = 1\}$. A família paramétrica $\{f_\theta : \theta \in \Theta\}$ é chamada de família Multinomial. Sejam $x = (x_0, ..., x_k), \eta = (\log p_0, \log p_1, ..., \log p_k)$, e $h(x) = [n!/(x_0! x_1! \cdots x_k!)] I_B(x)$. Então,

$$f_\theta(x_0, x_1, ..., x_k) = exp\{\eta^\tau x\} h(x), \quad x \in \mathcal{R}^{k+1}. \tag{2.10}$$

Por isso, a família multinomial é uma família exponencial natural com parâmetro natural η. No entanto, a representação (2.10) não fornece uma família exponencial de posto completo, uma vez que, não existe um conjunto aberto de $\mathcal{R}^{k+1}$ contido no espaço de parâmetro natural. Uma reparametrização conduz a uma família exponencial de posto completo. Usando o fato que $\Sigma_{i=0}^{k} X_i = n$ e $\Sigma_{i=0}^{k} p_i = 1$, nós obtemos que

$$f_\theta(x_0, x_1, ..., x_k) = exp\{\eta_*^\tau x_* - \zeta(\eta_*)\} h(x), \quad x \in \mathcal{R}^{k+1}\}, \tag{2.11}$$

onde $x_* = (x_1, ..., x_k)$, $\eta_* = (\log(p_1/p_0), ..., \log(p_k/p_0))$, e $\zeta(\eta_*) = -n \log p_0$. O espaço de η_*-parâmetro é o $\mathcal{R}^k$. Assim, a família de densidade dada por (2.11) é uma família exponencial de posto completo.

Se $X_1, ..., X_m$ são vetores aleatórios independentes com função densidade de probabilidade sobre famílias exponenciais, então a função densidade de probabilidade de $(X_1, ..., X_m)$ é novamente uma família exponencial. O resultado a seguir, resume algumas propriedades úteis de famílias exponenciais.

Teorema 2.1. *Seja $\mathcal{P}$ uma família exponencial natural dada por (2.8).*
(i) Seja $T = (Y, U)$ e $\eta = (\vartheta, \varphi)$, onde Y e ϑ tem a mesma dimensão. Então, Y tem a função densidade de probabilidade

$$f_\eta(y) = exp\{\vartheta^\tau y - \zeta(\eta)\}$$

com relação a medida σ-finita dependendo de φ. Em particular, Y tem uma função densidade

de probabilidade sobre uma família exponencial natural. Além disso, a distribuição condicional de Y dada por $U = u$ tem a função densidade de probabilidade (com relação à uma medida σ-finita dependendo em u)

$$f_{\vartheta,u}(y) = exp\{\vartheta^\tau y - \zeta_u(\vartheta)\},$$

a qual é uma família exponencial natural indexada por ϑ.
(ii) Se η_0 é um ponto interior do espaço de parâmetro natural, então a m.g.f. ψ_{η_0} de $P_{\eta_0} \circ T^{-1}$ é finita em uma vizinhança de 0 e é dada por

$$\psi_{\eta_0}(t) = exp\{\zeta(\eta_0 + t) - \zeta(\eta_0)\}.$$

Além disso, se f é uma função boreliana satisfazendo $\int |f| dP_{\eta_0} < \infty$, então a função

$$\int f(\omega) exp\{\eta^\tau T(\omega)\} h(w) d\nu(w)$$

é infinitamente diferenciável em uma vizinhança de η_0, e as derivadas podem ser calculadas por diferenciação sobre o sinal da integral.

Demonstração. Vamos começar provando a primeira parte de (i). De fato, fixe $\eta_0 = (\vartheta_0, \varphi_0)$ no espaço paramétrico. Defina uma medida auxiliar $d\mu(t) = \exp(\eta_0^\tau t - \zeta(eta_0)) d\lambda(t)$. Desta forma, temos que

$$dP_\eta(t) = \exp((\eta - \eta_0)^\tau t - \zeta(\eta) + \zeta(\eta_0)) d\mu(t),$$

ou seja,

$$dP_\eta(y, u) = \exp((\vartheta - \vartheta_0)^\tau y + (\varphi - \varphi_0)^\tau u - \zeta(\eta) + \zeta(\eta_0)) d\mu(y, u).$$

Estamos no cenário do Corolário 1.3. De fato, fazendo $b(y) = \exp\{(\vartheta - \vartheta_0)^\tau y - \zeta(\eta) + \zeta(\eta_0)\}$ e $a(u) = \exp\{(\varphi - \varphi_0)^\tau u\}$, temos que

$$dP_\eta^Y(y) = b(y) \int a(u) P^{U|Y}(du, y) d\mu^Y(y) = \exp\{(\vartheta - \vartheta_0)^\tau y - \zeta(\eta) + \zeta(\eta_0)\} \int a(u) P^{U|Y}(du, y) d\mu^Y(y).$$

Considere agora a medida σ-finita

$$d\lambda_\vartheta(y) = \exp\{-\vartheta_0^\tau y + \zeta(\eta_0)\} \int a(u) P^{U|Y}(du, y) d\mu^Y(y).$$

Temos que Y pertence à família exponencial natural com relação à medida λ_ϑ, pois,

$$dP_\eta^Y(y) = \exp\{\vartheta^\tau y - \zeta(\eta)\} d\lambda_\vartheta(y),$$

o que prova a primeira parte de (i). Para a segunda parte, vamos aplicar novamente o Corolário

1.3 para obter

$$P_\eta^{Y|U}(dy,u) = \frac{\exp\{(\vartheta-\vartheta_0)^\tau y - \zeta(\eta)+\zeta(\eta_0)\}\mu^{Y|U}(dy,u)}{\int \exp\{(\vartheta-\vartheta_0)^\tau y - \zeta(\eta)+\zeta(\eta_0)\}\mu^{Y|U}(dy,u)}.$$

Defina então a medida σ-finita

$$d\nu_u(y) = \exp\{-\vartheta_0^\tau y + \zeta(\eta_0)\}\mu^{Y|U}(dy,u).$$

Temos então que a distribuição condicional de Y dado $U=u$ pertence à família exponencial natural, pois,

$$dP_\eta^{Y|U}(dy,u) = \exp\{\vartheta^\tau y - \zeta_u(\eta)\}d\nu_u(y),$$

onde

$$\zeta_u(\eta) = \exp\{\zeta(\eta) - \log[\int \exp\{(\vartheta-\vartheta_0)^\tau y - \zeta(\eta)+\zeta(\eta_0)\}\mu^{Y|U}(dy,u)]\}.$$

Vamos agora provar a parte (ii). Comece relembrando que, por definição da família exponencial, para todo parâmetro η dentro do espaço paramétrico, temos pela definição de $\zeta(\cdot)$, que

$$\int \exp\{\eta^\tau T(\omega)\}d\lambda(\omega) < \infty.$$

Assim, dado η_0 pertencente ao interior do espaço paramétrico, existe uma bola em torno de η_0 inteiramente contida no espaço paramétrico. Defina a função $c(\eta) = \int \exp\{\eta^\tau T(\omega)\}d\lambda(\omega)$. Dado $i=1,\ldots,p$, e $h\in\mathbb{R}$ tal que $\eta_0 + he_i$ pertença à bola, para todo h suficientemente pequeno, onde $\{e_1,\ldots,e_p\}$ denota a base canônica no $\mathbb{R}^p$, temos que

$$c(\eta_0+he_i) = \int \exp\{\eta_0^\tau T(\omega)\}\exp\{hT_i(\omega)\}d\lambda(\omega) = \int \sum_{j=0}^{\infty}\frac{[hT_i(\omega)]^j}{j!}\exp\{\eta_0^\tau T(\omega)\}d\lambda(\omega).$$

Agora, observe que

$$|\sum_{j=0}^{\infty}\frac{[hT_i(\omega)]^j}{j!}| \leq \sum_{j=0}^{\infty}|\frac{[hT_i(\omega)]^j}{j!}| \leq \exp|hT_i| \leq \exp\{hT_i\}+\exp\{-hT_i\},$$

com $\int \exp\{hT_i\}\exp\{\eta_0^\tau T(\omega)\}d\lambda(\omega) < \infty$ e $\int \exp\{-hT_i\}\exp\{\eta_0^\tau T(\omega)\}d\lambda(\omega) < \infty$, pois tanto $he_i - \eta_0$ quanto $he_i+\eta_0$ pertencem ao espaço paramétrico. Logo, pelo teorema da convergência dominada, temos que

$$c(\eta_0+he_i) = \sum_{j=0}^{\infty}\int \frac{[hT_i(\omega)]^j}{j!}\exp\{\eta_0^\tau T(\omega)\}d\lambda(\omega),$$

ou seja, a função "parcial" pode ser expandida em série de Taylor, o que significa que $c(\eta_0+\cdot)$

possui derivadas parciais de todas as ordens. Isto por sua vez implica que $c(\eta_0 + \cdot)$ é de classe C^∞ em uma vizinhança de η_0.

Em particular, $c(\eta_0 + \cdot)$ é contínua, e como η_0 pertence ao interior do espaço paramétrico, podemos tomar uma bola compacta contida no espaço paramétrico. Logo, a função $c(\eta_0 + \cdot)$ assume máximo dentro dessa bola, o que significa que $c(\eta_0 + \cdot)$ é finita em uma vizinhança da origem, o que conclui a primeira afirmação do item (ii).

A fórmula para ψ_{η_0} é consequência imediata da definição da função $\zeta(\cdot)$.

Para provar a última afirmação, defina a função $d(\eta) = \int f(\omega) \exp\{\eta^\tau T(\omega)\} d\lambda(\omega)$ (note que a função h foi absorvida pela medida λ que estamos usando no lugar de ν). Repetindo os passos usados no início da demonstração, obtemos que

$$d(\eta_0 + he_i) = \sum_{j=0}^{\infty} \int \frac{[hT_i(\omega)]^j}{j!} f(\omega) \exp\{\eta_0^\tau T(\omega)\} d\lambda(\omega),$$

donde segue que $d(\cdot)$ é de classe C^∞, e além disso, como a expressão acima é a série de Taylor da função parcial de d, temos que a k-ésima derivada parcial é exatamente o termo $j = k$ da expansão, daí

$$\frac{\partial^k d(\eta_0)}{\partial x_i^k} = \int \frac{[hT_i(\omega)]^k}{k!} f(\omega) \exp\{\eta_0^\tau T(\omega)\} d\lambda(\omega),$$

e isso mostra que a derivada foi obtida derivando sob o sinal da integral. Como a função é de classe C^∞ e obtemos as fórmulas para as derivadas parciais, as fórmulas para as derivadas em geral seguem das derivadas parciais. □

Note que, usando o teorema 2.1(ii) no resultado do exemplo 2.3 obtemos que a função geradora de momento da distribuição binomial $B_i(p, n)$ é

$$\begin{aligned} \psi_\eta(t) &= exp\{n \log(1 + e^{(n+t)} - n \log(1 + e^\eta)\} \\ &= \left(\frac{1 + e^\eta e^t}{1 + e^\eta}\right)^n \\ &= (1 - p + pe^t)^n, \end{aligned}$$

sendo $p = e^\eta/(1 + e^\eta)$.

Definição 2.3. [Família Localização-escala]. Sejam P uma medida de probabilidade sobre $(\mathcal{R}^k, \mathcal{B}^k)$, $\mathcal{V} \subset \mathcal{R}^k$, e $\mathcal{M}_k$ uma coleção de matrizes simétricas positivas de ordem k. A família

$$\{P_{(\mu,\Sigma)} : \mu \in \mathcal{V}, \Sigma \in \mathcal{M}_k\} \tag{2.12}$$

é chamada *Família de localização-escala* (sobre $\mathcal{R}^k$), onde

$$P_{(\mu,\Sigma)}(B) = P\left(\Sigma^{-1/2}(B - \mu)\right), \quad B \in \mathcal{B}^k,$$

$\Sigma^{-1/2}(B-\mu) = \{\Sigma^{-1/2}(x-\mu) : x \in B\} \subset \mathcal{R}^k$, e $\Sigma^{-1/2}$ é a inversa da matriz "raiz quadrada"$\Sigma^{1/2}$ satisfazendo $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$. Os parametros μ e $\Sigma^{1/2}$ são chamados de *Parâmetro Localização* e *Parâmetro Escala*, respectivamente.

Exemplo 2.6. A família $\{P_{(\mu,I_k)} : \mu \in \mathcal{R}^k\}$ é chamada uma *família localização*, onde I_k é uma matriz identidade de ordem k. A família $\{P_{(0,\Sigma)} : \Sigma \in \mathcal{M}^k\}$ é chamada *família escala*. Em alguns casos, nós consideramos uma família de localização escala da forma $\{P_{(\mu,\sigma^2 I_k)} : \mu \in \mathcal{R}^k, \sigma > 0\}$. Se $X_1, ..., X_k$ são i.i.d. com uma distribuição comum na família de localização-escala $\{P_{(\mu,\sigma^2)} : \mu \in \mathcal{R}, \sigma > 0\}$, então a distribuição conjunta do vetor $(X_1, ..., X_k)$ está na família localização-escala $\{P_{(\mu,\sigma^2 I_k)} : \mu \in \mathcal{V}, \sigma > 0\}$ com $\mathcal{V} = \{(x, ..., x) \in \mathcal{R}^k : x \in \mathcal{R}\}$.

2.2 Estatística, Suficiência e Completude

2.2.1 Estatísticas, Suficiência e Suficiência Minimal

Definição 2.4. Uma função mensurável de X, $T(X)$, é chamada *Estatística* se $T(X)$ é um valor conhecido sempre que X é conhecido, isto é, a função T é uma função conhecida.

Claro que, X em si é uma Estatística, mais conhecida como Estatística trivial. Vejamos outros exemplos.

Exemplo 2.7. Sejam $X = (X_1, ..., X_n)$ com componentes aleatórias *i.i.d.* e $X_{(i)}$ o i-ésimo menor valor de $X_1, ..., X_n$. As estatísticas $X_{(1)}, ..., X_{(n)}$ são chamadas de *Estatísticas de Ordem*. Suponha que X_i tem uma função densidade acumulada F, tendo uma função densidade de probabilidade de Lebesgue f. Então, o p.d.f de Lebesgue de $X_{(1)}, ..., X_{(n)}$ é

$$g(x_1, x_2, ..., x_n) = \begin{cases} n! f(x_1) f(x_2) \cdots f(x_n), & x_1 < x_2 < \cdots < x_n \\ 0 & caso\ \ contrario \end{cases}$$

O p.d.f. de lebesgue de $X_{(i)}$ e $X_{(j)}$, $1 \leq i < j \leq n$, é

$$g_{i,j}(x, y) = \begin{cases} n![F(x_1)]^{i-1}[F(y) - F(x)]^{j-i-1}[1 - F(y)]^{n-j} f(x) f(y), & x < y \\ 0 & caso\ \ contrario \end{cases}$$

e o p.d.f. de Lebesgue de $X(i)$ é

$$g_i(x) = \frac{n!}{(i-1)!(n-i)!}[F(x)]^{i-1}[1 - F(x)]^{n-i} f(x).$$

Definição 2.5. [Suficiência]. Seja X uma amostra de uma população desconhecida $P \in \mathcal{P}$, onde $\mathcal{P}$ é uma família de populações. Uma Estatística $T(X)$ é dita ser *Suficiente* para $P \in \mathcal{P}$ (ou para

$\theta \in \Theta$ quando $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ é uma família paramétrica) se, e somente se, a distribuição condicional de X, dado T, é conhecida (não depende de P ou θ).

O conceito de suficiência depende da família $\mathcal{P}$. Se T é suficiente para $P \in \mathcal{P}$, então T é suficiente para $P \in \mathcal{P}_0 \subset \mathcal{P}$, mas não necessariamente suficiente para $P \in \mathcal{P}_1 \supset \mathcal{P}$.

Exemplo 2.8. Suponha que $X = (X_1, ..., X_n)$ e $X_1, ..., X_n$ são independentes e identicamente distribuídos a partir da distribuição binomial com o p.d.f. (com relação à medida de contagem)

$$f_\theta(z) = \theta^z(1-\theta)^{1-z} I_{\{0,1\}}(z) \;\; ; \;\; z \in \mathcal{R}, \;\; \theta \in (0,1).$$

Para qualquer realização x de X, x é uma sequência de n uns e zeros. Considere a estatística $T(X) = \Sigma_{i=1}^n X_i$, que é o número de uns em X. Antes de mostrar que T é suficiente, nós podemos argumentar intuitivamente que T contém todas as informações sobre Θ, uma vez que, θ é a probabilidade de uma ocorrencia de um em x. Dado $T = t$ (o número de uns em x), o que é deixado no conjunto de dados x são as informações redundantes sobre as posições dos t uns. Uma vez que as variáveis aleatórias são discretas, não é difícil calcular a distribuição condicional de X dado $T = t$. Note que,

$$P(X = x|T = t) = \frac{P(X = x, T = t)}{P(T = t)}$$

e $P(T = t) = \binom{n}{t}\theta^t(1-\theta)^{n-t} I\{0, 1, ..., n\}(t)$. Seja x_i a i-ésima componente de x. Se $t \neq \Sigma_{i=1}^n x_i$, então $P(X = x, T = t) = 0$. Se $t = \Sigma_{i=1}^n x_i$, então

$$P(X = x, T = t) = \prod_{i=1}^n P(X_i = x_i) = \theta^t(1-\theta)^{n-t} \prod_{i=1}^n I\{0,1\}(x_i).$$

Seja $B_t = \{(x_1, ..., x_n) : x_i = 0, 1, \Sigma_{i=1}^n x_i = t\}$. Então,

$$P(X = x|T = t) = \frac{1}{\binom{n}{t}} I_{B_t}(x).$$

é uma função densidade de probabilidade conhecida. Isso mostra que $T(X)$ é suficiente para $\theta \in (0,1)$, de acordo com a definição (2.6), com a família $\{f_\theta : \theta \in (0,1)\}$.

