Uma introdução à estimação não-paramétrica da densidade

Carlos Tenreiro

CMUC, Departamento de Matemática, Universidade de Coimbra

Estimação funcional: generalidades

Estimação não-paramétrica da densidade

1956, Annals of Mathematical Statistics

832 MURRAY ROSENBLATT

REMARKS ON SOME NONPARAMETRIC ESTIMATES OF A DENSITY FUNCTION[1]

BY MURRAY ROSENBLATT[2]

University of Chicago

1. Summary. This note discusses some aspects of the estimation of the density function of a univariate probability distribution. All estimates of the density function satisfying relatively mild conditions are shown to be biased. The asymptotic mean square error of a particular class of estimates is evaluated.

Rosenblatt, 1956

3. The difference quotient of the sample distribution function. An obvious estimate of $f(y)$ is the difference quotient

$$S(y; X_1, \cdots, X_n) = f_n(y) = \frac{F_n(y+h) - F_n(y-h)}{2h}$$

of the sample distribution function $F_n(y)$, where $h = h_n$ is a function of the sample size n and approaches zero as $n \to \infty$. The asymptotic behavior of this estimate as $n \to \infty$ is examined in terms of its mean square error. Fix and Hodges have used an estimate of this form in their discussion of a nonparametric discrimination problem [1].

$\boxed{\text{Fix, E., Hodges, J.L., 1951} \quad \text{Akaike, H., 1954}}$

Rosenblatt, 1956

Estimação funcional
Estimação NP da densidade
Abordagens paramétricas
O estimador de Fourier de f
Modelos P e NP
Medidas de discrepância
O estimador da janela móvel
O papel de h_n
Não existência de estimadores cêntricos de f
Outros estimadores do núcleo
Aplicação aos testes de ajustamento
O estimador do histograma
O estimador do núcleo
Bibliografia

4. A class of estimates of the density function. The discussion of the previous section suggests that the following class of estimates will be of interest. Let $w_n(u)$ be a nonnegative function such that

$$\int_{-\infty}^{\infty} w_n(u)\, du = 1.$$

The sequence of functions $\{w_n(u)\}$ is chosen so that the total mass concentrates in the neighborhood of zero as $n \to \infty$; that is, given any $\epsilon > 0$,

$$\int_{|u|<\epsilon} w_n(u)\, du \to 1$$

as $n \to \infty$. Corresponding to each sequence of weight functions $\{w_n(u)\}$ of this type, there is an estimate

$$\boxed{f_n(y) = \int_{-\infty}^{\infty} w_n(y-u)\, dF_n(u) = \frac{1}{n}\sum_{j=1}^{\infty} w_n(y - X_j).}$$

Rosenblatt, 1956

Estimação funcional
Estimação NP da densidade
Abordagens paramétricas
O estimador de Fourier de f
Modelos P e NP
Medidas de discrepância
O estimador da janela móvel
O papel de h_n
Não existência de estimadores cêntricos de f
Outros estimadores do núcleo
Aplicação aos testes de ajustamento
O estimador do histograma
O estimador do núcleo
Bibliografia

Note that all estimates of this form are themselves density functions; that is'

$$f_n(y) \geqq 0,$$

and

$$\int_{-\infty}^{\infty} f_n(y)\, dy = 1.$$

An estimate $f_n(y)$ with any desired regularity properties can be obtained by choosing a weight function $w_n(u)$ with these same regularity properties. Thus, $f_n(y)$ will be analytic if $w_n(u)$ is.

As an example, consider

$$w_n(u) = \frac{1}{h} w\left(\frac{u}{h}\right),$$

where $h = h_n \to 0$ as $n \to \infty$, and

$$\int_{-\infty}^{\infty} w(u)\, du = 1.$$

XVIII Congresso da SPE, S. Pedro do Sul 29 de Setembro de 2010

Estimação não-paramétrica da densidade

Estimação funcional
Estimação NP da densidade
Abordagens paramétricas
O estimador de Fourier de f
Modelos P e NP
Medidas de discrepância
O estimador da janela móvel
O papel de h_n
Não existência de estimadores cêntricos de f
Outros estimadores do núcleo
Aplicação aos testes de ajustamento
O estimador do histograma
O estimador do núcleo
Bibliografia

- Observamos n cópias independentes

$$X_1, \ldots, X_n$$

duma variável real X com distribuição de densidade desconhecida f:

$$\mathrm{P}(a < X < b) = \int_a^b f(x)\, dx$$

para todo $-\infty < a < b < +\infty$.

- Pretendemos estimar a densidade f através dum estimador f_n:

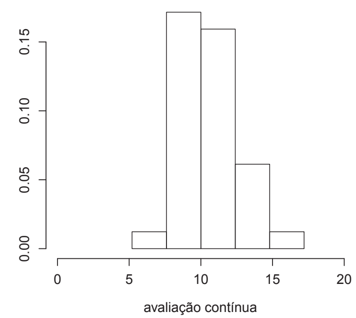
$$f_n(x) = f_n(x; X_1, \ldots, X_n)$$

- Não é assumida qualquer forma funcional para f.

XVIII Congresso da SPE, S. Pedro do Sul 29 de Setembro de 2010

Estimação não-paramétrica da densidade

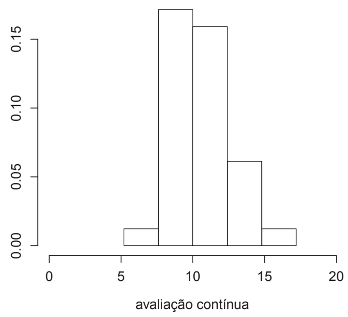
Estimação funcional
Estimação NP da densidade
Abordagens paramétricas
O estimador de Fourier de f
Modelos P e NP
Medidas de discrepância
O estimador da janela móvel
O papel de h_n
Não existência de estimadores cêntricos de f
Outros estimadores do núcleo
Aplicação aos testes de ajustamento
O estimador do histograma
O estimador do núcleo
Bibliografia

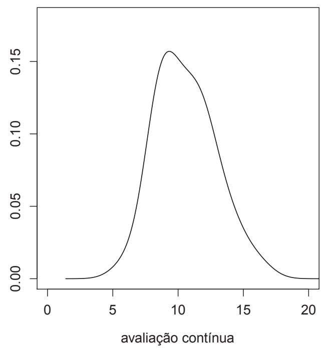


- O histograma desperdiça informação, substituindo cada observação pelo ponto médio da classe.
- Apesar da densidade subjacente ser assumida contínua, o histograma não reflecte esta característica.
- O histograma depende fortemente do tamanho escolhido para as classes e também da posição inicial dessas classes.

XVIII Congresso da SPE, S. Pedro do Sul 29 de Setembro de 2010

Estimação não-paramétrica da densidade





- ☐ Rosenblatt (1956) e Parzen (1962) desenvolvem uma abordagem alternativa introduzindo o estimador núcleo.
- ☐ As duas primeiras dificuldades anteriores são ultrapassadas.
- ☐ Sobra a última ...

O estimador do núcleo

- ☐ O estimador do núcleo depende de dois "parâmetros" a fixar pelo utilizador:
  - O núcleo K: $\int K(x)dx = 1$
  - A janela h: $h = h_n$
- ☐ Estimador de Parzen-Rosenblatt:

$$\boxed{\begin{aligned} f_n(x) &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i) \end{aligned}}$$

onde

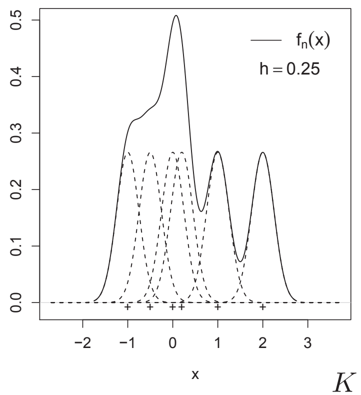
$$K_h(x) = K(x/h)/h$$

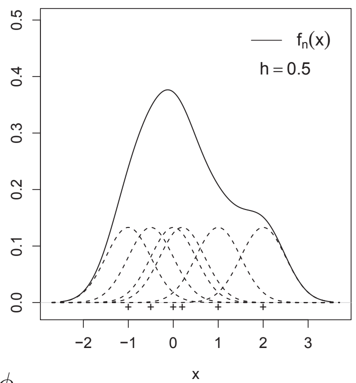
O estimador do núcleo

- ☐ A função

$$x \to K_h(x - X_i)$$

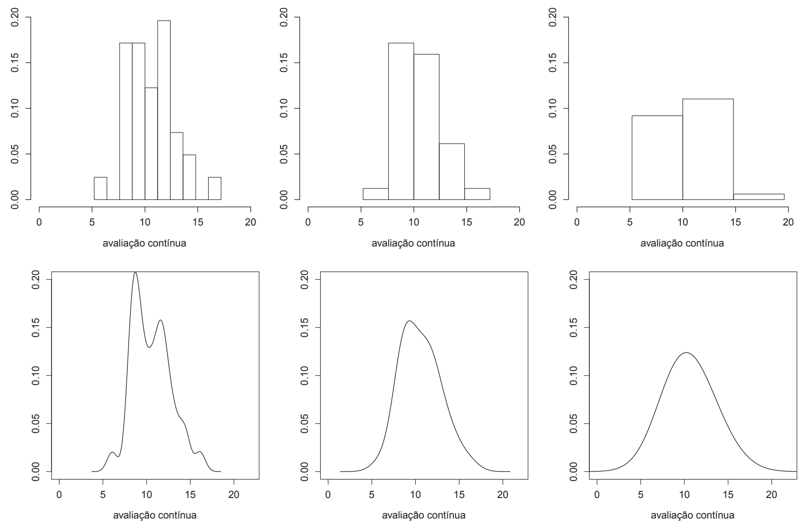
representa a contribuição da observação X_i para o estimador.





$K = \phi$

Influência da janela h_n



Abordagens paramétricas e não-paramétricas

- Abordagem não-paramétrica:
  - Não é assumida qualquer forma funcional para f.
  - A ideia de base da inferência não-paramétrica é a de impor condições o menos restritivas possível sobre f.
- Abordagem paramétrica:
  - Assumimos que f pertence a uma família paramétrica de distribuições, como a normal, a lognormal ou a gama.
  - Estimamos os parâmetros desconhecidos usando, por exemplo, o estimador da máxima verosimilhança.

Modelo normal

Estimação funcional
Estimação NP da densidade
▷ Abordagens paramétricas
O estimador de Fourier de f
Modelos P e NP
Medidas de discrepância
O estimador da janela móvel
O papel de h_n
Não existência de estimadores cêntricos de f
Outros estimadores do núcleo
Aplicação aos testes de ajustamento
O estimador do histograma
O estimador do núcleo
Bibliografia

$$\boxed{f(\cdot) = \phi_\sigma(\cdot - \mu) \text{ para algum } \mu \in \mathbb{R} \text{ e } \sigma > 0.}$$

- Estimador paramétrico:

$$f_n(x) = \phi_{\hat{s}}(x - \bar{x})$$

com $\bar{x} = \frac{1}{n}\sum_{i=1}^n X_i$ e $\hat{s}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$.

- Se μ e $\sigma^2 > 0$ são as verdadeiras média e variância de X então

$$f_n(x) \xrightarrow{p} \phi_\sigma(x - \mu)$$

- $f_n(x) \xrightarrow{p} f(x)$ apenas quando f é uma densidade normal.

XVIII Congresso da SPE, S. Pedro do Sul 29 de Setembro de 2010

O sistema de Pearson (1895)

Estimação funcional
Estimação NP da densidade
▷ Abordagens paramétricas
O estimador de Fourier de f
Modelos P e NP
Medidas de discrepância
O estimador da janela móvel
O papel de h_n
Não existência de estimadores cêntricos de f
Outros estimadores do núcleo
Aplicação aos testes de ajustamento
O estimador do histograma
O estimador do núcleo
Bibliografia

- Uma densidade de probabilidade pertence ao sistema de Pearson se é solução da equação diferencial

$$\frac{d(\log f(x))}{dx} = \frac{x - a}{b_0 + b_1 x + b_2 x^2}.$$

- Ao sistema de Pearson pertencem as distribuições normal, beta, gama, Student,...
- Os parâmetros a, b_0, b_1, b_2 podem ser expresso em termos dos primeiros quatro momentos de f podendo ser estimados pelo método dos momentos:

$$a = b_1 = -\mu_3(\mu_4 + 3\mu_2^2)/A, \quad b_0 = -\mu_2(2\mu_2\mu_4 - 3\mu_3^2)/A,$$

$$b_2 = -(2\mu_2\mu_4 - 3\mu_3^2 - 6\mu_2^3)/A, \quad A = 10\mu_2\mu_4 - 18\mu_2^3 - 12\mu_3^2$$

XVIII Congresso da SPE, S. Pedro do Sul 29 de Setembro de 2010

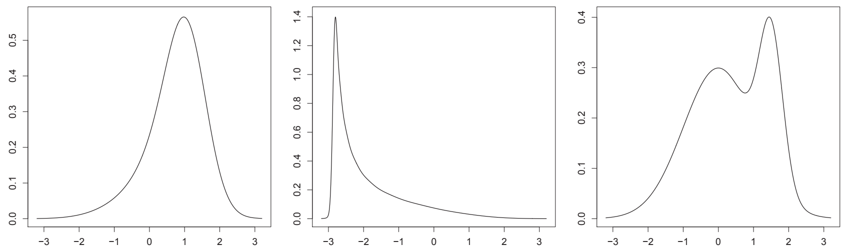
Estimação da densidade via modelos de mistura

Estimação funcional
Estimação NP da densidade
▷ Abordagens paramétricas
O estimador de Fourier de f
Modelos P e NP
Medidas de discrepância
O estimador da janela móvel
O papel de h_n
Não existência de estimadores cêntricos de f
Outros estimadores do núcleo
Aplicação aos testes de ajustamento
O estimador do histograma
O estimador do núcleo
Bibliografia

- Fraley e Raftery (2002) propõem ajustar aos dados cada um dos modelos

$$f(x) = \sum_{i=1}^{G} \alpha_i \phi_{\sigma_i}(x - \mu_i),$$

com $G = 1, \ldots, K$.



