# ESTIMAÇÃO DE MÁXIMA VEROSSIMILHANÇA NÃO PARAMÉTRICA PELOS MÉTODOS DE GRENANDER E DE MÁXIMA VEROSSIMILHANÇA PENALIZADA

Este exemplar corresponde à redação final da Tese devidamente corrigida e defendida pela Srta. PATRICIA CRISTINA GIMENEZ e aprovada pela Comissão Julgadora.

Campinas, 20 de setembro de 1993.

Prof. Dr. MAURO S. DE F. MARQUES

Orientador

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação (IMECC), UNICAMP, como requisito parcial para obtenção do Título de Mestre em ESTATÍSTICA.

UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística e Ciência da Computação - IMECC

# ESTIMAÇÃO DE MÁXIMA VEROSSIMILHANÇA NÃO PARAMÉTRICA PELOS METODOS DE GRENANDER E DE MÁXIMA VEROSSIMILHANÇA PENALIZADA

Patricia Cristina Giménez

Prof. Dr. Mauro S. de F. Marques Marques, Mauro S

(Orientador)

Campinas - São Paulo

1993

**AGRADECIMENTOS**

Ao meu orientador Prof. Dr. Mauro S. de Freitas Marques, pela sua dedicação e trabalho durante a orientação desta dissertação.

Ao Prof. Dr. Luiz San Martin, pelas correções e sugestões.

Ao Pedro, pelo auxílio na parte matemática.

Ao Departamento de Matematica da Facultad de Ciencias Exactas y Naturales da Universidad Nacional de Mar del Plata, pelo incentivo e apoio.

À minha família, pela compreensão e carinho.

Aos meus amigos, pelo apoio.

Ao Ricardo, pelo incentivo, amizade e amor.

Ao CNPq pelo apoio financeiro.

ÍNDICE

# INTRODUÇÃO

A inferência de verossimilhança pode ser considerada como um processo de obtenção de informação sobre um vetor de parâmetros $\theta$ a partir dos dados observados, via a função de verossimilhança.

O método de máxima verossimilhança é usado extensivamente na teoria estatística, pois tem um forte apelo intuitivo.

O método não contradiz os fatos (representados pelos dados) que nós realmente observamos e objetiva escolher o valor do parâmetro que dá chance mais provável para os fatos que ocorreram, ocorram novamente.

As suposições sobre o parâmetro a estimar, determinam a classificação do problema de sua estimação em dois casos. No caso paramétrico, o parâmetro tem sua forma funcional especificada exceto por um número finito de valores $\theta_1, \ldots, \theta_m$. Se isto não ocorrer nosso parâmetro pertence a um espaço de dimensão infinita, e temos então, um problema de estimação não paramétrica. Neste caso, geralmente são feitas apenas suposições de ordem qualitativa sobre o parâmetro.

Entretanto a metodologia tradicional de verossimilhança não é satisfatória em geral no caso não paramétrico, seja porque o estimador de máxima verossimilhança não existe, seja porque não é consistente.

Para contornar situações como esta foram desenvolvidos métodos alternativos, como o método de *"Sieves"* (Grenander (1981)) e o método de máxima verossimilhança penalizada (Good e Gaskins (1971)).

Estes métodos de estimação de máxima verossimilhança não paramétrica são aplicados em problemas de diversas áreas.

As facilidades computacionais hoje disponíveis para os estatísticos têm tido um impacto significativo na investigação estatística, especialmente no desenvolvimento de procedimentos não paramétricos de análise de dados.

Em particular, a pesquisa teórica e aplicada em estimação não paramétrica de densidade tem tido uma influência notável em tópicos relacionados, tais como regressão não paramétrica, discriminação não paramétrica e reconhecimento de padrões não paramétrico.

A estimação não paramétrica de densidade tem sido mais efetiva e flexível do que os métodos paramétricos tradicionais tanto em análise exploratória como em análise confirmatória.

Um exemplo de uma aplicação útil embora não muito familiar poderia ser a estimação de densidades de probabilidade de funções discriminantes. Esta poderia ser usada, por exemplo, para escolher entre tratamentos quando temos uma função discriminante para sucesso e fracasso segundo cada tratamento. Outro tipo de aplicação é na decisão sobre se um "*bump*" (parte entre dois pontos de inflexão, que é côncava vista por baixo) numa função de densidade existe genuinamente na população (Good e Gaskins (1980)). A detecção de "*bumps*" é uma das principais atividades dos físicos experimentais, e precisa da ajuda dos estatísticos. Quando um "*bump*" não é devido a variação aleatória, é usualmente devido a uma nova partícula ou resonância.

Os métodos de estimação de máxima verossimilhança não paramétrica também têm sido amplamente usados em inferência em processos estocásticos, em situações onde se torna difícil encontrar uma justificativa para a utilização de um modelo paramétrico (como, por exemplo, aplicações em bioestatística).

Por exemplo, os tempos de aparição de metástase poderiam ser modelados por um processo de Poisson não homogêneo. O interesse seria estimar a função de intensidade do processo na qual só restrições qualitativas seriam impostas.

Estimadores não paramétricos têm sido estudados para estimação de taxas de risco em análise de sobrevivência, demografia, confiabilidade e outras áreas, para amostras incompletas ou censuradas (Padgett e McNichols (1984), e Lubecke e Padgett (1985)).

A engenharia é outra área de aplicação destes métodos, particularmente na área conhecida como identificação de sistemas. Um processo estacionário é observado sobre um longo período de tempo e o engenheiro procura reconstruir a *"caixa preta"* que produziu o processo. Na prática isto envolve a estimação de uma densidade espectral ou uma função de transferência, o que conduz à utilização dos métodos não paramétricos de estimação de densidade.

Neste trabalho, estudaremos os aspectos teóricos de dois métodos não paramétricos para obtenção de estimadores de máxima verossimilhança: o método de *"sieves"* e o método de máxima verossimilhança penalizada.

Os estimadores obtidos por estes métodos podem ser considerados como estimadores de máxima verossimilhança restritos (Izenman (1991)) pois restrições são impostas no espaço paramétrico e na função de verossimilhança.

Especificamente no método de *"sieves"* as restrições são impostas no espaço de parâmetros, enquanto no método de máxima verossimilhança penalizada é a função de verossimilhança que é modificada.

No Capítulo 1 damos a definição de experimento estatístico abstrato e apresentamos os problemas da estimação de máxima verossimilhança ordinária.

O Capítulo 2 contém uma descrição do método de máxima verossimilhança penalizada, estudo de existência e unicidade dos estimadores e apresentação detalhada de dois exemplos de aplicação: estimação de densidade e estimação da intensidade de um processo de Poisson não homogêneo.

O Capítulo 3 traz a descrição do método de *"sieves"* com resultados sobre existência e consistência de estimadores obtidos através de sua aplicação. É apresentada uma nova versão do teorema desenvolvido por Geman e Hwang (1982) sobre consistência dos estimadores, onde são impostas restrições de semicontinuidade fraca nos funcionais e compacidade fraca nos subconjuntos que definem o *"sieve"*.

Por último, no Capítulo 4 é realizado um estudo comparativo de ambos os métodos. Não encontramos na literatura comparações profundas entre os métodos. Geman e Hwang (1982) comentam que para o caso particular de estimação de densidade, o método de máxima verossimilhança penalizada seria um tipo de *"dual"* do método de *"sieves"*. Isto é devido a que o problema de maximizar a verossimilhança penalizada corresponderia à versão de multiplicadores de Lagrange de um problema de otimização com restrições que representaria o método de *"sieves"*, mas a recíproca não resulta imediata. Grenander (1981) observa que uma idéia relacionada com a noção de *"sieves"* é a de função de penalidade, a qual atribui uma penalidade aos elementos *"menos regulares"* do espaço de parâmetros.

Neste trabalho é mostrado que na verdade ambos métodos são equivalentes em algumas situações particulares. Esta equivalência permite derivar um resultado geral de consistência para os estimadores de máxima verossimilhança penalizada utilizando a nova versão do teorema sobre consistência dos estimadores para o método de *"sieves"* desenvolvida no Capítulo 3. Observa-se que estes resultados são aplicáveis aos exemplos discutidos ao longo deste estudo sobre estimação de densidade e de intensidade de um processo de Poisson não homogêneo.

No Apêndice são definidos os conceitos utilizados ao longo do texto.

# CAPÍTULO I

## O MÉTODO DE MÁXIMA VEROSSIMILHANÇA

Neste capítulo, descreveremos o que vem a ser um experimento estatístico abstrato. Posteriormente será descrito o método de máxima verossimilhança e serão apresentados alguns dos problemas da estimação de MV ordinária no caso não paramétrico.

### 1.1. - EXPERIMENTO ESTATÍSTICO ABSTRATO

O processo de inferir a partir dos dados observados sobre parâmetros desconhecidos é parte fundamental da lógica indutiva.

As observações constituem a base de um experimento estatístico: podem ser valores numéricos ou dados de alguma outra natureza, obtidos como resultado de um experimento estatístico.

Em geral, uma observação $\underline{X}$ será um elemento aleatório com valores num espaço mensurável $\left\{\Omega,\ \mathcal{F}\right\}$. Isto significa que existe um espaço de probabilidade $\left\{\mathcal{X},\ \mathcal{G},\ P\right\}$ e X é uma função mensurável de $\left\{\mathcal{X},\ \mathcal{G}\right\} \longrightarrow \left\{\Omega,\ \mathcal{F}\right\}$.

A característica básica de um elemento aleatório X é sua distribuição de probabilidade associada $P^X$ onde $P^X(A) = P(X \in A)$

definida em $\mathcal{F}$.

Nos problemas estatísticos a única coisa conhecida sobre $P^X$ é que ela pertence a uma classe de distribuições $\mathcal{P}$. O interesse do estatístico é conhecer esta classe para fazer inferências sobre $P^X$ baseado na observação X.

Então os objetos básicos que o estatístico deve estudar são o espaço de valores $\left\{\Omega, \mathcal{F}\right\}$ e a família de distribuições $\mathcal{P}$. O espaço de probabilidade $\left\{\mathcal{X}, \mathcal{G}, P\right\}$ onde a observação X está definida tem um papel auxiliar e pode ser construído de diferentes formas. O método natural de construir $\left\{\mathcal{X}, \mathcal{G}, P\right\}$ é o seguinte: fazer $\mathcal{X} = \Omega$, $\mathcal{G} = \mathcal{F}$, $P = P^X$ e definir $X(x) = x$. Assim X é um elemento aleatório com valores em $\Omega$ e distribuição $P^X$ definida em $\left\{\Omega, \mathcal{F}, P^X\right\}$.

A família de distribuições $\mathcal{P}$ pode sempre ser parametrizada por meio da escolha apropriada de um parâmetro e representada na forma

$$\mathcal{P} = \left\{P_\theta : \theta \in \Theta\right\}$$

$\Theta$ é chamado de <u>espaço paramétrico</u>.

Uma terna $\mathcal{E} = \left\{\Omega, \mathcal{F}, \left\{P_\theta : \theta \in \Theta\right\}\right\}$ é chamada um <u>experimento estatístico</u>, e se o experimento é construído como acima por meio da observação X, dizemos que o <u>experimento</u> é <u>gerado pela observação X</u>.

Assumimos que distribuições de probabilidade distintas $P_{\theta_1}$ e $P_{\theta_2}$, definidas em $\mathcal{F}$, correspondem a pontos distintos $\theta_1$ e $\theta_2$ de $\Theta$, e vice-versa.

Portanto, as inferências estatísticas feitas sobre as distribuições $\left\{P_\theta : \theta \in \Theta\right\}$ são equivalentes àquelas feitas sobre os valores dos parâmetros $\theta \in \Theta$.

Dizemos que o experimento $\mathcal{E} = \left\{\Omega, \mathcal{F}, \left\{P_\theta : \theta \in \Theta\right\}\right\}$ é <u>dominado</u> quando existe uma medida $\mu$ $\sigma$-finita, definida em $\Omega$ que *"domina"* a família de distribuições $\left\{P_\theta\right\}$, i.e. se $\forall\ \theta \in \Theta$ as medidas $P_\theta$ são absolutamente contínuas com respeito à medida $\mu$ ($P_\theta \ll \mu$), ou seja, se $\mu(A) = 0 \Rightarrow P_\theta(A) = 0,\ \theta \in \Theta$.

Neste caso, pelo teorema de Radon-Nikodym (ver Royden (1968)), existem densidades relativas a $\mu$, para cada $\theta \in \Theta$, i.e.

$$p(x, \theta) = \frac{dP_\theta}{d\mu}(x)$$

e

$$\forall\ A \in \mathcal{F},\ P_\theta(A) = \int_A p(x, \theta)\ \mu(dx) = \int_A p(x, \theta) d\mu.$$

A medida dominante $\mu$ pode ser escolhida de diversas maneiras. Para o propósito da investigação estatística essa escolha é irrelevante, já que se $\mu_1$ e $\mu_2$ são duas medidas dominantes distintas, as densidades $p_1(x, \theta)$ e $p_2(x, \theta)$ diferem só num fator independente de $\theta$; pois

$$\frac{dP_\theta}{d(\mu_1 + \mu_2)}(x) = p_1(x, \theta)\ \frac{d\mu_1}{d(\mu_1 + \mu_2)}(x) = p_2(x, \theta)\ \frac{d\mu_2}{d(\mu_1 + \mu_2)}(x).$$

Se X representa uma seqüência de n observações independentes identicamente distribuídas $X_1, \ldots, X_n$ com distribuição $P_\theta$, $\theta \in \Theta$ no espaço mensurável $\left\{\Omega,\ \mathcal{F}\right\}$ (i.e. uma amostra aleatória simples de tamanho n de uma população com distribuição $P_\theta$); se as medidas são absolutamente contínuas com respeito à medida $\sigma$-finita $\mu$ em $\mathcal{F}$ e se tivermos a densidade $p(x,\ \theta) = \dfrac{dP_\theta}{d\mu}$, então cada observação $X_i$ gera um experimento $\mathcal{E}_i = \left\{\Omega_i,\ \mathcal{F}_i\ \left\{P_{i\theta} : \theta \in \Theta\right\}\right\}$ onde $\left\{\Omega_i,\ \mathcal{F}_i,\ \left\{P_{i\theta} : \theta \in \Theta\right\}\right\}$ são réplicas independentes de $\left\{\Omega,\ \mathcal{F},\ \left\{P_\theta : \theta \in \Theta\right\}\right\}$. O experimento $\mathcal{E} = \mathcal{E}_1 \times\ldots\times \mathcal{E}_n = \left\{\Omega^n,\ \mathcal{F}^n,\ \left\{P^n_\theta\!:\ \theta \in \Theta\right\}\right\}$ onde $\Omega^n = \Omega_1 \times\ldots\times \Omega_n$, $\mathcal{F}^n = \mathcal{F}_1 \otimes\ldots\otimes \mathcal{F}_n$, $P^n_\theta = P_{1\theta} \times\ldots\times P_{n\theta}$ corresponde à observação X é chamado produto dos experimentos $\mathcal{E}_i$, $i = 1 \ldots n$.

Todas as medidas $P^n_\theta$ são absolutamente contínuas com respeito à medida $\mu^n = \mu \times\ldots\times \mu$ em $\mathcal{F}^n$ e possuem "*densidades*"

$$\frac{dP^n_\theta}{d\mu^n} = L_n(\underset{\sim}{x},\ \theta) = p(x_1,\ \theta) \ldots p(x_n,\ \theta), \qquad \underset{\sim}{x} = (x_1,\ \ldots,\ x_n) \in \Omega^n$$

se as $X_i$ são variáveis aleatórias reais, com distribuições contínuas então $\Omega = \mathbb{R}$, $\mathcal{F}$ é a $\sigma$-álgebra de Borel em $\mathbb{R}$, $\Omega^n = \mathbb{R}^n$, $\mathcal{F}^n$ é a $\sigma$-álgebra de Borel em $\mathbb{R}^n$, X é um vetor n-dimensional e, p e $p_n$ são as densidades das distribuições de $X_i$ e X.

## 1.2. - PROBLEMAS DA ESTIMAÇÃO DE MÁXIMA VEROSSIMILHANÇA ORDINÁRIA NO CASO NÃO PARAMÉTRICO

Seja $\mathcal{E} = \left\{\Omega,\ \mathcal{F},\ \left\{P_\theta\ :\ \theta \in \Theta\right\}\right\}$ um experimento estatístico e seja $P_\theta$ absolutamente contínua com respeito à medida $\mu$ em $\mathcal{F}$, e $\frac{dP_\theta}{d\mu}(x) = p(x,\ \theta)$. Seja X a observação que gera $\mathcal{E}$.

A função $p(X,\ \theta)$ vista como uma função de $\theta$ para X fixo é chamada <u>função de verossimilhança</u> correspondente a $\mathcal{E}$ e X, i.e. $p(X,\ \theta)$ é uma função aleatória de $\theta$ definida em $\Theta \subset \mathbb{R}^k$. $p(X,\ \theta)$ serve para medir o quanto o dado X suporta uma hipótese sobre $\theta$. Um dado X é mais consistente com um vetor $\theta_1$ do que com outro $\theta_2$ se a função de verossimilhança associada a $\theta_1$ for maior do que a associada a $\theta_2$.

A estatística $\hat{\theta}$ definida pela relação

$$p(X,\ \hat{\theta}) = \sup_{\theta\in\Theta} p(X,\ \theta)$$

quando existe, é chamada <u>estimador de máxima verossimilhança</u> (EMV) para o parâmetro $\theta$ (baseado na observação X). I.e. o vetor de parâmetros mais plausível é aquele de maior verossimilhança.

No caso de observações $X_1,\ \ldots\ X_n$ independentes e identicamente distribuídas, onde $X_i$ possui densidade $p(x;\ \theta)$ com respeito à medida $\mu$, o estimador de máxima verossimilhança $\hat{\theta}_n$ é a solução da equação

$$L_n(\underset{\sim}{X}, \hat{\theta}_n) = \prod_{i=1}^{n} p(X_i, \hat{\theta}_n) = \sup_{\theta} \prod_{i=1}^{n} p(X_i, \theta) =$$

$$= \sup_{\theta} \prod_{i=1}^{n} \frac{dP_\theta}{d\mu}(X_i).$$

Muitas vezes existe um único parâmetro que maximiza a verossimilhança em $\Theta$, sendo portanto o único valor mais plausível neste espaço paramétrico. Entretanto, o EMV pode não ser único ou a função de verossimilhança resultar não limitada dentro de um espaço dado de parâmetros.

Em geral, no caso não paramétrico, i.e. quando o espaço de parâmetros $\Theta$ é de dimensão infinita, o EMV não existe, como mostraremos nos seguintes exemplos de estimação de densidade e de estimação da intensidade de um processo de Poisson não homogêneo.

### 1.2.1 - ESTIMAÇÃO DE DENSIDADE

Suponhamos que $X_1, \ldots, X_n$ sejam observações de variáveis aleatórias, independentes e identicamente distribuídas (v.a.i.i.d) cuja função de densidade de probabilidade é $\theta$, num intervalo $\Omega = (a, b)$ limitado ou não.

Queremos estimar $\theta \in L^1(\Omega)$ que originou a amostra $x_1, \ldots x_n$.

A função de verossimilhança é dada por

$$L_n(\underset{\sim}{x}, \theta) = \prod_{i=1}^{n} \theta(x_i)$$

seja $H \subset L^1(\Omega)$ alguma classe específica de funções definidas em $\Omega$ e consideremos o seguinte problema de otimização:

$$\begin{cases} \text{maximizar } L_n(\theta), \text{ sujeito a} \\ \theta \in H, \int_\Omega \theta(t)\,dt = 1 \text{ e } \theta(t) \geq 0, \ \forall\ t \in \Omega \end{cases} \tag{1.1}$$

Aqui dt denota a medida de Lebesgue em $\Omega$.

Uma solução de (1.1) será dita um <u>estimador de máxima verossimilhança</u> (EMV), baseado na amostra $x_1, \ldots, x_n$ (correspondente a H). Se H for infinito o EMV em geral não existirá.

Essa solução não existente é idealizada por uma combinação de picos de Dirac nos pontos amostrais,

$$\text{i.e.} \qquad \theta^*(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)$$

$\theta^*$ satisfaz as restrições do problema (1.1), mas formalmente $\delta \notin L^1(\Omega)$, $\delta$ é uma função generalizada.