Encontrar uma estatística suficiente por meio da definição não é conveniente, uma vez que, envolve em advinhar uma estatística T que pode ser suficiente e calcular a distribuição condicional de X, dado $T = t$. Para famílias de populações tendo p.d.f's, uma maneira simples de encontrar estatísticas suficientes é usando o teorema da fatoração. Antes, veremos o seguinte lema.

Lema 2.1. *Se uma família $\mathcal{P}$ é dominada por uma medida σ-finita, então $\mathcal{P}$ é dominada por uma medida de probabilidade $Q = \Sigma_{i=1}^n c_i P_i$, onde $c_i's$ são constantes positivas com $\Sigma_{i=1}^\infty c_i = 1$ e $P_i \in \mathcal{P}$.*

Demonstração. Vamos começar provando o caso das medidas serem dominadas por uma medida finita. Suponha que $\mathcal{P}$ é dominada por uma medida finita ν. Seja $\mathcal{P}_0$ a família de todas as medidas da forma $\sum_{i=1}^{\infty} c_i P_i$, onde $P_i \in \mathcal{P}$ e $c_i \geq 0$, e $\sum_{i=1}^{\infty} c_i = 1$. Então é suficiente mostrar que existe um $Q \in \mathcal{P}$, tal que $Q(A) = 0$ implique $P(A) = 0$ para todo $P \in \mathcal{P}_0$ (na realidade a tal Q terá todas as constantes positivas). Seja $\mathcal{C}$ a classe de eventos C para o qual existe uma medida $P \in \mathcal{P}_0$ de modo que $P(C) > 0$ e $dP/d\nu > 0$ ν-q.t.p. em C. Então, existe uma sequência $\{C_i\} \subset \mathcal{C}$ tal que $\nu(C_i) \to \sup_{C \in \mathcal{C}} \nu(C)$.

Seja C_0 a união de todos C_i's e $Q = \sum_{i=1}^{\infty} c_i P_i$, onde P_i é a medida de probabilidade correspondente a C_i e c_i é uma sequência qualquer fixada tal que $c_i > 0$ e $\sum_{i=1}^{\infty} c_i = 1$.

Note que $C \in \mathcal{C}$. De fato, $Q(C_0) \geq c_i P_i(C_0) \geq c_i P_i(C_i) > 0$ (pois C_0 é a união de todos os C_i's). Além disso, $dQ/d\nu \geq c_i dP_i/d\nu > 0$.

Suponha agora que $Q(A) = 0$. Seja $P \in \mathcal{P}_0$ e $B = \{x : dP/d\nu > 0\}$. Uma vez que $Q(A \cap C_0) = 0$, temos que $\nu(A \cap C_0) = 0$, pois como $dQ/d\nu > 0$, se $\nu(A \cap C_0) > 0$, teríamos $Q(A \cap C_0) > 0$. Finalmente, $P(A \cap C_0) = 0$, pois P é dominada por ν. Assim, $P(A) = P(A \cap C_0^c \cap B)$. Se $P(A) = P(A \cap C_0^c \cap B) > 0$, então $\nu(C_0 \cup (A \cap C_0^c \cap B)) > \nu(C_0)$, o que contradiria $\nu(C_0) = \sup_{C \in \mathcal{C}} \nu(C)$ uma vez que a hipótese $P(A \cap C_0^c \cap B) > 0$ e a definição de B implicariam que $A \cap C_0^c \cap B$ está em $\mathcal{C}$, e portanto $C_0 \cup (A \cap C_0^c \cap B)$ estaria em $\mathcal{C}$. Assim, $P(A) = 0$ para todo $P \in \mathcal{P}_0$. O que conclui a demonstração para o caso de medida finita.

Para a medida σ-finita, escreva $\Omega = \cup_{j=1}^{\infty} \Omega_j$ com Ω_j disjuntos e $\nu(\Omega_j) < \infty$. Consideremos agora a medida restrita ν^{Ω_j} e a família de medidas restritas $\mathcal{P}^{\Omega_j} = \{P^{\Omega_j}\}$, dadas por $\nu^{\Omega_j}(A) = \nu(A \cap \Omega_j)$ e analogamente para as medidas em $\mathcal{P}$.

Aplicando o resultado anterior (e fixando as mesmas constantes c_i's para todos os j's) para às medidas P^{Ω_j} e ν^{Ω_j}, que são finitas, obtemos uma medida dominadora $Q^{\Omega_j} = \sum_{i=1}^{\infty} c_i P_i^{\Omega_j}$. Observe que $\nu = \sum_{j=1}^{\infty} \nu^{\Omega_j}$, e $P = \sum_{j=1}^{\infty} P^{\Omega_j}$. Assim, definindo $Q = \sum_{j=1}^{\infty} Q^{\Omega_j} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} c_i P_i^{\Omega_j} = \sum_{i=1}^{\infty} c_i P_i$, temos que Q está bem definida e é de probabilidade, e além disso, se $Q(A) = 0$, então para todo j, $Q^{\Omega_j}(A) = 0$, o que por sua vez implica que $P^{\Omega_j}(A) = 0$, e como isto vale para todo j, segue que $P(A) = 0$. Logo, Q domina a família $\mathcal{P}$. O que conclui a demonstração para o caso de ν σ-finita. □

Teorema 2.2. *[Teorema da Fatoração]. Suponha que X é uma amostra para $P \in \mathcal{P}$ e $\mathcal{P}$ é uma família de medidas de probabilidade em $(\mathcal{R}^n, \mathcal{B}^n)$ dominada por uma medida σ-finita ν. Então, $T(X)$ é suficiente para $P \in \mathcal{P}$ se, e somente se, existem funcões borel não-negativas h (que não depende de P) em $(\mathcal{R}^n, \mathcal{B}^n)$, e g_p (que depende de P) sobre a imagem de T, tal que*

$$\frac{dP}{d\nu}(x) = g_P(T(x))h(x) \tag{2.13}$$

Demonstração. ($\Rightarrow$)

Suponha que T é suficiente para $P \in \mathcal{P}$. Seja Q a medida de probabilidade do lema 2.1.

Então, como T é suficiente, para cada $A \in \mathcal{B}^n$, $P(A|T)$ não depende de P, e em particular $P_j(A \mid T) = P(A \mid T)$. Segue que para cada $B \in \sigma(T)$ temos

$$\begin{aligned} Q(A \cap B) &= \sum_{j=1}^{\infty} c_j P_j(A \cap B) \\ &= \sum_{j=1}^{\infty} c_j E_{P_j}[1_A 1_B] \\ &= \sum_{j=1}^{\infty} c_j E_{P_j}[1_B E_{P_j}[1_A \mid T]] \\ &= \sum_{j=1}^{\infty} c_j \int_B P(A|T) dP_j \\ &= \int_B \sum_{j=1}^{\infty} c_j P(A|T) dP_j \\ &= \int_B P(A|T) dQ, \end{aligned}$$

onde utilizamos o teorema da convergência monótona para passar o somatória sob o sinal da integral. Assim, $P(A|T) = E_Q(I_A|T)$ Q-quase-certamente, onde $P(A|T) = E_Q(1_A|T)$ denota a esperança condicional de 1_A dado T com relação a Q. Considere agora a restrição das medidas P e Q à σ-álgebra $\sigma(T)$. P continua sendo dominada por Q, e assim, seja $g_P(T)$ a derivada Randon-Nikodym dP/dQ sobre o espaço $(\mathcal{R}^n, \sigma(T), Q)$. Note que,

$$\begin{aligned} P(A) &= E_P[P(A \mid T)] \\ &= \int P(A|T) dP \\ &= \int E_Q(1_A|T) g_P(T) dQ \\ &= \int E_Q[1_A g_P(T)|T] dQ \\ &= E_Q[E_Q[1_A g_P(T) \mid T]] \\ &= E_Q[1_A g_P(T)] \\ &= \int_A g_P(T) \frac{dQ}{d\nu} d\nu \end{aligned}$$

para cada $A \in \mathcal{B}^n$. Assim, a equação 2.13 é verdadeira com $h = dQ/d\nu$.

($\Leftarrow$)

Suponha que, para todo $P \in \mathcal{P}$ temos

$$\frac{dP}{d\nu}(x) = g_P(T(x))h(x).$$

Como

$$\frac{dP}{dQ} = \frac{dP/d\nu}{dQ/d\nu},$$

segue que

$$\frac{dP}{dQ} = \frac{g_P(T)}{\sum_{i=1}^{\infty} c_i g_{P_i}(T)}.$$

Isto mostra que dP/dQ é mensurável com relação à $\sigma(T)$.

Seja $A \in \sigma(X)$ e $P \in \mathcal{P}$. Queremos mostrar que

$$P(A|T) = E_Q(1_A|T) \quad P - \text{q.c..} \tag{2.14}$$

Daí, segue a suficiência de T. De fato, como $E_Q(1_A|T)$ não varia com $P \in \mathcal{P}$, e a equação 2.14 implica que a distribuição condicional de X dado T é determinada por $E_Q(1_A|T)$, $A \in \sigma(X)$. Uma forma de ver isso é utilizar a probabilidade condicional regular.

Vamos agora provar 2.14. Tome $A \in \sigma(X)$ e $B \in \sigma(T)$. Então, relembrando que dP/dQ é $\sigma(T)$-mensurável, temos que

$$\begin{aligned}
\int_B 1_A dP &= \int_B 1_A \frac{dP}{dQ} dQ \\
&= \int_B E_Q[1_A \frac{dP}{dQ} \mid T] dQ \\
&= \int_B E_Q[1_A \mid T] \frac{dP}{dQ} dQ \\
&= \int_B E_Q[1_A \mid T] dP.
\end{aligned}$$

Isto mostra que $P(A \mid T) = E_Q[1_A \mid T]$ P-q.c.. □

Se P é uma família exponencial com p.d.f's dado por (2.4) e $X(w) = w$, então podemos aplicar o teorema 2.2 com $g_\theta(t) = exp\{[\mu(\theta)]^\tau t - \xi(\theta)\}$ e concluir que T é uma estatística suficiente para $\theta \in \Theta$. No exemplo 2.8, a distribuição conjunta de X é uma família exponencial com $T(X) = \Sigma_{i=1}^n X_i$. Assim, podemos concluir que T é suficiente para $\theta \in (0, 1)$, sem precisar calcular a distribuição condicional de X, dado T.

Exemplo 2.9. [Famílias de Truncamento]. Seja $\phi(x)$ uma função Borel positiva sobre $(\mathcal{R}, \mathcal{B})$ tal que $\int_a^b \phi(x)dx < \infty$ para algum a e b, $-\infty < a < b < \infty$. Seja $\theta = (a, b)$, $\Theta = \{(a, b) \in \mathcal{R}^2 : a < b\}$, e

$$f_\theta(x) = c(\theta)\phi(x)I(a, b)(x),$$

onde $c(\theta) = [\int_a^b \phi(x)dx]^{-1}$. Então $\{f_\theta : \theta \in \Theta\}$, chamado de Família de Truncamento, é uma família paramétrica dominada pela medida de Lebesgue sobre $\mathcal{R}$. Seja $X_1, ..., X_n$ variáveis aleatórias i.i.d. tendo o f.d.p. f_θ. Então a função densidade de probabilidade conjunta de

$X = (X_1, ..., X_n)$ é

$$\prod_{i=1}^{n} f_\theta(x_i) = [c(\theta)]^n I_{(a,\infty)}(x_{(1)}) I_{(-\infty,b)}(x_{(n)}) \prod_{i=1}^{n} \phi(x_i), \tag{2.15}$$

onde $x_{(i)}$ é o i-ésimo menor valor de $x_1, ...x_n$. Sejam

$$T(X) = (X_{(1)}, X_{(n)}), \quad g_\theta(t_1, t_2) = [c(\theta)]^n I_{(a,\infty)}(t_1) I_{(-\infty,b)}(t_2) \quad e \quad h(x) = \Pi_{i=1}^{n} \phi(x_i).$$

Por 2.15 e o pelo teorema 2.2, T(X) é suficiente para $\theta \in \Theta$.

Antes de introduzir o próximo conceito, nós precisaremos da seguinte notação. Se uma informação vale exceto para resultados em um evento A satisfazendo $P(A) = 0$ para todo $P \in \mathcal{P}$, então dizemos que o enuciado vale quase certamente $\mathcal{P}$, isto é, $\mathcal{P}$- q.c..

Definição 2.6. [Suficiência Mínima] Seja T uma estatística suficiente para $P \in \mathcal{P}$. Dizemos que T é uma *Estatística Suficiente Mínima* se, somente se, dada alguma outra Estatística Suficiente S para $P \in \mathcal{P}$, existir uma função mensurável ψ tal que $T = \psi(S)$, $\mathcal{P}$-q.c..

Se ambos T e S são estatísticas suficientes mínimas, então por definição existe uma função mensurável injetiva ψ talque $T = \psi(S)$, $\mathcal{P}$-q.c.. Assim, a estatística suficiente mínima é única no sentido em que duas estatísticas que são funções mensuráveis injetivas uma da outra, podem ser tratadas como uma estatística.

Exemplo 2.10. Sejam $X_1, ..., X_n$ variáveis aleatórias i.i.d. de P_θ, a distribuição uniforme $U(\theta, \theta + 1)$, $\theta \in \mathcal{R}$. Suponha que $n > 1$. A f.d.p. conjunta de Lebesgue de $(X_1, ..., X_n)$ é

$$f_\theta(x) = \prod_{i=1}^{n} I_{(\theta,\theta+1)}(x_i) = I_{(x_{(n)}-1, x_{(n)})}(\theta) \ ; \ x = (x_1, ..., x_n) \in \mathcal{R}^n.$$

onde $x_{(i)}$ denota o i-ésimo menor valor de $x_1, ..., x_n$. Pelo teorema 2.2, $T = (X_{(1)}, X_{(n)})$ é suficiente para θ. Note que,

$$x_{(1)} = sup\{\theta : f_\theta(x) > 0\} \quad e \quad x_{(n)} = 1 + inf\{\theta : f_\theta(x) > 0\}.$$

Se $S(X)$ é uma estatística suficiente para θ, então pelo teorema 2.2, existem funções borel h e g_θ tal que $f_\theta(x) = g_\theta(S(x))h(x)$. Para x com $h(x) > 0$,

$$x_{(1)} = sup\{\theta : g_\theta(S(x)) > 0\} \quad e \quad x_{(n)} = 1 + inf\{\theta : g_\theta(S(x)) > 0\}.$$

Assim, existe uma função mensurável ψ tal que $T(x) = \psi(S(x))$ quando $h(x) > 0$. Uma vez que $h > 0$, $\mathcal{P}$-q.c., nós concluímos que T é suficiente mínima.

Teorema 2.3. *Seja $\mathcal{P}$ uma família de distribuição em $\mathcal{R}^k$.*
(i) Suponha que $\mathcal{P}_0 \subset \mathcal{P}$ e $\mathcal{P}_0$-q.c. implica $\mathcal{P}$-q.c.. Se T é suficiente para $P \in \mathcal{P}$ e suficiente mínima para $P \in \mathcal{P}_0$, então T é suficiente mínima para $P \in \mathcal{P}$.
(ii) Suponha que P contém p.d.f.'s $f_0, f_1, f_2, ...$ com relação a uma medida σ-finita. Seja $f_\infty = \Sigma_{i=0}^{\infty} c_i f_i(x)$, onde $c_i > 0$ $\forall i$ e $\Sigma_{i=0}^{\infty} c_i = 1$ e seja $T_i(X) = f_i(x)/f_\infty(x)$ quando $f_\infty(x) > 0$, $i = 0, 1, 2, ...$ Então, $T(X) = (T_0, T_1, T_2, ...)$ é suficiente mínima para $P \in \mathcal{P}$.
(iii) Suponha que P contém p.d.f's f_P com relação a uma medida σ-finita e que exista uma estatística suficiente $T(X)$ tal que, para alguns possíveis valores x e y de X, $f_P(x) = f_P(y)\phi(x, y)$ para todo P implique $T(x) = T(y)$, onde ϕ é uma função mensurável. Então $T(X)$ é suficiente mínima para $P \in \mathcal{P}$.

Demonstração. (i) Se S é suficiente para $P \in \mathcal{P}$, então é também suficiente para $P \in \mathcal{P}_\prime$ e, portanto, $T = \psi(S)$, $\mathcal{P}$-q.c. vale para uma função mensurável ψ. O resultado segue da hipótese de que $\mathcal{P}_0$-q.c. implica $\mathcal{P}$-q.c..
(ii) Note que $f_\infty > 0$, $\mathcal{P}$-q.c.. Seja $g_i(T) = T_i$, $i = 0, 1, 2, ...$ Então $f_i(x) = g_i(T(x))f_\infty(x)$, $\mathcal{P}$-q.c.. Pelo teorema 2.2, T é suficiente para $P \in \mathcal{P}$. Suponha que $S(X)$ seja outra estatística suficiente. Pelo teorema 2.2 existem funções borelianas h e $\tilde{g}_i$ tal que $f_i(x) = \tilde{g}_i(S(x))h(x)$, $i = 0, 1, 2, ...$ Então, $T_i(x) = \tilde{g}_i(S(x))/\sum_{j=0}^{\infty} c_j \tilde{g}_j(S(x))$ para $x's$ satisfazendo $f_\infty(x) > 0$. Pela definição 2.6, T é suficiente minimal para $P \in \mathcal{P}$. A prova para o caso onde f_∞ é substituído por f_o é idêntica.
(iii) Ver referência [4] □

Exemplo 2.11. Seja $P = \{f_\theta : \theta \in \Theta\}$ uma família exponencial com p.d.f.'s f_θ dado por (2.6) e $X(w) = w$. Suponha que exista $\Theta_0 = \{\theta_0, \theta_1, ..., \theta_p\} \subset \Theta$ tal que os vetores $\eta_i = \eta(\theta_i) - \eta(\theta_0)$, $i = 1, ..., p$ são linearmente independentes em $\mathcal{R}^p$. (Isso é verdade se a família é de posto completo). Nós temos que mostrar que $T(X)$ é suficiente para $\theta \in \Theta$. De fato, seja $\mathcal{P}_0 = \{f_\theta : \theta \in \Theta_0\}$. Note que o conjunto $\{x : f_\theta(x) > 0\}$ não depende de θ. Segue do teorema 2.3(ii), com $f_\infty = f_{\theta_0}$, que

$$S(X) = (exp\{\eta_1^\tau T(x) - \xi_1\}, ..., exp\{\eta_p^\tau T(x) - \xi p\})$$

é suficiente mínima para $\theta \in \Theta_0$, onde $\xi_i = \xi(\theta_i) - \xi(\theta_0)$. Uma vez que η_i's são linearmente independente, existe uma função mensurável injetiva ψ tal que $T(X) = \psi(S(X))$, $\mathcal{P}_0$-q.c.. Por isso, T é suficiente mínima para $\theta \in \Theta_0$. Observe que $\mathcal{P}_0$-q.c. implica $\mathcal{P}$-q.c.. Assim, pelo teorema 2.3(i), T é suficiente mínima de $\theta \in \Theta$.