- R: Package 'mclust' (2010)

XVIII Congresso da SPE, S. Pedro do Sul 29 de Setembro de 2010

Estimação da densidade por projecção

- Suponhamos que f está definida no intervalo $[0,1]$.
- Consideremos as funções trigonométricas

$$q_0(x) = 1; \quad q_{2k-1}(x) = \sqrt{2}\cos(2\pi kx); \quad q_{2k}(x) = \sqrt{2}\sin(2\pi kx),$$

$k = 1, 2, \ldots$ que formam uma base de $L^2([0,1])$.

- Qualquer densidade de quadrado integrável pode ser representada na forma

$$f(x) = \sum_{j=0}^{\infty} \theta_j q_j(x),$$

onde o coeficiente de Fourier θ_j é dado por

$$\theta_j = \int_0^1 f(x) q_j(x) dx = \mathrm{E}(q_j(X)).$$

Estimação da densidade por projecção

Estimação funcional
Estimação NP da densidade
Abordagens paramétricas
▷ O estimador de Fourier de f
Modelos P e NP
Medidas de discrepância
O estimador da janela móvel
O papel de h_n
Não existência de estimadores cêntricos de f
Outros estimadores do núcleo
Aplicação aos testes de ajustamento
O estimador do histograma
O estimador do núcleo
Bibliografia

- O coeficiente de Fourier $\theta_j = \mathrm{E}(q_j(X))$ pode ser estimado sem-viés por

$$\hat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} q_j(X_i).$$

- Um possível estimador para f é dado pela série truncada

$$\boxed{f_n(x) = \sum_{j=0}^{m} \hat{\theta}_j q_j(x)}$$

onde $m = m(n) \to \infty$, dito estimador de Fourier de f.

(Cencov, 1962)

XVIII Congresso da SPE, S. Pedro do Sul 29 de Setembro de 2010

Modelos paramétricos e não-paramétricos

- Um modelo para a estimação de f consiste na introdução de uma restrição da forma

$$f \in \mathscr{D},$$

onde $\mathscr{D}$ é uma família de densidades de probabilidade.

- Quando $\mathscr{D}$ pode ser indexada por um número finito de números reais, dizemos que o modelo é paramétrico.

  A restrição $f \in \mathscr{D}$ fixa uma forma funcional para f.

- Caso contrário, quando a família $\mathscr{D}$ é demasiado vasta para ser indexada por um número finito de números reais, o modelo diz-se não-paramétrico.

  A restrição $f \in \mathscr{D}$ traduz-se na especificação de condições de regularidade sobre f.

XVIII Congresso da SPE, S. Pedro do Sul 29 de Setembro de 2010

Medidas da qualidade dum estimador

- Uma medida local da qualidade dum estimador $f_n(x)$ de $f(x)$ é o erro quadrático médio definido por

$$\begin{aligned}\mathrm{EQM}(f_n(x)) &= \mathrm{E}\{f_n(x) - f(x)\}^2 \\ &= \mathrm{E}\{f_n(x) - \mathrm{E}f_n(x)\}^2 + \{\mathrm{E}f_n(x) - f(x)\}^2 \\ &= \mathrm{Var} f_n(x) + \mathrm{Viés} f_n(x)^2.\end{aligned}$$

- A mais usada medida global da qualidade de f_n como estimador de f é o erro quadrático médio integrado definido por

$$\begin{aligned}\mathrm{EQMI}(f_n) &= \mathrm{E}\int \{f_n(x) - f(x)\}^2 dx \\ &= \int \mathrm{Var} f_n(x)\, dx + \int \mathrm{Viés} f_n(x)^2 dx \\ &= \mathrm{IVAR}(f_n) + \mathrm{IVIES}(f_n).\end{aligned}$$

XVIII Congresso da SPE, S. Pedro do Sul 29 de Setembro de 2010

Medidas da qualidade dum estimador

☐ Outras medidas globais da qualidade do estimador f_n:

- Distância L_∞ (ou do supremo) entre f_n e f:

$$||f_n - f||_\infty = \sup_{-\infty<x<+\infty} |f_n(x) - f(x)|$$

- Distância L_2 entre f_n e f:

$$||f_n - f||_2 = \left(\int \{f_n(x) - f(x)\}^2 dx\right)^{1/2}$$

- Distância L_1 entre f_n e f:

$$||f_n - f||_1 = \int |f_n(x) - f(x)| dx$$

O estimador da janela móvel

☐ O estimador do núcleo proposto por Rosenblatt (1956) é uma extensão do estimador da janela móvel previamente considerado por Fix e Hodges (1951) e Akaike (1954) no contexto da análise discriminante não-paramétrica:

$$f_n(x) = \frac{F_n(x+h_n) - F_n(x-h_n)}{2h_n}$$

onde

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x).$$

☐ f_n é o estimador do núcleo associado ao núcleo uniforme:

$$K(u) = \frac{1}{2} I(|u| \le 1)$$

O estimador da janela móvel

☐ Como

$$2nh_n f_n(x) \sim B\Big(n, F(x+h_n) - F(x-h_n)\Big)$$

concluímos que

$$\text{Viés} f_n(x) = q_n(x) - f(x)$$

$$\text{Var} f_n(x) = \frac{1}{2nh_n} q_n(x)(1 - 2h_n q_n(x)),$$

onde

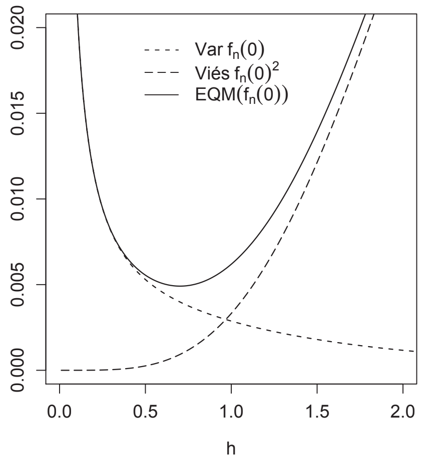
$$q_n(x) = \frac{F(x+h_n) - F(x-h_n)}{2h_n} \to f(x), \text{ se } h_n \to 0$$

Se f contínua em x, e se $h_n \to 0$ e $nh_n \to +\infty$, então

$$\text{EQM}(f_n(x)) \to 0$$

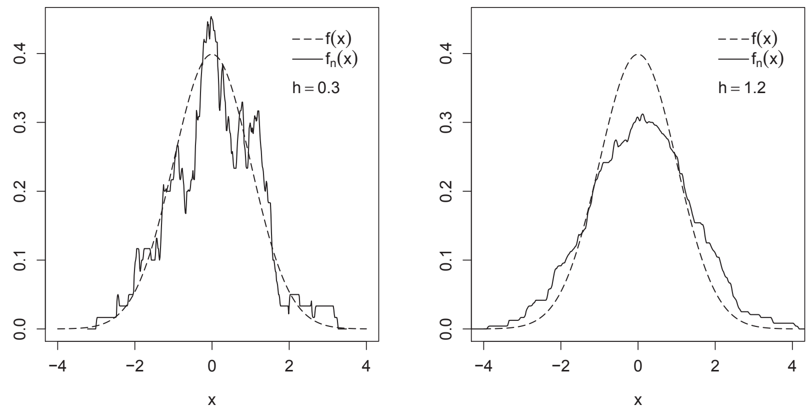
O papel do parâmetro de suavização h_n

☐ f é a densidade normal standard, $x = 0$ e $n = 100$.



O papel do parâmetro de suavização h_n

☐ f é a densidade normal standard, $x = 0$ e $n = 100$.



Não existência de estimadores cêntricos de f

Dado $n \in \mathbb{N}$, não existe um estimador não-negativo $\gamma_n(\cdot; X_1, \ldots, X_n)$ de f tal que

$$\mathrm{E}\gamma_n(x; X_1, \ldots, X_n) = f(x),$$

para todo o $f \in \mathscr{C}$ e $x \in \mathbb{R}$, onde $\mathscr{C}$ é a família das densidades de probabilidade contínuas em $\mathbb{R}$.

(Rosenblatt, 1956)

☐ O problema do viés do estimador é dos problemas mais importantes na estimação não-paramétrica da densidade.

☐ Será possível controlar ou reduzir este viés?

Estimador de Nadaraya-Watson

☐ Função de regressão: $r(x) = \mathrm{E}(Y|X = x)$

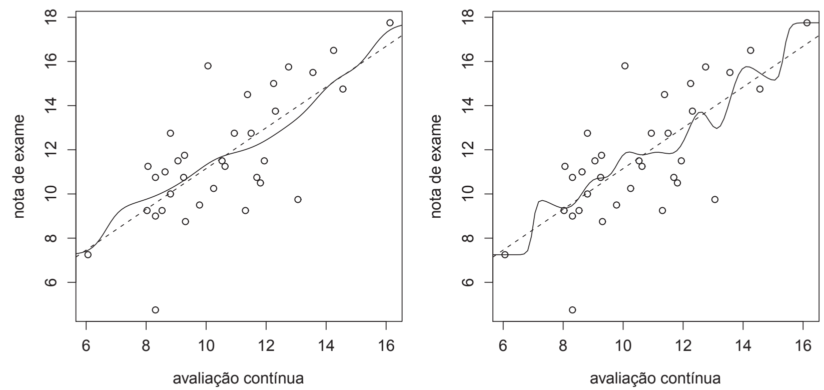
☐ Nadaraya 1964, Watson 1964:

$$\begin{aligned} r_n(x) &= \sum_{i=1}^{n} Y_i K_h(x - X_i) \Big/ \sum_{i=1}^{n} K_h(x - X_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} Y_i \, W_{i,h}(x) \end{aligned}$$

com

$$W_{i,h}(x) = \frac{K_h(x - X_i)}{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)}$$

Estimador de Nadaraya-Watson



$K = \phi$

Estimação da função de distribuição e da taxa de quebra

- Função de distribuição: $F(x) = \int_{-\infty}^{x} f(u)du$
- Taxa de quebra (failure rate, hazard function): $t(x) = f(x)/(1 - F(x))$
- Tiago de Oliveira 1963, Watson e Leadbetter 1964, Nadaraya 1964:

$$F_n(x) = \int_{-\infty}^{x} f_n(u)du = \frac{1}{n}\sum_{i=1}^{n} \bar{K}_h(x - X_i),$$

onde

$$\bar{K}(u) = \int_{-\infty}^{u} K(y)dy.$$

Tiago de Oliveira, 1963

Estimação funcional
Estimação NP da densidade
Abordagens paramétricas
O estimador de Fourier de f
Modelos P e NP
Medidas de discrepância
O estimador da janela móvel
O papel de h_n
Não existência de estimadores cêntricos de f
▷ Outros estimadores do núcleo
Aplicação aos testes de ajustamento
O estimador do histograma
O estimador do núcleo
Bibliografia

J. TIAGO DE OLIVEIRA

ESTATÍSTICA DE DENSIDADES

RESULTADOS ASSINTÓTICOS

SEPARATA DA REVISTA DA FACULDADE DE CIÊNCIAS DE LISBOA
2.ª Série — A — Vol. IX — Fasc. 1.º — Págs. 111 a 206

LISBOA
1963

Tiago de Oliveira, 1963

Estimação funcional
Estimação NP da densidade
Abordagens paramétricas
O estimador de Fourier de f
Modelos P e NP
Medidas de discrepância
O estimador da janela móvel
O papel de h_n
Não existência de estimadores cêntricos de f
▷ Outros estimadores do núcleo
Aplicação aos testes de ajustamento
O estimador do histograma
O estimador do núcleo
Bibliografia

II.2—**Estimação de funções de distribuição por médias.**

Vamos estudar o problema da estimação de funções de distribuição, de modo análogo ao que já tratámos na estimação de densidades.

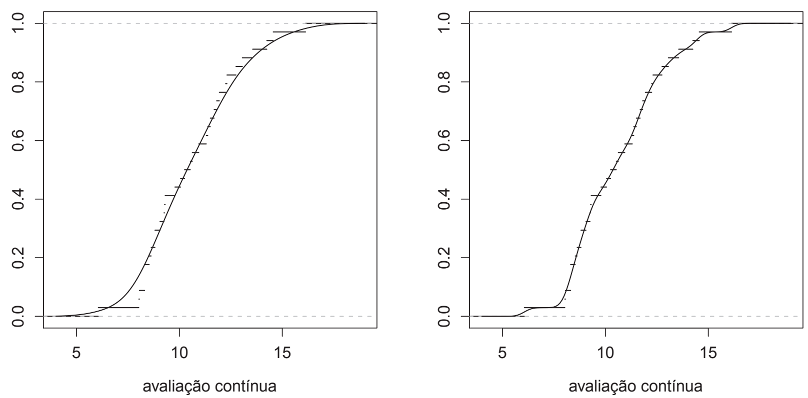
Seja K uma função de distribuição (contínua) e consideremos a sucessão de estatisticas

$$\boxed{F_n^*(x) = \frac{1}{n}\sum_i K\left(\frac{x - X_i}{\delta_n}\right)}$$

para x fixado e em que $\delta_n \to 0$. É óbvio que $K\left(\frac{x-y}{\delta_n}\right)$ converge completamente para $H(x-y)$, em que H é a função de distribuição da variável aleatória certa colocada na origem das coordenadas (função de salto de Heaviside).

Estimador do núcleo da FD

Estimação funcional
Estimação NP da densidade
Abordagens paramétricas
O estimador de Fourier de f
Modelos P e NP
Medidas de discrepância
O estimador da janela móvel
O papel de h_n
Não existência de estimadores cêntricos de f
▷ Outros estimadores do núcleo
Aplicação aos testes de ajustamento
O estimador do histograma
O estimador do núcleo
Bibliografia



$K = \phi$

Estimação de outros parâmetros funcionais e aplicações

Estimação funcional
Estimação NP da densidade
Abordagens paramétricas
O estimador de Fourier de f
Modelos P e NP
Medidas de discrepância
O estimador da janela móvel
O papel de h_n
Não existência de estimadores cêntricos de f
▷ Outros estimadores do núcleo
Aplicação aos testes de ajustamento
O estimador do histograma
O estimador do núcleo
Bibliografia

- Derivadas da densidade

$$f_n^{(k)}(x) = \frac{1}{nh^{k+1}} \sum_{i=1}^{n} K^{(k)}\left(\frac{x - X_i}{h}\right)$$

- Derivadas da regressão
- Funcionais da densidade e regressão e suas derivadas
- Função quantil
- Densidade condicional
- Análise discriminante não-paramétrica
- Bootstrap suave
- Testes de ajustamento

Aplicação aos testes de ajustamento

Estimação funcional
Estimação NP da densidade
Abordagens paramétricas
O estimador de Fourier de f
Modelos P e NP
Medidas de discrepância
O estimador da janela móvel
O papel de h_n
Não existência de estimadores cêntricos de f
Outros estimadores do núcleo
▷ Aplicação aos testes de ajustamento
O estimador do histograma
O estimador do núcleo
Bibliografia

- Observamos n cópias independentes da variável X com distribuição de densidade f desconhecida:

$$X_1, \ldots, X_n$$

- Pretendemos testar

$$\boxed{H_0 : f = f_0 \quad \text{contra} \quad H_1 : f \neq f_0}$$

onde f_0 é uma densidade de probabilidade fixa à partida.