Mas sabemos que $\delta$ pode representar-se como limite em certo sentido de uma seqüência de funções *"verdadeiras"* $\varphi_n$, cada uma das quais é zero fora de uma vizinhança de raio $\varepsilon_n$ ($\varepsilon_n \xrightarrow[n\to\infty]{} 0$) do ponto zero, e verifica

$$\int_\Omega \varphi_n(t)dt = 1 \quad \text{e} \quad \varphi_n(t) \geq 0\ \forall\ t \in \Omega.$$

Então, se nos restringirmos a $H \subset L^1(\Omega)$, onde H tem a propriedade de que os picos de Dirac podem ser aproximados pontualmente por uma seqüência de funções $\delta_n$, para as quais $\lim_{n \to \infty} \varphi_n(t) = +\infty$ para cada $t \in \Omega$, então a verossimilhança será não limitada e o EMV não existirá.

A maioria dos espaços de dimensão infinita de $L^1(\Omega)$ (funções contínuas, funções diferenciáveis, funções infinitamente diferenciáveis, polinômios, etc.) tem esta propriedade.

O fato do EMV não paramétrico não existir implica que o enfoque de máxima verossimilhança paramétrico para grandes dimensões deve necessariamente conduzir a estimadores pouco suaves e a problemas de tipo numérico.

As observações anteriores levam à seguinte crítica à estimação de máxima verossimilhança em geral: para problemas de pequenas dimensões existe pouca flexibilidade e a solução é fortemente influenciada pela eleição de H, enquanto para problemas de grandes dimensões a solução será pouco suave e o processo de otimização causará problemas numéricos.

O exemplo do histograma como EMV mostra estes problemas. Consideremos a partição de $\Omega = (a, b)$, $\pi$: $a = t_1 < t_2 < \dots < t_{m+1} = b$ onde $T_i = [t_i, t_{i+1})$, $i = 1 \dots m$ onde os $T_i$'s tem igual comprimento $h = \dfrac{b - a}{m}$

Seja

$$I(T_i)(x) = \begin{cases} 1, & x \in T_i \\ 0, & x \notin T_i \end{cases}$$

$\nu_i$: número de pontos amostrais pertencentes a $T_i$.

Mostra-se que o estimador de histograma da densidade de probabilidade

$$\theta^*(x) = \sum_{i=1}^{m} \frac{\nu_i}{nh} \, I(T_i)(x) \tag{1.2}$$

é o EMV para a amostra $x_1, \ldots, x_n$ correspondente ao subespaço H de $L^1(\Omega)$, onde $H = \left\{\omega : \omega = \sum_{i=1}^{m} y_i \, I(T_i) : y_i \in \mathbb{R}\right\}$ (ver Tapia e Thompson (1978)).

Para uma amostra fixa $x_1, \ldots, x_n$ o histograma $\theta^*$ dado por (1.2) tem a seguinte desvantagem:

$$\theta^*(x) \longrightarrow +\infty \quad \text{para } x \in \left\{x_1, \ldots, x_n\right\}$$

e

$$\theta^*(x) \longrightarrow 0 \quad \text{para } x \notin \left\{x_1, \ldots, x_n\right\}$$

quando o número de intervalos $T_i$ cresce para infinito (i.e. quando $m \longrightarrow \infty$) e seu comprimento $h \longrightarrow 0$.

O estimador será razoável ou não dependendo da escolha apropriada dos intervalos $T_i$, $i = 1, \ldots, m$. Portanto para m grande o EMV será muito pouco suave.

## 1.2.2 - ESTIMAÇÃO DA INTENSIDADE DE UM PROCESSO DE POISSON NÃO HOMOGÊNEO

As trajetórias de um processo de Poisson são funções escada com saltos de magnitude 1, contínuas à direita com limite à esquerda em cada ponto $t \in [0, \infty)$. Podemos então tomar como espaço de observações $\Omega$

o espaço

$$\mathbb{D}[0,1] = \left\{ g:[0,\ 1] \longrightarrow \mathbb{R} : g \text{ é contínua à direita com limites à esquerda} \right\}$$

Claramente, este espaço contém muitos elementos além das trajetórias de um processo pontual. No entanto, sua adoção como espaço de observações é útil por sua estrutura de espaço métrico.

Nosso espaço mensurável será dado por $\left(\mathbb{D}[0,\ 1],\ \mathcal{B}\left(\mathbb{D}[0,\ 1]\right)\right)$ onde a $\sigma$-álgebra $\mathcal{B}\left(\mathbb{D}[0,\ 1]\right)$ é a $\sigma$-álgebra gerada pelos conjuntos abertos em $\mathbb{D}[0,\ 1]$, segundo a métrica de Skorohod, que torna o espaço $\mathbb{D}[0,\ 1]$ separável (ver Parthasarathy (1967) por exemplo).

A $\sigma$-álgebra $\mathcal{B}\left(\mathbb{D}[0,\ 1]\right)$ coincide com aquela gerada pelos conjuntos cilíndricos

$$I(t_1,\ \dots,\ t_n) = \left\{ x \in \Omega : \left(x(t_1),\ \dots,\ x(t_n)\right) \in B,\ B \in \mathcal{B}(\mathbb{R}^n) \right\}$$

(ver Shiryayev (1984)).

Neste espaço mensurável definimos as medidas de probabilidade $P_\theta$, $\theta \in \Theta$ de tal forma que:

$$P_\theta(A) = P_\theta\left\{ x : N(\cdot,\ x) \in A \right\},\ A \in \mathcal{B}\left(\mathbb{D}[0,\ 1]\right)$$

onde $\left\{ N(t),\ t \geq 0 \right\}$ é o processo de Poisson não homogêneo que tem $\theta$ como sua função intensidade. I.e. o parâmetro a estimar será $\theta \geq 0$, satisfazendo

$$P(N(t + h) - N(t) = 1) = \theta(t)h + O(h)$$

Nosso experimento estatístico será considerado o *"produto"* dos experimentos $\mathscr{E}_i = \left\{\Omega_i,\ \mathscr{F}_i,\ \left\{P_{i\theta},\ \theta \in \Theta\right\}\right\}$, $i \in \mathbb{N}$, pois estaremos amostrando trajetórias independentes do processo de Poisson em $[0,\ 1]$ onde $\Omega_i = \mathbb{D}[0,\ 1]$, $\mathscr{F}_i = \mathscr{B}\left(\mathbb{D}[0,\ 1]\right)$ e $P_{i\theta} = P_\theta$.

Adotaremos como espaço paramétrico $\Theta \subset H^p([0,\ 1])$, espaço de funções absolutamente contínuas em $[0,\ 1]$ cuja p-ésima derivada é quadrado integrável com respeito à medida de Lebesgue.

O que nos interessa, de fato, é a obtenção de estimadores de máxima verossimilhança para a função $\theta(t)$. Devemos portanto determinar a função que desejamos maximizar. Como nosso espaço de observações é um espaço abstrato, não temos nenhuma medida que seja diretamente equivalente à medida de Lebesgue em $\mathbb{R}^n$.

Sabemos que, para qualquer processo de Poisson com função intensidade $\theta$, $P_\theta \ll P_1$, onde $P_1$ é a medida de probabilidade induzida sobre $\left(\mathbb{D}[0,\ 1],\ \mathscr{B}\left(\mathbb{D}[0,\ 1]\right)\right)$ pela função amostral de um processo de Poisson homogêneo com $\theta \equiv 1$ (ver Karr (1987)).

Pelo teorema de Radon-Nikodym existe $p(x,\ \theta) = \dfrac{dP_\theta}{dP_1}(x)$ tal que $P_\theta(A) = \int_A p(\theta)dP_1$ para todo $A \in \mathscr{F} = B\left(\mathbb{D}[0,\ 1]\right)$ dada por:

$$p(x,\ \theta) = \frac{dP_\theta}{dP_1}(x) = \exp\left[\int_0^1 \ln\left[\theta(t)\right]\ dN(t,\ x) + \int_0^1 \left[1 - \theta(t)\right]\ dt\right].$$

(ver Grenander (1981)).

Se possuirmos uma amostra i.i.d de tamanho n de nosso espaço $\Omega$ e definirmos

$$\bar{N}_n(t) = \sum_{i=1}^{n} N(t,\ x_i)$$

teremos

$$L_n(\theta,\ \underset{\sim}{x}) = \prod_{i=1}^{n} p(x_i,\ \theta) = \exp\left[\int_0^1 \ln\left[\theta(t)\right]\, d\bar{N}_n(t) + n\int_0^1 \left[1 - \theta(t)\right]\, dt\right] =$$

$$= \exp\left[\sum_{k=1}^{\bar{N}_n(1)} \ln\left[\theta(\bar{\tau}_k^{(n)})\right] + n\int_0^1 \left[1 - \theta(t)\right]\, dt\right]$$

onde $\bar{\tau}_k^{(n)}$ representa o instante do k-ésimo salto de $\bar{N}_n$, (podemos assumir $\bar{\tau}_k^{(n)}$ distintos). Maximizar $L_n(\theta,\underset{\sim}{x})$ é equivalente a maximizar

$$\exp\left[\sum_{k=1}^{\bar{N}_n(1)} \ln\left[\theta(\bar{\tau}_k^{(n)})\right] - n\int_0^1 \theta(t)\ dt\right]$$

Podemos ver facilmente que se quisermos maximizar esta expressão com respeito a $\theta$, então ela assumiria valores arbitrariamente grandes escolhendo por exemplo uma função contínua $\theta$ que seja zero em todas partes exceto em cada ponto de salto de $\bar{N}_n(t)$. Nestes pontos, $\theta$ deveria ser grande, decrescendo para zero em ambos os lados de maneira a fazer a integral $n\int_0^1 \theta(t)dt$ constante. Assim, não existe estimador de máxima verossimilhança no modelo original e o princípio de máxima verossimilhança falha.

Estes exemplos mostram a necessidade de métodos alternativos de estimação quando $\Theta$ possui dimensão infinita.

Os dois métodos descritos neste trabalho são enfoques do problema de estimação de máxima verossimilhança não paramétrico, onde são impostas restrições em $\Theta$ e na função de verossimilhança.

Especificamente, no método de *"sieves"* as restrições são impostas em $\Theta$. Neste caso, o EMV existirá em cada um de uma seqüência de subconjuntos de $\Theta$ que serão definidos de forma apropriada no Capítulo 3. Enquanto no método de máxima verossimilhança penalizada é a função de verossimilhança que é modificada, e o problema da maximização da verossimilhança penalizada terá solução sobre todo $\Theta$ sob condições apropriadas que daremos no Capítulo 2.

# CAPÍTULO 2

# O MÉTODO DE MÁXIMA VEROSSIMILHANÇA PENALIZADA

Quando não podemos aplicar o método de MV diretamente ou quando ele leva a resultados insatisfatórios, o uso do método de máxima verossimilhança penalizada (MVP) pode ser eficaz na eliminação dos entraves da estimação de MV usual.

Como vimos nos exemplos 1.2.1 e 1.2.2 de estimação de densidade e da intensidade de um processo de Poisson não homogêneo, a função de verossimilhança resulta não limitada e o EMV não existe nestes casos, i.e. não existe $\hat{\theta}_n \in \Theta$ tal que $L_n(\hat{\theta}_n) \geq L_n(\theta)$, $\forall\ \theta \in \Theta$. Neste Capítulo, consideraremos uma versão modificada da verossimilhança *"a verossimilhança $\phi$-penalizada"* e mostraremos que o problema da maximização deste funcional sobre $\Theta$ terá solução sobre condições apropriadas no funcional $\phi$ e no espaço paramétrico $\Theta$.

Feito isto o método será aplicado aos problemas de estimação de densidade e de intensidade de um processo de Poisson não homogêneo.

## 2.1. - DESCRIÇÃO DO MÉTODO

**DEFINIÇÃO 2.1:** Dada a amostra $X_1, \ldots, X_n$ em $\Omega$ a <u>log-verossimilhança $\phi$-penalizada</u> de $\theta \in \Theta$ é definida por

$$\ell_{n,\lambda}(\theta) = \ell_n(\theta) - \lambda\ \phi(\theta), \quad \lambda > 0$$

onde

$$\ell_n(\theta) = \log L_n(\theta) = \log \prod_{i=1}^{n} \frac{\partial P_\theta}{\partial \mu}(x_i) = \log \prod_{i=1}^{n} p(x_i, \theta) =$$

$$= \sum_{i=1}^{n} \log p(x_i, \theta)$$

**DEFINIÇÃO 2.2:** Seja o seguinte problema de otimização, para cada $\lambda > 0$

$$\begin{cases} \text{maximizar } \ell_{n,\lambda}(\theta) \text{ , sujeito a} \\ \theta \in \Theta \text{ , } \Theta \subset H \end{cases} \qquad (2.1)$$

Uma solução de (2.1) será dita um estimador de máxima verossimilhança penalizada (EMVP) baseada na amostra $X_1, \ldots, X_n$, correspondente a $\Theta$ e à função de penalidade $\phi$, e será denotada por $\theta_\lambda$.

Um EMVP será então uma curva que maximize a log-verossimilhança penalizada sobre a classe de todas as curvas satisfazendo condições de regularidade para $\phi(\theta)$ estar definida.

Quando H for espaço de Hilbert, uma função de penalidade natural é $\phi(\theta) = \|\theta\|^2$, onde $\|\cdot\|$ denota a norma em H.

Neste caso, nos referimos ao funcional de verossimilhança penalizada ou ao estimador correspondente a H sem referência ao funcional de penalidade $\phi$.

Daqui para frente $\Theta \subset H$ será considerado fechado, convexo, não limitado e tal que exista pelo menos um $\theta \in \Theta$ tal que $p(x_i, \theta) \neq 0$, $i = 1, \ldots, n$.

Estamos interessados em penalizar a log-verossimilhança de uma forma apropriada que permita obter soluções do problema (2.1) trazendo informações sobre os dados. No Capítulo 4 mostraremos que existe uma seqüência apropriada de $\lambda$'s tal que as respectivas soluções do problema (2.1) formam uma seqüência consistente.

## 2.2. - CONSIDERAÇÕES GERAIS

Na literatura, o método de estimação de MVP é apresentado com uma ótica diferente.

O método foi introduzido por Good e Gaskins em 1971 no contexto da estimação não paramétrica de densidades e tem sido discutido por vários autores (ver por exemplo, Silverman (1982), Tapia e Thompson (1978)). EMVP's também tem sido usados para uma variedade de outras curvas e problemas de estimação de funções, por exemplo, a estimação de discriminantes (Silverman (1978)), regressão logística (Anderson e Blair (1982)), estimação de taxas de risco (Anderson e Senthilselvan (1980)), Antoniadis e Gregoire (1990)), e estimação da função intensidade de um processo de Poisson não homogêneo (Bartoszynski (1981)).

Estes autores introduzem o método de MVP como uma técnica de suavização de estimadores "*asperos*" ("*rough*").

A idéia de penalizar uma medida de bondade de ajustamento por um termo baseado na *"aspereza"* (*"roughness"*) foi sugerida no contexto de regressão.

EMVP's de regressão tem sido amplamente considerados nos últimos 20 anos por analistas numéricos interessados em splines (ver por exemplo, De Boor (1978)).

Segundo Good e Gaskins (1971) no caso de estimação de densidades uma aplicação ingênua do método de MV produziria um estimador que seria média de um conjunto de funções de Dirac com massa nas n observações, mas este não seria um estimador aceitável de uma função de densidade pois é muito *"áspero"*, na verdade o estimador assim obtido não é uma função no sentido usual, é uma função generalizada.

Good e Gaskins sugerem subtrair da log-verossimilhança uma penalidade para a *"aspereza"* (*"roughness"*) antes de maximizar.

A idéia básica é que existem dois objetivos na estimação de uma curva: maximizar a adequação aos dados medida pela verossimilhança (ou log-verossimilhança) e evitar curvas *"pouco suaves"* ou que mostrem uma variabilidade muito rápida.

A *"aspereza"* de uma curva $\theta$ pode ser medida por um funcional $\phi(\theta)$ em várias formas. Uma escolha típica de $\phi(\theta)$ é $\int \theta''(t)^2 dt$, que dará um valor grande se $\theta$ tiver uma grande curvatura local e o valor zero se $\theta$ for uma reta. Maximizar $\ell_{n,\lambda}(\theta)$ representa então um critério de compromisso entre os dois objetivos. Assim o possível problema de obter um valor grande de $\ell_n(\theta)$ e também valores excessivos de $\phi(\theta)$ é resolvido subtraindo de $\ell_n(\theta)$ um múltiplo de $\phi(\theta)$. O parâmetro $\lambda$ é chamado de parâmetro de suavização e controla o balanceamento entre suavidade e bondade de ajustamento. O grau de suavização cresce quando $\lambda$ cresce, resultando um EMVP menos áspero em termos de $\phi(\theta)$.

Muitas técnicas estatísticas envolvem a escolha de um funcional de penalidade e um parâmetro de suavização para passar de uma solução muito áspera para uma solução ultrasuavizada. Neste contexto os objetivos básicos são manter confiança nos dados ao mesmo tempo que são evitados estimadores pouco suaves, e estes objetivos são medidos por funções de distância. O compromisso com estes dois objetivos conduz à resolução de um problema de minimização de distâncias penalizado, mostrando (Titterington (1985)) que o método de MVP é um caso particular desta técnica mais geral de suavização.

A justificativa filosófica do enfoque de EMVP foi originariamente um recurso *ad hoc*, mas o método tem também uma interpretação Bayesiana.

A idéia de ver a estimação não paramétrica de uma curva no contexto Bayesiano data no mínimo de 1958. (ver também Leonard (1978)).

No caso de estimação de densidade, a densidade a priori de uma curva $\theta$ é tomada proporcional a $\exp\left\{-\lambda\ \phi(\theta)\right\}$ de forma que o parâmetro de suavização $\lambda$ é um parâmetro a priori. Pode então ser mostrado que a log-verossimilhança corresponde, a menos de uma constante, ao logaritmo da densidade posterior, e assim o EMVP $\hat{\theta}$ é a moda da densidade posterior sobre o espaço de curvas. Desafortunadamente existem algumas desvantagens neste enfoque, causadas pelo menos em parte pela natureza infinito - dimensional das curvas suaves. Embora a intenção seja escolher entre curvas para as quais $\phi(\theta)$ é finita, a distribuição posterior está inteiramente concentrada fora do espaço de tais curvas suaves, (para detalhes adicionais e uma possível resolução deste paradoxo ver Silverman (1985)).

Mesmo com dificuldades deste tipo é natural pensar a estimação de curvas em uma forma Bayesiana, pois a escolha do parâmetro de suavização depende efetivamente de quão suave o estatístico pensa que a curva é.

## 2.3. - TEORIA DE EXISTÊNCIA

Neste parágrafo daremos condições suficientes para a existência dos EMVP.

Vamos nos restringir ao caso onde o funcional de penalidade $\phi$ é, pelo menos, convexo, e $0 \leq \phi(\theta) < \infty \quad \forall\ \theta \in \Theta$.

O objetivo é achar condições suficientes para a existência de maximizadores de $\ell_{n,\lambda}(\theta)$, $\lambda > 0$ sobre subconjuntos convexos, fechados e não limitados, $\Theta \subset H$, onde H representa um espaço de Hilbert.

Seja $J_n(\theta) = -\ell_n(\theta)$ e $J_{n,\lambda}(\theta) = -\ell_{n,\lambda}(\theta)$

Assim, para cada $\lambda > 0$, o problema (2.1) resulta equivalente ao problema:

$$\begin{cases} \text{minimizar } J_{n,\lambda}(\theta), \text{ sujeito a} \\ \theta \in \Theta,\ \Theta \subset H \end{cases} \tag{2.2}$$

**TEOREMA 2.1:** Seja $\Theta$ um subconjunto convexo e fechado de um espaço de Hilbert H. Suponhamos que $J_{n,\lambda}:\Theta \longrightarrow \mathbb{R}$ é coercivo em $\Theta$ e fracamente semicontínuo inferior em $\Theta$, então $J_{n,\lambda}$ atinge seu ínfimo em $\Theta$.

**PROVA:** (Corolário A-2)

No contexto deste problema a pergunta é: que condições devem satisfazer os funcionais $J_n$ e $\phi$ para que:

(i) $J_{n,\lambda}$ seja fracamente semicontínuo inferior em $\Theta$ ?

(ii) $J_{n,\lambda}$ seja coercivo em $\Theta$ ?

(i) pode ser respondido pelo seguinte Teorema:

**TEOREMA 2.2:** Suponhamos que $\Theta$ é subconjunto convexo e fechado de H.

Seja $J_n$ fracamente contínua em $\Theta$ e $\phi$ (limitada inferiormente) contínua e convexa em $\Theta$. Então $J_{n,\lambda}$ é fracamente semicontínua inferior.

**PROVA:** $\phi$ é contínua e convexa em $\Theta$ então é fracamente semicontínua inferior em $\Theta$ (Proposição A-3). Agora como $\lambda > 0$ e $J_n$ é fracamente semicontínua inferior (é fracamente contínua) segue que $J_{n,\lambda} = J_n + \lambda\phi$ é fracamente semicontínua inferior.