2.2.2 Estatística completa

Definição 2.7. Uma estatística $V(X)$ é dita ser *ancillary* se a distribuição não depende da população P e *ancillary de primeira ordem* se $E[V(X)]$ é independente de P.

Definição 2.8. [Completude]. Uma estatística $T(X)$ é dita ser *completa* por $P \in \mathcal{P}$ se, somente se,

para qualquer f borel, $E[f(T)] = 0 \; \forall \; P \in \mathcal{P}$ implica $F(T) = 0$, $\mathcal{P}$-q.c.. T é dita ser *limitadamente completa* se, somente se, a definição anterior é valida para qualquer f borel limitada.

Uma estatística completa é limitadamente completa. Se T é completa (ou limitadamente completa) e $S = \psi(T)$ para uma mensurável ψ, então S é completa (ou limitadamente completa).

Proposição 2.1. *Se P é de uma família exponencial de posto completo com p.d.f.'s dado por (2.8), então $T(X)$ é completa e suficiente para $\eta \in \Xi$.*

Demonstração. Ver referência [4] □

Exemplo 2.12. Suponha que $X_1, ..., X_n$ são variáveis aleatórias distribuidas identicamente e independentes tendo a distribuição $N(\mu, \sigma^2)$, $\mu \in \mathcal{R}$, $\sigma > 0$. No exemplo 2.6, a função densidade de probabilidade conjunta de $X_1, ..., X_n$ é $(2\pi)^{-n/2} exp\{\eta_1 T_1 + \eta_2 T_2 - \mu\zeta(n)\}$, onde $T_1 = \Sigma_{i=1}^n X_i$, $T_2 = -\Sigma_{i=1}^n X_i^2$, e $\eta = (\eta_1, \eta_2) = (\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2})$. Assim, a família de distribuição para $X = (X_1, ..., X_n)$ é uma família exponencial natural de posto completo ($\Xi = \mathcal{R} \times (0, \infty)$). Pela proposição 2.1, $T(X) = (T_1, T_2)$ é completo e suficiente para η. Uma vez que existe uma correspondência injetiva entre η e $\theta = (\mu, \sigma^2)$, T é também completa e suficiente para θ.

Teorema 2.4 (Teorema de Basu)**.** *Seja V e T duas estatísticas de X a partir de uma população $P \in \mathcal{P}$. Se V é auxiliar e T é limitadamente completo e suficiente para $P \in \mathcal{P}$, então V e T são independente com relação a qualquer $P \in \mathcal{P}$.*

Demonstração. Seja B um evento sobre a imagem de V. Uma vez que V é auxiliar, $P(V^{-1}(B))$ é uma constante. Sendo T suficiente, $E[I_B(V)|T]$ é uma função de T (independente de P). Visto que $E\{E[I_B(V)|T] - P(V^{-1}(B))\} = 0$ para todo $P \in \mathcal{P}$, $P(V^{-1}(B)|T) = E[I_B(V)|T] = P(V^{-1}(B))$, $\mathcal{P}$-q.c., pela completude limitada de T. Seja A um evento sobre a imagem de T. Então,

$$\begin{aligned} P(T^{-1}(A) \cap V^{-1}(B)) &= E\{E[I_A(T) I_B(V)|T]\} \\ &= E\{I_A(T) E[I_B(V)|T]\} \\ &= E\{I_A(T) P(V^{-1}(B))\} \\ &= P(T^{-1}(A)) P(V^{-1}(B)). \end{aligned}$$

Assim, T e V são independentes com relação a qualquer $P \in \mathcal{P}$. □

2.3 Inferencia Estatística

Vimos no início do capítulo que inferência estatística é o nome que se dá a um conjunto de técnicas e procedimentos que permitem dar ao pesquisador um grau de confiabilidade nas afirmações que faz para a população. Existem três tipos principais de procedimentos de inferência: *Estimadores pontuais, Testes de Hipóteses e conjuntos de confiança.*

Definição 2.9. Seja $\vartheta \in \tilde{\Theta} \subset \mathcal{R}$ um parâmetro a ser estimado. Um *estimador* é uma estatística cuja imagem é $\tilde{\Theta}$.

Definição 2.10. O *Viés* de um estimador $T(X)$ de um parâmetro de valor real ϑ de uma população desconhecida é definido como $b_T[P] = E[T(X)] - \vartheta$ (indicado por $b_T(\theta)$ quando P pertence a uma família paramétrica indexada por θ). Um estimador $T(X)$ é dito ser *não-viesado* para θ se, somente se, $b_T(P) = 0$ para qualquer $P \in \mathcal{P}$.

Definição 2.11. O erro quadrático médio (mse) de $T(X)$ como um estimador de θ é definido como

$$mse_T(P) = E[T(X) - \vartheta]^2 = [b_T(P)]^2 + Var(T(X)), \tag{2.16}$$

que é denotado por $mse_T(\theta)$ se P pertence a uma família paramétrica. Observe que $mse_T(P) = Var(T(X)) \Longleftrightarrow T(X)$ é não-viesado.

Iremos apresentar agora um procedimento estatístico baseado na análise de uma amostra, através da teoria de probabilidades, usado para avaliar determinados parâmetros que são desconhecidos numa população. Este método é conhecido como *Teste de Hipóteses* onde os elementos básicos serão apresentados no exemplo a seguir.

Exemplo 2.13. *Sejam $\mathcal{P}$ uma família de distribuição, $\mathcal{P}_0 \subset \mathcal{P}$ e $\mathcal{P}_1 = \{P \in \mathcal{P} : P \notin \mathcal{P}_0\}$. Um problema de teste de hipóteses tem como objetivo decidir qual das duas afirmações seguintes é verdadeira:*

$$H_0 : P \in \mathcal{P}_0 \quad \text{contra} \quad \mathrm{H}_1 : \mathrm{P} \in \mathcal{P}_1 \tag{2.17}$$

Aqui, chamamos H_0 de hipótese nula e H_1 de hipótese alternativa. Chamamos de teste de uma hipótese estatística a função decisão $T(X) : \chi \to \{0, 1\}$, em que 0 corresponde à ação de considerar a hipótese H_0 como verdadeira e 1 corresponde à ação de considerar a hipótese H_1 como verdadeira. O conjunto χ denota a imagem da amostra X de uma população $P \in \mathcal{P}$. A função de decisão $T(X)$ divide o espaço amostral χ em dois conjuntos

$$A_0 = \{(x_1, ..., x_n) \in \chi; T(x_1, ..., x_n) = 0\}$$

e

$$A_1 = \{(x_1, ..., x_n) \in \chi; T(x_1, ..., x_n) = 1\},$$

onde $A_0 \cup A_1 = \chi$ e $A_0 \cap A_1 = \emptyset$. Chamamos A_0 de região de aceitação de H_0 e A_1 de região de rejeição de H_0, também chamada de região crítica.

Para testar a hipótese H_0 contra a hipótese H_1 dada em (2.17), existem apenas dois tipos de erros estatísticos que podemos cometer: Rejeitar H_0 quando H_0 é verdadeira (chamado erro do tipo I); e aceitar H_0 sendo H_0 falsa.

Definição 2.12. *O poder do teste com região crítica A_1 para testar $H_0 : P \in \mathcal{P}_0$ contra $H_1 : P \in \mathcal{P}_1$ é dado por:*

$$\alpha_T(P) = P(T(X) = 1) \qquad P \in \mathcal{P}_0 \tag{2.18}$$

Note que $1 - \alpha_t(P) = P(T(X) = 0), P \in \mathcal{P}_1$.

Teorema 2.5 (Cramér-Rao)**.** *Seja $X = (X_1, ..., X_n)$ uma amostra de $P \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$, onde Θ é um conjunto aberto em $\mathcal{R}^k$. Suponha que $T(X)$ seja um estimador com $E(T(X)) = g(\theta)$ sendo uma função diferenciável de θ; P_θ tem uma função densidade f_θ com relação à uma medida ν para todo $\theta \in \Theta$; e f_θ é diferenciável como uma função de θ e além disso satisfaz*

$$\frac{\partial}{\partial\theta}\int h(x)f_\theta(x)d\nu = \int h(x)\frac{\partial}{\partial\theta}f_\theta(x)d\nu, \qquad \theta \in \Theta, \tag{2.19}$$

para $h(x) \equiv 1$ e $h(x) = T(x)$. Então,

$$Var(T(X)) \geq \left[\frac{\partial}{\partial\theta}g(\theta)\right]^\tau [I(\theta)]^{-1}\frac{\partial}{\partial\theta}g(\theta), \tag{2.20}$$

onde

$$I(\theta) = E\left\{\frac{\partial}{\partial\theta}\log f_\theta(x)\left[\frac{\partial}{\partial\theta}\log f_\theta(x)\right]^\tau\right\} \tag{2.21}$$

é assumida ser positiva definida para qualquer $\theta \in \Theta$.

Demonstração. A prova desse teorema para o caso de uma variável ($k = 1$) é simplesmente elegante e uma inteligente aplicação da inequação de Cauchy-Schwarz que diz que para quaisquer duas variáveis aleatórias X e Y,

$$[Cov(X, Y)]^2 \leq (VarX)(VarY). \tag{2.22}$$

Assim, quando $k = 1$, a equação (2.20) se reduz a

$$Var(T(X)) \geq \frac{[g'(\theta)]^2}{E[\frac{\partial}{\partial\theta}\log f_\theta(X)]^2}$$

e portanto basta mostrar que

$$E\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^2 = Var\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)$$

e

$$g'(\theta) = Cov\left(T(X), \frac{\partial}{\partial\theta}\log f_\theta(X)\right).$$

Primeiro note que,

$$\begin{aligned}\frac{\partial}{\partial\theta}E[h(x)] &= \frac{\partial}{\partial\theta}\int T(X)f_\theta(x)d\nu \\ &= \int h(x)\left[\frac{\partial}{\partial\theta}f_\theta(x)\right]d\nu \\ &= E\left[h(x)\frac{\frac{\partial}{\partial\theta}f_\theta(x)}{f_\theta(x)}\right] \\ &= E\left[h(x)\frac{\partial}{\partial\theta}\log f_\theta(x)\right], \end{aligned} \tag{2.23}$$

Assim, da condição (2.23) com $h(x) = 1$, temos que

$$E\left[\frac{\partial}{\partial\theta}\log f_\theta(x)\right] = \frac{\partial}{\partial\theta}E[1] = 0.$$

Da condição (2.23) com $h(x) = T(X)$, segue que

$$E\left[T(X)\frac{\partial}{\partial\theta}\log f_\theta(x)\right] = \frac{\partial}{\partial\theta}E[T(X)] = g'(\theta).$$

Uma vez que,

$$Cov\left(T(X), \frac{\partial}{\partial\theta}\log f_\theta(x)\right) = E\left(T(X)\frac{\partial}{\partial\theta}\log f_\theta(x)\right) - E(T(X))E\left(\frac{\partial}{\partial\theta}\log f_\theta(x)\right) = g'(\theta),$$

e

$$Var\left[\frac{\partial}{\partial\theta}\log f_\theta(x)\right] = E\left[\frac{\partial}{\partial\theta}\log f_\theta(x) - E\left[\frac{\partial}{\partial\theta}\log f_\theta(x)\right]\right]^2 = E\left[\frac{\partial}{\partial\theta}\log f_\theta(x)\right]^2.$$

o teorema fica provado para o caso de uma variável (k=1).

□

Definição 2.13. *Sejam $X = (X_1, ..., X_n)$ uma amostra de $P \in \mathcal{P}$ e $T_n(X)$ um estimador pontual de ϑ para cada n.*

(i) $T_n(X)$ é chamado consistente para ϑ se, somente se, $T_n(X) \to_p \vartheta$ com relação a qualquer $P \in \mathcal{P}$.

(ii) Seja $\{a_n\}$ uma sequência de constantes positivas divergindo para ∞. $T_n(X)$ é chamada a_n-consistente para ϑ se, somente se, $a_n[T_n(X) - \vartheta] = O_p(1)$ com relação a qualquer $P \in \mathcal{P}$.

(iii) $T_n(X)$ é chamado fortemente consistente para ϑ se, somente se, $T_n(X) \to_{q.c.} \vartheta$ com relação a qualquer $P \in \mathcal{P}$.

(iv) $T_n(X)$ é chamado L_r-consistente para ϑ se, somente se, $T_n(X) \to_{L_r} \vartheta$ com relação a

qualquer $P \in \mathcal{P}$, para algum fixo $r > 0$.

Consistência é na realidade um conceito relativo a uma sequência de estimadores, $\{T_n, n = n_0, n_0 + 1, ...\}$, mas diremos simplesmente "Consistência de T_n", apenas para simplificar. L_2-consistência é também chamado de *consistência em mse.*

Definição 2.14. (i) Sejam $\xi, \xi_1, \xi_2, ...$ variáveis aleatórias e $\{a_n\}$ uma sequência de números positivos onde ou $a_n \to \infty$ ou $a_n \to a > 0$. Se $a_n\xi_n \to_d \xi$ e $E|\xi| < \infty$, então $E\xi/a_n$ é chamado uma *esperança assintótica* de ξ_n.
(ii) Seja T_n um estimador pontual de ϑ para cada n. Uma esperança assintótica de $T_n - \vartheta$, se existir, é chamado um *viés assintótico* de T_n e denotado por $\tilde{b}_{T_n}(P)$ (ou $\tilde{b}_{T_n}(\theta)$ se P pertence a uma família paramétrica). Se $\lim_{n\to\infty} \tilde{b}_{T_n}(P) = 0$ para cada $P \in \mathcal{P}$, então T_n é dito ser *assintóticamente não-viesado.*

O seguinte resultado mostra que a esperança assintótica é essencialmente única.

Proposição 2.2. *Seja $\{\xi_n\}$ uma sequência de variáveis aleatórias. Suponha que tanto $E\xi/a_n$ quanto $E\eta/b_n$ sejam esperanças assintóticas de ξ_n definida em (2.10)(i). Então, um dos três casos deve ocorrer:*

(a) $E\xi = E\eta = 0$;

(b) $E\xi \neq 0$, $E\eta = 0$ e $b_n/a_n \to 0$; ou $E\xi = 0$, $E\eta \neq 0$ e $a_n/b_n \to 0$;

(c) $E\xi \neq 0$, $E\eta \neq 0$ e $(E\xi/a_n)/(E\eta/b_n) \to 1$.

Demonstração. Ver referência [4] □

Definição 2.15. Sejam T_n um estimador de ϑ para cada n e $\{a_n\}$ uma sequência de números positivos onde ou $a_n \to \infty$ ou $a_n \to a > 0$. Assuma que $a_n(T_n - \vartheta) \to_d Y$ com $0 < EY^2 < \infty$.
(i) O erro quadrático médio assintótico de T_n, denotado por $amse_{T_n}(P)$ ou $amse_{T_n}(\theta)$ se P é de uma família paramétrica indexada por θ, é definido como a esperança assintótica de $(T_n - \vartheta)^2$, isto é, $amse_{T_n}(P) = EY^2/a_n^2$. A variância assintótica de T_n é definida como $\sigma^2_{T_n}(P) = Var(Y)/a_n^2$.
(ii) Seja T'_n outro estimador de ϑ. A *eficiência relatica assintótica* de T'_n com relação a T_n é definido como $e_{T'_n,T_n}(P) = amse_{T_n}(P)/amse_{T'_n}(P)$.
(iii) T_n é dito ser *assintoticamente mais eficiente* que T'_n se, e somente se, $\limsup_n e_{T'_n,T_n}(P) \leq 1$ para qualquer P e < 1 para algum P.

Proposição 2.3. *Sejam T_n um estimador de ϑ para cada n e $\{a_n\}$ uma sequência de números positivos onde ou $a_n \to \infty$ ou $a_n \to a > 0$. Suponha que $a_n(T_n - \vartheta) \to_d Y$ com $0 < EY^2 < \infty$. Então,*
(i) $EY^2 \leq \liminf_n E[a_n^2(T_n - \vartheta)^2]$ e
(ii) $EY^2 = \lim_{n\to\infty} E[a_n^2(T_n - \vartheta)^2]$ se, somente se, $\{a_n^2(T_n - \vartheta)^2\}$ é uniformente integrável.