- O teste do qui-quadrado pode ser visto como um teste baseado do estimador do histograma:

$$\chi_n^2 = n \int \frac{\{\hat{f}_n(x) - \mathrm{E}_0 \hat{f}_n(x)\}^2}{\mathrm{E}_0 \hat{f}_n(x)}\, dx$$

Aplicação aos testes de ajustamento

Estimação funcional
Estimação NP da densidade
Abordagens paramétricas
O estimador de Fourier de f
Modelos P e NP
Medidas de discrepância
O estimador da janela móvel
O papel de h_n
Não existência de estimadores cêntricos de f
Outros estimadores do núcleo
▷ Aplicação aos testes de ajustamento
O estimador do histograma
O estimador do núcleo
Bibliografia

- Bickel e Rosenblatt (1973) propõem estatísticas de teste baseadas no estimador do núcleo da densidade:

$$\boxed{\begin{array}{c} n \int \{f_n(x) - f_0(x)\}^2 dx \\[2ex] \text{e} \quad n \int \{f_n(x) - \mathrm{E}_0 f_n(x)\}^2 dx \end{array}}$$

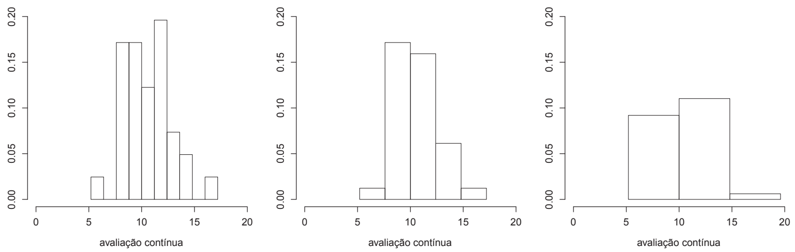
discrepância "pequena" → Aceitar H_0

discrepância "grande" → Rejeitar H_0

Estimação funcional
▷ O estimador do histograma
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Influência da origem da partição
Escolha prática da janela
O polígono de frequências
O estimador do núcleo
Bibliografia

O estimador do histograma

O estimador do histograma



- Definição do estimador
- Propriedades locais e globais
- A influência da origem da partição
- Escolha prática da janela
- O polígono de frequências

O estimador do histograma

□ Δ_n é uma partição de $\mathbb{R}$ definida a partir de uma origem $a_{n,0}$ por

$$\Delta_{n,j} =]a_{n,j}, a_{n,j+1}],$$

onde

$$a_{n,j+1} = a_{n,j} + h,\ j \in \mathbb{Z},$$

e

$$h = h_n > 0.$$

□ Para $x \in \Delta_{n,j}$ o estimador do histograma é definido por

$$\hat{f}_n(x) = \frac{F_n(a_{n,j+1}) - F_n(a_{n,j})}{h} = \frac{1}{nh}\sum_{i=1}^{n} I(X_i \in \Delta_{n,j})$$

O estimador do histograma

□ Para $x \in \mathbb{R}$,

$$\hat{f}_n(x) = \sum_{j\in\mathbb{Z}} \frac{\nu_j}{nh} I(x \in \Delta_{n,j})$$

com

$$\nu_j = \nu_{n,j} = \sum_{i=1}^{n} I(X_i \in \Delta_{n,j}) \sim B(n, p_j)$$

e

$$p_j = p_{n,j} = \int_{\Delta_{n,j}} f(y)dy.$$

□ O estimador do histograma é um estimador próprio da densidade (bona fide density estimator):

- $\hat{f}_n(x) \geq 0$
- $\int \hat{f}_n(x)dx = 1$

Viés do estimador

□ As propriedades do histograma variam em função do intervalo da partição.

□ Para $x \in \Delta_{n,j}$,

$$\mathrm{E}\hat{f}_n(x) = \frac{\mathrm{E}(\nu_j)}{nh} = \frac{p_j}{h} = \frac{1}{h}\int_{\Delta_{n,j}} f(y)dy$$

$$\mathrm{Vi\acute{e}s}\hat{f}_n(x) = \frac{1}{h}\int_{\Delta_{n,j}} \{f(y) - f(x)\}dy$$

Se f é contínua em $x \in \mathbb{R}$ e $h = h_n \to 0$ então $\hat{f}_n$ é assintoticamente sem-viés:

$$\mathrm{E}\hat{f}_n(x) \to f(x).$$

Viés do estimador

- Suponhamos f verifica uma condição de Lipschitz de ordem $\alpha \in]0,1]$ numa vizinhança de $x \in \mathbb{R}$: existem $\delta > 0$ e $M > 0$ (possivelmente dependentes de x) tais que

$$|f(y) - f(z)| \leq M|y-z|^\alpha,$$

para todo o $y, z \in]x-\delta, x+\delta[$.

- Para n suficientemente grande

$$|\text{Viés}\hat{f}_n(x)| \leq \frac{1}{h}\int_{\Delta_{n,j}} |f(y)-f(x)|dy \leq Mh^\alpha,$$

Para $x \in \mathbb{R}$, se $f \in \mathscr{L}_x(\alpha)$ e $h_n \to 0$ então

$$\text{Viés}\hat{f}_n(x) = O\left(h_n^\alpha\right).$$

Fórmula de Taylor

Seja g uma função real de variável real admitindo derivada até à ordem $p-1$ numa vizinhança do ponto $x \in \mathbb{R}$. Então, sendo u_n uma qualquer sucessão de números reais convergente para zero temos

$$g(x+u_n) = \sum_{\ell=0}^{p-1} \frac{u_n^\ell}{\ell!} g^{(\ell)}(x) + R_n(x),$$

onde

$$R_n(x) = o(u_n^{p-1}) \qquad \text{[resto infinitesimal]}$$

Além disso, se g admite derivada de ordem p contínua numa vizinhança do ponto x temos

$$R_n(x) = u_n^p \int_0^1 \frac{(1-t)^{p-1}}{(p-1)!} g^{(p)}(x+tu_n)dt \qquad \text{[resto integral]}$$

Viés do estimador

$$\text{Viés}\hat{f}_n(x) = \frac{1}{h}\int_{\Delta_{n,j}} \{f(y)-f(x)\}dy$$

$$f(y) = f(x) + (y-x)\int_0^1 f'(x+t(y-x))dt$$

Para $x \in \mathbb{R}$, se $f' \in \mathscr{L}_x(\alpha)$ e $h_n \to 0$ então

$$\text{Viés}\hat{f}_n(x) = \frac{h_n}{2} f'(x)\left(1 - 2(x-a_{n,j_x})/h_n\right) + O\left(h_n^{1+\alpha}\right),$$

onde $]a_{n,j_x}, a_{n,j_x}+h_n]$ é o elemento da partição Δ_n que contém x.

Variância do estimador

- Para $x \in \Delta_{n,j}$,

$$\text{Var}\hat{f}_n(x) = \frac{\text{Var}(\nu_j)}{(nh)^2} = \frac{p_j(1-p_j)}{nh^2}$$

com

$$\frac{p_j}{h} \to f(x)$$

Se f é contínua em $x \in \mathbb{R}$ e $h_n \to 0$ então

$$nh_n \,\text{Var}\hat{f}_n(x) \to f(x).$$

Além disso, se $nh_n \to +\infty$ então

$$\text{Var}\hat{f}_n(x) \to 0.$$

Erro quadrático médio

Para $x \in \mathbb{R}$, se f é derivável com $f' \in \mathscr{L}_x(\alpha)$ então,

$$\mathrm{EQM}(\hat{f}_n(x)) = \frac{1}{nh_n} f(x) + \frac{h_n^2}{4} f'(x)^2 \left(1 - 2(x - a_{n,j_x})/h_n\right)^2 + O(n^{-1}) + O\left(h_n^{2+\alpha}\right).$$

- A maior ordem de convergência é obtida quando

$$h_n = c\, n^{-1/3}, \text{ com } c > 0,$$

caso em que

$$\hat{f}_n(x) - f(x) = O_p(n^{-1/3}).$$

Convergência L_∞

- Pretendemos obter condições sobre f e h_n que assegurem a convergência

$$||\hat{f}_n - f||_\infty \xrightarrow{qc} 0.$$

- Utilizamos a desigualdade

$$||\hat{f}_n - f||_\infty \leq ||\hat{f}_n - \mathrm{E}\hat{f}_n||_\infty + ||\mathrm{E}\hat{f}_n - f||_\infty$$

- Para $x \in \mathbb{R}$,

$$\mathrm{E}\hat{f}_n(x) - f(x) = \frac{1}{h} \int_{\Delta_{n,j_x}} \{f(y) - f(x)\} dy$$

- É necessário impor condições globais de regularidade sobre f.

Convergência L_∞

- Assumiremos que f é uniformemente contínua em $\mathbb{R}$, isto é, para todo o $\epsilon > 0$, existe $\delta > 0$ tal que

$$|f(x) - f(y)| < \epsilon,$$

para todos os números reais x, y que satisfazem

$$|x - y| < \delta.$$

Se $f \in \mathscr{U}$ e $h_n \to 0$ então

$$||\mathrm{E}\hat{f}_n - f||_\infty \xrightarrow{qc} 0.$$

Convergência L_∞

- Para $x \in \Delta_{n,j}$,

$$\hat{f}_n(x) = \frac{F_n(a_{n,j+1}) - F_n(a_{n,j})}{h}$$

- Assim

$$||\hat{f}_n - \mathrm{E}\hat{f}_n||_\infty \leq \frac{2}{h} ||F_n - F||_\infty$$

onde

$$\limsup \sqrt{\frac{2n}{\log\log n}}\, ||F_n - F||_\infty \leq 1\, q.c.$$

Se $f \in \mathscr{U}$, $h_n \to 0$ e $nh_n^2/\log\log n \to +\infty$ então

$$||\hat{f}_n - f||_\infty \xrightarrow{qc} 0.$$

Convergência L_∞

Estimação funcional
O estimador do histograma
Definição
Propriedades locais
▷ Propriedades globais
EQMI
Janela óptima
Influência da origem da partição
Escolha prática da janela
O polígono de frequências
O estimador do núcleo
Bibliografia

- Uma análise mais fina do termo do termo aleatório $||\hat{f}_n - \mathrm{E}\hat{f}_n||_\infty$ permite melhorar o resultado anterior.
- As condições anteriores sobre h_n podem ser ainda enfraquecidas:

Se $h_n \to 0$ então as seguintes proposições são equivalentes:

i) $nh_n / \log n \to +\infty$;

ii) $\forall\, f \in \mathscr{U} \quad ||\hat{f}_n - f||_\infty \xrightarrow{p} 0$;

iii) $\forall\, f \in \mathscr{U} \quad ||\hat{f}_n - f||_\infty \xrightarrow{qc} 0$.

(Geffroy, 1974)

Convergência L_1

Estimação funcional
O estimador do histograma
Definição
Propriedades locais
▷ Propriedades globais
EQMI
Janela óptima
Influência da origem da partição
Escolha prática da janela
O polígono de frequências
O estimador do núcleo
Bibliografia

- Pelo Teorema de Glick (1974) a convergência pontual quase certa ou em probabilidade dum estimador próprio da densidade implica a convergência L_1.
- É, no entanto, possível fazer bem melhor:

As proposições seguintes são equivalentes:

i) $h_n \to 0$, $nh_n \to +\infty$;

ii) $\forall\, f \in \mathscr{F} \quad ||\hat{f}_n - f||_1 \xrightarrow{p} 0$;

iii) $\forall\, f \in \mathscr{F} \quad ||\hat{f}_n - f||_1 \xrightarrow{qc} 0$.