(ii) é mais difícil de verificar pois condições que garantam a coercividade de $J_{n,\lambda}$ em conjuntos convexos gerais $\Theta$ não são fáceis de identificar.

Mas para $J_n$ e $\phi$ bem comportadas e conjuntos $\Theta$ satisfazendo algumas condições adicionais o problema da coercividade de $J_{n,\lambda}$ é tratado mais facilmente (ver O'Sullivan (1983)).

**TEOREMA 2.3:** Se $J_n(\theta)$ for convexa em $\Theta$, duas vezes Gâteaux diferenciável em $\Theta$ e $\phi(\theta) = \|\theta\|_H^2$, então $J_{n,\lambda}(\theta)$ é uniformemente convexa em $\Theta$.

**PROVA:** $J''_{n,\lambda}(\theta)(\eta, \eta) = J''_n(\theta)(\eta, \eta) + 2\lambda \|\eta\|_H^2 \geq 2\lambda \|\eta\|_H^2$ pois $J''_n(\theta)(\eta, \eta) \geq 0$ por ser $J_n$ convexa (Proposição A-1, i)). Então $J''_{n,\lambda}(\theta)(\eta, \eta)$ é uniformemente definida positiva relativa a $\Theta$ com constante $2\lambda$, logo $J_{n,\lambda}(\theta)$ é uniformemente convexa em $\Theta$ com constante $\lambda$ (Proposição A-2).

**TEOREMA 2.4:** Seja $\Theta$ subconjunto fechado e convexo em H. Suponhamos que $J_n(\theta)$ seja contínua, duas vezes Gâteaux diferenciável e convexa em $\Theta$, e $\phi(\theta) = \|\theta\|_H^2$. Então o problema (2.2) tem solução única.

**PROVA:** $J_{n,\lambda}(\theta)$ é contínua em $\Theta$ pois $J_n(\theta)$ e $\phi(\theta) = \|\theta\|_H^2$ são contínuas em $\Theta$. Então por Teorema 2.3 é uniformemente convexa em $\Theta$, e por Teorema A-6, $J_{n,\lambda}(\theta)$ tem um único minimizador em $\Theta$.

Geralmente $\phi(\theta)$ representa uma norma em algum espaço ou bem pode ser mostrado que resolver o problema de MVP com penalidade $\phi(\theta)$ é análogo a resolver o problema utilizando uma outra função de penalidade que seja equivalente a uma norma.

Este é o caso para os EMVP duma função de densidade: O estimador de De Montricher, Tapia e Thompson (Exemplo 2.4.1) é obtido maximizando a verossimilhança penalizada com $\phi(\theta) = \int_a^b \left\{\theta^{(s)}(t)\right\}^2 dt = \|\theta\|^2_{H_0^s(a,b)}$, onde $H_0^s(a, b)$ é definido no Apêndice (Proposição A-5).

O primeiro EMVP de Good e Gaskins (Exemplo 2.4) corresponde à função de penalidade

$$\phi_1(\theta) = \int_{-\infty}^{\infty} \frac{\theta'(t)}{\theta(t)}\, dt = 4\int_{-\infty}^{\infty} \left[\frac{d\theta^{1/2}}{dt}\right]^2 dt, \text{ com } \theta^{1/2} \in H^1(-\infty, \infty)$$

mostraremos na seguinte seção que penalizar utilizando $\phi_1(\theta)$ é equivalente a penalizar utilizando $\phi(\theta) = 2\int_{-\infty}^{\infty} \theta'(t)^2 dt$, $\theta \in H^1(-\infty, \infty)$, que por sua vez é equivalente a penalizar com

$$\phi_\alpha(\theta) = \int_{-\infty}^{\infty} \theta'(t)^2 dt + \alpha\int_{-\infty}^{\infty} \theta(t)\, dt = \|\theta\|^2_\alpha,$$

onde $\|\cdot\|_\alpha$ define uma norma equivalente à norma original em $H^1(-\infty, \infty)$.

De forma análoga o segundo EMVP de Good e Gaskins (Exemplo 2.4.1) considera o funcional $\phi : H^2(-\infty, \infty) \longrightarrow \mathbb{R}$ definido por

$$\phi(\theta) = \frac{1}{2}\left[\alpha \int_{-\infty}^{\infty} \theta'(t)^2 dt + \beta \int_{-\infty}^{\infty} \theta''(t)^2 dt\right],$$

para algum $\alpha \geq 0$ e $\beta > 0$, e é mostrado que o mesmo estimador é obtido utilizando a penalidade

$$\phi_\gamma(\theta) = \phi(\theta) + \gamma \int_{-\infty}^{\infty} \theta(t)^2 dt = \|\theta\|_\gamma^2,$$

onde $\|\cdot\|_\gamma$ define uma norma equivalente à norma em $H^2(-\infty, \infty)$.

Também, Zucker e Karr (1990) aplicaram o método de MVP à estimação da função de regressão num modelo generalizado de Cox para a função de risco, utilizando uma penalidade $\phi(\theta) = \int_0^1 \left\{\theta^{(s)}(t)\right\}^2 dt = \|\theta\|^2_{H_0^s(0,1)}$, onde $H_0^s(0,1)$ é definido no Apéndice (Proposição A-5).

Antoniadis e Gregoire (1990), utilizaram na estimação da função de risco a partir de dados de sobrevivência censurados,

$$\phi(\theta) = \int_0^T \left\{\theta'(t)\right\}^2 dt = \|\theta\|^2_{H_0^s(0,\ T)}$$

Bartoszynski (1981) estimou a função intensidade de um processo de Poisson não homogêneo com uma penalidade

$$\phi(\theta) = \|\theta\|^2_{H_0^s(0,\ T)}.$$

Na seguinte seção mostraremos a existência de solução única para o problema (2.2) para os casos de estimação de densidade e estimação da intensidade de um processo de Poisson não homogêneo.

## 2.4 - EXEMPLOS DE APLICAÇÃO

### 2.4.1 - ESTIMAÇÃO DE DENSIDADE

Como já vimos no Capítulo 1 o EMV de uma função de densidade $\theta$, baseado numa amostra $x_1, \ldots, x_n$, não existe no caso não paramétrico.

Por isso a necessidade de resolver o seguinte problema de otimização; para $\lambda > 0$

$$\begin{cases} \text{minimizar } J_{n,\lambda}(\theta) = J_n(\theta) + \lambda\phi(\theta), \text{ sujeito a} \\ \theta \in \Theta \ , \ \Theta \subset H \end{cases} \tag{2.3}$$

onde

$$\Theta = \left\{ \theta \in H, \int_\Omega \theta(t)dt = 1 \text{ e } \theta(t) \geq 0 \quad \forall\, t \in \Omega \right\}$$

$H \subset L^1(\Omega)$, $H$ de dimensão infinita

$$J_n(\theta) = - \sum_{i=1}^{n} \log \theta(x_i)$$

Para o problema (2.3) fazer sentido, para $x_1, \ldots, x_n$ em $\Omega$ deve existir no mínimo um $\theta \in H$ tal que:

$$\int_\Omega \theta(t)dt = 1, \ \theta(t) \geq 0 \quad \forall\, t \in \Omega \quad \text{e} \quad \theta(x_i) > 0, \ i = 1, \ldots, n$$

Praticamente todos os métodos de estimação de densidade têm a propriedade de que o estimador no limite quando a quantidade de suavização decresce é uma soma de picos nas observações, mas o que acontece quando a quantidade de suavização cresce depende exatamente de qual método está sendo usado.

Entre as funções $\phi(\theta)$ consideradas na literatura aparecem:

$$\phi_1(\theta) = \int \frac{\{\theta'(t)\}^2}{\theta(t)}\, dt$$

$$\phi_2(\theta) = \alpha_1 \int \left\{\theta'(t)\right\}^2 dt + \alpha_2 \int \left\{\theta''(t)\right\}^2 dt$$

$$\phi_3(\theta) = \int \left\{ \left( \frac{d}{dt} \right)^3 \log \theta(t) \right\}^2 dt$$

$$\phi_4(\theta) = \int \left\{ \nabla \log \theta(t) \right\}^2 dt,$$

onde $\nabla$ é algum operador diferenciável.

As funções $\phi_1$ e $\phi_2$ foram introduzidas pór Good e Gaskins (1971) e a existência e unicidade do estimador $\hat{\theta}_\lambda$ foi provada por De Montricher, Tapia e Thompson (1975).

Klonias (1982) derivou propriedades de consistência para $\hat{\theta}_\lambda$ quando $\lambda \longrightarrow 0$ utilizando $\phi_1$. As funções $\phi_3$ e $\phi_4$ foram discutidas por Silverman (1982) que provou existência, consistência e normalidade assintótica de $\theta_\lambda$. Os resultados de consistência obtidos por estes autores resultam casos particulares de um resultado geral sobre consistência do EMVP que provaremos no Capítulo 4.

Os EMVP's para densidades com uma penalidade apropriada têm uma propriedade muito atrativa. A ultrasuavização ($\lambda \longrightarrow \infty$) tem características interessantes.

Com $\phi_3$, o estimador no limite quando $\lambda \longrightarrow \infty$ é a densidade normal com média e variância dadas pelos estimadores usuais de máxima verossimilhança a partir dos dados (Silverman (1982)).

Com $\phi_4$, o estimador no limite é um membro de certa família exponencial, dependendo dos dados (Silverman (1982)).

Estas observações dão uma forte justificativa para o uso do método de penalidade com uma penalidade $\phi(\theta)$ apropriada.

Como um dos objetivos dos métodos não paramétricos é investigar o efeito de relaxar suposições paramétricas, é interessante que o caso limite de um estimador de densidade não paramétrico seja um modelo paramétrico.

Estas observações fornecem indicações para a eleição do funcional de penalidade. Esta eleição tem sido feita em forma *ad hoc*, quer por razões de conveniência matemática quer por razões computacionais. É possível definir outras funções de penalidade segundo outras percepções de famílias de densidades exponenciais "*infinitamente suaves*". A propriedade fundamental, já que $\phi(\theta)$ com medida da aspereza da curva pode ser interpretada também como a distância entre o verdadeiro parâmetro $\theta$ e o estimador infinitamente suave, é que $\phi(\theta)$ deve ser zero se e somente se $\theta$ pertence à família desejada. Por exemplo, $\phi(\theta) = \int_0^\infty \left\{ \left( \log \theta'' \right) \right\}^2 dt$ produz densidades exponenciais no caso limite ($\lambda \longrightarrow \infty$) (Titterington (1985)).

As funções de densidade de probabilidade são por definição não negativas. Para cumprir com esta restrição devemos trabalhar com a avaliação pontual de funcionais. Como queremos que a coleção de funções de interesse seja um conjunto fechado, precisamos que a avaliação pontual seja uma operação contínua. Os espaços de Hilbert com Kernel reprodutivo (RKHS) são espaços de funções ricos ou grandes o suficiente (de dimensão infinita) caracterizados pela propriedade de que a avaliação pontual é operação contínua (ver Apêndice).

**PROPOSIÇÃO 2.1:** Se H for um RKHS, $\Theta \subset \left\{\theta \in H : \theta(x_i) \geq 0\right\}$ for fechado e convexo e tal que contenha no mínimo um $\theta$ tal que $\theta(x_i) > 0$, $i = 1, \ldots, n$, então o problema (2.3) com $\phi(\theta) = \|\theta\|^2$ tem solução única.

**PROVA:** Como H é RKHS, a avaliação pontual é operação contínua, e como a norma também é funcional contínuo então $J_{n,\lambda}(\theta)$ é funcional contínuo.

Um cálculo direto mostra que a segunda variação de Gâteaux de $J_{n,\lambda}$ é uniformemente definida positiva relativa a $\Theta$,

$$J''_{n,\lambda}(\theta)\ (\eta,\ \eta) = \sum_{i=1}^{n} \frac{\eta(x_i)^2}{\theta(x_i)^2} + 2\lambda \langle\eta,\ \eta\rangle \geq 2\lambda\ \|\eta\|^2,$$

para cada $\theta \in \Theta$, $\forall\ \eta \in T(\theta)$.

Então pelo Teorema 2.4 $J_{n,\lambda}$ tem um único minimizador em $\Theta$.

**TEOREMA 2.5:** Seja H um RKHS tal que a integração sobre $\Omega$ é um funcional contínuo e suponhamos que exista no mínimo um $\theta \in H$ tal que

$$\int_{\Omega} \theta(t)dt = 1, \ \theta(t) \geq 0 \quad \forall\ t \in \Omega \ \text{ e } \ \theta(x_i) > 0, \ i = 1, \dots, n.$$

Então o EMVP correspondente a H(i.e. com $\phi(\theta) = \|\theta\|^2$) existe e é único.

**PROVA:** Segue da Proposição 2.1 pois $\Theta$ definido em (2.3) é um subconjunto fechado e convexo de $\left\{\theta \in H : \theta(x_i) \geq 0, \ i = 1, \dots, n\right\}$.

## O ESTIMADOR DE DE MONTRICHER, TAPIA E THOMPSON

**TEOREMA 2.6:** O EMVP correspondente a $H_0^s(a, b)$ existe e é único.

**PROVA:** $H_0^s(a, b)$ é RKHS (Proposição A-5), $Q(\theta) = \int_a^b \theta(t)dt$ é contínuo em $H_0^s(a, b)$ (Proposição A-6), e obviamente existem $\theta \in H_0^s(a, b)$ que satisfazem $\int_{\Omega} \theta(t)dt = 1$, $\theta(t) \geq 0$ $\forall t \in \Omega$ e $\theta(x_i) > 0$, $i = 1, \dots, n$.

Então pelo Teorema 2.5, o EMVP correspondente a H existe e é único; seja $\theta_*$. Tapia e Thompson (1978) mostraram que a solução $\theta_*$ é um spline polinomial (monospline) de grau 2s.

### O PRIMEIRO EMVP DE GOOD E GASKINS

Good e Gaskins (1971) propõem para o mesmo problema a função de penalidade

$$\phi_1(\theta) = \int_{-\infty}^{\infty} \frac{\theta'(t)}{\theta(t)}\, dt$$

que também pode ser escrita como

$$\phi_1(\theta) = 4 \int_{-\infty}^{\infty} \left( \frac{d\theta^{1/2}}{dt} \right)^2 dt$$

Aqui $H = H^1(-\infty, \infty)$ e $\theta^{1/2} \in H^1(-\infty, \infty)$.

Eles consideram que a penalidade pode ser tomada proporcional à informação de Fisher relativa à locação considerada como um parâmetro.

Esta penalidade pode ser vista como uma medida da facilidade em detectar pequenas mudanças em $\theta$, ou também como uma medida da variabilidade de $\theta^{1/2}$. Observe que neste caso o funcional integração não será contínuo em $H^1(-\infty, \infty)$.

O EMVP correspondente à função de penalidade $\phi_1$ será a solução do seguinte problema:

$$\begin{cases} \text{minimizar } J^{(1)}_{n,\lambda}(\theta) = -\sum_{i=1}^{n} \log\, \theta(x_i) + \lambda\phi_1(\theta), \text{ sujeito a} \\ \theta^{1/2} \in H^1(-\infty, \infty), \displaystyle\int_{-\infty}^{\infty} \theta(t)dt = 1 \text{ e } \theta(t) \geq 0 \quad \forall\, t \in (-\infty, \infty) \end{cases} \tag{2.4}$$

Para evitar a restrição $\theta(t) \geq 0$, Good e Gaskins sugerem trabalhar com $\theta^{1/2}$ no lugar de $\theta$(a restrição de não negatividade é em geral impossível de se manter quando trabalhamos com algoritmos que tratam com densidades contínuas, não lineares por partes).

Seja $u = \theta^{1/2}$, então (2.4) fica:

$$\begin{cases} \text{minimizar } -\sum_{i=1}^{n} \log\, u(x_i)^2 + \lambda\, 4 \displaystyle\int_{-\infty}^{\infty} u'(t)^2 dt, \text{ sujeito a} \\ u \in H^1(-\infty, \infty) \text{ e } \displaystyle\int_{-\infty}^{\infty} u(t)^2 dt = 1 \end{cases} \tag{2.5}$$

Se $u^*$ é solução de (2.5), então $\theta^* = (u^*)^2$ é aceita como solução de (2.4).

Mas este artifício nem sempre pode ser usado. O Lema seguinte esclarece este ponto.

**LEMA 2.1:** Seja $H \subset L^2(\Omega)$ e J um funcional definido em H. Consideremos

$$\underline{\text{Problema I}} \begin{cases} \text{minimizar } J(\theta^{1/2}); \text{ sujeito a} \\ \theta^{1/2} \in H, \displaystyle\int_{\Omega} \theta(t)dt = 1 \text{ e } \theta(t) \geq 0 \quad \forall\, t \in \Omega \end{cases}$$

e

$$\text{Problema II}\begin{cases}\text{minimizar } J(u); \text{ sujeito a}\\ u \in H,\ \int_\Omega u^2(t)dt = 1\end{cases}$$

Seja $u^*$ solução do Problema II.Então $\theta^* = (u^*)^2$ é solução do Problema I se e somente se $|u^*| \in H$ e $J(u^*) = J(|u^*|)$.

**PROVA:** Seja $\theta^* = |u^*|^2$ e suponhamos que $|u^*| \notin H$ e $J(u^*) = J(|u^*|)$ então $J(\theta^{*1/2})$ não está definido e não faz sentido.

Agora se $|u^*| \in H$ e $J(u^*) = J(|u^*|)$ então $\forall\ \theta \geq 0\ /\ \theta^{1/2} \in H$ temos $J(\theta^{1/2}) \geq J(u^*) = J(|u^*|) = J(\theta^{*1/2})$, então $\theta^*$ é solução do problema I.

As condições do Lema valem quando $\theta^*$(solução de II) é não negativa. $H^1(-\infty, \infty)$ e $J_{n,\lambda}$ em (2.4) e (2.5) satisfazem as condições do Lema. Se $\theta^* \in H^s(-\infty, \infty)$ ou $H^s_0(a, b)$ então $|\theta^*|$ pertence ao mesmo espaço se e somente se $s \leq 1$. Portanto, o enfoque alternativo é válido no caso anterior com $H^1(-\infty, \infty)$ ou $H^1_0(a, b)$.

Mas o problema (2.5) pode não ter solução única. Note-se que se $u^*$ é uma solução, então $-u^*$ também é. Acrescentando a condição $u \geq 0$ e multiplicando por $\frac{1}{2}$ o funcional objetivo, temos o seguinte problema de otimização:

$$\begin{cases} \text{minimizar } J_{n,\lambda}(\theta) = -\sum_{i=1}^{n} \log \theta(x_i) + \lambda\, \phi(\theta), \text{ sujeito a} \\ \theta \in H^1(-\infty, \infty)\,, \displaystyle\int_{-\infty}^{\infty} \theta(t)^2 dt = 1 \text{ e } \theta(t) \geq 0 \quad \forall\, t \in (-\infty, \infty). \end{cases} \tag{2.6}$$

onde $\phi(\theta) = \frac{1}{2}\phi_1(\theta^2) = 2\int_{-\infty}^{\infty} \theta'(t)^2 dt.$

**PROPOSIÇÃO 2.2:**

(i) se $\theta$ é solução de (2.4), então $\theta^{1/2}$ é solução de (2.5) e (2.6).

(ii) se $u$ é solução de (2.5), então $|u|$ é solução de (2.6) e $u^2$ de (2.4).

(iii) se $\theta$ é solução de (2.6), então $\theta$ é solução de (2.5) e $\theta^2$ de (2.4).

**PROVA:** Segue do Lema 2.1 e do fato que $\theta \geq 0 \Rightarrow \phi(\theta^{1/2}) = \frac{1}{2}\ \phi_1(\theta)$ e $J^{(1)}_{n,\lambda}(\theta) = 2\ J_{n,\lambda}(\theta^{1/2})$.

**COROLÁRIO 2.1:** Se (2.6) tem solução única, então (2.4) tem solução única. Embora a solução de (2.5) possa não ser única, a raíz quadrada de uma solução será a única solução de (2.4).