Demonstração. (i) Sabemos que,

$$\min\{a_n^2(T_n-\vartheta)^2,t\} \to_d \min\{Y^2,t\}$$

para qualquer $t>0$. Uma vez que, $\min\{a_n^2(T_n-\vartheta)^2,t\}$ é limitado por t, segue que

$$\lim_{n\to\infty} E(\min\{a_n^2(T_n-\vartheta)^2,t\}) = E(min\{Y^2,t\}).$$

Então,

$$\begin{aligned} EY^2 &= \lim_{n\to\infty} E(min\{Y^2,t\}) \\ &= \lim_{t\to\infty}\lim_{n\to\infty} E(\min\{a_n^2(T_n-\vartheta)^2,t\}) \\ &= \liminf_{t,n} E\min\{a_n^2(T_n-\vartheta)^2,t\} \\ &\leq \liminf_{n} E[a_n^2(T_n-\vartheta)]. \end{aligned} \tag{2.24}$$

onde a terceira equação segue do fato de que $E\min\{a_n^2(T_n-\vartheta)^2,t\}$ é não-decrescente em t para qualquer n fixado. □

Teorema 2.6. *Seja g uma função em $\mathcal{R}^k$ diferenciável em $\theta\in\mathcal{R}^k$ e considere U_n um k-vetor de estatística satisfazendo $a_n(U_n-\theta)\to_d Y$ para um k-vetor aleatório Y com $0<E||Y||^2<\infty$ e $\{a_n\}$ uma sequência de números positivos com $a_n\to\infty$. Seja $T_n=g(U_n)$ um estimador de $\vartheta=g(\theta)$. Então o amse e a variância assintótica de T_n, são respectivamente, $E\{[\nabla g(\theta)]^\tau Y\}^2/a_n^2$ e $[\nabla g(\theta)]^\tau Var(Y)\nabla g(\theta)/a_n^2$.*

2.4 O método de Máxima Verossimilhança

O método de máxima verossimilhança introduzido nesta seção é o método mais popular para estimar os parâmetros de um modelo estatístico. Essas estimativas são obtidas a partir da maximização da função verossimilhança. O conceito de função verossimilhança, enuciado a seguir, é central na teoria da verossimilhança.

Definição 2.16. Sejam $X_1,...,X_n$ uma amostra aleatória de tamanho n da variável aleatória X com função densidade $f(x|\theta)$, com $\theta\in\Theta$, onde Θ é o espaço de parâmetros. A função de verossimilhança de θ correspondente à amostra aleatória observada é dada por

$$L(\theta;x)=\prod_{i=1}^{n} f(x_i|\theta). \tag{2.25}$$

Definição 2.17. O estimador de máxima verossimilhança de θ é o valor $\hat{\theta}\in\Theta$ que maximiza a função de verossimilhança $L(\theta;x)$.

O Logaritmo natural da função de verossimilhança de θ é denotado por

$$l(\theta; x) = \log L(\theta; x). \tag{2.26}$$

Uma vez que, $\log x$ é uma função estritamente crescente e $f(x|\theta)$ pode ser considerado positivo, temos sem perda de generalidade que, $\hat{\theta}$ é um estimador de máxima verossimilhança se, somente se, maximiza a função $l(\theta; x)$. Além disso, no caso uniparamétrico onde Θ é um intervalo da reta e $l(\theta; x)$ é derivável, o estimador de máxima verossimilhança pode ser encontrado como a raiz da *equação de verossimilhança*

$$l'(\theta; x) = \frac{\partial l(\theta; x)}{\partial \theta} = 0. \tag{2.27}$$

Note que, os valores de θ satisfazendo a equação (2.27) pode ser um mínimo local ou global, um máximo local ou global ou simplesmente pontos estacionários. Para se concluir que a solução da equação (2.27) é um ponto de máximo, é necessário verificar se

$$l''(\theta; x) = \frac{\partial^2 \log L(\theta; x)}{\partial \theta^2} \mid_{\theta=\hat{\theta}} \; < 0. \tag{2.28}$$

Além disso, o extremo pode ocorrer na fronteira de Θ ou quando $||\theta|| \to \infty$. Por isso, é importante analisar a função de verossimilhança por inteiro para encontrar o seu máximo.

Exemplo 2.14. Sejam $X_1, ..., X_n$ uma amostra aleatória da distribuição da variável aleatória X $N(\mu, 1)$. Nesse caso, a função de verossimilhança é dada por

$$L(\mu, x) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2},$$

com $\Theta = \{\mu; -\infty < \mu < \infty\}$. Como

$$l(\mu; x) = -n \log \sqrt{2\pi} - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2,$$

segue da equação (2.27) que a equação de verossimilhança é dada por

$$\sum_{i=1}^{n}(x_i - \hat{u}) = 0,$$

logo o estimador de máxima verossimilhança de μ é dada por

$$\hat{u} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

Capítulo 3

Geometria Diferencial de Modelos Estatísticos

O presente capítulo é dedicado à introdução de estruturas geométricas-diferenciáveis fundamentais de modelos estatísticos. O espaço tangente, a métrica Riemanniana e as α-conexões serão introduzidas numa variedade estatística.

3.1 Variedades de Modelos Estatísticos

Um *modelo estatístico* é um conjunto de distribuições de probabilidade para o qual acreditamos que a verdadeira distribuição pertence. É um subconjunto de todas as possíveis distribuições de probabilidade. Trataremos uma família parametrizada de distribuição de probabilidade como um modelo estatístico. Seja $S = \{p(x, \theta)\}$ um modelo estatístico, onde x é uma variável aleatória pertencente ao espaço amostral X, e $p(x, \theta)$ é a função densidade de probabilidade de x, parametrizada por θ, com relação a uma medida comum dominante P sobre X. Aqui, consideraremos θ como uma parâmetro real n-dimensional $\theta = (\theta^1, \theta^2, ..., \theta^n)$ pertencente a algum subconjunto aberto Θ do espaço real n-dimensional $\mathbb{R}^n$.

Exemplo 3.1. O modelo normal é uma família de distribuição de probabilidade tendo a seguinte função densidade,

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

onde o espaço amostral X é o R^1 com a medida de Lebesgue $dP = dx$ e o parâmentro θ é bidimensional. Podemos por $\theta = (\theta^1, \theta^2) = (\mu, \sigma)$, pois μ e σ são, geralmente, os parâmetros usados para especificar uma distribuição normal. O conjunto de parâmetro Θ é o semi-plano,

$$\Theta = \{(\mu, \sigma) | -\infty < \mu < \infty, 0 < \sigma\}.$$

Assim, o conjunto S é composto por todas as distribuições normais, e cada distribuição normal $N(\mu, \sigma^2)$ em S é especificado pelo parâmetro bidimensional $\theta = (\mu, \sigma)$.

Uma *variedade n-dimensional* S é um espaço de Hausdorff que é localmente homeomorfo a um espaço euclidiano n-dimensional $\mathcal{R}^n$. Considere $\phi : U \subset S \to \mathcal{R}^n$ o homeomorfismo de um subconjunto aberto de S com $\mathcal{R}^n$. Dado $p \in U$, o mapeamento $\phi(p) = \theta = (\theta^1, ..., \theta^n) \in \mathcal{R}^n$ é chamada *função coordenada* sobre a vizinhança coordenada U (Figura 2.1).

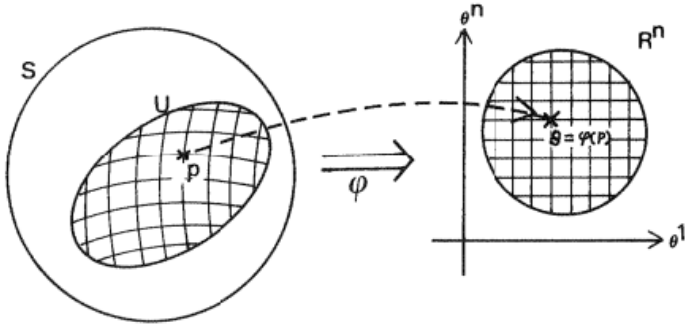


Fig. 2.1

Nós introduzimos um *sistema de coordenadas* em U de modo que cada ponto $p \in U$ é dado em coordenadas $\theta = (\theta^1, ..., \theta^n)$ ou $\theta = (\theta^i)$, i=1,...,n. As coordenadas de θ definem um nome para o ponto p. Nós podemos obter as curvas coordenadas em U mapeando-as em $\mathcal{R}^n$ por ϕ^{-1}.

Seja ψ uma outra função coordenada onde $\psi(p) = \xi = (\xi^1, ..., \xi^n)$. As coordenadas $\xi = (\xi^i), i = 1, ..., n$, definem outro nome para o mesmo ponto p. Assim, dados dois sistemas de coordenadas, cada ponto tem dois nomes ou duas coordenadas θ e ξ. Chamamos de *transformações de coordenadas* as correspondências injetivas entre as coordenadas θ e ξ dadas por:

$$\xi = \psi \circ \phi^{-1}(\theta), \qquad \theta = \phi \circ \psi^{-1}(\xi)$$

que podem ser escritas na forma de componente como

$$\xi^i = \xi^i(\theta^i, ..., \theta^n), \qquad \theta^i = \theta^i(\xi^1, ..., \xi^n), \quad i = 1, ..., n.$$

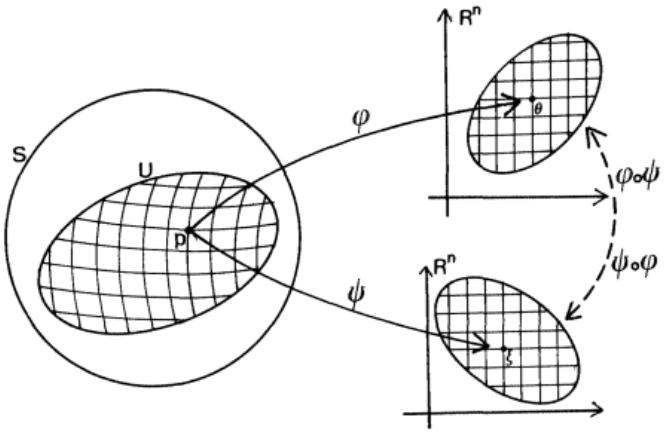


Fig. 2.2

Definição 3.1. A transformação de θ para ξ é dita ser um *difeomorfismo* quando as n funções $\xi^i(\theta^1, ..., \theta^n)$ são diferenciáveis com relação a $\theta^1, ..., \theta^n$ e o jacobiano da transformação

$$\det |\frac{\partial \xi^i}{\partial \theta^j}|$$

não se anula em U, onde det denota determinante da matriz cujo elemento a_{ij} é $\frac{\partial \xi^i}{\partial \theta^j}$. Neste caso, a transformação inversa de ξ para θ é também um difeomorfismo.

Quando nós nos referirmos a *estrutura diferenciável* de uma variedade, estaremos tratando apenas de sistemas de coordenadas que estão mutuamente ligados por difeomorfismo. Mais precisamente, uma estrutura diferenciável local é introduzida em U definindo um sistema de coordenadas. A mesma estrutura diferenciável é introduzida por qualquer um dos sistemas de coordenadas admissíveis ligados por difeomorfismos.

Nós temos até agora tratado a estrutura local de uma variedade S pela restrinção a um conjunto aberto U. Ao menos que S seja homeomorfo a R^n, não existem funções coordenadas que cubram toda S. Neste caso, considere uma cobertura aberta $U = \{U_i\}$ de S, com $\bigcup U_i = S$ de modo que a função coordenada ϕ_i é definida sobre cada conjunto aberto U_i. Sempre que dois conjuntos abertos U_i e U_j coincidirem, um ponto $p \in U_i \bigcap U_j$ tem os dois conjuntos de coordenadas $\theta = \phi_i(p)$ e $\xi = \phi_j(p)$. Por tanto, nós podemos definir a transformação de coordenadas de $\theta = \phi_i(p)$ para $\xi = \phi_j(p)$ nos pontos p pertencentes a ambos U_i e U_j. Quando todas as transformações coordenadas são difeomorfismo, a estrutura diferenciável é introduzida em S pela cobertura aberta U juntamente com as funções coordenadas ϕ_i definidas em U_i. Um espaço Hausdorff metrizável é chamado uma *variedade diferenciável* quando se tem essa tal cobertura. O par (U_i, ϕ_i) definido

por uma vizinhança coordenada e uma função coordenada é chamada de *gráfico*, e a coleção de (U_i, ϕ_i)'s é chamada de *atlas*.

Voltando a família de distribuição de probabilidade $S = \{p(x, \theta)\}$ de um modelo estatístico, considere a aplicação $\phi : S \to \mathcal{R}^n$ dado por

$$\phi[p(x, \theta)] = \theta.$$

Quando esta aplicação desempenha o papel de uma função coordenada, o vetor θ é utilizado como as coordenadas ou nome da distribuição $p(x, \theta)$ e, portanto, podemos introduzir uma estrutura diferencial em S por esta função coordenada. Assim, S é uma variedade diferenciável. Seja $\xi = (\xi^1, ..., \xi^n)$ outra parametrização do modelo S de modo que θ e ξ estão conectados pelo difeomorfismo $\xi = \xi(\theta)$ e $\theta = \theta(\xi)$. Então ξ define outro sistema de coordenadas em S. Qualquer sistema de coordenadas admissível pode ser usado para analisar as propriedades geométricas de S. Observe que as coordenadas são nada mais do que um nome ligado a cada ponto (distribuição) $p \in S$. As propriedades geométricas intrínsecas são independentes da nomenclatura.

As condições de regularidade a seguir são necessárias na teoria geométrica que serão estudadas mais na frente.

1. Todas as $p(x, \theta)$'s tem um suporte comum de modo que $p(x, \theta) > 0$ para todo $x \in X$, onde X é o suporte.

2. Seja $l(x, \theta) = \log p(x, \theta)$. Para todo θ fixado, as n funções em x

$$\frac{\partial}{\partial \theta^i} l(x, \theta), \qquad i = 1, 2, ..., n$$

   são linearmente independentes.

3. Os momentos das variáveis aleatórias $\frac{\partial}{\partial \theta^i} l(x, \theta)$ existem até certas ordens.

4. As derivadas parciais $\partial / \partial \theta^i$ e a integração com relação a medida P sempre podem ser trocados como

$$\frac{\partial}{\partial \theta^i} \int f(x, \theta) dP = \int \frac{\partial}{\partial \theta^i} f(x, \theta) dP$$

   para todas as funções $f(x, \theta)$.

3.2 Espaço Tangente

Iremos agora definir o espaço tangente de uma variedade S no ponto p e logo em seguida mostraremos alguns exemplos. A princípio, podemos dizer que um espaço tangente no ponto p de uma variedade S é, a grosso modo, um espaço vetorial obtido pela linearização local de S sobre p (ver Fig. 2.3). Ele é composto por vetores tangentes a curvas suaves passando por p.

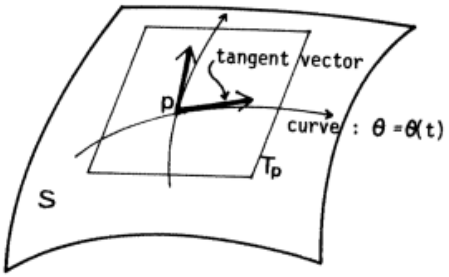


Fig. 2.3

Definiremos uma curva $c = c(t)$ como uma aplicação contínua de um intervalo fechado $[a, b] \in \mathcal{R}$ em S, onde $c(t)$ é a imagem de $t \in [a, b]$. Se nós usarmos um sistema de coordenada $\theta = \phi(p)$, a imagem do ponto t, $c(t)$, é dada pelas coordenadas

$$\theta(t) = \{\theta^1(t), ..., \theta^n(t)\}$$

e diremos que a equação $\theta = \theta(t)$ é a representação paramétrica da curva c.

Definição 3.2. Uma curva $c = c(t)$ é dita ser *suficientemente suave* quando $\theta(t)$ é diferenciável até uma certa ordem.

Considere F o conjunto de todas as funções reais suaves em S e seja θ um sistema de coordenadas. Dados uma curva suave $c = c(t)$ e uma função suave $f = f(\theta^1, ..., \theta^n) \in F$ no sistema de coordenadas θ, podemos definir a função $f \circ c : [a, b] \to \mathcal{R}$, a qual é escrita como $f\{\theta(t)\}$ na expressão de coordenadas.

Denotemos por $D_C(f)$ a derivada dessa função,

$$D_C(f) = \frac{d(f \circ c)}{dt} = \frac{df\{\theta(t)\}}{dt} = \sum_{i=1}^{n} \frac{d\theta^i}{dt} \frac{\partial}{\partial \theta^i} f. \tag{3.1}$$

Isto é, obviamente, a derivada de f ao longo da curva c ou na direção da tangente de c. Assim, um operador derivada direcional D_C está associado com cada curva, e intuitivamente, D_C depende apenas do "vetor tangente $d\theta^i/dt$"da curva c. Além disso, o operador D_C satisfaz as seguintes

condições em cada ponto da curva:

(1) D_C é uma aplicação linear de F em $\mathcal{R}$

(2) $D_C(fg) = [D_C(f)]g + f[D_C(g)]$, para $f, g \in F$.

Por outro lado, uma aplicação D_C que satisfaz as condições acima referidas é sempre derivado como o operador derivada direcional da curva. O conjunto dessas aplicações D_C formam um espaço vetorial n-dimensional, desde que S seja suficientemente suave (c^∞-variedade). Chamaremos este conjunto de **espaço tangente** T_p de S em p.