(Abou-Jaoudé, 1976)

EQMI

Estimação funcional
O estimador do histograma
Definição
Propriedades locais
Propriedades globais
▷ EQMI
Janela óptima
Influência da origem da partição
Escolha prática da janela
O polígono de frequências
O estimador do núcleo
Bibliografia

- A variância e o viés globais do histograma podem ser facilmente quantificadas a partir das medidas

$$\mathrm{IVAR}(\hat{f}_n) = \int \mathrm{Var}\hat{f}_n(x)\,dx$$

$$\mathrm{IVIES}(\hat{f}_n) = \int \mathrm{Viés}\hat{f}_n(x)^2 dx$$

$$\begin{aligned}\mathrm{EQMI}(\hat{f}_n) &= \mathrm{E}\int \{\hat{f}_n(x) - f(x)\}^2 dx \\ &= \mathrm{IVAR}(\hat{f}_n) + \mathrm{IVIES}(\hat{f}_n)\end{aligned}$$

EQMI e EQMIA

Estimação funcional
O estimador do histograma
Definição
Propriedades locais
Propriedades globais
▷ EQMI
Janela óptima
Influência da origem da partição
Escolha prática da janela
O polígono de frequências
O estimador do núcleo
Bibliografia

Suponhamos que f possui derivada de terceira ordem, com f'' absolutamente contínua e f, f', f'', f''' de quadrado integrável. Se $h_n \to 0$ então

$$\mathrm{IVAR}(\hat{f}_n) = \frac{1}{nh_n} + O(n^{-1}),$$

$$\mathrm{IVIES}(\hat{f}_n) = \frac{h_n^2}{12} R(f') + O\left(h_n^3\right)$$

e

$$\mathrm{EQMI}(\hat{f}_n) = \boxed{\frac{1}{nh_n} + \frac{h_n^2}{12} R(f')} + O(n^{-1}) + O\left(h_n^3\right).$$

EQMIA

(Freedman e Diaconis, 1981)

Janela assintoticamente óptima

Se f é tal que $R(f') > 0$, o valor de h_n que minimiza o EQMIA é dado por

$$h_{\text{EQMIA}} = \big(6/R(f')\big)^{1/3}\, n^{-1/3}.$$

☐ Nas condições do resultado anterior, a maior ordem de convergência para zero do erro quadrático médio integrado é obtida quando tomamos

$$h_n = c\,n^{-1/3}, \text{ com } c > 0,$$

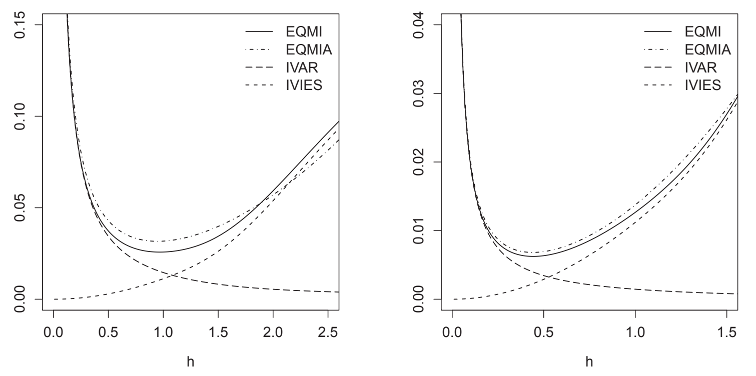
obtendo-se neste caso

$$\text{EQMI}(\hat{f}_n) = O(n^{-2/3}).$$

EQMI vs. EQMIA

N(0,1)

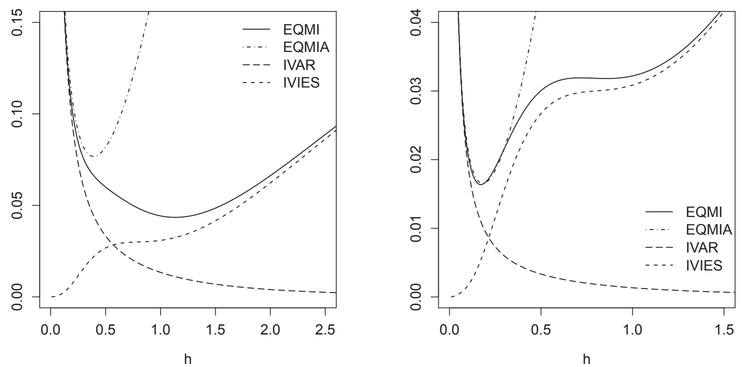
$n = 50$ $n = 500$



EQMI vs. EQMIA

LN(0,1)

$n = 50$ $n = 500$

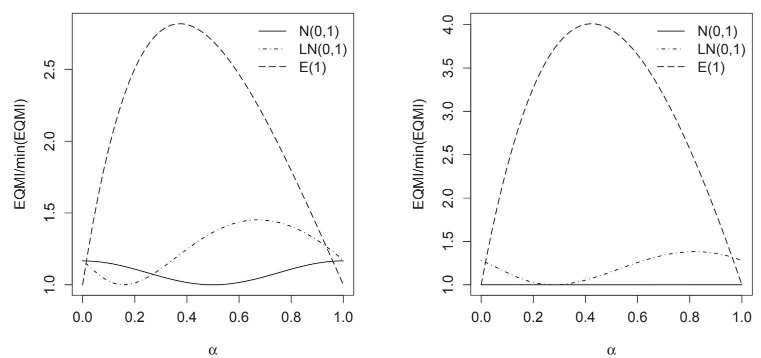


Influência da origem da partição

☐ $h = h_{\text{EQMIA}}$

☐ $a_{n,0} = \alpha\, h$, $\alpha \in [0, 1[$.

$n = 10$ $n = 50$



Influência da origem da partição

- No caso da densidade exponencial $E(1)$, se

$$a_{n,0} = \alpha\, h,\ \alpha \in]0,1[$$

o histograma não é assintoticamente sem-viés em $x = 0$:

$$\mathrm{E}\hat{f}_n(x) = \frac{1}{h}\int_0^{\alpha h} e^{-t}dt = \frac{1-e^{-\alpha h}}{h} \to \alpha$$

- Além disso,

$$\boxed{\mathrm{IVIES}(\hat{f}_n) = \alpha(\alpha-1)h_n + O(h_n^2)}$$

Escolha prática de h_n

- Uma das mais conhecidas regras práticas para a escolha do tamanho da janela num histograma foi proposta por Sturges (1926):

$$\hat{h}_{\mathrm{ST}} = \frac{\text{amplitude amostral}}{1+\log_2 n}.$$

- Apesar do seu valor histórico, esta regra não é baseada num critério que tenha em conta a qualidade do histograma como estimador da densidade desconhecida f.

- Tendo em conta o que vimos atrás, esta janela conduzirá a um estimador com pouca variabilidade mas com viés elevado que produzirá estimativas demasiadamente suaves (*oversmoothing*).

Janela com distribuição de referência normal

- Para uma vasta classe de densidades conhecemos a janela óptima no sentido da minimização do erro quadrático médio integrado assintótico:

$$h_{\mathrm{EQMIA}} = \left(6/R(f')\right)^{1/3}\, n^{-1/3}.$$

- Scott (1979) propõe que se utilize a regra anterior onde $R(f')$ é calculada tomando para f a densidade da distribuição $N(0,\sigma^2)$:

$$h_{\mathrm{EQMIA}} = 3.49\,\sigma\, n^{-1/3}.$$

$$\downarrow$$

$$\boxed{\hat{h}_{\mathrm{NR}} = 3.49\,\hat{s}\, n^{-1/3}}$$

Janela com distribuição de referência normal

- No caso da distribuição normal com função de distribuição F

$$\sigma = \frac{F^{-1}(3/4) - F^{-1}(1/4)}{F^{-1}_{N(0,1)}(3/4) - F^{-1}_{N(0,1)}(1/4)} \approx \frac{\mathrm{A\hat{I}Q}}{1.349}$$

- Silverman (1986) propõe que o desvio-padrão σ seja estimado por

$$\sigma \approx \min\left(\hat{s}\,,\ \frac{\mathrm{A\hat{I}Q}}{1.349}\right)$$

$$\downarrow$$

$$\boxed{\hat{h}_{\mathrm{SIL}} = \min(3.49\,\hat{s}, 2.59\,\mathrm{A\hat{I}Q})\, n^{-1/3}}$$

O método de validação cruzada

- O método de validação cruzada (*least-square cross-validation*) é proposto por Rudemo (1982) e Bowman (1984).
- O erro quadrático integrado, EQI, é tomado como medida da discrepância entre $\hat{f}_n$ e a densidade desconhecida f:

$$\mathrm{EQI}(h) = ||\hat{f}_n - f||_2^2 = \int \{\hat{f}_n(x) - f(x)\}^2 dx$$

- A ideia é escolher a janela h que minimiza $\mathrm{EQI}(h)$...
- Mas $\mathrm{EQI}(h)$ depende de f ...

$$\mathrm{EQI}(h) \;=\; R(\hat{f}_n) - 2\int \hat{f}_n(x) f(x)\,dx + R(f)$$

?

O método de validação cruzada

Estimação funcional
O estimador do histograma
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Influência da origem da partição
▷ Escolha prática da janela
O polígono de frequências
O estimador do núcleo
Bibliografia

$$\mathrm{EQI}(h) \;=\; R(\hat{f}_n) - \boxed{2\int \hat{f}_n(x) f(x)\,dx} + \boxed{R(f)}$$

$$\downarrow$$

$$\mathrm{E}(\hat{f}_n(X))$$

$$\downarrow$$

$$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{n,-i}(X_i)$$

- Na prática toma-se para janela o valor $\hat{h}_{\mathrm{CV}}$ que minimiza

$$\boxed{\mathrm{CV}(h) = R(\hat{f}_n) - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{n,-i}(X_i)}$$

O método de validação cruzada

Estimação funcional
O estimador do histograma
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Influência da origem da partição
▷ Escolha prática da janela
O polígono de frequências
O estimador do núcleo
Bibliografia

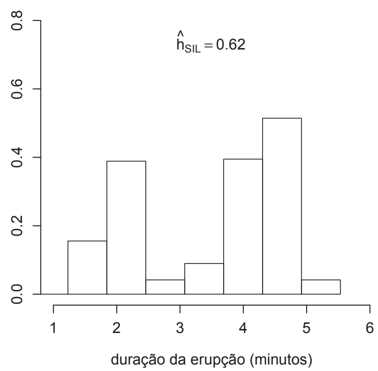
- Contrariamente ao método das distribuições de referência, o método de validação cruzada produz janelas assintoticamente "boas" independentemente da distribuição dos dados.

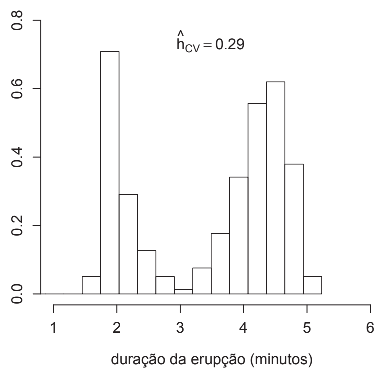
> Sob condições pouco restritivas sobre o f, $\hat{h}_{\mathrm{CV}}$ é assintoticamente óptima no sentido em que
>
> $$\frac{\mathrm{EQI}(\hat{h}_{\mathrm{CV}})}{\mathrm{EQI}(\hat{h}_{\mathrm{EQI}})} \xrightarrow{qc} 1,$$
>
> onde
>
> $$\hat{h}_{\mathrm{EQI}} = \mathrm{argmin}_{h>0}\mathrm{EQI}(h).$$
>
> (Stone, 1985)

Aplicação a um conjunto de dados

Estimação funcional
O estimador do histograma
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Influência da origem da partição
▷ Escolha prática da janela
O polígono de frequências
O estimador do núcleo
Bibliografia

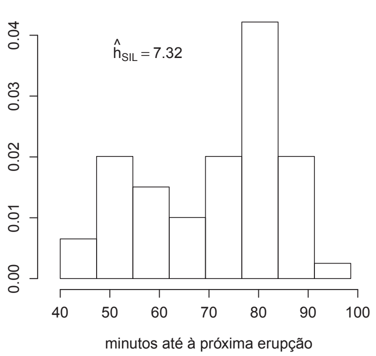
- Dados relativos a 272 erupções sucessivas do Old Faithful Geyser no Yellowstone National Park nos EUA.
- $a_{n,0} = 0,\ \hat{h}_{\mathrm{SIL}} = 0.62,\ \hat{h}_{\mathrm{CV}} = 0.29$

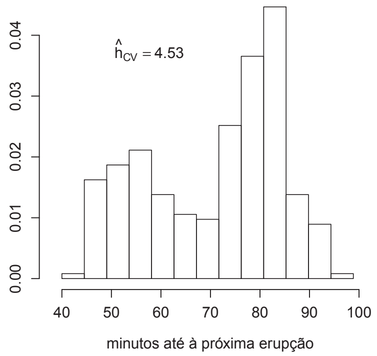




Aplicação a um conjunto de dados

- Dados relativos a 272 erupções sucessivas do Old Faithful Geyser no Yellowstone National Park nos EUA.
- $a_{n,0} = 40$, $\hat{h}_{\mathrm{SIL}} = 7.32$, $\hat{h}_{\mathrm{CV}} = 4.53$





O polígono de frequências

Para $x \in \mathbb{R}$, se f é derivável com $f' \in \mathscr{L}_x(1)$ então,

$$\mathrm{EQM}(\hat{f}_n(x)) = \frac{1}{nh_n} f(x) + \frac{h_n^2}{4} f'(x)^2 \left(1 - 2(x - a_{n,j_x})/h_n\right)^2 + O(n^{-1}) + O\left(h_n^3\right).$$

- Para $h_n = c\,n^{-1/3}$, com $c > 0$, temos

$$\mathrm{EQM}(\hat{f}_n(x)) = O(n^{-2/3})$$

- Para $h_n = c\,n^{-1/5}$, com $c > 0$, temos

$$\boxed{\mathrm{EQM}(\hat{g}_n(x)) = O(n^{-4/5})}$$

Ordem de convergência óptima

- Será ainda possível melhorar a ordem de convergência anterior?
- Existirá algum estimador da densidade cujo erro quadrático médio convirja para zero com uma ordem superior a $n^{-4/5}$?
- Sendo $\mathscr{C}_{2\alpha}$ o conjunto das densidades f com derivada absolutamente contínua e cuja derivada de segunda ordem satisfaz $||f''||_\infty \leq \alpha$, temos

$$\boxed{\limsup_{n\to\infty} \sup_{f \in \mathscr{C}_{2\alpha}} n^{4/5}\,\mathrm{EQM}(\hat{g}_n(x)) < \infty}$$

- Farrel (1972) estabelece que não existe um estimador da densidade cujo erro quadrático médio no ponto x convirja para zero, uniformemente no conjunto $\mathscr{C}_{2\alpha}$, com uma ordem superior a $n^{-4/5}$.