Vejamos que (2.6) tem solução única. Junto com o problema (2.6) consideremos:

$$\begin{cases} \text{minimizar } J_{n,\lambda}(\theta), \text{ sujeito a} \\ \theta \in H^1(-\infty, \infty), \displaystyle\int_{-\infty}^{\infty} \theta(t)^2 dt = 1 \text{ e } \theta(x_i) \geq 0,\ i = 1, \dots, n \end{cases} \tag{2.7}$$

Dado $\lambda$ em (2.6) e $\alpha > 0$ consideremos o seguinte problema:

$$\begin{cases} \text{minimizar } J^{(\alpha)}_{n,\lambda}(\theta) = -\displaystyle\sum_{i=1}^{n} \log \theta(x_i) + \lambda\, \phi_\alpha(\theta), \text{ sujeito a} \\ \theta \in H^1(-\infty, \infty)\ , \displaystyle\int_{-\infty}^{\infty} \theta(t)^2 dt = 1 \text{ e } \theta(x_i) \geq 0\ ,\ i = 1, \dots, n \end{cases} \tag{2.8}$$

onde $\phi_\alpha(\theta) = 2 \displaystyle\int_{-\infty}^{\infty} \theta'(t)dt + \alpha \int_{-\infty}^{\infty} \theta(t)^2 dt.$

O estudo de (2.8) começa com o estudo do seguinte problema:

$$\begin{cases} \text{minimizar } J^{(\alpha)}_{n,\lambda}(\theta), \text{ sujeito a} \\ \theta \in H^1(-\infty, \infty) \text{ e } \theta(x_i) \geq 0,\ i = 1, \dots, n \end{cases} \tag{2.9}$$

**PROPOSIÇÃO 2.3:** (2.9) tem solução única. Se $\theta_\alpha$ é esta solução, então:

(i) $\theta_\alpha(t) > 0,\ \forall\ t \in (-\infty, \infty).$

(ii) $\|\theta_\alpha\|_{L^2} \geq \left(\frac{n}{4\alpha}\right)^{1/2}$.

**PROVA:** $H^1(-\infty, \infty)$ é um RHKS (Proposição A-7). Também $\|\theta\|_\alpha^2 = \phi_\alpha(\theta) = 2\int_{-\infty}^{\infty} \theta'(t)^2 dt + \alpha\int_{-\infty}^{\infty} \theta(t)^2 dt$ da uma norma equivalente à norma em $H^1(-\infty, \infty)$. Então a existência de $\theta_\alpha$ segue da Proposição 2.1 com $\Theta = \left\{\theta \in H^1(-\infty, \infty): \theta(x_i) \geq 0,\ i = 1, \ldots, n\right\}$.

O resto da prova pode ser encontrado em Tapia e Thompson (1978), De Montricher et al (1975) mostraram que esta solução é um spline exponencial com nós só nos pontos amostrais.

**PROPOSIÇÃO 2.4:** O problema (2.7) tem solução única.

**PROVA:** Seja $B = \left\{\theta \in H^1(-\infty, \infty) : \int_{-\infty}^{\infty} \theta(t)^2 dt \leq 1 \text{ e } \theta(x_i) \geq 0,\ i = 1, \ldots, n\right\}$. Claramente B é fechado e convexo.

Então pela Proposição 2.2, $J_{n,\lambda}^{(\alpha)}$ dado em (2.8) terá um único maximizador em B, $u_\alpha$. Agora por ii) da Proposição 2.3 se escolhermos $0 < \alpha < \frac{1}{4}$, $\theta_\alpha$ a única solução de (2.9) será tal que $\|\theta_\alpha\|_{L^2(-\infty,\infty)} > 1$. Mostraremos que para esta amplitude de $\alpha$, $\|\mu_\alpha\|_{L^2}^2 = 1$.

Consideremos $\theta_\gamma = \gamma\theta_\alpha + (1 - \gamma)\ u_\alpha$.

Seja $J^{(\alpha)} = J^{(\alpha)}_{n,\lambda}$ por simplicidade de notação.

Sabemos que $J^{(\alpha)}$ é funcional estritamente convexo e $J^{(\alpha)}(\theta_\alpha) \leq J^{(\alpha)}(u_\alpha)$, pois $\theta_\alpha$ é a única solução de (2.9), então para $0 < \gamma < 1$, como $J^{(\alpha)}$ é estritamente convexo

$$J^{(\alpha)}(\theta_\gamma) \leq \gamma\, J^{(\alpha)}(\theta_\alpha) + (1 - \gamma)\, J^{(\alpha)}(u_\alpha).$$

Portanto $J^{(\alpha)}(\theta_\gamma) \leq J^{(\alpha)}(u_\alpha)$.

Agora suponhamos que $\|u_\alpha\|_{L^2} < 1$ e consideremos

$$g(\gamma) = \|\theta_\gamma\|_{L^2}$$

Temos
$$g(0) = \|u_\alpha\|_{L^2} < 1$$

$$g(1) = \|\theta_\alpha\|_{L^2} > 1$$

então para algum $0 < \gamma_0 < 1$, $g(\gamma_0) = 1$ e

$$J^{(\alpha)}(u_\alpha) \geq J^{(\alpha)}\left(\theta_{\gamma_0}\right),$$

o que contradiz o fato de $u_\alpha$ ser o único minimizador de $J^{(\alpha)}_{n,\lambda}$ em B.

$$\text{Logo } \|u_\alpha\|_{L^2(-\infty,\infty)} = 1.$$

Isto mostra que $u_\alpha$ é a única solução de (2.8) para $0 < \alpha < \frac{1}{4}$. Mas como o termo $\alpha\int_{-\infty}^{\infty} \theta(t)^2 dt$ é constante igual a $\alpha$ sobre o conjunto de restrições em (2.7) e (2.8), então os problemas (2.7) e (2.8) têm as mesmas soluções para qualquer $\alpha > 0$.

Isto prova a Proposição pois foi mostrado que (2.8) tem única solução para ao menos um $\alpha$.

**PROPOSIÇÃO 2.5:** O problema (2.6) tem uma única solução positiva.

**PROVA:** Vejamos que $\tilde{\theta}$, a única solução de (2.7), tem estas propriedades.

$$\text{Seja } G(\theta) = \ell_{n,\lambda}(\theta) \quad \text{e} \quad g(\theta) = \lambda\int_{-\infty}^{\infty} \theta(t)^2 dt \ , \ \theta \in H^1(-\infty, \infty).$$

Claramente $\tilde{\theta}(x_i) > 0$, $i = 1, \ldots, n$.

Então, pela teoria de multiplicadores de Lagrange, $\exists\ \alpha \ /\ \tilde{\theta}$ satisfaz

$$G'(\theta) - \alpha\, g'(\theta) = 0 \quad \text{e} \quad g(\theta) = \lambda \tag{2.10}$$

usando gradientes em $L^2(-\infty, \infty)$ no sentido de distribuições (2.10) é equivalente a

$$4\lambda\ \theta' + 2\ \alpha\ \lambda\ \theta = \sum_{i=1}^{n} \frac{\delta_i}{\theta(x_i)} \quad \text{e} \quad g(\theta) = \lambda \tag{2.11}$$

onde $\delta_i$ é a distribuição tal que $\int_{-\infty}^{\infty} \theta(t)\ \delta_i(t) = \theta(x_i)$, $i = 1, \ldots, n$.

Como (2.7) tem solução única, então (2.11) deve ter solução única em $H^1(-\infty, \infty)$. Seja $\tilde{\theta}$.

Se $\alpha \leq 0$ então uma solução de (2.11) seria uma soma de funções trigonométricas e poderia possivelmente não satisfazer $g(\theta) = \lambda$, i.e. não pertencer a $L^2(-\infty, \infty)$.

Seja então $\alpha > 0$. Observemos que $G - \alpha g = J^{(\alpha)}_{n,\lambda}$, então se $\theta$ satisfaz (2.10) deve tembém ser solução de (2.9) para este $\alpha$, portanto tem as propriedades da Proposição 2.3, i.e. $\theta$ é um spline exponencial com nós nos pontos $x_1, \ldots, x_n$.

**PROPOSIÇÃO 2.6:** O primeiro EMVP não paramétrico de Good e Gaskins existe e é único.

É o estimador correspondente a

$$\phi(\theta) = \int_{-\infty}^{\infty} \frac{\theta'(t)^2}{\theta(t)}\, dt$$

e

$$\Theta = \left\{ \theta : \theta \geq 0 \quad e \quad \theta^{1/2} \in H^1(-\infty, \infty) \right\}.$$

É positivo, e é um spline exponencial com nós só nos pontos $x_1, \ldots, x_n$.

**PROVA:** Pela Proposição 2.5, (2.6) tem solução única $\theta$ e pela Proposição 2.2 $\theta^2$ é solução de (2.4).

## O SEGUNDO ESTIMADOR DE MÁXIMA VEROSSIMILHANÇA PENALIZADA DE GOOD E GASKINS

Ao aplicar a função de penalidade $\phi_1(\theta) = 4\int(\theta')^2 dt$, Good e Gaskins acharam que algumas vezes a curva de densidade estimada tinha grandes porções lineares. Isto não foi surprendente pois a curvatura depende também da 2ª derivada. Como já tinham sido desenvolvidos métodos para $\phi(\theta) = \int(\theta'')^2 dt$, Good e Gaskins sugeriram uma penalidade mais geral, da forma

$$\phi(\theta) = \gamma\int_{-\infty}^{\infty} \theta'(t)^2 dt + \delta\int_{-\infty}^{\infty} \theta''(t)^2 dt, \text{ para algum } \gamma \geq 0 \text{ e } \delta > 0$$

$$\phi : H^2(-\infty, \infty) \longrightarrow \mathbb{R} \tag{2.12}$$

<u>O segundo EMVP de Good e Gaskins</u> será então a solução de

$$\begin{cases} \text{minimizar } - \sum_{i=1}^{n} \log\, \theta(x_i) + \lambda\phi(\theta^{1/2}), \text{ sujeito a} \\ \theta^{1/2} \in H^2(-\infty, \infty)\,,\ \int_{-\infty}^{\infty} \theta(t)dt = 1 \text{ e } \theta(t) \geq 0 \quad \forall\ t \in (-\infty, \infty) \end{cases} \tag{2.13}$$

como no primeiro caso, para evitar a restrição de não negatividade, Good e Gaskins sugerem resolver o seguinte problema

$$
\begin{cases}
\text{minimizar } -\displaystyle\sum_{i=1}^{n} \log \theta(x_i)^2 + \lambda\phi(\theta), \text{ sujeito a} \\
\theta \in H^2(-\infty, \infty) \text{ e } \displaystyle\int_{-\infty}^{\infty} \theta(t)^2 dt = 1
\end{cases}
\tag{2.14}
$$

se $u \in H^2(-\infty, \infty)$ for positivo em alguns pontos e negativo em outros, então em geral $|u| \notin H^2(-\infty, \infty)$.

Pelo Lema 2.1 é possível obter a solução do problema (2.13) a partir de uma solução do problema (2.14) se e somente se estas são não negativas. Mas, é provado (De Montricher et al (1975)) que as soluções do problema (2.14) não são necessariamente não negativas.

Segue então, que o segundo EMVP de Good e Gaskins não pode ser obtido considerando o problema (2.14).

Se usarmos $\theta_*^2$, onde $\theta_*$ é uma solução do problema (2.14) como estimador da função de densidade que originou a amostra aleatória $x_1, \ldots, x_n$, então claramente $\theta_*^2$ será não negativa, sua integral será igual a um e por isso será uma densidade de probabilidade. Mas o estimador obtido desta forma não será no sentido estrito da nossa definição um EMVP. Será chamado o <u>pseudo-estimador de máxima verossimilhança penalizada de Good e Gaskins</u>.

Mostra-se (De Montricher et al (1975)) que o segundo EMVP de Good e Gaskins e o pseudo-EMVP de Good e Gaskins, existem, são únicos e são distintos.

## 2.4.2- ESTIMAÇÃO DA INTENSIDADE DE UM PROCESSO DE POISSON NÃO HOMOGÊNEO

Como foi mostrado em 1.2.2. é necessário a utilização de um método alternativo para a estimação de $\theta_0$.

O EMVP será aquele que minimize

$$J_{n,\lambda}(\theta) = - \sum_{k=1}^{\bar{N}_n(1)} \ln\left[\theta(\bar{\tau}_k^{(n)})\right] + n \int_0^1 \theta(t)dt + \lambda\phi(\theta)$$

sobre um subconjunto $\Theta$ convexo e fechado de um espaço de Hilbert de funções, real e separável $\Theta$.

Para evitar a restrição de não negatividade ($\theta \geq 0$) o funcional de penalidade será escolhido de tal forma que penalizar a intensidade sob restrições de não negatividade ou penalizar a raiz quadrada da intensidade e elevar a solução ao quadrado sejam problemas de otimização equivalentes.

Assumiremos que $u = \theta^{1/2}$, o parâmetro desconhecido, pertence a $\Theta$.

O EMVP de u é obtido minimizando sobre $\Theta$ um funcional da forma

$$J_{n,\lambda}(u) = - \sum_{k=1}^{\bar{N}_n(1)} \ln\left[u^2(\bar{\tau}_k^{(n)})\right] + n \int_0^1 u^2(t)dt + \lambda\phi(u),\ \lambda > 0 \qquad (2.15)$$

onde $\phi : \Theta \longrightarrow \mathbb{R}^+$ é um funcional de penalidade.

Consideremos as seguintes suposições baseadas em noções previas sobre o comportamento da intensidade desconhecida $\theta_0$ e que permitem que o problema de otimização seja tratável.

i) $H = H_0^1\left(\left[0,\ 1\right]\right)$

ii) O funcional de penalidade $\phi$ é definido por

$$\phi(u) = \int_0^1 \left\{u'(s)\right\}^2 ds = \|u\|^2_{H_0^1[0,1]}$$

iii) A *"entropia"*

$$H(u_0) = - \int_0^1 \left\{u_0^1(s)\ \ln\left[u_0^2(s)\right] - u_0^2(s)\right\} ds$$

é finita, e $u_0$, a raíz quadrada da verdadeira intensidade e limitada inferiomente.

$J_{n,\lambda}(u)$ é um funcional par e então o problema de otimização anterior pode possivelmente não ter solução única. Segundo os argumentos de De Montricher, Tapia e Thompson (1975) podemos acrescentar a restrição de não negatividade $u > 0$ ao problema anterior e concluir sem perda de generalidade que se existir um minimizador de (2.15), seu valor absoluto será uma solução do problema com restrições e seu quadrado será a solução do problema de minimização de

$$G_{n,\lambda}(\theta) = - \sum_{k=1}^{\bar{N}_n(1)} \ell n\ \theta\left(\bar{\tau}_k^{(n)}\right) + n \int_0^1 \theta(s)ds + \lambda\phi(\sqrt{\theta}),$$

sujeito a $\theta \geq 0$, $\theta^{1/2} \in H^1$.

**PROPOSIÇÃO 2.7:** O funcional $G_{n,\lambda}(u)$ tem um único minimizador em $\left\{u \in H : u \geq 0\right\}$.

**PROVA:** $\left\{u \in H : u \geq 0\right\}$ é subconjunto convexo e fechado de $H_0^1([0, 1])$ $G(u) = G_{n,\lambda}(u)$ é funcional contínuo em $\left\{u \in \Theta : u \geq 0\right\}$ pois a norma é funcional contínuo, o logaritmo é contínuo e como $H_0^1[0, T]$ é RKHS a avaliação pontual também é operador contínuo.

Um cálculo direto mostra que:

$$G''(u)\ (\eta,\ \eta) = 2\lambda \int_0^1 \left\{\eta'(s)\right\}^2 ds + 2\eta \int_0^1 \eta^2(s) ds +$$

$$+ 2 \sum_{k=1}^{\bar{N}_n(1)} \frac{\eta^2(\bar{\tau}_k^{(n)})}{u(\bar{\tau}_k^{(n)})} \geq 2\lambda \left\|\eta\right\|_H^2$$

Portanto $G(u)$ é uniformemente convexa em $\left\{u \in H : u \geq 0\right\}$, e pelo Teorema A-6, $G(u)$ tem um único minimizador em $\left\{u \in H : u \geq 0\right\}$.

Antoniadis e Gregorie (1990) caracterizam o estimador como um spline hiperbólico com nós nos pontos de salto de $\bar{N}_n$.

# CAPÍTULO 3

# O MÉTODO DE "SIEVES"

Neste capítulo descrevemos o método de *"sieves"* proposto por Grenander (1981). O método impõe restrições no espaço paramétrico. Assim o EMV existirá em cada subconjunto de uma seqüência crescente de subconjuntos do espaço paramétrico.

Resultados sobre existência e consistência dos estimadores obtidos por este método serão apresentados.

## 3.1 - DESCRIÇÃO DO MÉTODO

A idéia básica consiste em restringir a estimação a subconjuntos do espaço paramétrico. O processo de otimização (maximização da função de verossimilhança no nosso caso) é repetido em cada subconjunto de uma seqüência crescente de subconjuntos de $\Theta$, indexados pelo tamanho amostral, obtendo-se então uma seqüência de estimadores.

A seqüência particular $\left\{S_m\right\}$ de subconjuntos de $\Theta$ é chamada um *"sieve"* e o processo de estimação resultante é o *"método de sieves"*. O *"sieve"* deve ser escolhido da seguinte maneira:

i) Para cada subconjunto $S_m$ do espaço paramétrico $\Theta$ deve existir um estimador de máxima verossimilhança.

ii) $S_m \subseteq S_{m+1}$, $\forall$ m = 1, 2, ....

iii) $\bigcup_{m=1}^{\infty} S_m$ deve ser denso em $\Theta$.

*"Sieves"* diferentes produzem diferentes estimadores.

Na prática, o *"sieve"* precisa ser cuidadosamente escolhido para explorar as propriedades estruturais específicas do problema.

Seria desejável saber como construir o *"sieve"* para produzir uma taxa de convergência ótima para a seqüência de estimadores de *"sieves"*, mas esta questão só foi estudada em alguns casos particulares.

Geman e Hwang (1982) estabeleceram alguns resultados sobre a consistência dos estimadores obtidos por este método. Antoniadis (1988) aplicou o método de *"sieves"* ao problema de estimação da média de um processo gaussiano com valores num espaço de Banach real separável. Utilizou uma seqüência crescente de *"sieves"* natural em termos das propriedades geométricas do espaço de parâmetros.

O método também tem sido aplicado a processos estocásticos para a estimação de funções dependentes do tempo como a média de um processo de Wiener (Grenander (1981), Geman e Hwang (1982)), o coeficiente *"drift"* de um processo de difusão linear (Nguyem e Pham (1982)), e a função de risco no modelo de intensidade multiplicativo para processos pontuais (Karr (1982)).

Outras aplicações podem ser encontradas em Mkeague (1986) e Kallianpur e Selukar (1990).

Os detalhes e a versatilidade do método são melhor ilustrados nos seguintes exemplos.

**EXEMPLO 3.1: ("Sieve de Histograma").**

Seja $X_1, \ldots, X_n$ uma amostra i.i.d de uma distribuição absolutamente contínua com função de densidade de probabilidade desconhecida $\theta_0(x)$.

O *"sieve"* de histograma para o problema de estimação de $\theta_0(x)$ é definido por

$$S_m = \left\{ \begin{array}{l} \theta : \theta \text{ é uma função de densidade de probabilidade,} \\ \text{constante em } \left[\frac{k-1}{m}, \frac{k}{m}\right), \; k = 0, \pm 1, \pm 2, \ldots \end{array} \right\},$$

e tal que $m = m_n$ cresça com o tamanho amostral.

$\left\{S_m\right\}$ constitui um *"sieve"* e o estimador do método de *"sieves"* será a solução do problema:

$$\left\{ \begin{array}{l} \text{maximizar } \prod_{i=1}^{n} \theta(x_i), \text{ sujeito a} \\ \theta \in S_m \end{array} \right. \tag{3.1}$$

A solução deste problema é a função

$$\hat{\theta}_{n,m}(x) = \frac{m}{n} \# \left\{ x_i : \frac{k-1}{m} \le x_i < \frac{k}{m} \right\} \text{ para } x \in \left[\frac{k-1}{m}, \frac{k}{m}\right)$$

i.e. o histograma com comprimento de intervalo $1/m_n$.

Pode ser provado (Geman (1982)) que se $m_n \uparrow \infty$ suficientemente lento, então $\hat{\theta}_{n,m_n} = \hat{\theta}_n$ é consistente, no sentido que

$$\int_{-\infty}^{\infty} |\hat{\theta}_n(x) - \theta_0(x)|\,dx \xrightarrow[n \to \infty]{} 0 \qquad \text{quase certamente.}$$

**EXEMPLO 3.2: ("Sieve de convolução")**

Para o mesmo problema do exemplo 3.1 um *"sieve"* diferente e mais interessante é o *"sieve de convolução"*:

$$S_m = \left\{ \theta : \theta(x) = \int \frac{m}{\sqrt{2\pi}} e^{-\frac{m^2}{2}(x-y)^2} F(dy),\ \text{F função de distribuição arbitrária} \right\}$$

O estimador do método de sieves será a solução do problema (3.1) para este *"sieve"* $\left\{S_m\right\}$ particular.