Dado um sistema de coordenadas θ, podemos considerar n curvas coordenadas $c_1, c_2, ..., c_n$ passando por um ponto p_0. Por exemplo, a primeira curva coordenada c_1 é a curva onde o valor da primeira coordenada θ^1 muda enquanto todas as outras coordenadas são fixas. Portanto, a curva c_1 é representada por

$$\theta_1(t) = (\theta_0^1 + t, \theta_0^2, ..., \theta_0^n)$$

onde $\theta_0 = (\theta_0^1, ..., \theta_0^n)$ é a coordenada de p_0. Então, o vetor tangente D_{C_1} de c_1 é nada mais que a derivada parcial com relação a θ^i,

$$D_{C_1}(f) = \frac{d}{dt} f[\theta_1(t)] = \frac{\partial}{\partial \theta^1} f.$$

Por isso, denotaremos o vetor tangente D_{C_1} por $\dfrac{\partial}{\partial \theta^1}$ ou por ∂_1. De forma similar, o vetor tangente D_{C_i} da curva coordenada c_i será denotado por ∂_i (Fig. 2.4).

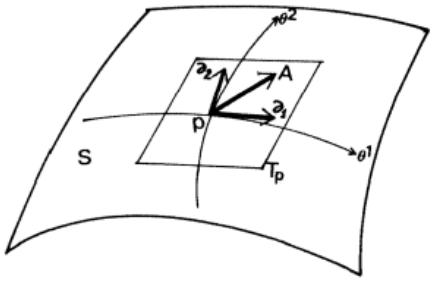


Fig. 2.4

Uma vez que D_{C_i} é simplesmente a derivada parcial $\frac{\partial}{\partial \theta^i}$, consideraremos ∂_i como a abreviação de $\frac{\partial}{\partial \theta^i}$. Note que, os n vetores ∂_i são linearmente independentes, formando assim uma base para o espaço tangente T_p.

Denotamos por $\{\partial_i\}$ a *base natural* associada ao sistema de coordenadas θ. Assim, qualquer

vetor tangente $A \in T_p$ pode ser escrito como uma combinação linear de ∂_i

$$A = \sum_{i=1}^{n} A^i \partial_i \quad ,$$

onde A^i são as componentes de A com relação a bse natural $\{\partial_i\}$.

A partir de agora, adotaremos a convenção da soma de Einstein. Portanto, $A^i\partial_i$ implica automaticamente $\sum_{i=1}^{n} A^i\partial_i$. O vetor tangente θ de uma curva $\theta(t)$ expressa em coordenadas é, de fato, dada por $\dot{\theta} = \dot{\theta}^i\partial_i$ (que implica $\sum \dot{\theta}^i\partial_i$), onde $(\dot{})$ denota d/dt,

$$\dot{\theta}f = \frac{d}{dt}f[\theta(t)] = \dot{\theta}^i\frac{\partial}{\partial\theta^i}f \tag{3.2}$$

Por isso, $\dot{\theta}^i$ são as componentes do vetor tangente $\dot{\theta}$ da curva $\theta(t)$.

Uma representação mais familiar de um vetor tangente no caso da variedade $S = \{p(x,\theta)\}$ de um modelo estatístico será dado a seguir. Suponha

$$l(x,\theta) = \log p(x,\theta) \tag{3.3}$$

e considere as n derivadas parciais $\partial_i l(x,\theta)$, i=1,2,...,n. Assumiremos que as derivadas parciais são funções linearmente independentes em x para cada θ fixado. Assim, podemos construir o seguinte espaço vetorial n-dimensional gerado pelas n funções $\partial_i l(x,\theta)$ em x,

$$T_\theta^{(1)} = \{A(x)|A(x) = A^i\partial_i l(x,\theta)\}$$

isto é, $A(x) \in T_\theta^{(1)}$ pode ser escrito como combinação linear de $\partial_i l(x,\theta)$, onde A^i são as componentes de $A(x)$ com relação a base $\partial_i l(x,\theta)$. Uma vez que x é uma variável aleatória, $T_\theta^{(1)}$ é o espaço linear das variáveis aleatórias gerado por $\partial_i l(x,\theta)$. Existe um isomorfismo natural entre os dois espaços vetoriais T_θ e $T_\theta^{(1)}$ dada pela seguinte correspondência

$$\partial_i \in T_\theta \longleftrightarrow \partial_i l(x,\theta) \in T_\theta^{(1)}.$$

Obviamente, um vetor tangente $A = A^i\partial_i \in T_\theta$ corresponde a uma variável aleatória $A(x) = A^i\partial_i l(x,\theta) \in T_\theta^{(1)}$, possuindo as mesmas componentes A^i.

Definição 3.3. O espaço $T_\theta^{(1)}$ é chamado a *1-representação do espaço tangente.*

Seja $E[\cdot]$ a esperança com relação a distribuição $p(x,\theta)$,

$$E[f(x)] = \int f(x)p(x,\theta)dP. \tag{3.4}$$

Diferenciando a identidade $\int p(x,\theta)dP = 1$ com relação a θ^i, temos:

$$0 = \partial_i \int p(x,\theta)dP = \int \partial_i p(x,\theta)dP = \int p(x,\theta)\partial_i l(x,\theta)dP = E[\partial_i l(x,\theta)]$$

Por isso, para toda variável aleatória $A(x) \in T_\theta^{(1)}$, $E[A(x)] = 0$.

Temos utilizado até agora um sistema de coordenadas $\theta = (\theta^i)$. No entanto, podemos utilizar outro sistema de coordenadas $\xi = (\xi^\alpha), \alpha = 1, ..., n$, para especificar uma distribuição em S. Existe um difeomorfismo entre θ e ξ dado por $\xi = \xi(\theta)$, $\theta = \theta(\xi)$, ou na forma de componentes

$$\xi^\alpha = \xi^\alpha(\theta^1, ..., \theta^n), \qquad \theta^i = \theta^i(\xi^1, ..., \xi^n),$$

$$i = 1, ..., n; \qquad \alpha = 1, ..., n.$$

Aqui, o índice i é usado para denotar as componentes de θ, enquanto que o índice α é usado para denotar as componentes de ξ. É conveniente usar diferentes letras de índices para denotar as componentes com relação a diferentes sistemas de coordenadas. Assim, usaremos $i, j, k, etc.$ para representar as quantidades com relação a θ, e $\alpha, \beta, \gamma, etc.$ para as quantidades com relação a ξ.

As matrizes Jacobianas das transformações de coordenadas acima são escritas como

$$B_i^\alpha(\theta) = \frac{\partial \xi^\alpha}{\partial \theta^i}, \qquad \bar{B}_\alpha^i(\xi) = \frac{\partial \theta^i}{\partial \xi \alpha}.$$

Diferenciando a identidade $\theta[\xi(\theta)] = \theta$ ou $\theta^i[\xi^1(\theta), ..., \xi^n(\theta)] = \theta^i$ com relação a θ^j, temos

$$\frac{\partial \theta^i}{\partial \xi^\alpha}\frac{\partial \xi^\alpha}{\partial \theta^j} = \bar{B}_\alpha^i B_j^\alpha = \delta_j^i$$

onde δ_j^i é o delta de Kronecker, que é igual a 1 quando $i = j$ e é igual a 0 quando $i \neq j$. De forma análoga, temos

$$B_i^\alpha \bar{B}_\beta^i = \delta_\beta^\alpha.$$

Por isso, segue que a inversa da matriz jacobiana (B_i^α) é a matriz jacobiana $(\bar{B}_\alpha^i)$. Sejam $\{\partial_i\}$ e $\{\partial_\alpha\}$ as bases naturais do espaço tangente com relação a θ e ξ, respectivamente. Então, as relações

$$\partial_\alpha = \bar{B}_\alpha^i \partial_i \ , \qquad \partial_i = B_i^\alpha \partial_\alpha \tag{3.5}$$

são validas, pelo fato de serem derivadas parciais. Representando o mesmo vetor A nessas duas bases $A = A^i\partial_i = A^\alpha\partial_\alpha$, temos as respectivas componentes A^i e A^α. A partir das relações 3.5 pode-se mostrar que as componentes estão relacionadas por

$$A^i = \bar{B}_\alpha^i A^\alpha \quad , \quad A^\alpha = B_i^\alpha A^i. \tag{3.6}$$

A 1-representação $A(x)$ de A é invariante para qualquer sistemas de coordenadas

$$A(x) = A^i \partial_i l(x, \theta) = A^\alpha \partial_\alpha l(x, \xi),$$

e somente suas componentes mudam em uma maneira contravariante com a mudança de base.

3.3 Métrica Riemanniana e Informação de Fisher

Definição 3.4. Quando o produto interno $<A, B>$ de dois vetores tangentes $A, B \in T_\theta$ é definido, a variedade S é chamada um *espaço Riemanniano*.

Iremos introduzir o produto interno na variedade de um modelo estatístico de maneira natural. Sejam $A(x), B(x)$ as 1-representações de A, B. Então, seu produto interno é definido por

$$<A, B> = E[A(x)B(x)] \tag{3.7}$$

Por isso, o produto interno é a covariância das variáveis aleatórias $A(x), B(x)$, pois $E[A(x)] = E[B(x)] = 0$. Em particular, o produto interno dos vetores bases ∂_i e ∂_j são

$$g_{ij}(\theta) = <\partial_i, \partial_j> = E[\partial_i l(x, \theta) \partial_j l(x, \theta)]. \tag{3.8}$$

Os n^2 valores de $g_{ij}(\theta)$, $i, j = 1, 2, ..., n$, juntos formam um objeto geométrico chamado **tensor métrico**. O produto interno de dois vetores $A = A^i \partial_i$ e $B = B^j \partial_j$ pode ser expresso como

$$<A, B> = <A^i \partial_i, B^j \partial_j> = A^i B^j g_{ij}$$

na forma de componente. Portanto, o produto interno é determinado mediante o tensor métrico g_{ij}. O tensor métrico $g_{\alpha\beta}$ no sistema de coordenadas $\xi = (\xi^\alpha)$ é dado por

$$g_{\alpha\beta} = <\partial_\alpha, \partial_\beta> = <\overline{B}^i_\alpha \partial_i, \overline{B}^j_\beta \partial_j> = \overline{B}^i_\alpha \overline{B}^j_\beta <\partial_i, \partial_j> = \overline{B}^i_\alpha \overline{B}^j_\beta g_{ij}.$$

Definição 3.5. Uma quantidade geométrica T é chamada um *tensor covariante de ordem 2*, quando é expressa por n^2 componentes t_{ij} em um sistema de coordenadas $\theta = (\theta^i)$ e transformam-se em

$$t_{\alpha\beta} = \overline{B}^i_\alpha \overline{B}^j_\beta t_{ij}$$

quando é expressa em outro sistema de coordenadas $\xi = (\xi^\alpha)$.

Exemplo 3.2. O tensor métrico é, de fato, um tensor covariante de ordem 2.

Definição 3.6. Dois vetores tangentes A, B são ditos *ortogonais*, quando o seu produto interno é igual a zero, $<A, B> = A^i B^j g_{ij} = 0$.

Naturalmente, A, B são ortogonais quando suas 1-representações A(x) e B(x) não estão correlacionados, isto é, suas covariâncias se anulam. Duas curvas $\theta_1(t)$ e $\theta_2(t)$ passando por um ponto $\theta_0 = \theta_1(0) = \theta_2(0)$, em $t = 0$, são ditas ortogonais neste ponto, quando seus vetores tangentes $\dot{\theta}_1(0)$ e $\dot{\theta}_2(0)$ são ortogonais neste ponto, $<\dot{\theta}_1, \dot{\theta}_2> = \dot{\theta}_1^i \dot{\theta}_2^j g_{ij} = 0$, isto é,

$$cov[\dot{l}\{x, \theta_1(t)\}, \dot{l}\{x, \theta_2(t)\}] = 0, \quad \text{para t} = 0.$$

Definição 3.7. *O comprimento $|A|$ de um vetor tangente A é dado por*

$$|A|^2 = <A, A> = A^i A^j g_{ij}.$$

Note que, esta definição de comprimento é obviamente a variância da 1-representação $A(x)$

$$|A|^2 = E[\{A(x)\}^2].$$

Seja p e p' dois pontos em S, os quais estão "infinitesimanente próximos", sendo θ e $\theta + d\theta$ suas respectivas coordenadas. Considere $\overrightarrow{pp'} = d\theta^i \partial_i$ um vetor "infinitesimal"em T_θ, o quadrado da distância $ds = \left|\overrightarrow{pp'}\right|$ entre duas distribuições $p = p(x, \theta)$ e $p' = p(x, \theta + d\theta)$ é dado pela forma quadrática

$$ds^2 = \left|\overrightarrow{pp'}\right|^2 = \left\langle \overrightarrow{pp'}, \overrightarrow{pp'} \right\rangle = g_{ij} d\theta^i d\theta^j.$$

A matriz (g_{ij}) é conhecida na estatística como a *matriz informação de fisher* e iremos denotar a sua inversa por (g^{ij}).

Seja $c : \theta(t)$ uma curva suave ligando o ponto $\theta_0 = \theta(t_0)$ ao ponto $\theta_1 = \theta(t_1)$. Então a distância s de θ_0 a θ_1 ao longo da curva c é obtida integrando a distância infinitesimal entre $\theta(t)$ e $\theta(t + dt) = \theta(t) + \dot{\theta} dt$,

$$ds^2 = g_{ij}[\theta(t)] \dot{\theta}^i \dot{\theta}^j dt^2$$

de modo que,

$$s = \int ds = \int_{t_0}^{t_1} \sqrt{g_{ij} \dot{\theta}^i \dot{\theta}^j} dt.$$

Definição 3.8. Dentre todas as curvas que ligam os pontos θ_0 e θ_1, chamaremos de *geodésica Riemanniana* aquela curva cuja distância é mínima.

3.4 Conexão Afim

Nesta seção introduziremos o conceito de conexão afim sobre o espaço S. Para isso, precisaremos das definições de campo de vetores e derivada covariante. A coleção $A = \{A(\theta) | \theta \in S\}$ é chamada um *campo vetorial*, quando cada ponto $\theta \in S$ está associado a um vetor $A(\theta) \in T_\theta$. Mais formalmente, podemos definir um campo vetorial da seguinte forma:

Definição 3.9. Um campo vetorial é uma aplicação de S em T_θ, a qual atribui a cada ponto $\theta \in S$ um vetor $A(\theta) \in T_\theta$.

Exemplo 3.3. O i-ésimo vetor da base ∂_i no sistema de coordenadas θ é um campo vetorial, que atribiu para cada $\theta \in S$ a sua derivada parcial $\partial_i \in T_\theta$.

Um campo vetorial A, que pode ser representado na forma de componente como $A = A^i(\theta)\partial_i(\theta)$, é dito ser suave quando as componentes $A^i(\theta)$ são funções suaves em θ. Iremos denotar por $T(S)$ o conjunto de todos os campos vetoriais suaveis de S.

Uma vez que dois espaços tangentes T_θ e $T_{\theta'}$ são diferentes quando $\theta \neq \theta'$, não existe uma forma direta para comparar os vetores $A(\theta) \in T_\theta$ e $A(\theta') \in T_{\theta'}$. A comparação direta de suas componentes $A^i(\theta)$ e $A^i(\theta')$ não tem sentido pois os vetores bases $\partial_i(\theta)$ e $\partial_i(\theta')$ são diferentes. A fim de comparar dois vetores pertencentes a dois espaços vetorias diferente, é necessário estabelecer uma correspondência injetiva entre os espaços. O que iremos fazer agora é obter uma correspondencia afim entre os espaços tangentes adjacentes T_θ e $T_{\theta'}$, onde $\theta' = \theta + d\theta$ é "infinitesimante" próximo a θ. Uma vez estabelecida essa tal correspondência para qualquer dois pontos adjacentes, ela poderá ser estendida ao longo de uma curva $\theta(t)$ para obtermos uma correspondência entre os espaços tangentes $T_{\theta(t_0)}$ e $T_{\theta(t_1)}$ nos pontos distantes $\theta(t_0)$ e $\theta(t_1)$, embora a correspondência dependa, na maioria das vezes, da curva que liga esse dois pontos.

Considere $m : T_{\theta+d\theta} \to T_\theta$ uma aplicação linear, a qual se reduz a função identidade quando $d\theta \to 0$. Uma vez que $d\theta$ é pequeno, o vetor base $\partial'_j = \partial(\theta + d\theta) \in T_{\theta+d\theta}$ é aplicado para um vetor $m(\partial'_j)$ próximo a $\partial_j(\theta)$.

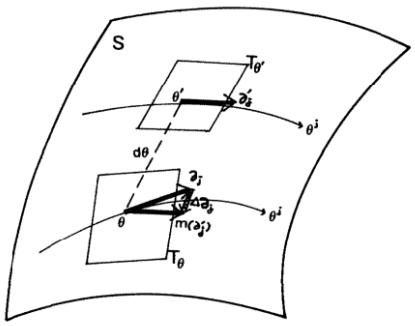


Fig. 2.5

Expandindo a diferença $\Delta\partial_j$ entre $m(\partial'_j)$ e ∂_j, isto é,

$$\Delta\partial_j = m[\partial'_j] - \partial_j(\theta) \in T_\theta,$$

com relação a $d\theta$ e negligenciando os termos de ordem superior, o vetor $\Delta\partial_j$ é expresso como

$$\Delta\partial_j = d\theta^i \Gamma_{ij}^k(\theta)\partial_k$$

onde $d\theta^i\Gamma_{ij}^k$ são as componentes de $\Delta\partial_j \in T_\theta$. Por isso, a aplicação m é determinada por n^3 funções $\Gamma_{ij}^k(\theta)$, $i,j,k = 1,...,n$, em θ. Uma vez que m é linear, podemos estabelecer uma correspondência entre vetores de T_θ e $T_{\theta+d\theta}$ por,

$$A^i\partial_i' \in T_{\theta+d\theta} \mapsto A^i m[\partial_i'] = (A^k + d\theta^i\Gamma_{ij}^k A^j)\partial_k \in T_\theta.$$

Uma correspondência afim entre T_θ e $T_{\theta+d\theta}$ é obtida a partir da aplicação m considerando que a origem de $T_{\theta+d\theta}$ é mapeado para o ponto $d\theta^i\partial_i \in T_\theta$. Por esta correspondência afim, temos que um ponto $A^i\partial_i \in T_{\theta+d\theta}$ é mapeado para um ponto em T_θ, onde a sua imagem é dada pelo vetor

$$(A^k + d\theta^k + d\theta^i A^j\Gamma_{ij}^k)\partial_k.$$

As n^3 funções $\Gamma_{ij}^k(\theta)$ em θ são chamadas de *coeficientes da conexão afim*.