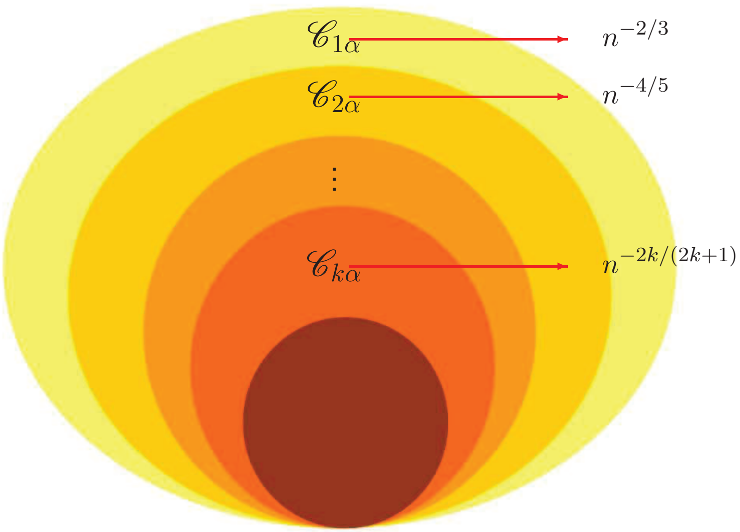
Ordem de convergência óptima

- Para $k \in \mathbb{N}$, seja $\mathscr{C}_{k\alpha}$ o conjunto das densidades f com derivada absolutamente contínua de ordem $k-1$ e cuja derivada de ordem k satisfaz $||f^{(k)}||_\infty \leq \alpha$.
- Farrel (1972) estabelece que não existe um estimador da densidade cujo erro quadrático médio no ponto x convirja para zero, uniformemente no conjunto $\mathscr{C}_{k\alpha}$, com uma ordem superior a

$$n^{-2k/(2k+1)}$$

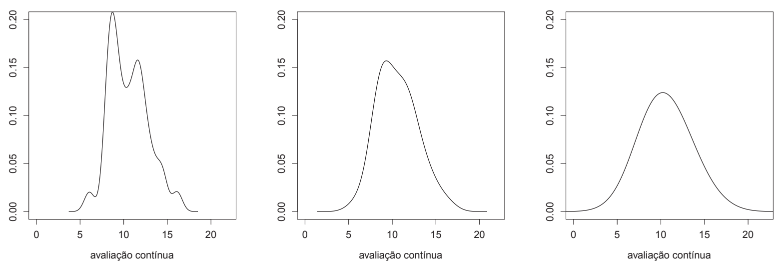
$$\boxed{\mathscr{C}_{k\alpha} \longleftrightarrow n^{-2k/(2k+1)}}$$

Ordem de convergência óptima



O estimador do núcleo

O estimador do núcleo



- Definição do estimador
- Propriedades locais e globais
- A escolha do núcleo
- Escolha prática da janela
- Núcleos de fronteira
- O caso multivariado

O estimador do núcleo

Estimador de Parzen-Rosenblatt:

$$\boxed{f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)}$$

onde

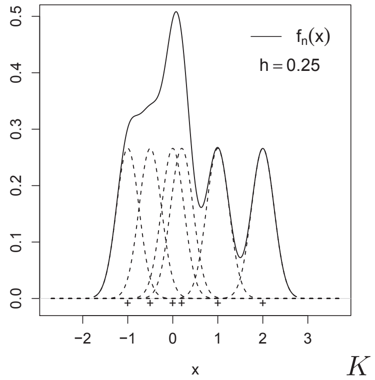
$$K_h(x) = K(x/h)/h$$

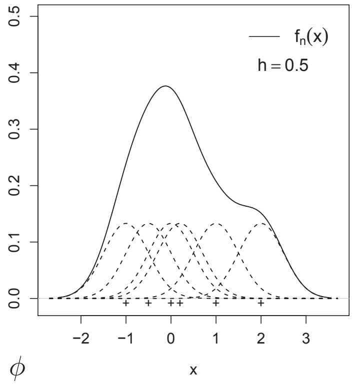
Núcleo	Janela
K	$h = h_n$
$\int K(u)du = 1$	$h \to 0$

O estimador do núcleo

- Estimador de Parzen-Rosenblatt:

$$\boxed{f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)}$$



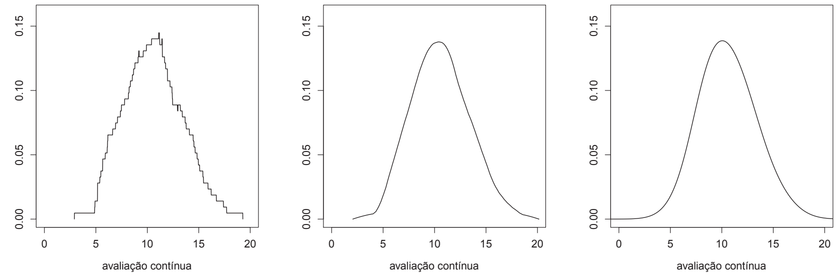


$K = \phi$

O estimador do núcleo

Estimação funcional
O estimador do histograma
O estimador do núcleo
▷ Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

- Quando K é uma densidade de probabilidade, habitualmente tomada simétrica relativamente à origem, f_n é também uma densidade de probabilidade.
- As propriedades de regularidade do núcleo são transferidas para f_n:



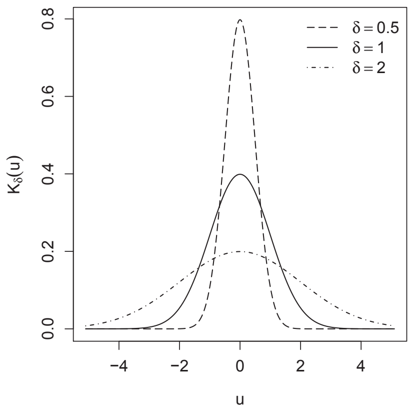
Mudança de escala

Estimação funcional
O estimador do histograma
O estimador do núcleo
▷ Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

- Para um valor $\delta > 0$, a função definida a partir de K por

$$K_\delta(u) = K(u/\delta)/\delta$$

é também um núcleo.



Mudança de escala

Estimação funcional
O estimador do histograma
O estimador do núcleo
▷ Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

$$f_n(x; h, K) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

$$\downarrow$$

$$K \longrightarrow K_\delta$$

$$\downarrow$$

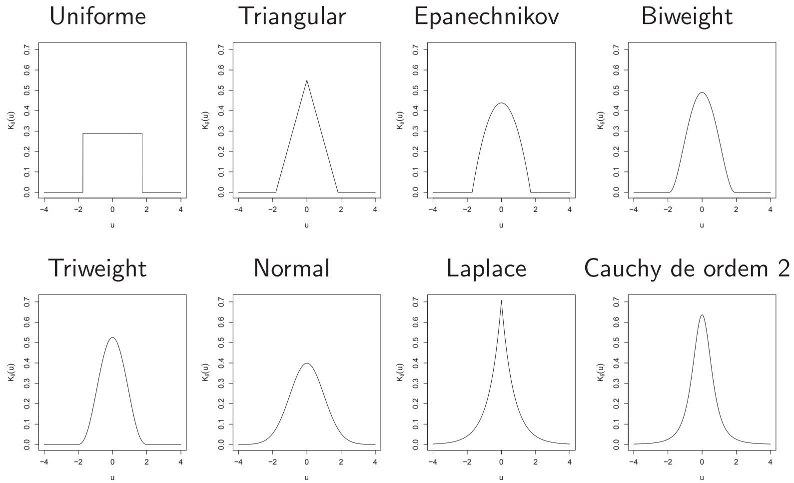
$$f_n(x; h, K_\delta) = \frac{1}{n}\sum_{i=1}^{n} K_{\delta h}(x - X_i) = f_n(x; \delta h, K)$$

- Quando comparamos núcleos é habitual tomar a escala δ de modo que

$$m_2(K_\delta) = \int u^2 K_\delta(u)du = 1$$

Exemplos de núcleos

Estimação funcional
O estimador do histograma
O estimador do núcleo
▷ Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia



Exemplos de núcleos

Estimação funcional
O estimador do histograma
O estimador do núcleo
▷ Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

Núcleo	$K(u)$
Uniforme	$\frac{1}{2}I(\lvert u\rvert \leq 1)$
Triangular	$(1-\lvert u\rvert)I(\lvert u\rvert \leq 1)$
Epanechnikov	$\frac{3}{4}(1-u^2)I(\lvert u\rvert \leq 1)$
Biweight	$\frac{15}{16}(1-u^2)^2 I(\lvert u\rvert \leq 1)$
Triweight	$\frac{35}{32}(1-u^2)^3 I(\lvert u\rvert \leq 1)$
Normal	$(2\pi)^{-1/2}\exp(-u^2/2)$
Laplace	$\exp(-\lvert u\rvert)/2$
Cauchy de ordem 2	$2\pi^{-1}(1+u^2)^{-2}$

Viés do estimador

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
▷ Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

- Para $x \in \mathbb{R}$,

$$\mathrm{E}f_n(x) = \int K_h(x-y)f(y)dy = K_h * f(x)$$

Se o núcleo K é limitado, f é contínua em $x \in \mathbb{R}$, $h = h_n \to 0$ e uma das condições seguintes é satisfeita

a) f é limitada;

b) $\lim_{|x|\to+\infty} xK(x) = 0$; ⟶ Núcleo de Parzen-Rosenblatt

então

$$K_h * f(x) \to f(x).$$

Além disso, se f é limitada e $f \in \mathscr{U}$ então

$$||K_h * f - f||_\infty \to 0.$$

(Bochner, 1955)

Viés do estimador

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
▷ Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

Se f é contínua em $x \in \mathbb{R}$ e $h = h_n \to 0$ então

$$\mathrm{E}f_n(x) \to f(x).$$

$$\begin{aligned} \text{Viés} f_n(x) &= K_{h_n} * f(x) - f(x) \\ &= \int K(z)(f(x - zh_n) - f(x))dz, \end{aligned}$$

- Se $f \in \mathscr{L}_x(\alpha)$ concluímos que

$$\boxed{\text{Viés} f_n(x) = O(h_n^\alpha)}$$

Viés do estimador

- Se $f' \in \mathscr{L}_x(\alpha)$ concluímos que

$$\text{Viés} f_n(x) = -h_n f'(x) \boxed{\int yK(y)} dy + O(h_n^{\alpha+1})$$

- Se $m_1(K) := \int uK(u)du = 0$, o viés do estimador do núcleo é de ordem inferior ao do histograma.

Se f'' é limitada e contínua em $\mathbb{R}$ e $h_n \to 0$ então

$$\text{Viés} f_n(x) = \frac{h_n^2}{2} m_2(K) f''(x) + o(h_n^2).$$

Erro quadrático médio

Se f' e f'' são limitadas e contínuas em $\mathbb{R}$ e $h_n \to 0$ então

$$\text{EQM}(f_n(x)) = \frac{1}{nh_n} R(K)\, f(x) + \frac{h_n^4}{4} m_2^2(K)\, f''(x)^2.$$
$$+O(n^{-1}) + o(h_n^4).$$

- A maior ordem de convergência é obtida quando

$$h_n = c\, n^{-1/5}, \text{ com } c > 0,$$

caso em que

$$f_n(x) - f(x) = O_p(n^{-2/5}).$$

Convergência L_∞

- Pretendemos obter condições sobre f e h_n que assegurem a convergência

$$||f_n - f||_\infty \xrightarrow{qc} 0.$$

- Utilizamos a desigualdade

$$||f_n - f||_\infty \le ||f_n - \text{E} f_n||_\infty + ||\text{E} f_n - f||_\infty$$

onde

$$||\text{E} f_n - f||_\infty = ||K_h * f - f||_\infty \to 0$$

sempre que $f \in \mathscr{U}$.

Convergência L_∞

- No caso do estimador da janela móvel (núcleo uniforme $K(u) = \frac{1}{2} I(|u| \le 1)$)

$$f_n(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}$$

podemos usar os argumentos usados para o histograma:

$$||f_n - \text{E} f_n||_\infty \le \frac{1}{h} ||F_n - F||_\infty$$

Se $f \in \mathscr{U}$, $h_n \to 0$ e $nh_n^2 / \log\log n \to +\infty$ então

$$||f_n - f||_\infty \xrightarrow{qc} 0.$$

(Nadaraya, 1965)

Convergência L_∞

- Uma análise mais fina do termo do termo aleatório $||f_n - \mathrm{E}f_n||_\infty$ permite melhorar o resultado anterior.
- As condições anteriores sobre h_n podem ser ainda enfraquecidas:

Se $h_n \to 0$ então as seguintes proposições são equivalentes:

i) $nh_n/\log n \to +\infty$;

ii) $\forall\, f \in \mathscr{U} \quad ||f_n - f||_\infty \xrightarrow{p} 0$;

iii) $\forall\, f \in \mathscr{U} \quad ||f_n - f||_\infty \xrightarrow{qc} 0$.

(Bertrand-Retali, 1974)

Convergência L_1

As proposições seguintes são equivalentes:

i) $h_n \to 0,\ nh_n \to +\infty$;

ii) $\forall\, f \in \mathscr{F} \quad ||f_n - f||_1 \xrightarrow{p} 0$;

iii) $\forall\, f \in \mathscr{F} \quad ||f_n - f||_1 \xrightarrow{qc} 0$;

iv) $\exists\, f \in \mathscr{F} \quad ||f_n - f||_1 \xrightarrow{qc} 0$.

(Abou-Jaoudé, 1977; Devroye, 1983)

- O estimador do núcleo converge para todo o $f \in \mathscr{F}$, ou então não converge para nenhum $f \in \mathscr{F}$.
- Não existem casos intermédios.

EQMI e EQMIA

Sejam K um núcleo de quadrado integrável com $m_1(K) = 0$ e $\int z^2|K(z)|dz < \infty$, e f com derivada de segunda ordem contínua e f e f'' de quadrado integrável. Se $h_n \to 0$ então

$$\mathrm{IVAR}(f_n) = \frac{1}{nh_n} R(K) + O(n^{-1}),$$

$$\mathrm{IVIES}(f_n) = \frac{h_n^4}{4} m_2^2(K)\, R(f'') + o\left(h_n^4\right)$$

e

$$\mathrm{EQMI}(f_n) = \boxed{\frac{1}{nh_n} R(K) + \frac{h_n^4}{4} m_2^2(K)\, R(f'')} + O(n^{-1}) + o\left(h_n^4\right).$$

EQMIA

(Rosenblatt, 1956; Epanechnikov, 1969)