Pode ser mostrado (Geman (1982)) que $\hat{\theta}$ tem a forma

$$\hat{\theta}_{n,m}(x) = \sum_{i=1}^{n} p_i \frac{m_n}{\sqrt{2\pi}} \exp\left\{-\frac{m_n^2}{2}(x-y_i)^2\right\}$$

para alguns $y_1, \ldots, y_n$ e $p_1, \ldots, p_n$ satisfazendo $p_i \geq 0$, $1 \leq i \leq n$, $\sum_{i=1}^{n} p_i = 1$. Pode-se mostrar que $\left\{y_1, \ldots, y_n\right\} \neq \left\{x_1, \ldots, x_n\right\}$ com probabilidade um.

Observe que o estimador de kernel com kernel gaussiano $\beta(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{m}{\sqrt{2\pi}} \exp\left\{- \frac{m^2}{2} (x - x_i)^2\right\}$ está em $S_m$ mas $\beta$ não é solução de máxima verossimilhança em $S_m$.

Prova-se (Geman (1982)) que se $m_n \uparrow \infty$ suficientemente lento então a seqüência de estimadores $\hat{\theta}_{n,m} = \hat{\theta}_n$ é consistente.

**EXEMPLO 3.3: (Splines para regressão não paramétrica)**

Sejam X e Y variáveis aleatórias e seja $(x_1, y_1), \ldots, (x_n, y_n)$ uma amostra i.i.d. da distribuição bivariada de (X, Y).

O estimador de mínimos quadrados da função de regressão $E(Y/X = x)$ minimiza

$$\sum_{i=1}^{n} \left\{y_i - \theta(x_i)\right\}^2$$

O mínimo é zero e é atingido por qualquer função que interpole os pontos amostrais $(x_1, y_1), \ldots, (x_n, y_n)$.

Exceto em alguns casos muito especiais este conjunto não converge em nenhum sentido útil à verdadeira regressão.

Então, o método de "*sieves*" realiza a otimização (neste caso, minimização da soma dos quadrados dos erros) dentro dos subconjuntos

$$S_m = \left\{ \theta : \theta \text{ é absolutamente contínua e } \int \left| \frac{d}{dx} \theta(x) \right|^2 \leq m \right\}$$

A solução $\hat{\theta}_{n,m}$ é determinada de forma única e é um spline polinomial de primeiro grau (i.e. contínuo e linear por partes, com discontinuidades em $\frac{d\hat{\theta}_{n,m}}{dx}$ em $x_1, \dots, x_n$).

É possível provar que se $m_n$ cresce suficientemente lento então a seqüência de estimadores $\hat{\theta}_{n,m}$ é consistente para $E(Y/X = x)$ numa métrica apropriada.

Especificamente,

Se $m_n \longrightarrow \infty$ de tal forma que $m_n = O(n^{1/4-\varepsilon})$ para algum $\varepsilon > 0$,

então

$$\sup_{\hat{\theta}_{m_n} \text{ EMV em } S_{m_n}} \int \left| \hat{\theta}_{m_n}(x) - \theta_0(x) \right|^2 F(dx) \longrightarrow 0 \quad \text{quase certamente.}$$

Outro *"sieve"* para o problema de regressão é

$$S_m = \left\{ \begin{array}{l} \theta : \theta \text{ tem p-1 derivadas absolutamente contínuas} \\ \text{e} \int \left| \frac{d^p}{dx^p} \theta(x) \right|^2 dx \le m \end{array} \right\}.$$

Neste caso a solução $\hat{\theta}_{n,m}$ é um spline polinomial de grau 2p-1.

Estes dois *"sieves"* considerados para o problema de regressão não paramétrica são exemplos de um tipo usual de *"sieve"* da forma $S_m = \left\{ \theta \in \Theta : \phi(\theta) \le m \right\}$, onde $\phi : \Theta \longrightarrow \mathbb{R}$ é algum funcional, geralmente contínuo e convexo.

*"Sieves"* deste tipo têm sido usados também para a estimação da função média de um processo gaussiano. Consideremos o problema de estimar $\theta(t)$, $t \in [0, 1]$ a partir de uma amostra de tamanho n de um processo $X(t) = \int_0^t \theta(s)\, ds + W(t)$, $t \in [0, 1]$, onde $W(t)$, $t \geq 0$ é um processo de Wiener com variância um por unidade de tempo, onde $\theta \in L^2[0,1]$ .

Um *"sieve"* proposto para este problema é dado por

$$S_m = \left\{ \theta \in L^2[0, 1] : \sum_{r=1}^{\infty} a_r^{\,2} < \theta, \phi_r >^2 \leq m \right\}$$

onde $< \cdot, \cdot >$ denota o produto interno em $L_2[0, 1]$ , $\left\{\phi_r, \ r \geq 1\right\}$ é um sistema ortonormal completo em $L_2[0, 1]$ e $\sum_{r\geq 1} a_r^{-2} < \infty$.

Este *"sieve"* foi estudado por Geman e Hwang (1982). Para o mesmo problema Grenander (1981) propos o *"sieve"* de Sturm-Liouville

$$S_m = \left\{ \theta \in \Theta : \int_0^1 \left\{ p(t) \left[\theta'(t)\right]^2 + q(t)\ \theta^2(t)\right\} dt \leq m \right\}$$

onde p e q são positivas, com primeira derivada contínua.

Chow e Grenander (1985) consideraram a estimação da densidade espectral de um processo Gaussiano estacionário $\left\{X(t),\ t = 1, 2, \ldots\right\}$ com média zero e covariância $r_t = E\left(X(s)\ X(s + t)\right) = \int_{-\pi}^{\pi} e^{it\lambda} \ f(\lambda)d\lambda$, onde f é a densidade espectral.

O *"sieve"* utilizado foi da forma

$$S_m = \left\{ f \in L^1\left[-\pi, \pi\right],\ f = \frac{1}{g}\ e \int_{-\pi}^{\pi} \left(\frac{d}{d\lambda}\, g\right)^2 d\lambda \le m \right\}.$$

Várias técnicas de estimação não paramétrica, como por exemplo os métodos de séries ortogonais, "*kernels*", "*splines*", máxima verossimilhança penalizada, etc., estão relacionados com o método de sieves.

## 3.2 - EXISTÊNCIA E CONSISTÊNCIA DOS ESTIMADORES

Embora o método de "*sieves*" produza uma variedade de estimadores segundo o "*sieve*" utilizado, certas propriedades (existência e consistência, no mínimo) podem ser provadas desde uma ótica unificada, sob a condição que o "*sieve*" cresça suficientemente lento com o tamanho amostral.

Em continuação, apresentaremos os principais resultados a este respeito obtidos por Geman e Hwang (1982), também apresentados por Grenander (1981).

Primeiramente descrevemos a notação e suposições básicas para a exposição dos teoremas.

Dado H espaço métrico, $\Theta \subset H$, um experimento estatístico abstrato $\left\{\Omega,\ \mathcal{F},\ \left\{P_\theta : \theta \in \Theta\right\}\right\}$ dominado por uma medida $\sigma$-finita $\mu$ e $p(x, \theta)$ a derivada de Radon-Nikodym $\dfrac{dP_\theta(x)}{d\mu}$,

$$H(\theta_0, \theta) = E_{\theta_0}\left[\ln p(X, \theta)\right], \quad \theta \in \Theta$$

é a entropia formal da distribuição de probabilidade $P_{\theta_0}$, onde $\theta_0$ designa o verdadeiro valor do parâmetro.

OBS: Para uma amostra $X_1, X_2, \ldots, X_n$ i.i.d. de $P_{\theta_0}$, seja

$$Y_i = \ln\left[p(X_i, \theta)\right]$$

Temos, então que

$$\bar{Y} = \frac{1}{n} \ln L_n(\underset{\sim}{X}, \theta) = \frac{1}{n} \sum_{i=1}^{n} \ln p(X_i, \theta)$$

onde $L_n$ é a função de verossimilhança.

Como $E(Y_i) = H(\theta_0, \theta)$ ,então pela lei forte dos grandes números

$$\frac{1}{n} \sum_{i=1}^{n} \ln p(X_i, \theta) \xrightarrow[n \to \infty]{} H(\theta_0, \theta), \text{ quase certamente} \tag{3.2}$$

Temos também que

$$H(\theta_0, \theta) < H(\theta_0, \theta_0)$$

(ver Rényi (1970)).

Para $\theta \in S_m$, $B_m(\theta, \varepsilon) = \left\{\beta : \beta \in S_m, \ d(\theta, \beta) < \varepsilon\right\}$ onde

$d(\cdot,\cdot)$ é uma métrica de $\Theta$.

Para $g : \Theta \longrightarrow \mathbb{R}$, $\forall\ B \subset \Theta$, seja

$$g[\,B\,] = \sup_{\beta \in B} g(\beta)$$

$$\mathbb{M}^n_m = \left\{\theta \in S_m : L_n(\underset{\sim}{x}, \theta) = L_n\left[\underset{\sim}{x}, S_m\right]\right\}$$

onde $L_n(\underset{\sim}{x}, \theta) = \prod_{i=1}^{n} \dfrac{dP_\theta}{d\mu}(x_i)$ e $\underset{\sim}{x} = (x_1, \ldots, x_n)$ representa a amostra.

I.e. $\mathbb{M}^n_m$ é o conjunto de estimadores de máxima verossimilhança em $S_m$ dado o tamanho amostral n.

$$\mathbb{A}_m = \left\{\theta : \theta \in S_m,\ H(\theta_0, \theta) = H\left[\theta_0, S_m\right]\right\}$$

representa o conjunto de entropia máxima em $S_m$.

Para $C_m \subseteq \Theta$, $C_m \longrightarrow \left\{\theta\right\}$ quando $m \longrightarrow \infty$ se $\sup_{\beta \in C_m} d(\theta, \beta) \longrightarrow 0$ quando $m \longrightarrow \infty$. Para $D \subseteq \Theta$, $C_m \longrightarrow D$ quando $m \longrightarrow \infty$ se $\sup_{\theta \in C_m} \inf_{\beta \in D} d(\theta, \beta) \longrightarrow 0$ quando $m \longrightarrow \infty$.

O seguinte Teorema trata da existência de uma seqüência $m_n$ para a qual o conjunto de máxima verossimilhança $\mathbb{M}^n_{m_n}$ é consistente.

**TEOREMA 3.1:** Suponhamos que um *"sieve"* $\left\{S_m\right\}$ seja escolhido de tal forma que:

1) Para todo $m \in \mathbb{N}$, todo $\theta \in S_m$ e todo $\varepsilon > 0$,

$$p\left[x,\ B_m(\theta,\ \varepsilon)\right] \text{ é } \mathcal{F}\text{-mensurável;}$$

e, para todo $m \in \mathbb{N}$ e todo $\theta \in S_m$, exista $\varepsilon > 0$ tal que

$$E_{\theta_0}\left[\ln p\left[x,\ B_m(\theta,\ \varepsilon)\right]\right] < \infty$$

2) $p(x,\ \cdot)$ é semicontinua superior em $\theta$, para $\theta \in S_m$,

i.e. $\lim_{\varepsilon \downarrow 0} p\left[x,\ B_m(\theta,\ \varepsilon)\right] = p(x,\ \theta)$ $\mu$-quase certamente.

3) $S_m$ é compacto para todo $m \in \mathbb{N}$.

4) $\mathbb{A}_m \longrightarrow \theta_0$ quando $m \longrightarrow \infty$.

Então, para todos $n,\ m \in \mathbb{N}$ e quase todo $\underset{\sim}{x}$, $\mathbb{M}^n_m(\underset{\sim}{x})$ é não vazio e para uma seqüência $m_n$ crescendo suficientemente lenta , $\mathbb{M}^n_{m_n} \longrightarrow \theta_0$ P-quase certamente.

**OBS:** Como, em geral $H(\theta_0,\ \theta_m) \longrightarrow H(\theta_0,\ \theta_0)$ implica $\theta_m \longrightarrow \theta_0$, quando $m \longrightarrow \infty$, se o *"sieve"* $\left\{S_m\right\}$ for escolhido de tal forma que exista $\theta_m \in S_m$ com $H(\theta_0,\ \theta_m) \longrightarrow H(\theta_0,\ \theta_0)$, então $\mathbb{A}_m \longrightarrow \theta_0$. A condição 4 garante, então que $\bigcup_m S_m$ é denso em $\Theta$.

**PROVA:** (ver Geman e Hwang (1982)).

Este Teorema da condições para um sieve ser suficientemente regular para produzir estimadores consistentes. Mas não diz nada sobre a velocidade com a qual $m_n$ deve tender ao infinito.

O método não tem aplicação se não pudermos identificar uma seqüência $m_n$ que garanta a consistência qualquer que seja o parâmetro.

O seguinte Teorema fornece um meio para a identificação de uma tal seqüência $m_n$.

**TEOREMA 3.2:** Seja H espaço métrico, $\Theta \subset H$, e suponhamos que um *"sieve"* $\left\{S_m\right\}$ seja escolhido de tal forma que:

1) Para todos $m, n \in \mathbb{N}$, $M_m^n$ é não vazio P-q.c.

2) a) Se para alguma seqüência $\theta_m \in S_m$

$$H(\theta_0, \theta_m) \longrightarrow H(\theta_0, \theta_0),$$

então $\theta_m \longrightarrow \theta_0$, quando $m \longrightarrow \infty$.

b) existe uma seqüência $\theta_m \in S_m$ tal que

$$H(\theta_0, \theta_m) \longrightarrow H(\theta_0, \theta_0).$$

Para cada $\delta > 0$ e cada m, seja

$$D_m = \left\{\theta \in S_m : H(\theta_0, \theta) \leq H(\theta_0, \theta_m) - \delta\right\}$$

onde $\left\{\theta_m\right\}$ é a seqüência de 2) b).

Dados $\ell$ conjuntos $O_1, \ldots, O_\ell$ em $S_m$ tais que $p\left[\ \cdot, O_k\right]$ seja $\mathcal{F}$-mensurável para cada k, seja

$$\rho_m = \sup_k \inf_{t \geq 0} E_{\theta_0}\left[\exp\left\{ t \ln \left[ \frac{p\left[x, O_k\right]}{p(x, \theta_m)}\right]\right\}\right]$$

considere a seqüência $\left\{m_m\right\}$ : $m_n \longrightarrow \infty$ e suponha que para cada $\delta > 0$ seja possível encontrar $O_1^m, \ldots, O_{\ell_m}^m$, $m \in \mathbb{N}$ tais que:

i) $D_m \subseteq \bigcup_{k=1}^{\ell_m} O_k^m$;

ii) $p\left[\,\cdot, O_k^m\right]$ é $\mathcal{F}$-mensurável;

iii) $\sum_{n=1}^{\infty} \ell m_n \left(\rho_{m_n}\right)^n < \infty$

Então, $\mathbb{M}_{m_n}^n \longrightarrow \theta_0$ P-quase certamente.

**OBS:** A métrica d a ser escolhida para o espaço H e em termos da qual é expressa a consistência dos estimadores, geralmente é sugerida pela condição 2.

Ela deve ser tal que se $H(\theta_0, \theta_m) \longrightarrow H(\theta_0, \theta_0)$ então $d(\theta_m, \theta_0) \longrightarrow 0$.

**PROVA:** (ver Geman e Hwang (1982)).

A aplicação do Teorema para a determinação da seqüência $m_n$ não é simples. Gomes (1991) estudou diferentes seqüências $m_n$ para exemplos simulados utilizando os "sieves" de histograma e de Fourier para a estimação da função intensidade de um processo de Poisson não homogêneo.

Note que no Teorema 3.1 a única suposição feita sobre o espaço H é que ele seja um espaço métrico com métrica d, em termos da qual será expressa a consistência dos estimadores.

Em continuação, apresentaremos uma nova versão deste Teorema necessaria para os resultados do Capítulo 4.

Suposições adicionais serão impostas sobre o espaço H e pediremos também semicontinuidade fraca superior das funções $H(\theta_0, \theta)$ e $L_n(\underset{\sim}{x}, \theta)$, e que os $S_m$ sejam limitados.

**TEOREMA 3.3:** Seja H um espaço de Banach reflexivo, $\Theta \subset H$, e suponhamos que um *"sieve"* $\left\{S_m\right\}$ verifique as seguintes condições:

1) Para todo $m \in \mathbb{N}$, todo $\theta \in S_m$, todo $\varepsilon > 0$ e todo subconjunto finito $\left\{\varphi_1, \ldots, \varphi_k\right\} \subset H^*$

$$p\left[x, U^m\varphi_1, \ldots, \varphi_k(\theta, \varepsilon)\right] \text{ é } \mathcal{F}\text{-mensurável,}$$

onde

$$U^m\varphi_1, \ldots, \varphi_k(\theta, \varepsilon) = \left\{\theta' \in S_m : |\varphi_i(\theta') - \varphi_i(\theta)| < \varepsilon,\ i = 1, \ldots, k\right\}$$

é uma vizinhança fraca do ponto $\theta \in S_m$;

Para todo $m \in \mathbb{N}$, todo $\theta \in S_m$ e toda $U^m\varphi_1, \ldots, \varphi_k(\theta, \varepsilon)$

$$\lim_{\varepsilon \to 0} p\left[x, U^m\varphi_1, \ldots, \varphi_k(\theta, \varepsilon)\right] = p(x, \theta)$$

i.e. $p(x, \theta)$ é fracamente semicontínua superior em $\theta$, para $\theta \in S_m$.

2) Para todo $m \in \mathbb{N}$, todo $\theta \in S_m$, existe $U^m\varphi_1, \dots, \varphi_k(\theta, \varepsilon)$ tal que $E_{\theta_0}\left[\ln p\left[x, U^m\varphi_1, \dots \varphi_k(\theta, \varepsilon)\right]\right] < \infty$;

3) $S_m$ é limitado para todo $m \in \mathbb{N}$;

4) $\mathbb{A}_m \longrightarrow \theta_0$, quando $m \longrightarrow \infty$.

Então para todos $n$, $m \in \mathbb{N}$, $\mathbb{M}_m^n \neq \phi$ P-q.c. e para uma seqüência $m_n$ que cresça de forma suficientemente lenta, $\mathbb{M}_{m_n}^n \longrightarrow \theta_0$ P-q.c.

**PROVA:** As condições 1) e 2) implicam que $H(\theta_0, \theta)$ e $L_n(\underset{\sim}{x}, \theta)$ são fracamente semicontínuas superiores em $\theta$, $\theta \in S_m$.

$S_m$ é fracamente compacto por ser um conjunto limitado num espaço de Banach reflexivo (ver Goldstein (1967)). Então $\mathbb{A}_m \neq \phi$ e $\mathbb{M}_m^n \neq \phi$ já que uma função fracamente semicontínua superior definida num conjunto fracamente compacto assume seu máximo.

Primeiramente mostra-se que, quando $n_m$, definido de tal forma que

$$n_{m_0} = \min\left\{k : m_k \geq m_0\right\}$$

cresce de forma suficientemente rápida, $\mathbb{M}_{m_n}^n$ eventualmente fica contído em

$$B_m\left(\mathbb{A}_m, \frac{1}{m}\right) = \left\{\theta : \theta \in S_m,\ d(\theta, \mathbb{A}_m) < \frac{1}{m}\right\}$$,

onde $d(\theta, \mathbb{A}_m) = \inf\limits_{\alpha \in \mathbb{A}_m} d(\theta, \alpha)$. A condição 4), então completa a demonstração.

Seja $\theta \in S_m - B_m\left(\mathbb{A}_m, \frac{1}{m}\right)$, então existe $c_\theta > 0$ tal que

$$H\left(\theta_0, \theta_m\right) - H\left[\theta_0, S_m\right] < - c_\theta\ , \quad \text{i.e.}$$

$$H\left(\theta_0, \theta_m\right) - H\left(\theta_0, \alpha_m\right) < - c_\theta\ ,$$

fixando $\alpha_m \in \mathbb{A}_m$.

