Neural Network Interest Group

Título/*Title*:
Estimação de FDPs

Autor(es)/*Author(s)*:
J.P. Marques de Sá

Relatório Técnico/*Technical Report* No. 2 /2005

FEUP/INEB, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal

Título/*Title*:

Estimação de FDPs

(Tutorial Baseado no livro de Luc Devroye, Gábor Lugosi (2001), Springer-Verlag)

Autor(es)/*Author(s)*:

J.P. Marques de Sá

Relatório Técnico/*Technical Report* No. 2 /2005

Publicado por/*Published by*: NNIG. http://paginas.fe.up.pt/~nnig/

Dezembro de 2005

Estimação de FDPs

Texto Tutorial

baseado no livro de

Luc Devroye, Gábor Lugosi (2001). Springer-Verlag

J.P. Marques de Sá

INEB, Dezembro 2005

Índice

1 Variação Total

1.1 Definição

Dada uma estimativa $f_n(x \mid X_i,\ldots,X_n)$ baseada em $X_i,\ldots,X_n$ i.i.d. de $f(x)$, como avaliar a qualidade de estimação de probabilidades de f_n?

Critério da **variação total**:

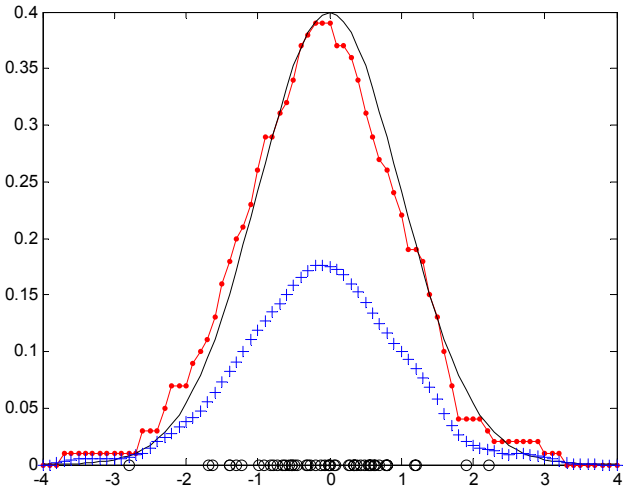
$$\sup_A \left| \int_A f_n - \int_A f \right|$$

A é qualquer conjunto de Borel

Trata-se do supremo, para qualquer conjunto de Borel, das probabilidades (áreas) calculadas pelas duas fdp's.

(Nota: é supremo e não max porque, como se sabe, os conjuntos de Borel podem ser construídos com intervalos abertos)

Exemplo:



tmax = exptv(1,0,2,50,-4,4,0.1,5) = 1.7646

1.2 Distância L_1

Medimos a semelhança entre duas fdp's pela sua distância L_1: $\int |f - g|$

Há várias razões para usar L_1:

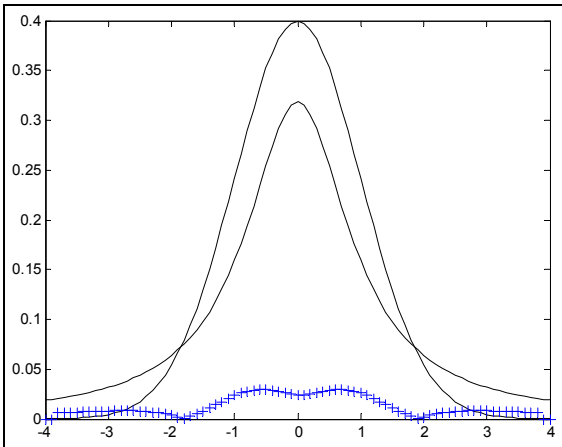
1-

Teorema (identidade de Scheffé). Sejam f e g duas funções definidas em $\Re^d$ tal que $\int f = \int g = 1$. Seja $\mathcal{B}$ a classe de todos os conjuntos de Borel em $\Re^d$. Então:

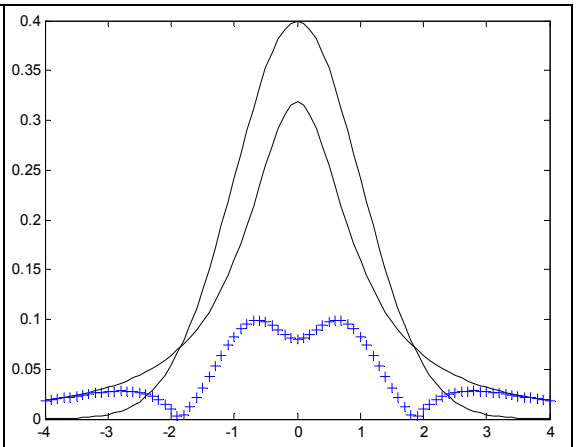
$$\sup_{B \in \mathcal{B}} \left| \int_B f_n - \int_B f \right| = \frac{1}{2} \int |f - g|$$