Janela assintoticamente óptima

Se f é tal que $R(f'') > 0$, o valor de h_n que minimiza o $\mathrm{EQMIA}(K, h_n)$ é dado por

$$h_{\mathrm{EQMIA}} = c_K R(f'')^{-1/5}\, n^{-1/5},$$

onde

$$c_K = \left(R(K)/m_2^2(K)\right)^{1/5}.$$

- A maior ordem de convergência ocorre quando tomamos

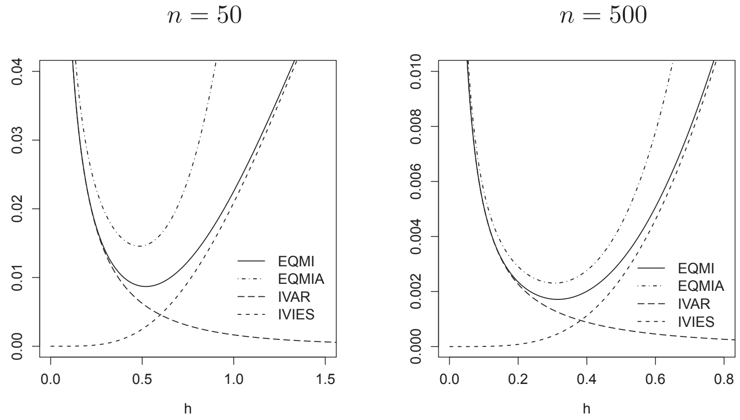
$$h_n = c\, n^{-1/5}, \text{ com } c > 0,$$

obtendo-se neste caso

$$\mathrm{EQMI}(\hat{f}_n) = O(n^{-4/5}).$$

EQMI vs. EQMIA

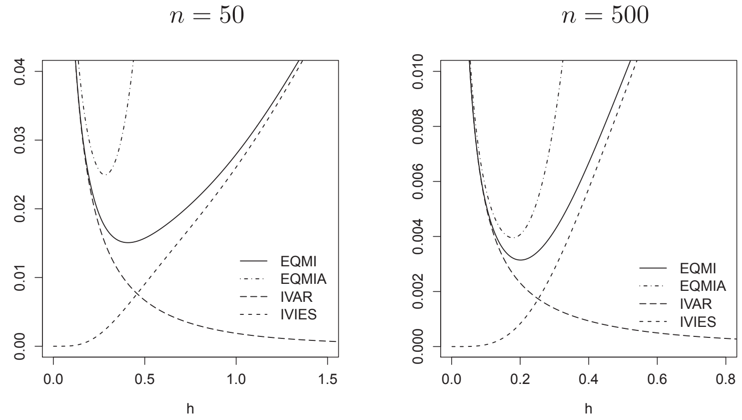
N(0,1)



EQMI vs. EQMIA

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
▷ Janela óptima
Escolha do núcleo
Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

$\frac{3}{4}$N(0,1)+$\frac{1}{4}$N($\frac{3}{2}$,$\frac{1}{9}$)



XVIII Congresso da SPE, S. Pedro do Sul — 29 de Setembro de 2010

Eficiência relativamente ao polígono de frequências

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
▷ Janela óptima
Escolha do núcleo
Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

- Comparando os EQMIA com janela óptima associados ao polígono de frequências e ao estimador do núcleo obtemos a relação
$$\mathrm{EQMIA}(h_{\mathrm{EQMIA}}) = \Psi(K)\,\mathrm{EQMIA}(K, h_{\mathrm{EQMIA}}),$$
onde
$$\Psi(K) = \frac{1}{3}\left(\frac{49}{15R(K)^4\, m_2^2(K)}\right)^{1/5}.$$

$\frac{n}{n_K}$

$K(u)$	$(\Psi(K))^{5/4}$	$K(u)$	$\boxed{(\Psi(K))^{5/4}}$
Uniforme	1.180	Triangular	1.251
Epanechnikov	1.269	*Biweight*	1.261
Triweight	1.252	Normal	1.207
Laplace	0.963	Cauchy de ord. 2	0.856

XVIII Congresso da SPE, S. Pedro do Sul — 29 de Setembro de 2010

Núcleo óptimo

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
▷ Escolha do núcleo
Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

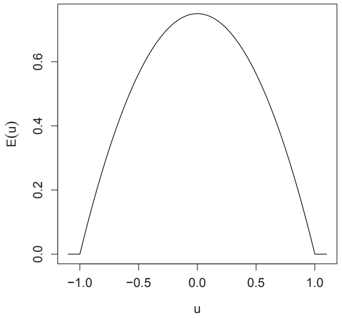
- A escolha do núcleo óptimo é feita a partir do EQMIA com escolha óptima da janela
$$\mathrm{EQMIA}(K, h_{\mathrm{EQMIA}}) = \boxed{\frac{5}{4}\, R(K)^{4/5} m_2(K)^{2/5}}\, R(f'')^{1/5}\, n^{-4/5}$$
- Denotando por $\mathscr{N}$ a classe dos núcleos limitados, simétricos e não negativos K com $\int u^2 K(u)du < \infty$, pretendemos minimizar
$$\int K(u)^2 du$$
sujeito às condições
$$K \in \mathscr{N} \text{ e } m_2(K) = a^2,$$
onde $a \neq 0$ é fixado à partida.

XVIII Congresso da SPE, S. Pedro do Sul — 29 de Setembro de 2010

Núcleo óptimo

Núcleo óptimo:

$$\boxed{E(x) = \frac{3}{4}(1 - x^2)\, I(|x| \le 1).}$$



(Hodges e Lehmann, 1956; Bartlett, 1963; Epanechnikov, 1969)

Eficiência de outros núcleos

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
▷ Escolha do núcleo
Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

- Os erros quadráticos médios integrados associados aos estimadores com núcleos K e E e janelas assintoticamente óptimas, estão relacionados pela igualdade

$$\mathrm{EQMIA}(E, h_{\mathrm{EQMIA}}(E)) = \frac{\Phi(E)}{\Phi(K)}\,\mathrm{EQMIA}(K, h_{\mathrm{EQMIA}}(K)).$$

com

$$\Phi(K) = R(K)^{4/5} m_2(K)^{2/5}.$$

- O quociente

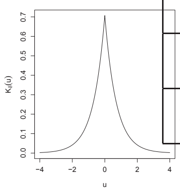
$$\boxed{\left(\frac{\Phi(E)}{\Phi(K)}\right)^{5/4} = \frac{n}{n_K}}$$

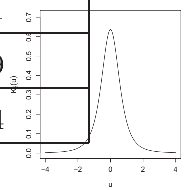
traduz a eficiência do estimador com núcleo K relativamente ao estimador com núcleo óptimo E.

Eficiência de outros núcleos

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
▷ Escolha do núcleo
Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

$K(u)$	$(\Phi(E)/\Phi(K))^{5/4}$
Uniforme	0.930
Triangular	0.986
Epanechnikov	1.000
Biweight	0.994
Triweight	0.987
Normal	0.951
Laplace	0.759
Cauchy de ordem 2	0.674

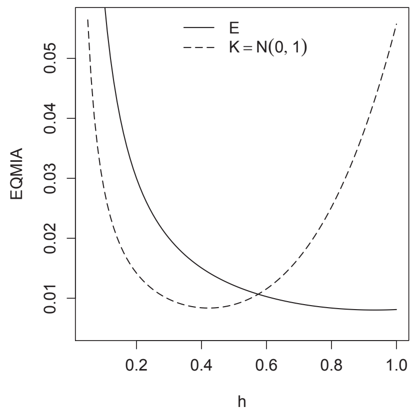




Núcleos equivalentes

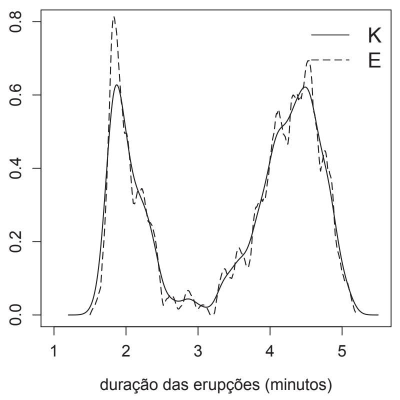
Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
▷ Escolha do núcleo
Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

N(0,1), $n = 100$



Núcleos equivalentes

K=N(0,1), $h = 0.1$



Núcleos equivalentes

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
▷ Escolha do núcleo
Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

- Uma primeira forma de atenuar o efeito da escolha do núcleo é a de substituir o núcleo K pela mudança de escala K_δ de modo que os núcleos E e K_δ tenham a mesma variabilidade.
- A escala δ é assim determinada pela condição

$$m_2(K_\delta) = m_2(E),$$

ou seja,

$$\delta = \left(\frac{m_2(E)}{m_2(K)}\right)^{1/2}$$

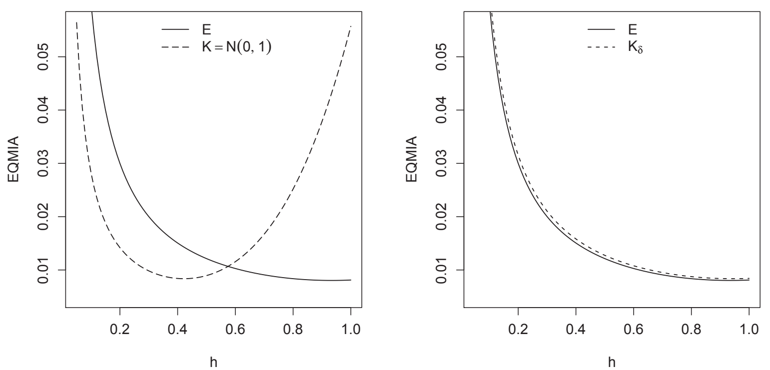
$$\boxed{(E, h) \longleftrightarrow (K_\delta, h)}$$

Núcleos equivalentes

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
▷ Escolha do núcleo
Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

$$\boxed{(E, h) \longleftrightarrow (K_\delta, h)}$$

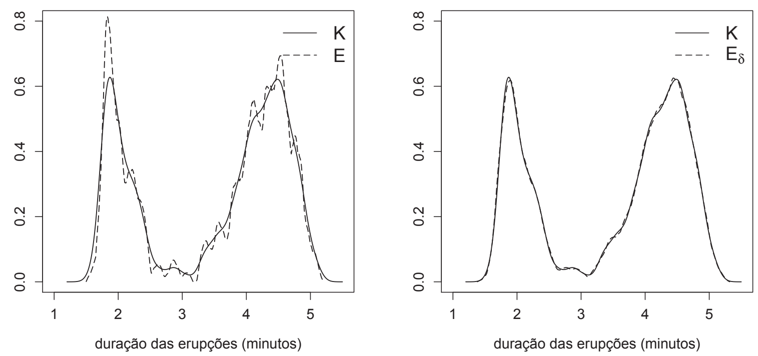
N(0,1), $n = 100$



Núcleos equivalentes

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
▷ Escolha do núcleo
Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

$K = N(0,1)$, $h = 0.1$



$$\delta = \left(\frac{m_2(K)}{m_2(E)}\right)^{1/2} = 2.236$$

Núcleos canónicos

$$K$$

$$\downarrow$$

$$K_{\delta(K)}$$

$$\uparrow$$

$$\delta(K) = \left(R(K)/m_2^2(K) \right)^{1/5}$$

$$\boxed{\mathrm{EQMIA}(K_{\delta(K)}, h) \approx \mathrm{EQMIA}(E_{\delta(E)}, h),}$$

(Marron e Nolan, 1989)

Redução do viés

- Viés do estimador do núcleo:

$$\text{Viés} f_n(x) = \int K(z)(f(x-zh) - f(x))dz$$

- Usando a fórmula de Taylor podemos obter o desenvolvimento

$$\text{Viés} f_n(x) = \boxed{\sum_{\ell=1}^{k-1} (-1)^\ell \frac{h^\ell}{\ell!} m_\ell(K) f^{(\ell)}(x)}$$
$$+ h^k \int\int_0^1 z^k K(z) \frac{(1-t)^{k-1}}{(k-1)!} f^{(k)}(x - tzh) dt dz,$$

onde

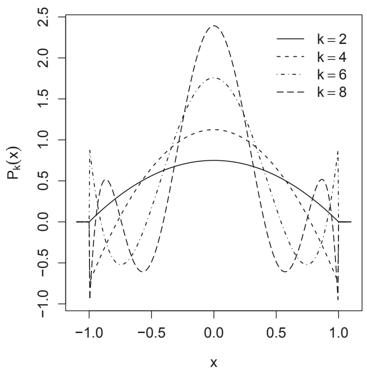
$$m_\ell(K) = \int u^\ell K(u) du,$$

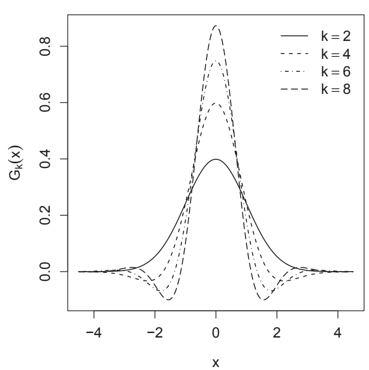
é o momento de ordem ℓ do núcleo K.

Núcleos de ordem k

- Maiores ordens de convergência podem ser obtidas se K for um núcleo de ordem k, isto é, se

$$m_0(K) = 1, m_j(K) = 0, \text{ para } j = 1, \ldots, k-1, \text{ e } m_k(K) \neq 0.$$





Análise do viés

Sejam K um núcleo simétrico de ordem k e f com derivada de ordem p contínua e limitada em $\mathbb{R}$. Para $x \in \mathbb{R}$, se $h_n \to 0$ então

$$\text{Viés} f_n(x) = \frac{h_n^r}{r!} m_r(K) f^{(r)}(x) + o(h_n^r),$$

onde

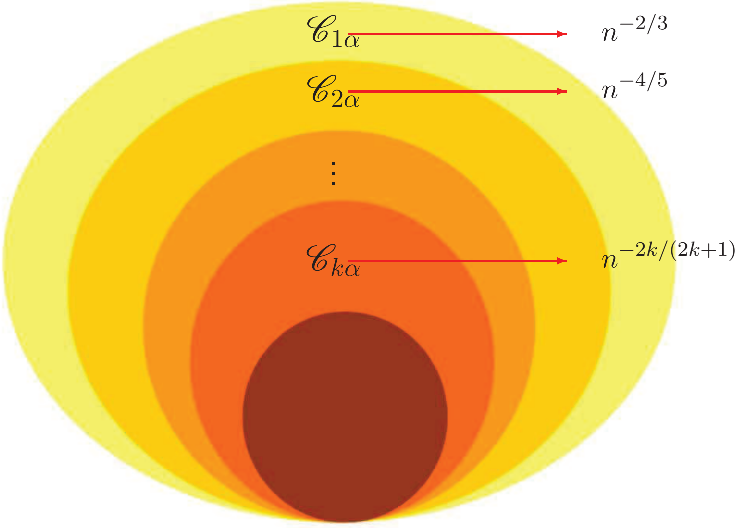
$$r = \min\{k, p\}.$$

Além disso, se $f^{(r)}$ é de quadrado integrável temos

$$\mathrm{IVIES}(f_n) = \frac{h_n^{2r}}{(r!)^2} m_r^2(K) R(f^{(r)}) + o(h_n^{2r}).$$

Ordem de convergência óptima

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
▷ Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia



Supernúcleos

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
▷ Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

- Ao fixarmos a ordem k do núcleo, estamos a limitar a ordem de convergência do viés do estimador do núcleo para densidades que possam admitir derivadas contínuas e limitadas para além da ordem k:

$$\boxed{\text{Viés} f_n(x) = O(h_n^{\min\{k,p\}})}$$

- Esta limitação seria evitada se o núcleo K admitisse momentos nulos de todas as ordens:

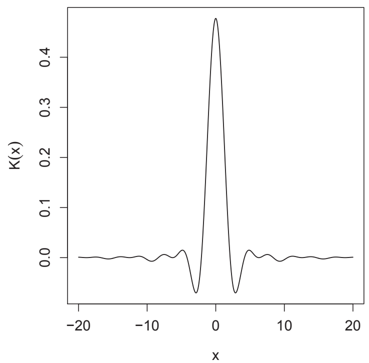
$$\boxed{\text{Viés} f_n(x) = O(h_n^{p})}$$

Supernúcleos

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
▷ Redução do viés
Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

- Os supernúcleos, que são caracterizados pelo facto de possuírem uma função característica constante numa vizinhança da origem, possuem a propriedade anterior.