Iremos denominar a diferença $\Delta\partial_j$ como a mudança intrínseca no vetor base $\partial_j(\theta)$ quando o ponto muda de θ para $\theta + d\theta$. Assim, a taxa da mudança intrínseca de ∂_j quando o ponto θ muda na direção de ∂_i, denotada por $\nabla_{\partial_i}\partial_j$, é dado pelo vetor

$$\nabla_{\partial_i}\partial_j = \Gamma_{ij}^k(\theta)\partial_k(\theta). \tag{3.9}$$

Note que $\nabla_{\partial_i}\partial_j$ é um campo vetorial, o qual será chamado de *derivada covariante* do campo vetorial ∂_j ao longo de ∂_i. Assim, podemos pensar que a derivada covariante é determinada a partir dos coeficientes $\Gamma_{ij}^k(\theta)$ da conexão afim. Pelo contrário, as derivadas covariantes determinam unicamente os coeficientes da conexão afim subjacente. De fato, tomando o produto interno de ambos os lados da equação (3.8) com ∂_m, temos

$$\begin{aligned} <\nabla_{\partial_i}\partial_j, \partial_m> &= \Gamma_{ij}^k \langle\partial_k, \partial_m\rangle \\ &= \Gamma_{ij}^k(\theta) g_{km}(\theta) \end{aligned} \tag{3.10}$$

Para facilitar a escrita, é conveniente definir a expressão covariante dos coeficientes da conexão afim por

$$\Gamma_{ijm}(\theta) = \Gamma_{ij}^k(\theta) g_{km}(\theta). \tag{3.11}$$

Assim, multiplicando (3.10) pela inversa (g^{mk}) da matriz (g_{km}), obtemos

$$\Gamma_{ij}^{k} = g^{km}\Gamma_{ijm}. \tag{3.12}$$

Depois dessa introdução intuitiva acima, iremos agora dar uma definição mais precisa de uma conexão afim.

Definição 3.10. Dados dois campos vetorias $A, B \in T(S)$, a derivada covariante de B ao longo de A é um campo vetorial C, denotado por $C = \nabla_A B$, onde o vetor $C(\theta) \in T_\theta$ é interpretado como a taxa de mudança intríseca no campo vetorial $B(\theta)$ quando o ponto θ muda na direção do vetor $A(\theta)$.

Dados os campos vetoriais $A, A', B, B' \in T(S)$, a derivada covariante deve satisfazer as condições de linearidade

$$\nabla_A(B + B') = \nabla_A B + \nabla_A B', \tag{3.13}$$

$$\nabla_{(A+A')}B = \nabla_A B + \nabla_{A'} B. \tag{3.14}$$

Dada uma função escalar suave $f : S \to \mathcal{R}$, temos que fA é também um campo vetorial cujo valor em θ é $f(\theta)A(\theta) \in T_\theta$. Além disso, a derivada covariante deve também satisfazer as seguintes condições:

$$\nabla_A(fB) = (Af)B + f\nabla_A B, \tag{3.15}$$

$$\nabla_{(A+A')}B = \nabla_A B + \nabla_{A'} B. \tag{3.16}$$

onde $Af = A^i(\theta)\partial_i f(\theta)$.

Definição 3.11. Uma conexão afim sobre S é uma derivada covariante ∇, isto é, uma aplicação de $T(S) \times T(S)$ em $T(S)$ satisfazendo as quatro condições acima.

Obviamente, os coeficientes da conexão afim são obtidos por

$$\Gamma_{ijk}(\theta) = <\nabla_{\partial_i}\partial_j, \partial_k>. \tag{3.17}$$

Por outro lado, quando Γ_{ijk} é dado, a derivada covariante $\nabla_A B$, onde $A = A^i(\theta)\partial_i$ e $B = B^i(\theta)\partial_i$, pode ser calculado como

$$\begin{aligned}
\nabla_A B &= \nabla_A(B^j\partial_j) \\
&= (AB^j)\partial_j + B^j\nabla_A\partial_j \\
&= (A^i\partial_i B^j)\partial_j + B^j A^i\Gamma_{ij}^k\partial_k \\
&= (A^i\partial_i B^k + A^i B^i\Gamma_{ij}^k)\partial_k
\end{aligned} \tag{3.18}$$

Portanto, as n^3 quantidades $\Gamma_{ijk}(\theta)$ definem uma conexão afim. Os coeficientes de uma conexão afim dependem do sistema de coordenadas θ. Dado outro sistema de coordenadas $\xi = (\xi^\alpha)$ com a matriz jacobiana $\overline{B}^i_\alpha = \dfrac{\partial\theta^i}{\partial\xi^\alpha}$, os coeficientes da conexão afim são calculados como

$$\begin{aligned}
\Gamma_{\alpha\beta\gamma}(\xi) &= <\nabla_{\partial_\alpha}\partial_\beta, \partial_\gamma> \\
&= <\nabla_{\overline{B}^i_\alpha\partial_i}(\overline{B}^j_\beta), \overline{B}^k_\gamma\partial_k> \\
&= \overline{B}^i_\alpha\overline{B}^k_\gamma<\nabla_{\partial_i}(\overline{B}^j_\beta\partial_j), \partial_k> \\
&= \overline{B}^i_\alpha\overline{B}^k_\gamma\{\overline{B}^j_\beta<\nabla_{\partial_i}\partial_j, \partial_k> + \partial_i\overline{B}^j_\beta<\partial_j, \partial_k>\} \\
&= \overline{B}^i_\alpha\overline{B}^j_\beta\overline{B}^k_\gamma\Gamma_{ijk} + g_{jk}\overline{B}^k_\gamma\partial_\alpha\overline{B}^j_\beta. \qquad (3.19)
\end{aligned}$$

Isso mostra como os coeficientes de uma conexão afim mudam sob transformação de coordenadas. Note que Γ_{ijk} não é um tensor. Além disso, embora a derivada covariante ∇ seja definida independentemente do sistema de coordenadas, a expressão Γ_{ijk} depende do sistema pois os campos de vetores naturais ∂_i são definidos com base no sistema de coordenadas $\theta = (\theta^i)$.

Uma vez que introduzido o conceito de uma conexão afim, podemos falar sobre a a linearidade (e portanto curvatura) de uma curva. Sejam $\theta(t)$ uma curva em S e B um campo de vetores definido sobre a curva tal que $B(\theta)$, em $\theta = \theta(t)$, é escrito como $B(t) = B^i(t)\partial_i$. O vetor tangente $\dot\theta = \dot\theta^i(t)\partial_i$ da curva é também um campo vetorial definido sobre a curva. (Ver fig. 2.6).

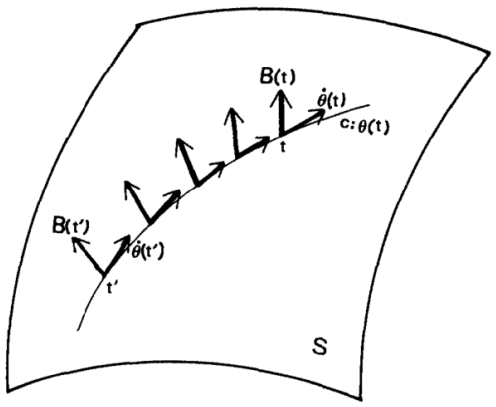


Fig. 2.6

A derivada covariante $\nabla_{\dot\theta}B$ indica como $B(t)$ muda ao longo da curva, isto é, a mudança intrínseca de B ao longo da curva. Quando não existem mundanças intrínsecas de $B(t)$, temos que $B(t)$ satisfaz a equação $\nabla_{\dot\theta}B = 0$, que é equivalente a

$$\dot B^k(t) + \dot\theta^i B^j\Gamma^k_{ij} = 0, \qquad (3.20)$$

onde a relação $\dot\theta f = \dot\theta^i\partial_i f = \frac{d}{dt}f[\theta(t)] = f$ é usada. Quando $B(t)$ é a solução da equação acima, o

vetor $B(t)$ em $T_{\theta(t)}$ é dito ser a *mudança paralela* de $B(t')$ em $T_{\theta(t')}$ ao longo da curva. Um vetor B em θ' pode ser paralelamente deslocado para qualquer ponto θ ao longo da curva que conecta esses dois pontos.

Quando o vetor tangente $\dot{\theta}$ de uma curva $\theta(t)$ pode mudar sua magnitude mas não sua direção, ele deve satisfazer a equação

$$\nabla_{\dot{\theta}}\dot{\theta} = c(t)\dot{\theta}.$$

Escolhendo um parâmetro apropriado t, essa equação reduz para a forma simples

$$\nabla_{\dot{\theta}}\dot{\theta} = 0, \tag{3.21}$$

que implica que o vetor tangente da curva não muda ao longo de toda a curva. Isso é uma generalização da reta na geometria euclidiana. Ela é chamada uma *geodésica* com relação à conexão afim. Note que, a equação (3.20) pode ser reescrito como

$$\ddot{\theta}^k(t) + \dot{\theta}^i(t)\dot{\theta}^j(t)\Gamma_{ij}^k\{\theta(t)\} = 0 \tag{3.22}$$

na forma de componente.

3.5 α-conexões estatísticas

Na seção anterior, vimos uma noção matemática geral de conexão afim. Agora, iremos introduzir uma conexão afim no espaço S de um modelo estatístico tal que representem as propriedades intrínsecas da família de distribuição probabilidade. A 1-representação do espaço tangente novamente fornece uma boa orientação para isso. Uma vez que a base natural $\partial_j(\theta + d\theta)$ de $T_{\theta+d\theta}$ é representada por uma variável aleatória

$$\partial_j l(x, \theta + d\theta) = \partial_j l(x, \theta) + \partial_i\partial_j l(x, \theta)d\theta^i,$$

para obter uma conexão afim, é necessário encontrar uma maneira de mapeá-lo para o espaço $T_\theta^{(1)}$ gerado pelas n funções $\partial_j l(x, \theta)$. Visto que a esperança em θ de $\partial_i\partial_j l(x, \theta)$ não se anula por causa da equação (3.7), $\partial_j l(x, \theta + d\theta)$ não pertence a $T_\theta^{(1)}$. Então, primeiro modificamos $\partial_j l(x, \theta + d\theta) \in T_{\theta+d\theta}^{(1)}$ de modo que a esperança se anule em θ. Isso pode ser feito adicionando $g_{ij}(\theta)d\theta^i$, obtendo assim

$$\partial_j l(x, \theta) + \{\partial_i\partial_j l(x, \theta) + g_{ij}(\theta)\}d\theta^i.$$

Uma vez que esta variável aleatória não pertence, em geral, a $T_\theta^{(1)}$, nós a projetamos para o espaço linear $T_\theta^{(1)}$. Assim, estabelecemos por esta projeção uma correspondência linear entre T_θ e $T_{\theta+d\theta}$. Visto que $\Delta\partial_j$ é a projeção de $\{\partial_i\partial_j l(x, \theta) + g_{ij}(\theta)\}d\theta^i$ em $T_\theta^{(1)}$, a conexão afim resultante é dada

por

$$\begin{aligned}\Gamma_{ijk}(\theta) &= E[\{\partial_i\partial_j l(x,\theta) + g_{ij}(\theta)\}\partial_k l(x,\theta)] \\ &= E[\partial_i\partial_j l(x,\theta)\partial_k l(x,\theta)]. \end{aligned} \tag{3.23}$$

Essa conexão é também obtida projetando $\partial_i\partial_j l(x,\theta)d\theta^i$ diretamente em $T_\theta^{(1)}$.

Existe uma outra possibilidade de modificar a variável aleatória $\partial_i\partial_j l(x,\theta)$, por causa de que a esperança em θ de

$$\partial_i\partial_j l(x,\theta) + \partial_i l(x,\theta)\partial_j l(x,\theta)$$

também se anula. Essa modificação nos leva para outra conexão afim cujos coeficientes são dados por

$$\Gamma_{ijk}(\theta) = E[\{\partial_i\partial_j l(x,\theta) + \partial_i l(x,\theta)\partial_j l(x,\theta)\}\partial_k l(x,\theta)]. \tag{3.24}$$

As duas definições acima sugerem que um número infinito de conexões afins pode ser introduzido usando uma média ponderada. Seja α um parâmetro escalar. Então, a modificação de $\partial_i\partial_j l(x,\theta)$ para

$$\partial_i\partial_j l(x,\theta) + \frac{1+\alpha}{2}g_{ij}(\theta) + \frac{1-\alpha}{2}\partial_i l\partial_j l$$

nos leva para a conexão afim cujos os coeficientes são dados por

$$\Gamma_{ijk}^{(\alpha)}(\theta) = E[\{\partial_i\partial_j l(x,\theta) + \frac{1-\alpha}{2}\partial_i l(x,\theta)\partial_j l(x,\theta)\}\partial_k l(x,\theta)]. \tag{3.25}$$

A representação (3.25) é chamada de α-*conexão*, e se reduz a equação (3.23) quando $\alpha = 1$, e para (3.24) quando $\alpha = -1$. A derivada covariante com relação a α-conexão será denotado por $\nabla^{(\alpha)}$.

Vamos definir um *tensor de terceira ordem* por

$$T_{ijk}(\theta) = E[\partial_i l(x,\theta)\partial_j l(x,\theta)\partial_k l(x,\theta)]. \tag{3.26}$$

A α-conexão pode ser escrita como

$$\Gamma_{ijk}^{(\alpha)} = \Gamma_{ijk}^{(1)} + \frac{1-\alpha}{2}T_{ijk}, \tag{3.27}$$

o qual é conveniente para calcular os coeficientes das α-conexões.

Exemplo 3.4. [A α-conexão da distribuição normal]. Diferenciando $l(x,\theta)$ duas vezes com relação a $\theta = (\mu,\sigma)$ para as distribuições normais, obtemos

$$\partial_1\partial_1 l(x,\theta) = \frac{-1}{\sigma^2}, \qquad \partial_1\partial_2 l(x,\theta) = -\frac{2(x-\mu)}{\sigma^3}, \qquad \partial_2\partial_2 l(x,\theta) = -\left\{\frac{3(x-\mu)^2}{\sigma^4}\right\} + \frac{1}{\sigma^2}.$$

Note que as variáveis aleatórias acima não pertencem a $T_\theta^{(1)}$, porque suas esperanças não se anulam.

Então, a fim de encontrar a contrapartida de $\partial_i l(x, \theta + d\theta) = \partial_i l + \partial_i \partial_j l d\theta^j \in T^{(1)}_{\theta+d\theta}$ em $T^{(1)}_{\theta}$, é necessário modificar os termos $\partial_i \partial_j l(x, \theta)$ de forma que suas esperanças se anulem. A 1-conexão é obtida a partir da equação (2.25), assim temos que

$$\Gamma^{(1)}_{111} = \Gamma^{(1)}_{112} = \Gamma^{(1)}_{221} = \Gamma^{(1)}_{122} = \Gamma^{(1)}_{212} = 0,$$

$$\Gamma^{(1)}_{121} = \Gamma^{(1)}_{211} = -2/\sigma^3, \quad \Gamma^{(1)}_{222} = -6/\sigma^3.$$

Calculando as componentes de T_{ijk}, obtemos

$$T_{111} = T_{221} = 0 \quad , \quad T_{112} = 2/\sigma^3 \quad , \quad T_{222} = 8/\sigma^3.$$

e todas as outras componentes são obtidas a partir da simetria de T_{ijk}, isto é,

$$T_{ijk} = T_{jik} = T_{kij} = \dots.$$

Por tanto, a α-conexão é obtida por (3.26) como

$$\Gamma^{(\alpha)}_{111} = \Gamma^{(\alpha)}_{212} = \Gamma^{(\alpha)}_{122} = \Gamma^{(\alpha)}_{221} = 0 \quad , \quad \Gamma^{(\alpha)}_{112} = (1-\alpha)/\sigma^3,$$

$$\Gamma^{(\alpha)}_{121} = \Gamma^{(\alpha)}_{211} = -(1+\alpha)/\sigma^3 \quad , \quad \Gamma^{(\alpha)}_{222} = -2(1+2\alpha)/\sigma^3.$$

3.6 Curvatura e Torção

Definição 3.12. Um (covariante) *tensor* Q de ordem k é uma aplicação multilinear de k campos vetoriais em $\mathcal{R}$,

$$Q : T(S) \times \cdots \times T(S) \longrightarrow \mathcal{R}.$$

Para k campos vetorias $A_1, ..., A_k \in T(S)$, seu valor é escrito como $Q(A_1, ..., A_k)$, onde Q é linear para qualquer A_i com quaisquer funções escalares como coeficientes. Dado um sistema de coordenadas θ, podemos ver o valor de Q para os campos de vetores base $\partial_1, ..., \partial_n$. Então, obtemos as n^k quantidades

$$Q_{i_1 \dots i_k} = Q(\partial_{i_1}, ..., \partial_{i_k}), \quad i_1, ..., i_k = 1, 2, ..., n.$$

Para campos vetorias $A_1 = A^i_1 \partial_i, ..., A_k = A^i_k \partial_i$, temos

$$Q(A_1, ..., A_k) = Q_{i_1 \dots i_k} A^{i_1}_1 \dots A^{i_k}_k.$$

Essa é a forma de componente do tensor operação. As n^k quantidades $Q_{i_k \dots i_k}$ são chamadas

as componetes do tensor Q com relação a base natural $\{\partial_i\}$. Em termos de outro sistema de coordenadas ξ, as componentes são dadas por

$$Q_{\alpha_i \ldots \alpha_k} = B^{i_1}_{\alpha} \ldots B^{i_k}_{\alpha} Q_{i_1 \ldots i_k}. \tag{3.28}$$

Assim, a equação 3.28 nos dá a regra de transformação de um tensor.