Então para $\varepsilon$ suficientemente pequeno,

$$E_{\theta_0}\left[\ln p\left[x, U^m\varphi_1, \ldots, \varphi_k(\theta, \varepsilon)\right]\right] - E_{\theta_0}\left[\ln p(x, \alpha_m)\right] < - c_\theta \tag{3.3}$$

Agora,

$$\frac{1}{n}\left\{\ln L_n\left[\underset{\sim}{x}, U^m\varphi_1,\ldots,\varphi_k(\theta, \varepsilon)\right] - \ln L_n(\underset{\sim}{x}, \alpha_m)\right\}$$

$$= \frac{1}{n}\left\{\ln \left[\sup_{\gamma \in U^m\varphi_1,\ldots,\varphi_k(\theta,\varepsilon)} \prod_{i=1}^{n} p(x_i, \gamma)\right] - \ln\left[\prod_{i=1}^{n} p(x_i, \alpha_m)\right]\right\}$$

$$\leq \frac{1}{n}\left\{\ln\left[\prod_{i=1}^{n} p\left[x_i, U^m\varphi_1, \ldots, \varphi_k(\theta, \varepsilon)\right]\right] - \ln\left[\prod_{i=1}^{n} p(x_i, \alpha_m)\right]\right\}$$

$$= \frac{1}{n}\left\{\sum_{i=1}^{n}\left[\ln p\left[x_i, U^m\varphi_1, \ldots, \varphi_k(\theta, \varepsilon)\right] - \ln p(x_i, \alpha_m)\right]\right\}$$

e como por (3.2)

$$\frac{1}{n}\sum_{i=1}^{n}\ln p(x_i,\ \theta) \xrightarrow[n\to\infty]{} E_{\theta_0}\left[\ln p(x,\ \theta)\right]$$

quase certamente, temos por (3.3)

$$\ln\left\{\frac{L_n\left[x,\ U^m\varphi_1,\ \ldots,\ \varphi_k(\theta,\ \varepsilon)\right]}{L_n(\underset{\sim}{x},\ \alpha_m)}\right\} < -\, n\, c_\theta$$

P-q.c. para n suficientemente grande, i.e.

$$\frac{L_n\left[\underset{\sim}{x},\ U^m\varphi_1,\ \ldots,\ \varphi_k(\theta,\ \varepsilon)\right]}{L_n(x,\ \alpha_m)} < e^{-n\, c_\theta}$$

P-q.c. para n suficientemente grande.

Para todo $\theta \in S_m - B_m\left(A_m,\ \frac{1}{m}\right)$, podemos tomar $c_\theta > 0$ e $\varepsilon_\theta > 0$ tais que, com probabilidade 1, a desigualdade acima seja válida para n grande. As bolas $U^m\varphi_1,\ \ldots,\ \varphi_k(\theta,\ \varepsilon)$ cobrem o conjunto $S_m - B_m\left(A_m,\ \frac{1}{m}\right)$, que é fracamente compacto, então podemos encontrar um conjunto finito $\theta_1^m,\ \ldots,\ \theta_{p_m}^m$ tal que $\left\{U^m\varphi_1,\ \ldots,\ \varphi_k(\theta_i^m,\ \varepsilon_i^m)\right\}$, $i = 1,\ \ldots,\ p_m$ cobre $S_m - B_m\left(A_m,\ \frac{1}{m}\right)$.

Então, com $c = \min\limits_{i=1\ldots p_m}\left\{c_{\theta_i}\right\}$, temos, com probabilidade 1,

$$\sup_{i=1\ldots p_m}\frac{L_n\left[\underset{\sim}{x},\ U^m\varphi_1,\ \ldots,\ \varphi_k(\theta,\ \varepsilon)\right]}{L_n(\underset{\sim}{x},\ \alpha_m)} < e^{-n\, c}$$

para todo $n \in \mathbb{N}$ suficientemente grande.

Portanto, para todo $m \in \mathbb{N}$, existe $k_m$ tal que, para todo $n \geq k_m$

$$P\left\{ \sup_{l=1\ldots p_m} \frac{L_n\left[\underset{\sim}{x},\ \bigcup \varphi_1^m,\ \ldots,\ \varphi_k(\theta,\ \varepsilon)\right]}{L_n(\underset{\sim}{x},\ \alpha_m)} \geq 1 \right\} \leq \frac{1}{m^2}\ .$$

Finalmente, para qualquer seqüência $\left\{n_m\right\}$ tal que $n_m \geq k_m$ para todo $m \in \mathbb{N}$

$$\left\{ \mathbb{M}_m^{n_m} \quad B_m\left(\mathbb{A}_m,\ \frac{1}{m}\right) \text{ i.o.} \right\}$$

$$\subseteq \left\{ \sup_{\theta \in S_m - B\left(\mathbb{A}_m,\ \frac{1}{m}\right)} \frac{L_n(x,\ \theta)}{L_n(\underset{\sim}{x},\ \alpha_m)} \geq 1 \quad \text{i.o.} \right\}$$

$$\subseteq \left\{ \sup_{l=1\ldots p_m} \frac{L_n\left[x,\ U^m\varphi_1,\ \ldots,\ \varphi_k(\theta_l^m,\ \varepsilon_l^m)\right]}{L_n(\underset{\sim}{x},\ \alpha_m)} \geq 1 \quad \text{i.o.} \right\}$$

e este último conjunto possui probabilidade zero, pelo Lema de Borel-Cantelli.

CAPÍTULO 4

# COMPARAÇÃO ENTRE O MÉTODO DE "SIEVES" E O MÉTODO DE MÁXIMA VEROSSIMILHANÇA PENALIZADA

Existem numerosas técnicas em análise numérica e estatística que estão muito relacionadas com o método de *"sieves"*, visto no Capítulo anterior.

Para problemas de regressão existe uma relação entre os splines polinomiais de mínimos quadrados e certo tipo de estimadores de *"sieves"*. Sendo p=1,2,... e $\Theta$ a coleção de funções com p-1 derivadas absolutamente contínuas, o *"sieve"*

$$S_m = \left\{ \theta \in \Theta : \int \left| \frac{d^p}{dx^p} \theta(x) \right|^2 dx \leq m \right\}$$

aplicado ao problema de minimização de $\sum_{i=1}^{n} \left\{ y_i - \theta(x_i) \right\}^2$ produz um spline polinomial de grau 2p-1. Este estimador muito usado em regressão não paramétrica é solução do problema de mínimos quadrados análogo ao problema de máxima verossimilhança penalizada

$$\text{minimizar} \sum_{i=1}^{n} \left\{ y_i - \theta(x_i) \right\}^2 + \lambda \int \left| \frac{d^p}{dx^p} \theta(x) \right|^2 dx \quad \text{sobre } \Theta$$

Os estimadores de séries ortogonáis truncadas e o estimador de "máxima verossimilhança admissível" introduzidos por Wegman (1975) também estão relacionados com o método de *"sieves"*. Se os coeficientes num estimador de séries ortogonais são escolhidos de forma a otimizar algum critério, então o estimador tem uma interpretação óbvia como um exemplo do método de *"sieves"*. Se relaxarmos a definição de *"sieve"* de maneira que ele possa depender da amostra aleatória, então o estimador de Wegman também permite esta interpretação.

Neste Capítulo mostraremos que para alguns problemas específicos de estimação o método de MVP é uma especie de *"dual"* do método de *"sieves"*, para uma classe particular de *"sieve"*.

Como foi visto no Capítulo 3 o método de *"sieves"* produz uma variedade de estimadores, e propriedades de existência e consistência têm sido provadas de um ponto de vista unificado por Grenander (1981) e Geman e Hwang (1982) (Teoremas 3.1 e 3.2), sob a condição que o *"sieve"* cresça suficientemente lento com o tamanho amostral.

No entanto, para o método de MVP o estudo da consistência dos estimadores tem tido um tratamento particular caso a caso. Antoniadis e Grégoire (1990) estudaram a consistência do EMVP da função de risco para dados de sobrevivência censurados. Utilizaram um modelo de intensidade multiplicativo e provaram que o estimador da raiz quadrada da intensidade é fracamente consistente em $L^2$ quando a seqüência de parámetros de suavização depende do tamanho amostral de tal forma que $\lambda_n \longrightarrow 0$ de uma maneira apropriada quando $n \longrightarrow \infty$. Sob as mesmas condições também foi provada a consistência na norma do supremo. Klonias (1982) estudou as propriedades de consistência do primeiro EMVP de Good e Gaskins de uma função de densidade. Mostrou a convergência quase certa do estimador nas normas de $L^1$, $L^2$, supremo e Sobolev, sob a condição de que o parâmetro de suavização dependa do tamanho amostral de tal forma que $\lambda_n \longrightarrow 0$ e $\lambda_n n \longrightarrow \infty$ quando $n \longrightarrow \infty$.

Zucker e Karr (1990) provaram resultados de consistência do estimador da função de regressão no contexto de um modelo de regressão de Cox generalizado sempre que $\lambda_n \longrightarrow 0$ a uma taxa controlada quando $n \longrightarrow \infty$.

Silverman (1982) provou resultados de consistência do estimador do logaritmo de uma densidade quando $\lambda_n \longrightarrow 0$ de uma forma apropriada.

Da observação destes resultados surge a pergunta: é possível demonstrar um resultado geral de consistência do EMVP quando a seqüência de parâmetros de suavização $\lambda_n \longrightarrow 0$ de alguma forma apropriada quando o tamanho amostral $n \longrightarrow \infty$ ?.

É o objetivo deste Capítulo demonstrar um resultado deste tipo utilizando a "dualidade" entre o método de *"sieves"* e o método de MVP.

O primeiro passo será então provar essa "dualidade" para um certo tipo de *"sieves"* que deverá relacionar necessariamente o parâmetro m do *"sieve"* com o parâmetro de suavização $\lambda$ do método de MVP.

Posteriormente utilizaremos o Teorema 3.3 sobre consistência do estimador de *"sieves"* para derivar resultados de consistência para o EMVP.

A seguir é descrita a notação necessária para a exposição dos resultados desde Capítulo.

Seja $\Theta \subset H$, onde H é espaço vetorial

$\ell_n : \Theta \longrightarrow \mathbb{R}$, definido por $\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log p(x_i, \theta)$

$\phi : \Theta \longrightarrow \mathbb{R}$, onde $0 \le \phi(\theta) < \infty \quad \forall\ \theta \in \Theta$.

$J_n(\theta) = - \ell_n(\theta)$

$\ell_{n,\lambda}(\theta) = \ell_n(\theta) - \lambda\ \phi(\theta), \quad \lambda > 0$

$J_{n,\lambda}(\theta) = - \ell_{n,\lambda}(\theta) = J_n(\theta) + \lambda\ \phi(\theta)$

Consideremos novamente o problema de minimização (sem restrições) definido no Capítulo 2.

$$\begin{cases} \text{minimizar } J_{n,\lambda}(\theta), \text{ sujeito a} \\ \theta \in \Theta,\ \Theta \subset H,\ \lambda > 0 \end{cases} \tag{2.2}$$

e formulemos o seguinte problema de minimização (com restrições)

$$\begin{cases} \text{minimizar } J_n(\theta), \text{ sujeito a} \\ \theta \in \Theta,\ \phi(\theta) \le m, \quad m \in \mathbb{N} \end{cases} \tag{4.1}$$

Note que o problema (4.1) é equivalente ao problema de maximização de $\ell_n(\theta)$ sujeito a $\theta \in S_m$, onde $S_m = \left\{ \theta \in \Theta\colon \phi(\theta) \le m \right\}$, $m \in \mathbb{N}$, define um *"sieve"*.

Portanto uma solução de (4.1) será um EMV do método de *"sieves"*, i.e. um elemento do conjunto $\mathbb{M}_m^n$ definido no Capítulo 3.

Uma solução de (2.2) será um EMVP. Notaremos por $\mathbb{N}_\lambda^n$ o conjunto destas soluções.

Como foi dito no Capítulo 1, estamos considerando a resolução destes problemas de otimização como método alternativo, nos casos em que não existe $\hat{\theta}_n \in \Theta$ que maximize a log-verossimilhança $\ell_n(\theta)$, ou equivalentemente, que minimize $J_n(\theta)$ sobre todo $\Theta$, i.e. estamos supondo que,

não existe $\hat{\theta}_n \in \Theta$ tal que $J_n(\hat{\theta}_n) \leq J_n(\theta)$, $\forall\ \theta \in \Theta$ (4.2)

A relação entre os problemas (2.2) e (4.1) será discutida neste Capítulo.

## 4.1.- O MÉTODO DE "SIEVES" COMO CASO PARTICULAR DO MÉTODO DE MVP

Seja o *"sieve"* $\left\{S_m\right\}_{m \in \mathbb{N}}$ definido por

$$S_m = \left\{\theta \in \Theta : \phi(\theta) \leq m\right\}$$

A seguinte proposição mostra que, quando $\Theta$ é convexo e $J_n$ e $\phi$ são funcionais convexos, um EMVP em $S_m$ é também um EMVP para algum parâmetro de suavização $\lambda_m$ e função de penalidade $\phi$.

**PROPOSIÇÃO 4.1:** Seja H espaço vetorial, $\Theta \subset$ H convexo, $\phi:\Theta \longrightarrow \mathbb{R}$ e $J_n:\Theta \longrightarrow \mathbb{R}$ funcionais convexos. Suponhamos que exista um $\theta \in \Theta$ tal que $\phi(\theta) < m$. Então, existe $\lambda_m \geq 0$, $\lambda_m \in \mathbb{R}$, tal que se o mínimo em (4.1) é atingido por $\theta^*$, então $\theta^* \in \mathbb{N}^n_{\lambda_m}$, i.e. $\theta^*$ é solução do problema (2.2) para $\lambda = \lambda_m$.

**PROVA:** Tomando $\phi_m(\theta) = \phi(\theta) - m$ o problema (4.1) é equivalente a

$$\begin{cases} \text{minimizar } J_n(\theta), \text{ sujeito a} \\ \theta \in \Theta, \ \phi_m(\theta) \le 0 \end{cases}$$

Como o funcional $J_n(\theta)$ é convexo e $\phi_m(\theta)$ também é convexo, pelo Teorema A.7, $\exists\ \lambda_m \ge 0$, $\lambda_m \in \mathbb{R}$, tal que

$$\inf_{\substack{\theta\in\Theta \\ \phi_m(\theta)\le 0}} J_n(\theta) = \inf_{\theta\in\Theta} \left\{ J_n(\theta) + \lambda_m\ \phi_m(\theta) \right\} = \inf_{\theta\in\Theta} \left\{ J_n(\theta) + \lambda_m\ \phi(\theta) - \lambda_m\ m \right\}$$

e se $\theta^*$ minimiza $J_n(\theta)$ em (4.1), então $\theta^*$ minimiza $J_n(\theta) + \lambda_m\ \phi(\theta)$ sobre $\Theta$. Como $\lambda_m\ m$ é independente de $\theta$, então $\theta^*$ minimiza $J_n(\theta)+\lambda_m\phi(\theta)$ sobre $\Theta$.

Para os casos de estimação de densidade e da intensidade de um processo de Poisson não homogêneo foi visto que o funcional $J_n(\theta)$ resulta contínuo e convexo em $\Theta$ (portanto, fracamente contínuo em $\Theta$) e que $\phi(\theta)$ pode ser tomada como uma norma no espaço de Hilbert H (i.e. $\phi(\theta)$ é contínua, convexa e coerciva).

Foi provado em 2.4.1 que $J_{n,\lambda}(\theta)$ tem nestes casos um único minimizador em $\Theta$.

Isto implica que existe um único EMV em $S_m = \left\{ \theta \in \Theta\text{: } \theta_H^2 \le m \right\}$

Para ver isto, seja então $m \in \mathbb{N}$, e suponhamos que existam $\theta_1$ e $\theta_2$ EMV em $S_m$ (i.e. $\theta_1$, $\theta_2 \in \mathbb{M}^n_m$ ); $\theta_1 \ne \theta_2$.

Pela Proposição 4.1, como $J_n$ é convexa e $\phi(\theta) = \theta^2$ é convexa, existe $\lambda_m \geq 0$ tal que $\theta_1 \in \mathbb{N}^n_{\lambda_m}$ e $\theta_2 \in \mathbb{N}^n_{\lambda_m}$ . Mas pelo Teorema 2.4 $J_{n,\lambda}$ tem um único minimizador em $\Theta$; então $\theta_1 = \theta_2$.

## 4.2.- O MÉTODO DE MVP COMO CASO PARTICULAR DO MÉTODO DE "SIEVES"

Consideremos o problema (2.2), para cada $\lambda > 0$ onde H é espaço de Hilbert, $\Theta$ é subconjunto fechado e convexo, $\phi$ é alguma função de penalidade apropriada que supomos contínua, convexa e coerciva, $J_n(\theta)$ é funcional fracamente contínuo e $J_{n,\lambda}(\theta)$ é coercivo em $\Theta$. Então $\mathbb{N}^n_\lambda \neq \phi$ $\forall\ \lambda > 0$, pelo Teorema 2.1.

Provaremos que se $\hat{\theta}$ for solução de (2.2) então $\hat{\theta}$ será solução de (4.1) para algum m.

Mostraremos que os conjuntos de restrições em (4.1) definem um *"sieve"*.

**PROPOSIÇÃO 4.2:** Sejam $J_n(\theta)$ e $\phi(\theta)$ funcionais definidos em $\Theta$. Suponhamos que exista $\lambda > 0$ e $\theta_\lambda \in \Theta$ tal que

$$J_n(\theta_\lambda) + \lambda\ \phi(\theta_\lambda) \leq J_n(\theta) + \lambda\ \phi(\theta), \quad \forall\ \theta \in \Theta,$$

então $\theta_\lambda$ é solução do problema

$$\begin{cases} \text{minimizar } J_n(\theta), \text{ sujeito a} \\ \theta \in \Theta,\ \phi(\theta) \leq \phi(\theta_\lambda) \end{cases} \tag{4.3}$$

**PROVA:** Suponhamos que exista $\theta^* \in \Theta$ com $J_n(\theta^*) < J_n(\theta_\lambda)$ e $\phi(\theta^*) \le \phi(\theta_\lambda)$

Como $\lambda > 0$ , $\lambda\ \phi(\theta^*) \le \lambda\ \phi(\theta_\lambda)$ , e então

$$J_n(\theta^*) + \lambda\ \phi(\theta^*) < J_n(\theta_\lambda) + \lambda\ \phi(\theta_\lambda)$$

o que contradiz a hipótese.

Portanto, se $\theta_\lambda \in \mathbb{N}^n_\lambda$ , então $\theta_\lambda$ é solução do problema (4.3).

**PROPOSIÇÃO 4.3:** Seja $\mu > \lambda > 0$, então $\phi(\theta_\mu) \le \phi(\theta_\lambda)$ para $\theta_\lambda \in \mathbb{N}^n_\lambda$ e $\theta_\mu \in \mathbb{N}^n_\mu$.

**PROVA:**

$$\mu > \lambda > 0 \ \Rightarrow \ \mu - \lambda > 0$$

$$J_n(\theta_\mu) + \mu\ \phi(\theta_\mu) \le J_n(\theta_\lambda) + \mu\ \phi(\theta_\lambda)$$

pois $\theta_\mu \in \mathbb{N}^n_\mu$

e

$$J_n(\theta_\lambda) + \lambda\ \phi(\theta_\lambda) \le J_n(\theta_\mu) + \lambda\ \phi(\theta_\mu)$$

pois $\theta_\lambda \in \mathbb{N}^n_\lambda$.

Portanto, somando as duas últimas desigualdades, temos

$$\mu\ \phi(\theta_\mu) + \lambda\ \phi(\theta_\lambda) \le \mu\ \phi(\theta_\lambda) + \lambda\ \phi(\theta_\mu)$$

i.e.

$$(\mu - \lambda)\ \phi(\theta_\mu) \le (\mu - \lambda)\ \phi(\theta_\lambda)$$

Logo,

$$\phi(\theta_\mu) \le \phi(\theta_\lambda).$$

**COROLÁRIO 4.1:** Seja $\left\{\lambda_m\right\}$ seqüência decrescente de números positivos tal que $\lambda_m \longrightarrow 0$, quando $m \longrightarrow \infty$. Então, $\left\{\phi\left(\theta_{\lambda_m}\right)\right\}$ é seqüência não decrescente quando $m \longrightarrow \infty$.

**PROVA:** Pela Proposição 4.3, $\lambda_i > \lambda_j$ implica $\phi\left(\theta_{\lambda_i}\right) \le \phi\left(\theta_{\lambda_j}\right)$, $i < j$.

**PROPOSIÇÃO 4.4:** $\forall\ m \in \mathbb{N},\ \exists\ \lambda^{(m)} > 0$ tal que $\phi\left(\theta_{\lambda^{(m)}}\right) > m$.

**PROVA:** Suponhamos que $\exists\ m_0$ tal que $\phi(\theta_\lambda) \le m_0$, $\forall\ \lambda > 0$ , e seja

$$T = \left\{\theta \in \Theta:\ \phi(\theta) \le m_0\right\}$$

T é limitado (por ser $\phi$ coerciva), convexo (por ser $\phi$ convexa) e fechado (por ser $\phi$ contínua), então é fracamente compacto (ver Tapia e Thompson (1978), por exemplo).