- Supernúcleo trapezoidal: $K(x) = \dfrac{\cos(x) - \cos(2x)}{\pi\, x^2}$



Janela com distribuição de referência normal

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
Redução do viés
▷ Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

- Para uma vasta classe de densidades conhecemos, a menos da quantidade $R(f'')$, a janela óptima no sentido da minimização do EQMIA:

$$h_{\text{EQMIA}} = c_K R(f'')^{-1/5}\, n^{-1/5}$$

onde

$$c_K = \big(R(K)/m_2^2(K)\big)^{1/5}$$

- Deheuvels (1977) e Deheuvels e Hominal (1980) propõem que se utilize a distribuição normal como distribuição de referência:

$$h_{\text{EQMIA}} = c_K \left(\frac{8\sqrt{\pi}}{3}\right)^{1/5} \sigma\, n^{-1/5} \approx 1.36\, c_K\, \sigma\, n^{-1/5}$$

Janela com distribuição de referência normal

$$\boxed{\hat{h}_{\mathrm{NR}} = 1.36\, c_K\, \hat{s}\, n^{-1/5}}$$

- Silverman (1986) sugere que se estime a escala da distribuição pelo estimador combinado

$$\hat{\sigma}_{\mathrm{SIL}} = \min\left(\hat{s}, \hat{\mathrm{AIQ}}/1.349\right)$$

dando origem à janela

$$\boxed{\hat{h}_{\mathrm{SIL}} = \min(1.36\, \hat{s}, 1.01\, \hat{\mathrm{AIQ}})\, c_k\, n^{-1/5}}$$

O método de validação cruzada

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
Redução do viés
▷ Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

- O método de validação cruzada (*least-square cross-validation*) é proposto por Rudemo (1982) e Bowman (1984).
- O erro quadrático integrado, EQI, é tomado como medida da discrepância entre f_n e a densidade desconhecida f:

$$\mathrm{EQI}(h) = ||f_n - f||_2^2 = \int \{f_n(x) - f(x)\}^2 dx$$

- A ideia é escolher a janela h que minimiza $\mathrm{EQI}(h)$...
- Mas $\mathrm{EQI}(h)$ depende de f ...

$$\mathrm{EQI}(h) = R(f_n) - 2\int f_n(x) f(x)\, dx + R(f)$$

?

O método de validação cruzada

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
Redução do viés
▷ Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

$$\mathrm{EQI}(h) = R(f_n) - \boxed{2\int f_n(x) f(x)\, dx} + \boxed{R(f)}$$

$$\downarrow$$

$$\mathrm{E}(f_n(X))$$

$$\downarrow$$

$$\frac{1}{n}\sum_{i=1}^{n} f_{n,-i}(X_i)$$

- Na prática toma-se para janela o valor $\hat{h}_{\mathrm{CV}}$ que minimiza

$$\begin{aligned} \mathrm{CV}(h) &= R(f_n) - \frac{2}{n}\sum_{i=1}^{n} f_{n,-i}(X_i) \\ &= \frac{R(K)}{nh} + \frac{1}{n(n-1)}\sum_{i\neq j}\left(\tfrac{n-1}{n}K_h * K_h - 2K_h\right)(X_i - X_j) \end{aligned}$$

O método de validação cruzada

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
Redução do viés
▷ Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

- Contrariamente ao método das distribuições de referência, o método de validação cruzada produz janelas assintoticamente "boas" independentemente da distribuição dos dados.

Sob condições pouco restritivas sobre o núcleo e sobre f, $\hat{h}_{\mathrm{CV}}$ é assintoticamente óptima no sentido em que

$$\frac{\mathrm{EQI}(\hat{h}_{\mathrm{CV}})}{\mathrm{EQI}(\hat{h}_{\mathrm{EQI}})} \xrightarrow{qc} 1,$$

onde

$$\hat{h}_{\mathrm{EQI}} = \mathrm{argmin}_{h>0}\mathrm{EQI}(h).$$

(Hall, 1983; Stone, 1984)

O método de validação cruzada

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
Redução do viés
▷ Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

Além disso, vale a convergência

$$\boxed{n^{1/10}}\left(\frac{\hat{h}_{\mathrm{CV}}}{h_{\mathrm{EQMI}}}-1\right)\xrightarrow{d} N(0,\sigma^2_{\mathrm{CV}}),$$

onde $\sigma^2_{\mathrm{CV}}>0$ depende de f e de K e h_{EQMI} é a janela óptima no sentido da minimização de EQMI, isto é,

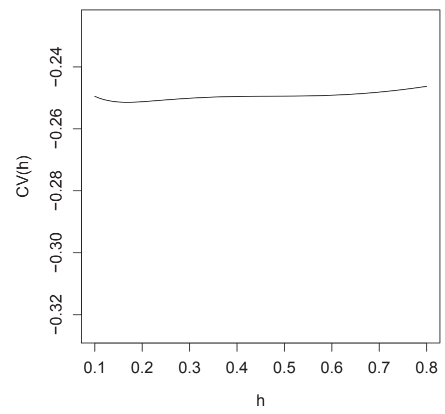
$$h_{\mathrm{EQMI}} = \text{argmin}_{h>0}\mathrm{EQMI}(h)$$

(Hall e Marron, 1987)

O método de validação cruzada

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
Redução do viés
▷ Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

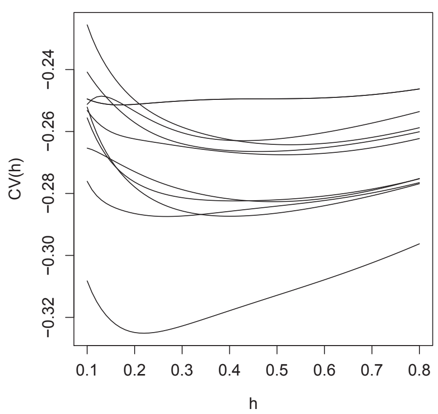
$$f=N(0,1),\ K=N(0,1)$$



O método de validação cruzada

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
Redução do viés
▷ Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

$$f=N(0,1),\ K=N(0,1)$$



Métodos plug-in

Estimação funcional
O estimador do histograma
O estimador do núcleo
Definição
Propriedades locais
Propriedades globais
EQMI
Janela óptima
Escolha do núcleo
Redução do viés
▷ Escolha prática da janela
Núcleos de fronteira
O caso multivariado
Bibliografia

- Os métodos *plug-in* directos partem da expressão para a janela óptima

$$h_{\mathrm{EQMIA}} = c_K\, R(f'')^{-1/5}\, n^{-1/5}$$

onde a funcional desconhecida é substituída por estimador convergente da mesma.

- Funcionais de interesse:

$$\theta_r = R(f^{(r)}) = (-1)^r \int f^{(2r)}(x)f(x)dx = (-1)^r\,\mathrm{E}(f^{(2r)}(X))$$

$$\hat{\theta}_r(g) = \frac{(-1)^r}{n^2}\sum_{i,j=1}^{n} L_g^{(2r)}(X_i-X_j)$$

g – nova janela; L – novo núcleo

Métodos plug-in

$$\boxed{\hat{h}_{\mathrm{PI}} = c_K\, \hat{\theta}_2(g)^{-1/5}\, n^{-1/5}}$$

$$\hat{\theta}_2(g) = \frac{1}{n^2}\sum_{i,j=1}^{n} L_g^{(4)}(X_i - X_j)$$

$$g = ?$$

Métodos plug-in

$$\boxed{\hat{h}_{\mathrm{PI}} = c_K\, \hat{\theta}_2(g)^{-1/5}\, n^{-1/5}}$$

$$\hat{\theta}_2(g) = \frac{1}{n^2}\sum_{i,j=1}^{n} L_g^{(4)}(X_i - X_j)$$

$$g_2 = \left(\frac{2L^{(4)}(0)}{m_2(L)\,\theta_3\, n}\right)^{1/7}$$

Métodos plug-in

$$\boxed{\hat{h}_{\mathrm{PI}} = c_K\, \hat{\theta}_2(g)^{-1/5}\, n^{-1/5}}$$

$$\hat{\theta}_2(g) = \frac{1}{n^2}\sum_{i,j=1}^{n} L_g^{(4)}(X_i - X_j)$$

$$g_2 = \left(\frac{2L^{(4)}(0)}{m_2(L)\,\theta_3\, n}\right)^{1/7}$$

$$\hat{\theta}_3(g) = \frac{-1}{n^2}\sum_{i,j=1}^{n} L_g^{(6)}(X_i - X_j)$$

$$g_3 = \left(\frac{2L^{(6)}(0)}{m_2(L)\,\theta_4\, n}\right)^{1/9}$$

(Sheather e Jones, 1991)

Métodos plug-in

$$\boxed{\frac{\hat{h}_{\mathrm{CV}}}{h_{\mathrm{EQMI}}} - 1 = O_p\left(n^{-1/10}\right)}$$

$$\downarrow$$

$$\boxed{\frac{\hat{h}_{\mathrm{PI}}}{h_{\mathrm{EQMI}}} - 1 = O_p\left(n^{-5/14}\right)}$$

$$\downarrow$$

Usando núcleos de ordem superior:

$$\boxed{\frac{\hat{h}_{\mathrm{PI}}}{h_{\mathrm{EQMI}}} - 1 = O_p\left(n^{-2/5}\right)}$$

$$1/10 = 0.1 \;\rightsquigarrow\; 5/14 \approx 0.36 \;\rightsquigarrow\; 2/5 = 0.4 \;\rightsquigarrow\; 1/2 = 0.5$$

Métodos plug-in

$$h_{\text{EQMIA}} = c_K\, R(f'')^{-1/5}\, n^{-1/5}$$

$$h_{\text{EQMIA}} = c_K\, \hat{\theta}_2(g_{\text{EQMA}})^{-1/5}\, n^{-1/5}$$

$$h_{\text{EQMIA}} = c_K\, \hat{\theta}_2(\alpha(h_{\text{EQMIA}}))^{-1/5}\, n^{-1/5}$$

com

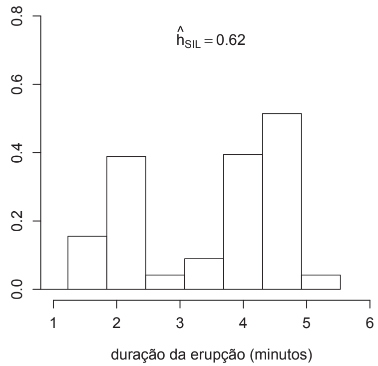
$$\alpha(h) = c_K^{-5/7}(2L^{(4)}(0)/m_2(L))^{1/7}(\theta_2/\theta_3)^{1/7}h^{5/7}$$

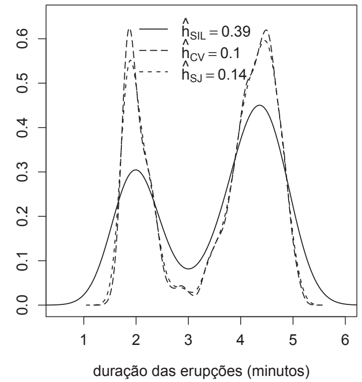
- Sheather e Jones (1991) propõem tomar a janela $\hat{h}_{\text{SJ}}$ que satisfaça a equação

$$\boxed{h = c_K\, \hat{\theta}_2(\hat{\alpha}(h))^{-1/5} n^{-1/5}}$$

Aplicação a um conjunto de dados

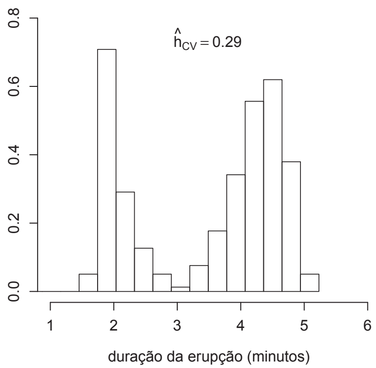
- Dados relativos a 272 erupções sucessivas do Old Faithful Geyser no Yellowstone National Park nos EUA.
- $\hat{h}_{\text{SIL}} = 0.39$, $\hat{h}_{\text{CV}} = 0.10$, $\hat{h}_{\text{SJ}} = 0.14$

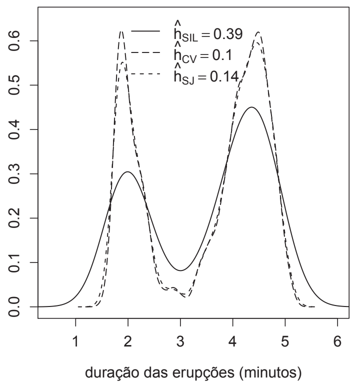




Aplicação a um conjunto de dados

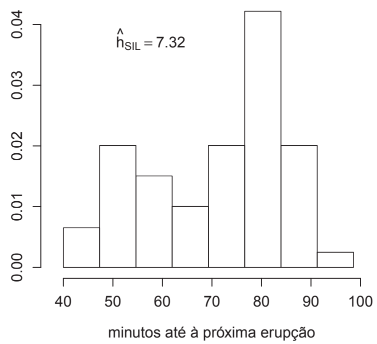
- Dados relativos a 272 erupções sucessivas do Old Faithful Geyser no Yellowstone National Park nos EUA.
- $\hat{h}_{\text{SIL}} = 0.39$, $\hat{h}_{\text{CV}} = 0.10$, $\hat{h}_{\text{SJ}} = 0.14$