Exemplo 3.5. A métrica Riemanniana é um tensor de ordem 2 cujas componentes são dadas por g_{ij}.

Exemplo 3.6. Uma conexão afim não é um tensor, pois ela não define uma aplicação multilienar e sua regra de transformação 3.19 não é equivalente a regra 3.28.

Iremos agora definir outros tipos de tensores.

Definição 3.13. Um tensor R de ordem $k+1$, de covariante k e ordem contravariante 1 é uma aplicação multilinear de k campos vetoriais em um campo vetorial,

$$Q : T(S) \times \cdots \times T(S) \longrightarrow T(S).$$

A aplicação dos campos de vetores bases $\partial_{i_1}, ..., \partial_{i_k}$, nos dá um vetor,

$$R(\partial_{i_1}, ..., \partial_{i_k}) = R^{j}_{i_1 \ldots i_k} \partial_j.$$

As $n^k + 1$ quantidades $R^{j}_{i_1 \ldots i_k}$ são chamadas as componentes do tensor R com relação ao sistema de coordenadas $\theta = (\theta^i)$ ou a base natural $\{\partial_i\}$. Para campos vetorias $A_1 = A^i_1 \partial_i, ..., A_k = A^i_k \partial_i$, temos

$$R(A_1, ..., A_k) = A^{i_1}_1 \ldots A^{i_k}_j R^{j}_{i_1 \ldots i_k} \partial_j.$$

Quando um produto interno é introduzido no espaço tangente T_θ, esta aplicação R define um tensor contravariante R' de ordem $k+1$ por

$$R'(A_1, ..., A_k) = <R(A_1, ..., A_k), B>.$$

As componentes de R' são dados por

$$\begin{aligned} R_{i_1 \ldots i_k j} &= R'(\partial_{i_1}, ..., \partial_{i_k}, \partial_j) \\ &= <R(\partial_{i_1}, ..., \partial_{i_k}, \partial_j)> \\ &= R^{m}_{i_1 \ldots i_k} g_{mj}. \end{aligned} \tag{3.29}$$

onde $R^{m}_{i_1 \ldots i_k} = R_{i_1 \ldots i_k j} g^{jm}$. Por isso, R' é uma versão covariante (versão de indice inferior) de R.

Definição 3.14. A torção é uma aplicação bilinear de $T(S)\times T(S)$ em $T(S)$ induzido pela conexão afim.

Portanto, a torção é um tensor de ordem 3. Dados dois campos vetorias $A, B \in T(S)$, a aplicação é definida por

$$S(A,B) = \nabla_A B - \nabla_B A - (AB - BA) \tag{3.30}$$

que é também um campo vetorial. Dado um sistema de coordendas θ, a torção é representada pelo *tensor torção* cujas componentes são definidas por $S_{ijk}(\theta) = <S(\partial_i, \partial_j), \partial_k>$. Uma vez que as derivadas parcias ∂_i e ∂_j comutam, $\partial_i\partial_j - \partial_j\partial_i = 0$, S_{ijk} é obtido de Γ_{ijk} por

$$S_{ijk}(\theta) = \Gamma_{ijk} - \Gamma_{jik}. \tag{3.31}$$

Esse S_{ijk} é chamado um tensor antissimétrico com relação a i e j.

Uma vez que os coeficientes $\Gamma^{(\alpha)}_{ijk}(\theta)$ da α-conexão são simétricos com relação aos dois primeiros índices i e j, como pode ser verificado a partir da definição 3.25, o tensor torção S_{ijk} se anula identicamente para qualquer α-conexão. Portanto, a variedade de um modelo estatístico é torção-livre.

Definição 3.15. A *curvatura Riemann-Christoffel* R é uma aplicação trilinear de $T(S) \times T(S) \times T(S)$ em $T(S)$ induzida por uma conexão afim.

Portanto é um tensor de ordem 4. Dados três campos de vetores $A, B, C \in T(S)$, a aplicação é definida por

$$R(A,B,C) = [\nabla_A, \nabla_B]C - \nabla_{[A,B]}C, \tag{3.32}$$

onde [,] implica alternação, por exemplo

$$[\nabla_A, \nabla_B] = \nabla_A\nabla_B - \nabla_B\nabla_A \ , \quad [A,B] = AB - BA.$$

Dado um sistema de coordenadas θ, a curvatura é representada pela componente de um tensor

$$R_{ijkm} = <R(\partial_i, \partial_j, \partial_k), \partial_m>, \tag{3.33}$$

chamado o *tensor curvatura de Riemann-Christoffel*. Por outro lado, $R(A,B,C)$ pode ser calculada a partir das componentes de A, B, C e do tensor curvatura. Assim, calculando 3.33, o tensor curvatura de Riemann-Christofell é obtido como

$$R_{ijkm} = (\partial_i\Gamma^S_{ij} - \partial_j\Gamma^S_{ik})g_{sm} + (\Gamma_{irm}\Gamma^r_{jk} - \Gamma_{jrm}\Gamma^r_{ik}). \tag{3.34}$$

Definição 3.16. Um espaço com uma conexão afim é dita ser um *plano*, quando a curvatura Riemann-Christoffel se anula identicamente, isto é, $R(A,B,C) = 0$ para qualquer $A, B, C \in T(S)$,

ou equivalentemente, $R_{ijkm}(\theta) = 0$ para qualquer θ.

Definição 3.17. Uma curva $\theta(t)$, $t \in [t_0, t_1]$ é chamada um *loop* quando ela é fechada, isto é, $\theta(t_0) = \theta(t_1)$ com $(t_0 \neq t_1)$.

Em um espaço plano, um vetor $A \in T(\theta)$ não sofre mudanças quando ele é deslocado em paralelo ao longo de um loop passando por θ. Isso implica que, quando um vetor $A \in T_\theta$ é deslocado em paralelo para o espaço tangente $T_{\theta'}$ em outro ponto θ', o vetor deslocado $A' \in T(\theta')$ é unicamente determinado, independente de qualquer curva que o vetor esteja sendo deslocado em paralelo. Por isso, podemos construir um campo vetorial A tal que, para qualquer $\theta \in S$, $A(\theta)$ é obtido a partir de $A(\theta_0)$ em θ_0 pelo seu descolamente paralelo. Esse campo vetorial é dito ser um *campo vetorial paralelo.*

Capítulo 4

Variedade de Pertubação e Medidas de Influência

Seja $p(Y|\theta)$ a função probabilidade de um $M(n) \times 1$ vetor aleatório $Y^T = (Y_1^T, ..., Y_n^T)$ parametrizada por um vetor de parâmetros desconhecido $\theta = (\theta_1, ..., \theta_q)^T$ em um subconjunto aberto Θ de $\mathcal{R}^q$. Além disso, cada Y_i é um $m_i \times 1$ vetor aleatório, onde $\sum_{i=1}^n m_i = M(n)$. Por exemplo, em estudos longitudinais, m_i pode representar o número de observações no i-ésimo grupo. Na base do modelo adotado $p(Y|\theta)$ e observações em $Y^T = (Y_1^T, ..., Y_n^T)$, podemos então levar a inferência estatística, como a estimação e testes de hipóteses.

Seja $\omega = (\omega_1, ..., \omega_p)^T$ um vetor de pertubação com ω variando em $\Omega \subset \mathbb{R}^p$. Se um vetor de pertubação, que é introduzido para perturbar $p(Y|\theta)$, tem um grande efeito, então é importante saber a causa (por exemplo, as observações influentes ou suposições de modelos invalidos) desse efeito tão grande. Portanto, é importante desenvolver métodos estatísticos para quantificar o efeito de perturbação de um modelo estatístico e identificar a causa potencial. Em Cook, um método geral foi desenvolvido para avaliar a influência local de perturbação a um modelo estastístico através da introdução de ω em $p(Y|\theta)$, denotado por $p(Y|\theta, \omega)$. A metodologia é baseada na curvatura direcional de um gráfico de influência, que é definida como

$$IG(\omega) = (\omega, f(\omega)^T), \tag{4.1}$$

onde $f : \mathcal{R}^p \to \mathcal{R}^1$ é uma função objetivo suficientemente suave (diferenciável até uma certa ordem). Considere a linha reta $\omega(t) : \omega(t) = \omega^0 + th$ no espaço euclidiano $\mathcal{R}^p$ e a linha levantada $IG_h(\omega(t))$ para qualquer vetor $h \neq 0$, onde ω^0 é um vetor coluna fixo em $\mathcal{R}^p$. O vetor tangente e o vetor normal da linha levantada são, respectivamente, dadas por $\binom{I_p,}{\nabla_f^T}$ e $(1 + \nabla_f^T \nabla_f)^{-1/2} \binom{-\nabla_f}{1}$, onde $\nabla_f = (\partial f(\omega)/\partial \omega_i)$ é avaliado em ω^0 e I_p é a matriz identidade de ordem p. A curvatura

normal do gráfico de influência é dado por

$$C_h = \frac{1}{(1+\nabla_f^T\nabla_f)^{1/2}}\frac{h^T H_f h}{h^T(I_p+\nabla_f\nabla_f^T)h}, \tag{4.2}$$

onde H_f denota a matriz $(\partial^2 f(\omega)/\partial\omega_i\partial\omega_j)$ avaliada em ω^0. O valor máximo de C_h e a direção correspondente tem sido amplamente utilizados para avaliar os efeitos do uso de $\omega(t)=\omega^0+th$ para perturbar um modelo estatístico. Definiremos a *curvatura normal conformal* em ω^0 na direção h como

$$B_h = \frac{1}{||H_f||_M}\frac{h^T H_f h}{h^T(I_p+\nabla_f\nabla_f^T)h}, \tag{4.3}$$

onde $||\cdot||_M$ denota a norma de uma matriz de modo que $||H_f||_M = \sqrt{tr[H_f]^2}$. Porém, C_h não é invariante escalar em qualquer ω com $\nabla_f \neq 0$, por causa de que a curvatura normal de $\hat{IG}(\omega) = (\omega^T, kf(\omega))^T$ é dado por

$$\hat{C}_h = \frac{1}{(1+k^2\nabla_f^T\nabla_f)^{1/2}}\frac{kh^T H_f h}{h^T(I_p+k^2\nabla_f\nabla_f^T)h} \neq C_h,$$

o que pode levar a conclusões ambíguas quando k varia. O mesmo problema também surge com B_h, isto é, B_h não é invariante escalar. Em particular, Fung e Kwan [3] argumentaram que as conclusões tiradas a partir do novo gráfico $\hat{IG}(\omega)$ devem ser as mesmas que aquelas tiradas a partir do gráfico antigo $IG(\omega) = (\omega^T, f(\omega))^T$.

A partir de agora, iremos denotar a função densidade por $p(Y|\theta,\omega)$ de modo que $\int p(Y|\theta,\omega)dY = 1$. Para avaliar a influência local de um modelo de perturbação, estamos interessados principalmente no comportamento de $p(Y|\theta,\omega)$ como uma função de ω em torno de ω^0, e não o vetor de parâmetros θ. Daqui em diante, θ é assumido ser conhecido ou fixado a um determinado valor (por exemplo, a estimativa de máxima verossimilhança) e $p(Y|\theta,\omega^0) = p(Y|\theta)$. Além disso, $P(Y|\theta,\omega)$ satisfaz as quatro condições de regularidade do capítulo 3 e ω^0 representa nenhuma perturbação.

O modelo perturbado $p(Y|\theta,\omega)$ é caracterizado por um conjunto de perturbações ω, que tem um estrutura geométrica natural. O modelo perturbado $M = \{p(Y|\theta,\omega) : \omega \in \Omega\}$ pode ser considerado como uma variedade p-dimensional. Dado um sistema de coordenada ω, e_i $(i = 1,...,p)$ é a base natural do espaço tangente T_ω de M associado ao sistema de coordenadas. Seja $T_\omega^{(1)}$ o espaço vetorial de M em ω, o qual é gerado por p funções $\partial_i l(\omega|Y,\theta)$, onde $l(\omega|Y,\theta) = \log p(Y|\theta,\omega)$. Existe um isomorfismo natural entre esses dois espaços vetorias tangentes T_ω e $T_\omega^{(1)}$. O espaço $T_\omega^{(1)}$ é chamado a 1-representação do espaço tangente de M. Para qualquer vetor $h = \sum_{i=1}^p h^i e_i \in T_\omega$, a 1-representação $h(Y)$ de h em $T_\omega^{(1)}$ é dado por $h(Y) = \sum_{i=1}^p h^i\partial_i l(\omega|Y,\theta)$, onde $\partial_i = \partial/\partial\omega_i$.

Definição 4.1. O produto interno de dois operadores base ∂_i e ∂_j é calculado como

$$g_{ij}(\omega) = <\partial_i, \partial_j> = E_\omega[\partial_i l(\omega|Y,\theta)\partial_j l(\omega|Y,\theta)], \tag{4.4}$$

onde E_ω denota a esperança tomado com relação à $p(Y|\theta,\omega)$. As p^2 quantidades $g_{ij}(\omega)$, $i,j = 1,2,...,p$, formam o tensor métrico.

A matriz métrica $G(\omega) = (g_{ij}(\omega))$ é uma matriz informação de Fisher esperado com relação ao vetor de perturbação ω. Os elementos de $G(\omega)$ medem as quantidades de perturbações que todas as componentes de um vetor de perturbação ω contribuem para o modelo estatístico. Temos que o (i,i)-ésimo elemento $g_(i,i)(\omega)$ indica a quantidade de perturbação introduzida pelo i-ésimo componente de ω. Já os elementos fora da diagonal de $G(\omega)$ representam a associação entre os diferentes componentes de ω. Por exemplo, seja $r_{ij}(\omega) = g_{ij}(\omega)/\sqrt{g_i i(\omega) g_j j(\omega)}$. Um grande valor absoluto de $r_{ij}(\omega)$ indica uma forte associação entre o i-ésimo e o j-ésimo componentes de $r_{ij}(\omega)$. Em particular, se $G(\omega)$ é uma matriz diagonal, então todas as componentes de ω são ortogonais uns aos outros no modelo perturbado. Além disso, se $G(\omega)$ não é positiva definida para um esquema de perturbação, então os p operadores ∂_i são linearmente dependentes. Assim, alguns componentes do vetor de perturbação são redundantes e portanto devem ser removidos.

A partir da discussão acima, uma pertubação apropriada para um modelo estatístico deve satisfazer no mínimo as seguintes condições:

1. $G(\omega)$ ser positiva definida em uma pequena vizinhança de ω^0;

2. Os elementos fora da diagonal de $G(\omega)$ em ω^0 devem ser tão pequenos quanto possível.

A condição 1 é necessária para evitar qualquer componente redundante de ω. A condição 2 é necessária para assegurar que podemos identificar facilmente a causa de um grande efeito. Por exemplo, se os diferentes componentes de ω estão altamente associados, então é difícil inferir se um grande efeito é causado por apenas uma simples componente ou por várias componentes de ω. Por tanto, uma perturbação apropriada requer que $G(\omega^0)$ seja

$$diag(g_{11}(\omega^0), ..., g_{pp}(\omega^0)).$$

Além disso, podemos sempre escolher um novo vetor de perturbação $\tilde{\omega}$, definido por

$$\tilde{\omega} = \omega^0 + c^{-1/2} G(\omega)^{1/2}(\omega - \omega^0), \tag{4.5}$$

tal que $G(\tilde{\omega})$ avaliado em ω^0 seja igual a cI_p, onde $c > 0$. Portanto, sem perda de generalidade, vamos assumir que uma perturbação apropriada ω satisfaz a equação $G(\omega^0) = cI_p$. No entanto, nem sempre é possível encontrar um vetor de perturbação de modo que $G(\omega) = cI_p \ \forall \omega \in \Omega$.

Iremos agora introduzir algumas quantidades geométricas para o modelo perturbado M baseado no tensor métrico.

O comprimento $||h||^2$ de um vetor tangente $h \in T_\omega$ é dado por

$$||h||^2 = <h, h> = \sum_{i,j} h^i h^j g_{ij}(\omega) = h^T G(\omega) h. \tag{4.6}$$

Seja $C : \omega(t) = (\omega_1(t), ..., \omega_p(t))$ uma curva suave na variedade M ligando dois pontos $\omega^1 = \omega(t_1)$ e $\omega^2 = \omega(t_2)$. A distância $S(\omega^1, \omega^2)$ de ω^1 a ω^2 ao longo da curva C é dada por

$$S(\omega^1, \omega^2) = \int_{t_1}^{t_2} \sqrt{\sum_{i,j} g_{ij}(\omega(t)) \frac{d\omega_i(t)}{dt} \frac{d\omega_j(t)}{dt}} dt. \tag{4.7}$$

Definimos o tensor assimetria T e uma família de conexões Γ^α para $\alpha \in \mathcal{R}$, respectivamente, como

$$T_{ijk}(\omega) = E_\omega[\partial_i l(\omega|Y, \theta) \partial_j l(\omega|Y, \theta) \partial_k l(\omega|Y, \theta)],$$

e

$$\Gamma^\alpha_{ijk}(\omega) = E_\omega[\partial_i \partial_j l(\omega|Y, \theta) \partial_k l(\omega|Y, \theta)] + 0,5(1 - \alpha) T_{ijk}(\omega).$$

Note que

$$\begin{aligned} \Gamma^\alpha_{ijk}(\omega) &= E_\omega[\partial_i \partial_j l(\omega|Y, \theta) \partial_k l(\omega|Y, \theta)] + 0,5(1 - \alpha) T_{ijk}(\omega) \\ &= E_\omega[\partial_i \partial_j l(\omega|Y, \theta) \partial_k l(\omega|Y, \theta)] + \frac{T_{ijk}(\omega)}{2} - \frac{\alpha T_{ijk}(\omega)}{2} \\ &= \Gamma^0_{ijk}(\omega) - \frac{\alpha T_{ijk}(\omega)}{2} \end{aligned}$$

onde Γ^0_{ijk} é o símbolo de Christoffel para a conexão Lévi-Civita do tensor métrico e

$$\Gamma^0_{ijk}(\omega) = \frac{1}{2}[\partial_i g_{jk}(\omega) + \partial_j g_{ik}(\omega) - \partial_k g_{ij}(\omega)].$$

Com as quantidades acima, o modelo de perturbação M é uma variedade estatística, que desempenha um papel importante na compreensão do comportamento do modelo perturbado.