Como $J_n(\theta)$ é fracamente contínuo, é fracamente semicontínuo inferior, então atinge seu mínimo sobre T (Teorema A.2), i.e.

$$\exists\ \theta^* \in T \text{ tal que } J_n(\theta^*) \le J_n(\theta) \quad \forall\ \theta \in T,$$

em particular,

$$J_n(\theta^*) \le J_n(\theta_\lambda) \quad \forall\ \lambda > 0$$

fazendo $\lambda \longrightarrow 0$

$$J_n(\theta^*) \le \underline{\lim}_{\lambda \to 0}\ J_n(\theta_\lambda) \tag{4.6}$$

Agora,

$$J_n(\theta_\lambda) + \lambda\, \phi(\theta_\lambda) \le J_n(\theta) + \lambda\, \phi(\theta),\ \forall\, \theta \in \Theta$$

e como $\left\{\phi(\theta_\lambda)\right\}$ é limitada

$$\overline{\lim_{\lambda \to 0}}\ J_n(\theta_\lambda) \le J_n(\theta) \quad \forall\, \theta \in \Theta \tag{4.7}$$

de (4.6) e (4.7),

$$J_n(\theta^*) \le \underline{\lim_{\lambda \to 0}}\, J_n(\theta_\lambda) \le \overline{\lim_{\lambda \to 0}}\ J_n(\theta_\lambda) \le J_n(\theta) \quad \forall\, \theta \in \Theta$$

i.e. existe $\theta^*$ com norma finita, tal que

$$J_n(\theta^*) \le J_n(\theta) \quad \forall\, \theta \in \Theta$$

mas isto é um absurdo, pela suposição (4.2)

Portanto, $\forall\, m \in \mathbb{N}$ deve existir $\lambda^{(m)} > 0$ tal que $\phi\left(\theta^{(m)}\right) > m$.

Seja agora, $\left\{\lambda_m\right\}$ uma seqüência decrescente de números positivos tal que $\lambda_m \longrightarrow 0$ quando $m \longrightarrow \infty$.

Definamos a seguinte seqüência de subconjuntos de $\Theta$:

$$S_m = \left\{\theta \in \Theta\colon\ \phi(\theta) \le \phi\left(\theta_{\lambda_m}\right)\right\},\ m \in \mathbb{N}$$

Na seguinte proposição provaremos que $\left\{S_m\right\}_{m \in \mathbb{N}}$ constitue um *"sieve"*.

**PROPOSIÇÃO 4.5:** $\left\{S_m\right\}_{m \in \mathbb{N}}$ constitue um *"sieve"*.

**PROVA:** (i) Para cada m, $\exists\ \lambda_m > 0$ e $\theta_{\lambda_m} \in \mathbb{N}^n_{\lambda_m}$, então $\theta_{\lambda_m} \in \mathbb{M}^n_m$ pela Proposição 4.2, i.e. $\theta_{\lambda_m}$ é EMV em $S_m$. Portanto, para cada $S_m \subset \Theta$ existe um EMV.

(ii) Se $m_1 < m_2$, então $\lambda_{m_1} > \lambda_{m_2}$ pois a seqüência $\left\{\lambda_m\right\}$ é decrescente, e $\phi\left(\theta_{\lambda_{m_1}}\right) \leq \phi\left(\theta_{\lambda_{m_2}}\right)$ pelo Corolário 4.2, então $S_{m_1} \subset S_{m_2}$.

(iii) $\overline{\bigcup_{m\in\mathbb{N}} S_m} = \Theta$.

Seja $\theta \in \Theta$, e $k = [\phi(\theta)] + 1$, então pela Proposição 4.4 existe $\lambda^{(k)} > 0$ tal que $\phi\left(\theta_{\lambda^{(k)}}\right) > k > \phi(\theta)$. Agora dado $\lambda^{(k)}$, $\exists\ \lambda_{m_k} \in \left\{\lambda_m\right\}$ tal que $\lambda_{m_k} < \lambda^{(k)}$ pois $\left\{\lambda_m\right\}$ é decrescente, e então $\phi\left(\theta_{\lambda_{m_k}}\right) \geq \phi\left(\theta_{\lambda^{(k)}}\right)$ pela Proposição 4.3.

Portanto $\phi\left(\theta_{\lambda_{m_k}}\right) > \phi(\theta)$ e então $\theta \in S_{m_k}$.

Logo $\overline{\bigcup_{m\in\mathbb{N}} S_m} = \Theta$.

Por (i), (ii), e (iii), $\left\{S_m\right\}_{m\in\mathbb{N}}$ constitui um *"sieve"*.

Conseqüentemente, temos provado que dada uma seqüência de problemas: $\begin{cases} \text{minimizar } J_n(\theta) + \lambda_m\, \phi(\theta) \\ \theta \in \Theta \end{cases}$, onde $\lambda_m \downarrow 0$ quando $m \longrightarrow \infty$, existe um *"sieve"* $\left\{S_m\right\}_{m\in\mathbb{N}}$, onde $S_m = \left\{\theta \in \Theta\colon \phi(\theta) \leq \phi\left(\theta_{\lambda_m}\right)\right\}$ tal que se $\theta_{\lambda_m} \in \mathbb{N}^n_{\lambda_m}$ então $\theta_{\lambda_m} \in \mathbb{M}^n_m$.

Podemos provar agora um resultado de consistência para os EMVP.

## 4.3.- CONSISTÊNCIA DOS EMVP

A seguir será utilizada a notação definida no Capítulo 3.

**TEOREMA 4.1:** Seja o problema de otimização (2.2), onde H é espaço de Hilbert, $\Theta$ é um subconjunto fechado e convexo de H, $\phi(\theta)$ é funcional contínuo, convexo e coercivo, e tal que $0 \le \phi(\theta) < \infty \quad \forall\ \theta \in \Theta$.

Suponhamos $\mathbb{N}^n_\lambda \neq \phi \quad \forall\ \lambda > 0$ . Seja $\left\{\lambda_m\right\}$ uma seqüência decrescente de números positivos tal que $\lambda_m \longrightarrow 0$ quando $m \longrightarrow \infty$, e $S_m = \left\{\theta \in \Theta\colon \phi(\theta) \le \phi\left(\theta_{\lambda_m}\right)\right\}$, $m \in \mathbb{N}$, tal que

1) Para todo $m \in \mathbb{N}$, todo $\theta \in S_m$ e toda vizinhança fraca $U^m\varphi_1, ..,\varphi_k(\theta, \varepsilon)$ do ponto $\theta \in S_m$

$$p\left[x,\ U^m\varphi_1, ..,\varphi_k(\theta, \varepsilon)\right] \text{ é } \mathcal{F}\text{-mensurável}$$

2) Para todo $m \in \mathbb{N}$, todo $\theta \in S_m$ e todo $U^m\varphi_1, ..,\varphi_k(\theta, \varepsilon)$

$$\lim_{\varepsilon \longrightarrow 0} p\left[x,\ U^m\varphi_1, ..,\varphi_k(\theta, \varepsilon)\right] = p(x, \theta) \quad \mu\text{-q.c.}$$

3) Para todo $m \in \mathbb{N}$ , todo $\theta \in S_m$ existe $U^m\varphi_1, ..,\varphi_k(\theta, \varepsilon)$ tal que

$$E_{\theta_0}\left[\ln p\left[x,\ U^m\varphi_1, ..,\varphi_k(\theta, \varepsilon)\right]\right] < \infty;$$

4) $A_m \longrightarrow \theta_0$, quando $m \longrightarrow \infty$.

Então existe uma subseqüência $\left\{\lambda_{m_n}\right\}$ com $\lambda_{m_n} \longrightarrow 0$ quando $n \longrightarrow \infty$ de uma forma apropriada tal que $\left\|\theta_{\lambda_{m_n}} - \theta_0\right\| \longrightarrow 0$ P-q.c., onde $\theta_{\lambda_{m_n}}$ é solução do problema (2.2) para $\lambda = \lambda_{m_n}$.

**OBS.** Note que H é espaço de Hilbert e portanto espaço de Banach reflexivo.

**PROVA:** Seja $\lambda_m > 0$ e $\theta_{\lambda_m} \in \mathbb{N}^n_{\lambda_m}$, pela Proposição 4.5 $\left\{S_m\right\}_{m \in \mathbb{N}}$ constitui um *"sieve"*.

Este *"sieve"* verifica as hipóteses 1) e 2) e 4) do Teorema 3.3. A condição 3) é satisfeita pois $S_m$ é limitado por ser $\phi$ coerciva (ver Tapia e Thompson, por exemplo).

Então existe uma subseqüência $m_n$ crescendo suficientemente lenta quando $n \longrightarrow \infty$ tal que

$$\mathbb{M}^n_{m_n} \longrightarrow \theta_0 \quad \text{P-q.c.}$$

Como $\theta_{\lambda_{m_n}} \in \mathbb{N}^n_{\lambda_{m_n}}$ então $\theta_{\lambda_{m_n}} \in \mathbb{M}^n_{m_n}$ pela Proposição 4.2, então,

$$\left\|\theta_{\lambda_{m_n}} - \theta_0\right\| \leq \sup_{\theta \in \mathbb{M}^n_{m_n}} \left\|\theta - \theta_0\right\| \longrightarrow 0$$

Portanto existe uma subseqüência $\theta_{\lambda_{m_n}} \in \mathbb{N}^n_{\lambda_{m_n}}$ tal que

$\left\|\theta_{\lambda_{m_n}} - \theta_0\right\| \longrightarrow 0$ quando $\lambda_{m_n} \longrightarrow 0$, quando $n \longrightarrow \infty$.

Como no Teorema 3.2 este teorema não diz nada sobre a velocidade com a qual $\lambda_{m_n}$ deve tender a zero.

Mostraremos a seguir que o *"sieve"* da Proposição 4.5 (aliás qualquer *"sieve"* da forma $S_m = \left\{\theta \in \Theta;\ \phi(\theta) \leq m\right\}$) satisfaz as condições do Teorema 3.3 para os exemplos discutidos antes, de estimação de densidade e de intensidade de um processo de Poisson não homogêneo.

## ESTIMAÇÃO DE DENSIDADE

Nosso espaço mensurável é dado por $([a,\ b],\ \mathcal{B}([a,\ b]))$, onde $\mathcal{B}([a,\ b])$ é a $\sigma$-álgebra de Borel em $[a,\ b]$.

$p(x,\ \theta) = \theta(x)$, $\theta \in \Theta$, $\Theta$ fechado e convexo.

Para todo $m \in \mathbb{N}$, todo $\theta \in S_m$, todo $\varepsilon > 0$ e todo subconjunto finito $\left\{\varphi_1,\ ..,\varphi_k\right\} \subset H^*$

$p\left[x,\ U^m\varphi_1,\ ..,\varphi_k(\theta,\ \varepsilon)\right]$ é $\mathcal{F}$-mensurável, pois

$$\left\{x :\ \sup_{x \in U^m\varphi_1,\ ..,\varphi_k(\theta,\varepsilon)} p(x,\ \theta) > y\right\}$$

$$= \bigcup_{x\ \in\ U\varphi_1,\ ..,\varphi_k(\theta,\varepsilon)} \left\{x:\ p(x,\ \theta) > y\right\}$$

e $\left\{x:\ p(x,\ \theta) = \theta(x) > y\right\}$ é aberto em $[a,\ b]$ $\forall\ y \in \mathbb{R}$, pois $\theta \in \Theta$ é contínua, com $\Theta \subset H = H_0^1[a, b]$ ou $H^1(-\infty, \infty)$ (para os casos do estimador de De Montricher, Tapia e Thompson e o primeiro EMVP de Good e Gaskins, respectivamente).

$p(x, \theta)$ será fracamente semicontínua superior em $\Theta$, se provarmos que $p(x, \theta)$ é contínua e côncava em $\Theta$, pois uma função contínua e côncava num conjunto fechado e convexo $\Theta$ é fracamente semicontínua superior (Proposição A.3). Como H é RKHS, e $\theta \in H$, então $p(x, \theta) = \theta(x)$ é contínua e côncava em $\Theta$.

Vejamos que $E_{\theta_0}\left[\ln p\left[x, U^m\varphi_1, ..,\varphi_k(\theta, \varepsilon)\right]\right] < \infty$

Como $\phi$ é coerciva, $S_m$ é limitado, então se $\theta \in S_m$ existe $C_m$ tal que $\theta_H \le C_m$.

Agora, como H é RKHS, $|\theta(x)| \le M_x \|\theta\|_H \le M_x C_m, \ \forall\ \theta \in S_m$.

Portanto $\ln p(x, \theta) = \ln \theta(x) \le \ln (M_x C_m), \ \forall\ \theta \in S_m$

Conseqüentemente,

$$E_{\theta_0}\left[\ln p\left[x, U^m\varphi_1, ..,\varphi_k(\theta, \varepsilon)\right]\right] < \infty$$

Provemos finalmente a condição 4) do Teorema.

Como

$$H(\theta_0, \theta) = E_{\theta_0}[\ln p(x, \theta)] < \infty,$$

e

$$\ln p(x, \theta_n) \le \ln (M_x C_m)$$

para $\theta_n \in S_m$, então

$$\lim_{n\to\infty} E_{\theta_0}[\ln p(x, \theta)] = E_{\theta_0}\left[\lim_{n\to\infty} \ln p(x, \theta_n)\right] = E_{\theta_0}[\ln p(x, \theta)]$$

pelo Teorema da Convergência Dominada e pela continuidade da função logarítmica e de $p(x, \theta)$, ou seja $H(\theta_0, \theta)$ é contínua em $\theta$.

Como $\bigcup_{m=1}^{\infty} S_m$ é denso em $\Theta$ e $\left\{S_m\right\}$ é crescente, existe $\theta_m \in S_m$ tal que $\theta_m \longrightarrow \theta_0$ quando $m \longrightarrow \infty$. Portanto,

$$\sup_{\theta\in S_m} H(\theta_0, \theta) \longrightarrow H(\theta_0, \theta_0), \quad \text{quando } m \longrightarrow \infty, \text{ e}$$

$\hat{\theta}_m \longrightarrow \theta_0$, quando $m \longrightarrow \infty$.

## ESTIMAÇÃO DA INTENSIDADE DE UM PROCESSO DE POISSON NÃO HOMOGÊNEO

Lembremos que o espaço de observações é dado por $\left(\mathbb{D}[0, 1], \mathcal{B}\left(\mathbb{D}[0, 1]\right)\right)$ e que

$$p(x, \theta) = \exp\left[\sum_{k=1}^{N(1,x)} \ln\left[\theta(\tau_k)\right] + \int_0^1 \left[1 - \theta(t)\right] dt\right]$$

onde $\tau_k$ representa o instante do k-ésimo salto de N.

$\theta \in \Theta$, onde $\Theta$ é subconjunto fechado e convexo do RKHS $H = H_0^1[0, 1]$.

Como no caso de densidade, para verificar a condição 1) do Teorema 3.3 só precissamos mostrar que $\left\{x\colon p(x, \theta) > y\right\}$ é subconjunto aberto em $\mathbb{D}[0, 1]$, mas, para ver isto é suficiente provar que

$$\left\{x: \sum_{k=1}^{N(1,x)} \ln \theta(\tau_k) > y\right\} \text{ é aberto em } \mathbb{D}[0,\ 1] \quad \forall\ y \in \mathbb{R},$$

pois a função exponencial é contínua e $\int_0^1 [1 - \theta(t)]\ dt$ é constante (por ser $\theta$ determinística).

Como a função logarítmica é uma função contínua e $\theta \in S_m$ também é contínua, se x for tal que

$$\sum_{k=1}^{N(1,x)} \ln \theta(\tau_k) > y,$$

então existe $\varepsilon_x > 0$ tal que

$$\sum_{k=1}^{N(1,x)} \ln \theta(\tau_k - \varepsilon_x) > y.$$

Conseqüentemente, para todo x' tal que

$$d_S(x,\ x') < \varepsilon_x$$

temos

$$p(x',\ \theta) = \sum_{k=1}^{N(1,x')} \ln \theta(\tau'_k) > y$$

Portanto, $\left\{x:\ p(x,\ \theta) > y\right\}$ é um subconjunto aberto de $\mathbb{D}[0,\ 1]$ e então $\left\{x: \sup_{\theta \in U^m \varphi_1,\ \ldots,\varphi_k(\theta,\varepsilon)} p(x,\ \theta) > y\right\}$ também é aberto.

Vejamos que $p(x,\ \theta)$ é fracamente semicontínua superior em $\Theta$. Um cálculo direto mostra que

$$p''(x,\ \theta)(\eta,\ \eta) =$$

$$= \left\{\exp\left[\sum_{k=1}^{N(1,x)} \ln \theta(\tau_k) + 1 - \int_0^1 \theta(s)\, ds\right]\right\}\left\{-\sum_{k=1}^{N(1,x)} \frac{\eta^2(\theta_k)}{\theta^2(\theta_k)}\right\} \le 0$$

$\forall\ \theta \in \Theta$, para cada $\eta \in T(\theta)$,

i.e. $p(x, \theta)$ é côncava em $\Theta$, e como $p(x, \theta)$ é contínua em $\Theta$ pois a função exponencial é contínua, logaritmo é contínua e a avaliação pontual e a integração são contínuas no RKHS $H_0^1[0, 1]$, então $p(x, \theta)$ é fracamente semicontínua superior em $\Theta$.

Isto completa a demostração para a condição 1) do Teorema 3.3

Como $\phi$ é coerciva, $S_m$ é limitado, então $\|\theta\|_H \le C_m \quad \forall\ \theta \in S_m$. Portanto, como H é RKHS, $|\theta(x)| \le M_x C_m \quad \forall\ \theta \in S_m$.

Então

$$\sum_{k=1}^{N(1,x)} \ln \theta(\tau_k) \le N(1, x) \ln(M_x C_m)$$

Portanto

$$\ln p(x, \theta) \le N(1, x) \ln(M_x C_m) + 1$$

$\forall\ x \in \Omega$ tal que $N(1, x) < \infty$.

Conseqüentemente

$$E_{\theta_0}\left[\ln p\left[x, U^m_{\varphi_1, \dots, \varphi_k}(\theta, \varepsilon)\right]\right] < \infty$$

pois $P(N(1) = \infty) = 0$.

Finalmente, como

$$H(\theta_0, \theta_n) = E_{\theta_0}\left[\ln p(x, \theta_n)\right] < \infty$$

e

$$\ln p(x, \theta_n) \le N(1, x) \ln(M_x C_m) + 1$$

$\forall\ \theta_n \in S_m$, então,

$$\lim_{n\to\infty} E_{\theta_0}\left[\ln p(x, \theta_n)\right] = E_{\theta_0}\left[\lim_{n\to\infty} \ln p(x, \theta_n)\right] = E_{\theta_0}\left[\ln p(x, \theta)\right]$$

pelo Teorema da Convergência Dominada e pela continuidade da função logarítmica e de $p(x, \theta)$, i.e. $H(\theta_0, \theta)$ é contínua em $\theta$.

Como $\bigcup_{n=1}^{\infty} S_m$ é denso em $\Theta$ e $\left\{S_m\right\}$ é crescente, provamos como no caso anterior de estimação de densidade que $\hat{\theta}_m \longrightarrow \theta_0$, quando $m \longrightarrow \infty$.

## 4.4.- CONCLUSÕES

É possível reformular o teorema desenvolvido por Geman e Hwang (1982) sobre a existência de uma seqüência de estimadores de "*sieves*" consistente. Na nova versão do teorema é pedida compacidade fraca dos subconjuntos do "*sieve*" (ou equivalentemente, que eles sejam limitados num espaço de Banach reflexivo) e semicontinuidade fraca superior dos funcionais de verossimilhança e entropia sobre esses subconjuntos. É mostrado que estas hipóteses não resultam restritivas demais. Os exemplos de aplicação mais estudados na literatura, sobre estimação de densidade e da intensidade de um processo de Poisson não homogêneo, verificam estas hipóteses.

Por outro lado, dada uma seqüência decrescente de números reais positivos $\left\{\lambda_m\right\}$, com $\lambda_m \longrightarrow 0$, quando $m \longrightarrow \infty$, temos uma seqüência de problemas de maximização da verossimilhança penalizada com parâmetro de suavização $\lambda_m$ e penalidade $\phi(\theta)$. Supondo a verossimilhança fracamente contínua, $\phi(\theta)$ contínua, convexa e coerciva, e a verossimilhança $\phi$-penalizada coerciva, cada um destes problemas da seqüência tem solução. É possivel então definir um *"sieve"* $\left\{S_m\right\}$, onde $S_m = \left\{\theta \in \Theta\colon \phi(\theta) \leq \phi\left(\theta_{\lambda_m}\right)\right\}$ e $\theta_{\lambda_m}$ é solução do problema da maximização da verossimilhança $\phi$-penalizada para o parâmetro de suavização $\lambda_m$.