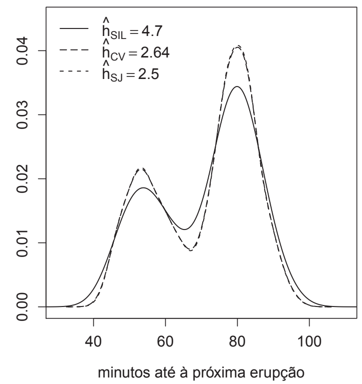




Aplicação a um conjunto de dados

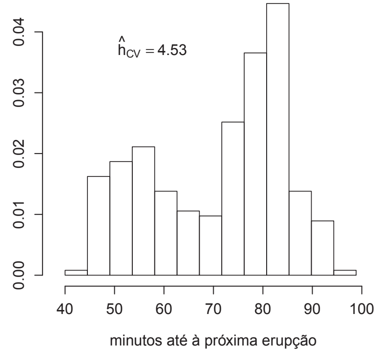
- Dados relativos a 272 erupções sucessivas do Old Faithful Geyser no Yellowstone National Park nos EUA.
- $\hat{h}_{\text{SIL}} = 4.70$, $\hat{h}_{\text{CV}} = 2.64$, $\hat{h}_{\text{SJ}} = 2.50$

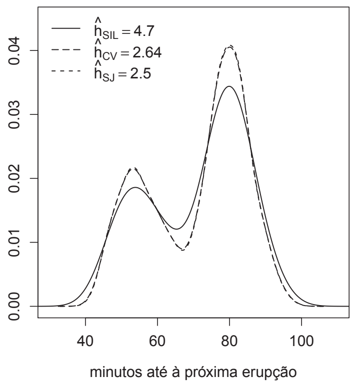




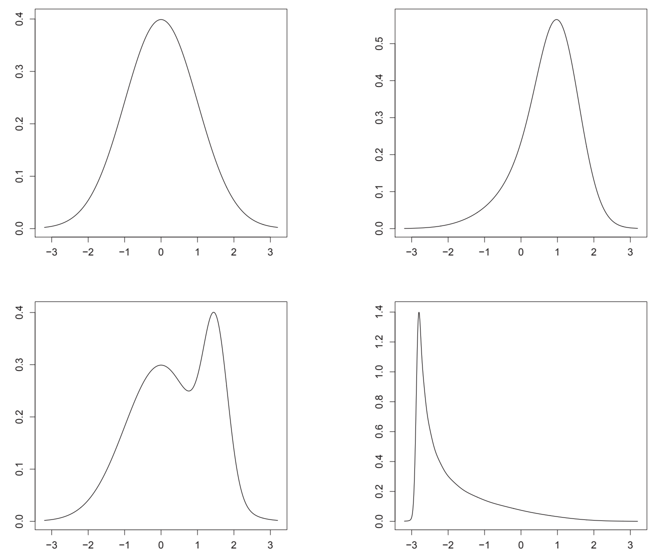
Aplicação a um conjunto de dados

- Dados relativos a 272 erupções sucessivas do Old Faithful Geyser no Yellowstone National Park nos EUA.
- $\hat{h}_{\mathrm{SIL}} = 4.70$, $\hat{h}_{\mathrm{CV}} = 2.64$, $\hat{h}_{\mathrm{SJ}} = 2.50$

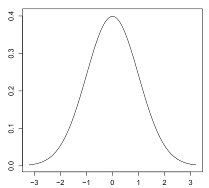


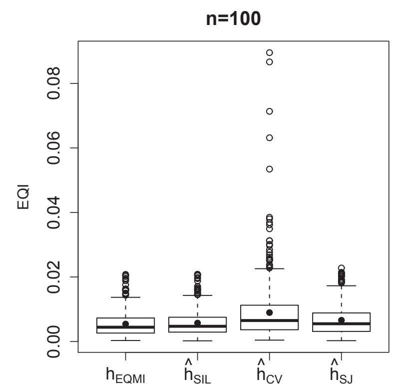


Resultados de simulação

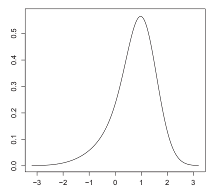


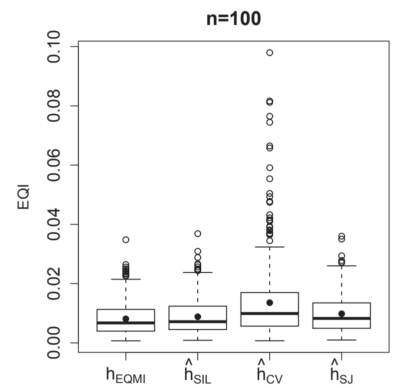
Resultados de simulação



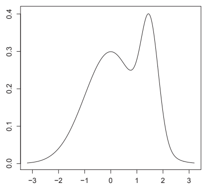


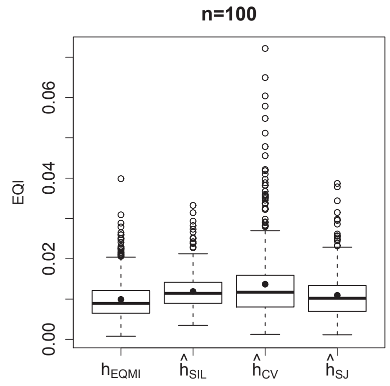
Resultados de simulação



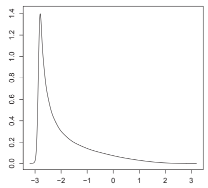


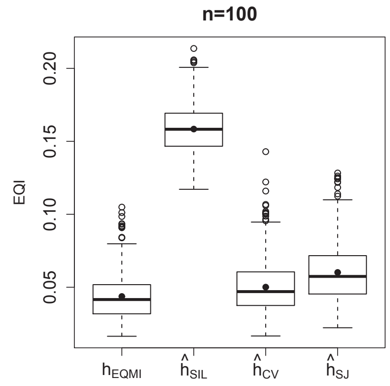
Resultados de simulação



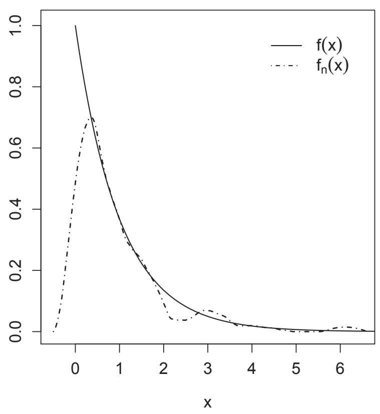


Resultados de simulação





Estimação em pontos fronteiros



Estimativa baseada numa amostra de tamanho 200 da distribuição $E(1)$ com $h = 0.5$ e núcleo de Epanechnikov.

Estimação em pontos fronteiros

□ Análise do viés em pontos fronteiros:

$$x = a + \alpha h, \alpha \in]0,1[$$

$$\boxed{\begin{aligned} \mathrm{E} f_n(x) &= m_{0,\alpha}(K) f(x) - h\, m_{1,\alpha}(K) f'(x) \\ &\quad + \frac{h^2}{2} m_{2,\alpha}(K) f''(x) + o(h^2), \end{aligned}}$$

$$m_{\ell,\alpha}(K) = \int_{-1}^{\alpha} z^\ell K(z) dz$$

Estimação em pontos fronteiros

- Análise do viés em pontos fronteiros:

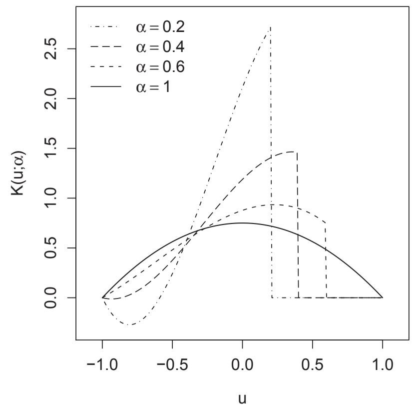
$$x = a + \alpha h, \alpha \in]0,1[$$

$$\boxed{\begin{aligned} \mathrm{E} f_n(x) &= m_{0,\alpha}(K) f(x) - h\, m_{1,\alpha}(K) f'(x) \\ &\quad + \frac{h^2}{2} m_{2,\alpha}(K) f''(x) + o(h^2), \end{aligned}}$$

$$m_{\ell,\alpha}(K) = \int_{-1}^{\alpha} z^\ell K(z) dz$$

$$\boxed{K(u) \longrightarrow K(u;\alpha)}$$

Núcleos de fronteira



Núcleos de fronteira baseados no núcleo de Epanechnikov.

(Gasser e Muller, 1979; Rice, 1984)

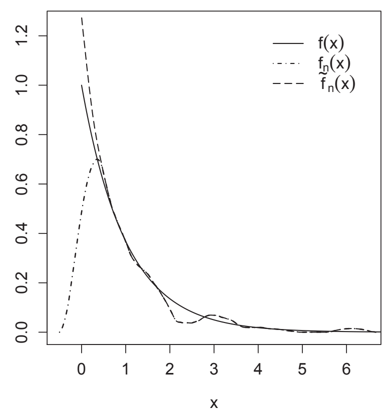
Estimador com núcleos de fronteira

$$\tilde{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K_{x,h}\left(\frac{x - X_i}{h}\right)$$

$$K_{x,h}(u) = \begin{cases} K(u;(x-a)/h), & a \le x < a+h \\ K(u), & x \ge a+h, \end{cases}$$

Estimador com núcleos de fronteira

$$\tilde{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K_{x,h}\left(\frac{x - X_i}{h}\right)$$



Estimador multivariado do núcleo

- Cacoullos (1966):

$$f_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

- Epanechnikov (1969):

$$f_n(x) = \frac{1}{nh_1 \dots h_d} \sum_{i=1}^{n} K\left(\frac{x_1 - X_{i1}}{h_1}, \dots, \frac{x_d - X_{id}}{h_d}\right)$$

- Deheuvels (1977):

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_H(x - X_i)$$

$$K_H(x) = |H|^{-1/2} K(H^{-1/2}x)$$

Estimador multivariado do núcleo

- No caso $H = diag(h_1^2, \dots, h_d^2)$

$$\text{EQMIA} = \frac{R(K)^d}{nh_1 \dots h_d} + \frac{1}{4} m_2(K) \sum_{i,j=1}^{d} h_i^2 h_j^2 \int f_{ii}(x) f_{jj}(x) dx$$

- Para

$$h_k = c_k \, n^{-1/(d+4)}$$

obtemos

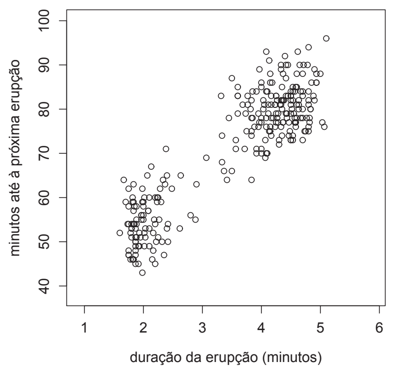
$$\text{EQMIA} = O\left(\boxed{n^{-4/(d+4)}}\right)$$

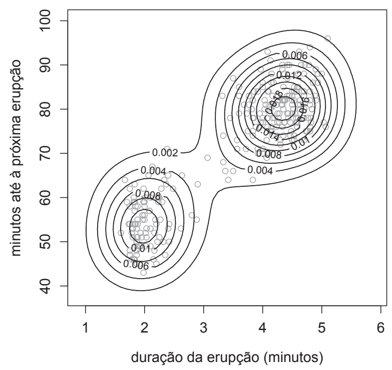
Maldição da dimensão

- Escolha da janela tomando como distribuição de referência a distribuição normal $N(0, \Sigma)$, com $\Sigma = diag(\sigma_1^2, \dots, \sigma_d^2)$:

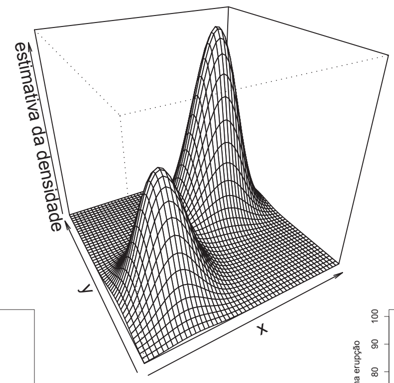
$$\hat{h}_k = \left(\frac{8}{d+5}\right)^{1/(d+4)} \hat{\sigma}_k \, n^{-1/(d+4)}$$

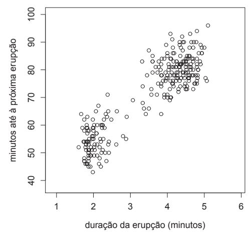
Estimador multivariado do núcleo

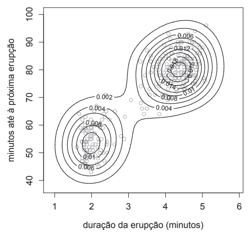




Estimador multivariado do núcleo







Estimação funcional
O estimador do histograma
O estimador do núcleo
▷ Bibliografia

Referências bibliográficas

Principal bibliografia sobre estimação funcional

Estimação funcional
O estimador do histograma
O estimador do núcleo
Bibliografia

- ☐ Prakasa Rao (1983)
  Nonparametric functional estimation. Academic Press.
- ☐ Devroye e Györfi (1985)
  Nonparametric density estimation: the L_1 view. Wiley.
- ☐ Bosq e Lecoutre (1987)
  Théorie de l'estimation fonctionnelle. Economica.

- ☐ Silverman (1986)
  Density estimation for statistics and data analysis. Chapman & Hall.
- ☐ Scott (1992)
  Multivariate density estimation. Wiley.
- ☐ Wand e Jones (1995)
  Kernel smoothing. Chapman & Hall.

Principal bibliografia sobre estimação funcional

Estimação funcional
O estimador do histograma
O estimador do núcleo
Bibliografia

- ☐ Thompson e Tapia (1990)
  Nonparametric function estimation, modeling, and simulation. SIAM.
- ☐ Härdle (1991)
  Smoothing techniques: with implementation in S. Springer.
- ☐ Simonoff (1996)
  Smoothing methods is statistics. Springer.
- ☐ Bowman e Azzalini (1997)
  Applied smoothing techniques for data analysis. Oxford University Press.
- ☐ Härdle, Müller, Sperlich e Werwatz (2004)
  Nonparametric and semiparametric models. Springer.
- ☐ Tsybakov (2004)
  Introduction à l'estimation non-paramétrique. Springer.
- ☐ Wasserman (2006)
  All of nonparametric statistics. Springer.