Definição 4.2. Uma variedade de perturbação estatística $(M, G(\omega), T(\omega))$ é a variedade M com uma métrica $G(\omega)$ e um 3-tensor covariante $T(\omega)$.

Agora vamos considerar uma curva suave específica, chamada uma α-geodésica.

Definição 4.3. Dizemos que $\omega(t)$ é uma α-geodésica com relação a uma conexão afim $\Gamma^\alpha_{ijk}(\omega)$ se satisfaz a equação

$$\frac{d^2\omega_i(t)}{dt^2} + \sum_{s,j,k} g^{is}(\omega(t)) \Gamma^\alpha_{jks}(\omega(t)) \frac{d\omega_j(t)}{dt} \frac{d\omega_k(t)}{dt} = 0 \tag{4.8}$$

onde g^{is} são os elementos da i-ésima linha e s-ésima coluna da matriz $G(\omega)^{-1}$.

A geodésica é uma extensão direta da reta $\omega(t) = \omega^0 + th$ no espaço Euclidiano. Em particular, à medida que movemos ao longo de uma geodésica, o vetor tangente da geodésica não muda de comprimento e nem direção. Se $\Gamma^{\alpha}_{ijk}(\omega) = 0$ para qualquer ω, então a variedade é um α-plano e a equação da geodésica para esse α é linear em $t : \omega(t) = \omega^0 + th$.

Algumas propriedades importantes relacionados com as quantidades geométricas acima são resumidos no seguinte lema.

Lema 4.1. *Sejam $\phi = (\phi^1, ..., \phi^p) = \phi(\omega)$ um novo sistema de coordenadas de M, $\partial_a = \partial/\partial\phi^a$, $B^a_i = \partial\phi^a/\partial\omega_i$ e $B^i_a = \partial\omega_i/\partial\phi^a$. Então as quantidades geométricas de M no sistema de coordenadas ϕ podem ser escritos como*

1. $g_{ab} = \sum_{i,j} B^i_a B^j_b g_{ij}$;
2. $T_{abc} = \sum_{i,j,k} B^i_a B^j_b B^k_c T_{ijk}$;
3. $\Gamma^{\alpha}_{abc} = \sum_{i,j,k} B^i_a B^j_b B^k_c \Gamma^{\alpha}_{ijk} + \sum_{i,j} g_{ij} B^i_c \partial_a B^j_b$.

Usamos os índices i, j, k para denotar as quantidades relacionadas ao sistema de coordenadas ω, e os índices a, b, c para as quantidades relacionadas ao sistema de coordenadas ϕ.

Introduziremos a seguir o conceito de medida de influência de 1ª e 2ª ordem e mostraremos alguns resultados importantes.

Sejam $f(\omega) : \mathcal{R}^p \to \mathcal{R}^1$ a função objeto e $\omega(t)$ uma curva suave em M com $\omega(0) = \omega^0$ e $d\omega(t)/dt|_{t=0} = h \in T_{\omega^0}$. Portanto, $f(\omega(t))$ é uma função de $\omega(t)$ definida na variedade de perturbação M. A partir de uma expansão de taylor segue que

$$f(\omega(t)) = f(\omega(0)) + \dot{f}_h(0)t + \frac{1}{2}\ddot{f}_h(0)t^2 + o(t^2). \tag{4.9}$$

A primeira e a segunda derivada de $f(\omega(t))$ em $t = 0$ são, respectivamente, dadas por

$$\dot{f}_h(0) = \sum_j \frac{\partial f(\omega^0)}{\partial\omega_j} h_j = \nabla_f^T h \quad e \quad \ddot{f}_h(0) = h^T H_f h + \nabla_f^T \frac{d^2\omega(0)}{dt^2}. \tag{4.10}$$

Se $\nabla_f \neq 0$, então o termo de primeira ordem $\dot{f}_h(0)$ caracteriza principalmente a influência local de um vetor de perturbação ω para um modelo. No entanto, se $\nabla_f = 0$, então segue de (4.9) e (4.10) que $\dot{f}_h(0) = 0$ e $\ddot{f}_h(0) = h^T H_f h$. Assim, devemos usar o termo de segunda ordem $\ddot{f}_h(0)$ para avaliar o comportamento local da função objeto quando $\nabla_f = 0$.

Definição 4.4. A medida de influência de 1ª ordem (FI) na direção de $h \in T_\omega$ é definida como

$$FI_{f,h} = FI_{f(\omega^0),h} = \frac{h^T \nabla_f \nabla_f^T h}{h^T G h}, \tag{4.11}$$

onde $G = G(\omega^0)$.

Teorema 4.1. *Temos os seguintes resultados:*

(i) $FI_{f,h} = \lim_{t\to 0} \frac{[f(\omega(t)) - f(\omega(0))]^2}{S(\omega(0),\omega(t))^2}$.

(ii) Se ϕ *é um difeomorfismo de* ω*, então* $FI_{f(\omega),h}$ *é invariante com relação a qualquer reparametrização correspondente a* ϕ *e* $FI_{kf,h} = k^2 FI_{f,h}$ *para todo* k*.*

Demonstração. Segue da equação (4.7) que

$$S(\omega(0), \omega(t))^2 = t^2 h^T G h + o(t^2).$$

Usando a regra de L'Hôpital e a equação (4.9), fica provado a parte (i).

Supondo que $\omega = \omega(\phi)$ e $\phi = \phi(\omega)$, as matrizes jacobianas das transformações de coordenadas são dadas por

$$\Phi = \frac{\partial \phi}{\partial \omega} \quad e \quad \Psi = \frac{\partial \omega}{\partial \phi}.$$

Diferenciando as identidades $\phi[\omega(\phi)]$ e $\omega[\phi(\omega)] = \omega$ com relação a ϕ e ω, respectivamente, temos que

$$\Psi\Phi = \Phi\Psi = I_p.$$

Asim temos que $G(\phi) = \Psi^T(G(\omega))\Psi$ e $\nabla_{f(\phi^0)} = \Psi^T \nabla_{f(\omega^0)}$, onde $\phi^0 = \omega^0$. Usando a definição (4.4), provamos (ii). □

Temos duas significâncias estatísticas neste teorema. A primeira é que, a parte (i) indica que a medida de primeira ordem está associada com a primeira derivada de $f(\omega(t))$ em M avaliada em $t = 0$. Se M é um espaço euclidiano, $h^T h = 1$ e $\omega(t) = th + \omega^0$, então $FI_{f,h}$ reduz ao quadrado da derivada direcional de f em ω^0 na direção de h, dado por

$$\lim_{t\to 0} \frac{[f(\omega^0 + th) - f(\omega^0)]^2}{t^2}.$$

Em segundo lugar, embora ω possa não ser uma perturbação apropriada, podemos sempre usar G para obter uma $\tilde{\omega}$ em (4.5), de modo que

$$FI_{f(\tilde{\omega}),h}|_{\tilde{\omega}=\omega^0} = \frac{h^T G^{-1/2} \nabla_f \nabla_f^T G^{-1/2} h}{h^T h}.$$

O valor máximo de $FI_{f,h}$ é igual a $\nabla_f^T G^{-1} \nabla_f$, que quantifica o grau de influência local de $\tilde{\omega}$ para um modelo estatístico, enquanto que o vetor direção correspondente $\tilde{h}_{max} = G^{-1/2}\nabla_f$ pode ser usado para a identificar as observações influentes.

Mesmo quando $\nabla_f \neq 0$, usamos $\ddot{f}_h(0)$ para avaliar a influência local de segunda ordem de ω em um modelo estatístico. A abordagem que utiliza as informações em $\ddot{a}_h(0)$ é denominada a

abordagem de segunda ordem. No entanto, dada uma curva qualquer $\omega(t)$ em M, $\ddot{f}_h(0)$ pode não ser geometricamente bem comportada. Em vez disso, consideraremos apenas a 0-geodésica $\omega(t)$ associada à conexão Lévi-Civita do tensor métrico $G(\omega)$, que é única e definida em um intervalo contendo 0 tal que $\omega(t) = \omega^0$, e $d\omega(t)/dt = h \in T_{\omega^0}$. Assim podemos obter uma versão covariante do teorema de Taylor:

$$f(\omega(t)) = f(\omega^0) + t\nabla_f^T h + \frac{1}{2}t^2 h^T \tilde{H}_f^0 h + o(t^2), \tag{4.12}$$

onde $\tilde{H}_f^0 = \tilde{H}_{f(\omega^0)}^0$ e o (i,j)-ésimo elemento de $\tilde{H}_{f(\omega)}^0$ é dado por

$$[\tilde{H}_{f(\omega)}^0]_{(i,j)} = \partial_i\partial_j f(\omega) - \sum_{s,r} g^{sr}(\omega)\Gamma_{ijs}^0(\omega)\partial_r f(\omega).$$

A matriz $\tilde{H}_{f(\omega)}^0$ é denominada a *covariante Hessian* de $f(\omega)$. Note que, $\tilde{H}_{f(\omega)}^0$ é uma matriz simétrica pelo fato de Γ_{ijk}^0 ser simétrico com relação a i e j. Em particular, $\tilde{H}_{f(\omega)}^0$ safisfaz a seguinte propriedade:

Lema 4.2. *Seja ϕ um difeomorfismo de ω com matriz jacobiana $\Psi = \partial\omega/\partial\phi$. Então, $\tilde{H}_{f(\phi)}^0 = \Psi^T \tilde{H}_{f(\omega)}^0 \Psi$.*

O lema (4.2) mostra que $\tilde{H}_{f(\omega)}^0$ é um 2-tensor, portanto é geometricamente bem comportado.

Definição 4.5. A *medida de influência de 2^a ordem* (SI) na direção de $h \in T_{\omega^0}$ é definida como

$$SI_{f,h} = SI_{f(\omega^0),h} = \frac{h^T \tilde{H}_f^0 h}{h^T G h}. \tag{4.13}$$

A *medida de influência de 2^a ordem padronizada* (SSI) na direção $h \in T_{\omega^0}$ é definida como

$$SSI_{f,h} = SSI_{f(\omega^0),h} = \frac{1}{||G^{-1}\tilde{H}_f^0||_M} \frac{h^T \tilde{H}_f^0 h}{h^T G h}. \tag{4.14}$$

Iremos agora estabelecer algumas propriedades de $SI_{f,h}$ e $SSI_{f,h}$.

Teorema 4.2. *Temos os seguintes resultados:*

(i) $SI_{f(\omega^0),h} = \lim_{t\to 0} \dfrac{2[f(\omega(t)) - f(\omega(0)) - t\nabla_f^T h]}{S(\omega(0),\omega(t))^2}$.

(ii) Suponha que ϕ seja um difeomorfismo de ω. Então $SI_{f(\omega^0),h}$ e $SSI_{f(\omega^0),h}$ são invariantes com relação a qualquer reparametrização correspondente a ϕ em ω^0. Além disso,

$$SI_{kf(\omega),h} = kSI_{f(\omega),h} \quad e \quad SSI_{kf(\omega),h} = SSI_{f(\omega),h} \tag{4.15}$$

para qualquer $k \neq 0$ *e* $\omega \in \Omega$.

(iii) Seja$\{(\lambda_i, u_i), i = 1, ..., p\}$ *os pares autovalor-autovetor* $(E - E)$ *de* H_f^0 *com relação a* G. *Então, para qualquer direção* h, *temos* $0 \leq SSI_{f,h} \leq 1$,

$$SI_{f,u_i} = \lambda_i \quad e \quad SSI_{f,u_i} = \hat{\lambda}_i = \frac{\lambda_i}{\sqrt{\sum_{j=1}^{p} \lambda_j^2}},$$

onde $\hat{\lambda}_i$ *é o autovalor normalizado.*

Demonstração. (i) Usando (4.7), (4.12) e a regra de L'Hôpital, segue o resultado como desejado.

(ii) Pelo fato de existir um difeomorfismo entre ω e ϕ tal que $\omega = \omega(\phi)$ e $\phi = \phi(\omega)$, temos que $\Psi\Phi = \Phi\Psi = I_p$. Além disso, por $G(\phi)$ ser um tensor métrico e $\tilde{H}^0_{f(\phi)}$ um 2-tensor, temos

$$G(\phi) = \Psi^T G(\omega)\Psi \quad e \quad \tilde{H}^0_{f(\phi^0)} = \Psi^T \tilde{H}^0_{f(\omega^0}\Psi.$$

Considere a geodésica $\omega(t)$ com $\omega(0) = \omega^0$ e $d\omega(0)/dt = h \in T_{\omega^0}$. Então $\phi(\omega(t))$ é uma geodésica na ϕ-coordenada de modo que $\phi(\omega^0) = \phi^0$ e $d\phi(\omega(0))/dt = \Phi h$. Se G é positiva definida, então segue dos lemas (4.1) e (4.2) que

$$SI_{f(\phi^0),\Phi h} = \frac{h^T\Phi^T\tilde{H}^0_{f(\phi)}\Phi h}{h^T\Phi^T G(\phi^0)\Phi h} = \frac{h^T\Phi^T\Psi^T\tilde{H}^0_{f(\omega^0)}\Psi\Phi h}{h^T\Phi^T\Psi^T G(\omega^0)\Psi\Phi h} = SI_{f(\omega^0),h}.$$

De forma análoga, temos $SSI_{f(\phi^0),\Phi h} = SSI_{f(\omega^0),h}$. Então, $SI_{f,h}$ e $SSI_{f,h}$ são invariantes com relação a reparametrização ϕ em ω^0. Para qualquer k, sabemos que $\tilde{H}^0_{kf(\omega)} = k\tilde{H}^0_{f(\omega)}$ e assim a equação (4.15) é verdadeira para qualquer ω e $k \neq 0$.

(iii) Usando a definição (4.5) podemos provar a parte (iii). □

O Teorema (4.2) tem as seguintes implicações. Em primeiro lugar, se ω é uma perturbação apropriada e $\nabla_f = 0$, então

$$SSI_{f,h} = B_h \quad e \quad SI_{f,h} = C_h.$$

Em geral, embora podemos escolher uma perturbação ω que não seja apropriada, podemos sempre usar G para obter uma perturbação apropriada $\tilde{\omega}$ em (4.5). Neste caso, a curvartura normal e a medida de influência de 2ª ordem levarão ao mesmo resultado quando $\nabla_f = 0$ e a perturbação escolhida for apropriada. Portanto, o método de diagnóstico proposto aqui pode ser considerado como uma extensão da abordagem de influência local de Cook[2] em uma definição mais geral.

Em segundo lugar, $SI_f(\omega), h$ e $SSI_{f(\omega),h}$ são invariante escalar mesmo quando $\nabla_f \neq 0$, enquanto que C_h e B_h não são (Fung and Kwan[3]). Essa generalização facilita novos métodos e

técnicas para fazer uma análise de sensibilidade de um modelo estatístico.

Mostraremos agora 4 passos fundamentais na avaliação de influência local de perturbação de um modelo paramétrico $p(Y|\theta)$:

Passo 1. Escolha um esquema de perturbação ω de modo que $\int p(Y|\theta,\omega)dY = 1$.

Passo 2. Dado o modelo perturbado, calcular as quantidades geométricas da variedade de perturbação.

Passo 3. Verificar se a perturbação ω é apropriada, isto é, $G(\omega^0) = cI_p$. Se sim, prossiga para a etapa 4. Caso contrário, encontre um novo esquema de perturbação e volte para a etapa 2.

Passo 4. Escolha uma função objeto $f(\omega)$. Se $\nabla_f = 0$ então utilize SI e SSI para avaliar a influência local de menores perturbações para um modelo. No entanto, se $\nabla_f \neq 0$, use FI, SI e SSI juntos.

Referências Bibliográficas

[1] Amari, S. *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statist. 28. Springer, Berlin. 1985.

[2] Cook, R. D. (1986). Assessment of local influence (with discussion). J. Roy. Statist. Soc. Ser. B 48 133?169. MR0867994

[3] Fung, W. and Kwan, C. (1997). A note on local influence based on normal curvature. J. Roy. Statist. Soc. Ser. B 59 839?843. MR1483218

[4] SHAO, J. *Mathematical Statistics*, spriger, 2ª edição

[5] Lee, S. and Tang, N. (2004). Local infuence analysis of nonlinear structural equation models, Psychometrika

[6] Poon, W. and Poon, Y. (1999). Conformal normal curvature and assessment of local influence. J. Roy. Stat. Soc. Ser. B Stat. Methodol. 61 51?61. MR1664096

[7] Durrett, R., *Probability: Theory and Examples 3rd ed.*, Conell University. 2005.

[8] Zhu, H. and Lee, S. (2001). Local infuence for incomplete data models. J. R. Stat. Soc. Ser. B Stat. Methodol.

[9] Zhu, H. and Zhang, H. (2004). A diagnostic procedure based on local infuence. Biometrika.

[10] Zhu, Z., He, X. and Fung, W. (2003). Local infuence analysis for penalized Gaussian likelihood estimators in partially linear models. Scand. J. Statist.