A nova versão do teorema desenvolvida no Capítulo 3 pode ser aplicada a este *"sieve"* particular para provar a existência de uma seqüência de EMVP consistente. É mostrada a existência de uma subseqüência $\lambda_{m_n}$ dependendo do tamanho amostral n, com $\lambda_{m_n} \longrightarrow 0$ quando $m_n \longrightarrow \infty$ de uma forma apropriada, tal que se $\theta_{\lambda_{m_n}}$ é EMVP para o parâmetro de suavização $\lambda_{m_n}$, então $d\left(\theta_{\lambda_{m_n}}, \theta_0\right) \longrightarrow 0$, onde $\theta_0$ é o verdadeiro valor do parâmetro.

Não encontramos na literaura resultados gerais de consistência para os EMVP. Os resultados de consistência dos EMVP, provados por Antoniadis e Grégoire (1990) para o problema de estimação da intensidade de um processo de Poisson não homogêneo e por Klonias (1982) para a estimação de densidades, resultam casos particulares do obtido neste trabalho. Os resultados por eles provados permanecem mais fortes no sentido que conseguem identificar a seqüência $\lambda_{m_n}$.

Supondo o funcional de log-verossimilhança côncavo, mostra-se que dado um *"sieve"* $\left\{S_m\right\}_{m \in \mathbb{N}}$, onde $S_m = \left\{\theta \in \Theta\colon \phi(\theta) \leq m\right\}$ com $\phi$ convexa, associado a cada m existe um $\lambda_m \geq 0$ tal que qualquer EMV em $S_m$ é maximizador da log-verossimilhanca $\phi$-penalizada com parâmetro de suavização $\lambda_m$.

Em geral os funcionais $\phi(\theta)$ podem ser considerados como normas em algum espaço. Para este caso, e considerando a log-verossimilhança côncava e contínua, o problema da maximização da verossimilhanca $\phi$-penalizada tem solução única, e isto implica na existência de um único EMV em cada $S_m$, onde $S_m = \left\{\theta \in \Theta: \phi(\theta) \le m\right\}$.

## 4.5.- SUGESTÕES

Para o método de MVP fica ainda o importante problema da identificação da seqüência de parâmetros de suavização $\lambda_n$ para a consistência dos estimadores. O próximo passo de pesquisa poderia ser a reformulação do Teorema 3.2, que fornece um meio para a identificação de tal seqüência, para permitir sua aplicação a um "*sieve*" da forma $S_m = \left\{\theta \in \Theta: \phi(\theta) \le \phi\left(\theta_{\lambda_m}\right)\right\}$. Mas, um resultado teórico deste tipo, como no Teorema 3.2, não forneceria nenhuma regra simples para o cálculo de tal seqüência. Então, posteriormente poderiam ser estudadas diferentes seqüências para exemplos simulados.

Outra proposta seria a aplicação de métodos de validação cruzada para a estimação do parâmetro de suavização.

O cálculo numérico dos estimadores para $\lambda$ fixo, poderia ser objeto de trabalhos futuros. Isto envolve a aplicação de métodos de otimização com restrições não lineares. Os métodos que têm sido usados incluem o método de Raleigh-Ritz (Good e Gaskins (1971)),onde a verossimilhança é desenvolvida em séries ortogonais. Outra sugestão é um procedimento de discretização onde o funcional de penalidade é substituído por uma soma de segundas diferenças ao quadrado e os EMVP são encontrados resolvendo um problema de otimização de dimensão finita (Tapia e Thompson (1978)).

Estes e outros métodos numéricos para a maximização dos funcionais de verossimilhança penalizados poderiam ser estudados e comparados.

Uma vantagem observada do método de MVP sobre os chamados "métodos de janela", e que seria de interesse estudar parece ser sua aplicabilidade a pequenas amostras.

A análise destes problemas deveria ser feita através de discussões de exemplos com dados reais e estudos de simulação.

# APÊNDICE

Ao longo deste trabalho são usadas as nocões topológicas de convergência, compacidade, continuidade, etc.

Quando falarmos destas nocões estaremos fazendo referência à topologia usual, a menos que seja feita referência explícita à topologia fraca.

Para uma definição dos conceitos envolvidos referimos, por exemplo, a Royden (1968) ou Kolmogórov e Fomín (1978).

As provas dos resultados apresentados neste apêndice sobre otimização em espaços de Hilbert e espaços de Hilbert com kernel reprodutivo podem ser encontradas, por exemplo, em Tapia e Thompson (1978) e os resultados sobre teoría de multiplicadores de Lagrange em Luenberguer (1968).

**DEFINIÇÃO A.1:** Seja H um espaço de Hilbert e $f:H \longrightarrow \mathbb{R}$. Dados $x,\eta_1,\ldots,\eta_n \in H$, a *n-ésima variação de Gâteaux de $f$ em $x$ nas direções* $\eta_1,\ldots,\eta_n$ é definida por

$$f^{(n)}(x)(\eta_1,\ldots,\eta_n) =$$

$$= \lim_{t \longrightarrow 0} t^{-1} \left[ f^{(n-1)}(x+t\eta_n)(\eta_1,\ldots,\eta_{n-1}) - f^{(n-1)}(x)(\eta_1,\ldots,\eta_{n-1}) \right]$$

com $f^{(0)}(x)=f(x)$

Dizemos que f *é m vezes Gâteaux diferenciável* em $S \subset H$ se as m primeiras variações de Gâteaux existirem em x e forem lineares em $\eta_i$ $\forall$ i e para cada $x \in S$ fixo.

Também dizemos que f *é continuamente diferenciável* em S se for Gâteaux diferenciável em S, $f^{(1)}(x) \in H^*$ $\forall\ x \in S$ e $f^{(1)}:S \subset H \longrightarrow H^*$ for um operador contínuo. No caso em que $f^{(1)}(x) \in H^*$, o representante de Riesz de $f^{(1)}(x)$ é denotado por $\nabla f(x)$ e é chamado o *gradiente* de f em x (i.e. $f^{(1)}(x)(\eta) = \langle \nabla f(x), \eta \rangle$ $\forall\ \eta \in H$ )

**DEFINIÇÃO A.2:** Consideremos $f:H \longrightarrow \mathbb{R}$ e um subconjunto convexo S de H. O *cone tangente a S en x* é definido por

$$T(x) = \left\{ \eta \in H: \exists\ t>0: x+t\eta \in S \right\}$$

Suponhamos que f seja duas vezes Gâteaux diferenciável em S. Então $f^{(2)}$ é dita *semidefinida positiva relativa a S* se para cada $x \in S$ tivermos

$$f^{(2)}(x)(\eta,\eta) \geq 0 \quad \forall\ \eta \in T(x) \tag{1}$$

Dizemos que $f^{(2)}$ é *definida positiva relativa a S* se a desigualdade estrita valer em (1) para $\eta \neq 0$. Finalmente, $f^{(2)}$ é dita *uniformemente definida positiva relativa a S* se existir C>0, tal que para cada $x \in S$, tivermos

$$f^{(2)}(x)(\eta,\eta) \geq C \left\| \eta \right\|^2, \quad \forall\ \eta \in T(x)$$

**PROPOSIÇÃO A.1:** Suponhamos que $f:\longrightarrow\mathbb{R}$ seja duas vezes Gâteaux diferenciável em um subconjunto convexo S de H.

Então,

(i) f é convexa $\Leftrightarrow$ $f^{(2)}$ é semidefinida positiva relativa a S.

(ii) f é estritamente convexa em S $\Leftrightarrow$ $f^{(2)}$ é definida positiva relativa a S.

**PROPOSIÇÃO A.2:** Suponhamos que $f:\longrightarrow\mathbb{R}$ seja duas vezes Gâteaux diferenciável em um subconjunto convexo S de H. Então, f é uniformemente convexa em S com constante C $\Leftrightarrow$ $f^{(2)}$ é uniformemente definida positiva relativa a S com constante 2C.

## EXISTÊNCIA E UNICIDADE DE SOLUÇÕES PARA PROBLEMAS DE OTIMIZAÇÃO EM ESPAÇOS DE HILBERT

Seja H espaço de Hilbert

Para $S \subset H$ e $f:\longrightarrow\mathbb{R}$ consideremos o seguinte problema de otimização:

$$\begin{cases} \text{minimizar } f(x), \text{ sujeito a} \\ x \in S \end{cases} \tag{2}$$

as soluções do problema (2) são chamadas minimizadores.

**TEOREMA A.1: (Unicidade).** Se f for estritamente convexa e S convexo, então o problema (2) tem no máximo um minimizador.

**TEOREMA A.2: (Existência).** Seja $f:H \longrightarrow \mathbb{R}$ funcional fracamente semicontínuo inferior e seja $S \subset H$ fracamente compacto, então o problema (2) tem pelo menos um minimizador.

**COROLÁRIO A.1:** Seja $f:H \longrightarrow \mathbb{R}$ um funcional fracamente semicontínuo inferior e $S \subset H$ um subconjunto fechado, limitado e convexo. Então o problema (2) tem pelo menos um minimizador.

**DEFINIÇÃO A.3:** Consideremos $S \subset H$ e $f: \longrightarrow \mathbb{R}$. Dizemos que f é *coercivo* em S se $\{x_n\} \subset S$ e $\|x_n\| \xrightarrow[n \to \infty]{} \infty$ implica que $f(x_n) \xrightarrow[n \to \infty]{} +\infty$.

**COROLÁRIO A.2:** Seja $f:H \longrightarrow \mathbb{R}$ um funcional fracamente semicontínuo inferior e $S \subset H$, fechado, convexo e não limitado. Suponhamos que f seja coercivo em S. Então, o problema (2) tem pelo menos um minimizador.

**PROPOSIÇÃO A.3:** Um funcional contínuo e convexo (côncavo) f, definido num subconjunto convexo e fechado S de H é fracamente semicontínuo inferior (superior).

**TEOREMA A.3:** Suponhamos que no problema (2) o funcional f seja convexo e contínuo em S, e o subconjunto S seja fechado, limitado e convexo. Então o problema (2) tem no mínimo um minimizador.

**TEOREMA A.4:** Seja S subconjunto convexo e fechado de H. Se $f:S \longrightarrow \mathbb{R}$

for convexo, contínuo e coercivo em S, então o problema (2) tem no mínimo um minimizador.

**PROPOSIÇÃO A.4:** Seja S subconjunto convexo e fechado de H. Se $f:S \longrightarrow \mathbb{R}$ for contínuo e uniformemente convexo em S, então f é coercivo em S.

**TEOREMA A.5:** Seja S subconjunto convexo e fechado de H. Se $f:S \longrightarrow \mathbb{R}$ for contínuo e uniformemente convexo em S, então o problema (2) tem um único minimizador.

**TEOREMA A.6:** Seja S subconjunto convexo e fechado de H. Suponhamos que $f:H \longrightarrow \mathbb{R}$ seja contínuo em S, duas vezes Gâteaux diferenciável em S e que a segunda variação de Gâteaux seja uniformemente definida positiva em S. Então o problema (2) tem um único minimizador.

## MULTIPLICADORES DE LAGRANGE

Seja o seguinte problema de otimização:

$$\begin{cases} \text{minimizar } J(\theta), \text{ sujeito a} \\ \theta \in \Theta,\ \phi(\theta) \le 0 \end{cases}$$

**TEOREMA A.7:** Seja H espaço vetorial $\Theta$ subconjunto convexo de H, J e $\phi$ funcionais convexos em $\Theta$. Suponhamos que exista $\theta \in \Theta$ tal que $\phi(\theta_1) < 0$. Seja

$$\mu_0 = \inf J(\theta) \tag{3}$$

sujeito a $\theta \in \Theta$ , $\phi(\theta) \le 0$, e suponhamos $\mu_0$ finito.

Então existe $\lambda \in \mathbb{R}$, $\lambda \ge 0$ tal que

$$\mu_0 = \inf_{\theta\in\Theta} \left\{ J(\theta) + \lambda\ \phi(\theta) \right\} \tag{4}$$

$\lambda$ é chamado é chamado multiplicador de Lagrange para o problema.

Se o ínfimo em (3) for atingido por um $\theta^* \in \Theta$, $\phi(\theta^*) \le 0$, então também será atingido por $\theta^*$ em (4).

## ESPAÇOS DE HILBERT COM KERNEL REPRODUTIVO

**DEFINIÇÃO A.4:** Um espaço de Hilbert H de funções definidas num conjunto T é dito um *espaço de Hilbert con kernel reprodutivo* (RKHS) se para cada $t \in T$ a avaliação pontual em t for operação contínua, i.e. existe $M_t$, tal que

$$|f(t)| \le M_t \|f\|, \quad \forall f \in H$$

**PROPOSIÇÃO A.5:** O espaço

$$H_0^s(a,b) = \left\{ f:\ f^{(j)} \in L^2(a,b),\ j=0,..,s \text{ e } f^{(j)}(a)=f^{(j)}(b)=0,\ j=0,..,s-1 \right\}$$

com produto interno

$$\langle f, g \rangle = \langle f^{(s)}, g^{(s)} \rangle_{L^2(a,b)}$$

é um RKHS.

**PROPOSIÇÃO A.6:** O funcional integração $Q(f) = \int_a^b f(t)\ dt$ é contínuo em $H_0^s(a, b)$.

**PROPOSIÇÃO A.7:** O espaço

$$H^s(-\infty, \infty) = \left\{ f: f^{(j)} \in L^2(-\infty, \infty) \text{ para } j=0,\dots,s \right\}$$

com produto interno

$$\langle f, g\rangle = \sum_{j=0}^{s} \langle f^{(j)}, g^{(j)}\rangle_{L^2(-\infty,\infty)}$$

é um RKHS.

**PROPOSIÇÃO A.8:** O espaço de Sobolev

$$H^s(-\infty, \infty) = \left\{ v \in L: (1 + w^2)^{s/2}\, \hat{v}(w) \in L^2(-\infty, \infty) \right\}$$

onde L é o espaço de distribuições com decrecimento polinomial no infinito e $\hat{v}$ denota a transformada de Fourier de v, é um RKHS se e somente se $s > 1/2$.

## FUNÇÕES SPLINE

Seja $\Delta = \left\{x_i\right\}_{i=0}^{k+1}$ $a=x_0 < x_1 < \dots < x_k < x_{k+1}=b$ uma partição do intervalo [a, b] em k+1 subintervalos, $I_i = [x_i, x_{i+1})$, $i=0,\dots,k$ e $I_k = [x_k, x_{k+1}]$ .

**DEFINIÇÃO A.5:** Uma função s(x) é chamada *spline polinomial de grau m* associado à partição Δ se :

(i) existem $s_0,\dots,s_k$ polinômios de grau m tais que $s(x)=s_i(x)$

para $x \in I_i$, $i=0,..,k$.

(ii) $s(x)$ tem $(m-1)$ derivadas contínuas em cada $x_i$.

**DEFINIÇÃO A.6:** Uma função $f(x)$ é chamada *monospline polinomial de grau m* associado à partição $\Delta$ se

$$f(x) = \frac{x^m}{m!} + s(x)$$

onde $s(x)$ é um spline polinomial de grau m associado à partição $\Delta$.

**DEFINIÇÃO A.7:** Uma função $f(x)$ é chamada *spline exponencial* com nós nos pontos $x_1,..,x_k$ se

(i) existem $u_0,..,u_k \in sp\{ e^{\alpha_1 x},..., e^{\alpha_m x} \}$ onde $\alpha_1 < \alpha_2 <..< \alpha_m$ com $f(x) = u_i(x)$ para $x \in I_i$, $i=0,..,k$.

(ii) $f(x)$ tem $(m-1)$ derivadas contínuas em cada $x_i$.

## REFERÊNCIAS

- Anderson, J.A., and Blair, V. (1982) - *Penalized maximum likelihood estimation in logistic regression and discrimination*, Biometrika 69, 123-136.

- Anderson, J.A. and Senthilselvan, A. (1980) - *Smooth estimates for hazard function*, J.R. Statist. Soc. B, 42, 322-327.

- Antoniadis, A. (1988) - *Parametric estimation for the mean of a Gaussian process by the method of Sieves*, J. of Multiv. Anal. 26, 1-15.

- Antoniadis, A., and Grégoire, G. (1990) - *Penalized likelihood estimation for rates with censored survival data*, Scand. J. Statist. 17, 43-63.

- Bartoszynski, R., Brown, B.W., Mc Bride, C.M., and Thompson, J.R. (1981) - *Some nonparametric techniques for estimating the intensity function of a cancer related nonstationary Poisson process*, Ann. Statist. 9, 1050-1060.

- Chow, Y., and Grenander, U. (1985) - *A sieve method for the spectral density*, Ann. Statist. 13, 998–1010.

- De Boor, C. (1978) - *A practical guide to Splines*, Springer - Verlag, New York.

- De Montricher, G.M., Tapia, R.A., and Thompson, J.R. (1975) - *Nonparametric maximum likehood estimation of probability density by penalty function methods*, Ann. Statist 3, 1329–1348.

- Geman, S. (1982) - *Method of Sieves*, Encyclopedia of Statistical Sciences, Volume 5.

- Geman, S., and Hwang, C.R. (1982) - *Nonparametric maximum likelihood estimation by the method of sieves*, Ann. Statist. 10, 401–414.

- Goldstein, A.A. (1967) - *Constructive Real Analysis*, Harper and Row, New York.

- Gomes, A.E. (1991) - *Estimação da função intensidade de processos de Poisson não homogêneos pelo método de Grenander*, Tese de Mestrado, IMECC-UNICAMP.

- Good, I.J, and Gaskins, R.A. (1971) - *Nonparametric roughness penalties for probability densities*. Biometrika 58, 255-277.

- Good, I.J., and Gaskins, R.A. (1980) - *Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data*. J. Amer. Statist, Assoc. 75, 42-72.

- Grenander, U. (1981) - *Abstract Inference*, Wiley, New York.

- Ibragimov, I.A., and Has'minskii, R.Z. (1981) - *Statistical Estimation - Asymptotic Theory*, Spinger-Verlag, New York.

- Izenman, A.J. (1991) - *Recent developments of the likelihood estimate*, Ann. Math. Statist. 20, 595-601.

- Kallianpur, G., and Selukar, R.S. (1989) - *Estimation of Hilbert space valued parameters by the method of sieves*, Technical Report nº 278, Center for Stochastic Processes, Univ. North Carolina, Chapel Hill, N.C. USA.

- Karr, A.F. (1987) - *Maximum likelihood estimation in the multiplicative intensity model via sieves*, Ann. Statist. 15, 473-490.

- Klonias, V.K. (1982) - *Consistency of two nonparametric maximum penalized likelihood estimators of the probability density function*, Ann. Statist. 10, 811-824.

- Kolmogorov, A.N., and Fomín, S.V. (1978) - *Elementos de la teoría de funciones y del análisis funcional*, Mir, Moscu.

- Kotz, S., and Johnson, N.L. (1988) - *Encyclopedia of Statistical Sciences*, Wiley, New York.

- Leonard, T. (1978) - *Density estimation, Stochastic process and prior information*, J. Roy. Statist. Soc. B, 40, 113-146.

- Lubecke, A.M., and Padgett, W.J. (1985) - *Nonparametric maximum penalized likelihood estimation of a density from arbitrarily right-censored observations*, Communications in Statistics, Theory and Methods 14, 257-271.

- Luenberguer, D.G. (1968) - *Optimization by Vector Space Methods*, Wiley, New York.

- Mckeague, I.W. (1986) - *Estimation for a semimartingale regression model using the method of sieves*, Ann. Statist. 14, 579-589.

- Nguyen, H.T., and Pham, T.D. (1982) - *Identification of nonstationary diffusion model by the method of sieves*, SIAM J. Control and Optimization 20, 603-611.

- O'Sullivan, F. (1983) - *The analysis of some penalized likelihood schemes*. Technical Report nº 726. University of Wisconsin.

- Padgett, W.J., and Mc Nichols, D.T. (1984) - *Nonparametric density estimation from censored data*, Communications in Statistics, Theory and Methods, 13, 1581-1611.

- Parthasarathy, K.R. (1967) - *Probability Measures on Metric Spaces*, Academic Press, New York.

- Rényi, A. (1970) - *Probability Theory*, North-Holland, Amsterdam.

- Royden, H.L. (1968) - *Real Analysis*, Macmillan, New York.

- Shiryayev, A.N. (1984) - *Probability*, Springer-Verlag, New York.

- Silverman, B.W. (1978) - *Density ratios, empirical likelihood and cot death*, Appl. Statist. 27, 26-33.