•

Exemplo:



```
>> x=-4:0.1:4;x=x';
>> f=probplot(1,x);
>> hold on;g=probplot(4,x);
>> xf=[x f];
>> t=totalvar(xf,4,-4,4,0.1,4);
>> plot(xf(:,1),t,'b+');
>> tmax=max(t)
tmax =
   0.0294
>> t=totalvar(xf,4,-4,4,0.1,80);
>> tmax=max(t)
tmax =
   0.1578
```



```
>> f=probplot(1,x);
>> hold on
>> g=probplot(4,x);
>> t=abs(f-g);
>> plot(x,t,'b+');
>> tt=integral(t)*0.1/2
tt =
   0.1733
```

Os valores de t fora de [-2,2] são <0 (antes de converter para abs)

2-

Se $\int |f - g| < 0.04$ então, pelo teorema anterior, as diferenças em probabilidade são no máximo 0.02. Pelo contrário a interpretação de $\int (f - g)^2 < 0.04$ ou de $\int f \log(f / g) < 0.04$ não é evidente.

1.3 Propriedades

1. A variação total é invariante para transformações monotónicas dos eixos.
2. A variação total decresce para um qualquer mapeamento arbitrário, tal como a distância L_1. Por exemplo, se f^+ e g^+ são as fdps de ||X|| e ||Y||, então

$$\int \left| f^+ - g^+ \right| \leq \int |f - g|$$

3. A *convolução* também diminui a variação total:

$$\int |f * K - g * K| \leq \int |K| \int |f - g|$$

A demonstração deste facto baseia-se na *desigualdade de Young*:

$$\int |f * K| \leq \int |f| \int |K|$$

2 Escolhendo uma Estimativa

Dadas duas estimativas f_n e g_n de $X_i, \ldots, X_n$ i.i.d. de $f(x)$, escolhemos uma delas usando L_1:

$$\arg\min\left(\int |f_n - f|, \int |g_n - f| \right)$$

Ou, de forma mais geral, queremos construir uma estimativa ϕ_n tal que:

$$\int |\phi_n - f| \approx \min\left(\int |f_n - f|, \int |g_n - f| \right)$$

Trata-se de um problema difícil.

2.1 Estimativa de Scheffé

Seja:

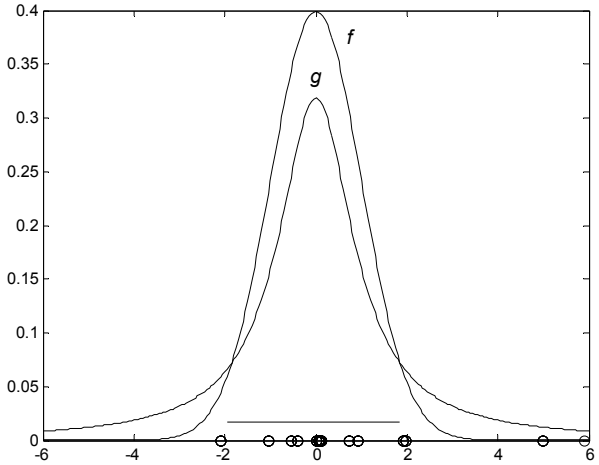
1. $\mu_n(A) = (1/n) \sum_{i=1}^{n} I_{x_i \in A}$ a *medida empírica* de A.
2. $A(f_n, g_n) = \{x : f_n(x) > g_n(x)\}$ o *conjunto de Scheffé* do par (f_n, g_n)

A *estimativa de Scheffé* é definida como:

$$f_n^* = \begin{cases} f_n & \text{se } \left| \int_A f_n - \mu_n(A) \right| < \left| \int_A g_n - \mu_n(A) \right| \\ g_n & \text{noutro caso} \end{cases}$$

Esta definição não é simétrica em f_n e g_n!

Exemplo:



>> s = scheffeset(f,g); [-1.95,1.95] >> c = empmeasure(a,x,s) c = 0.6667 (10 pontos em 15) >> d = distemp(f,a,x,s) d = 0.2690 >> d = distemp(g,a,x,s) d = 0.0178 Escolhe g	>> s = scheffeset(g,f); [-6,-1.9]∪[1.9,6] >> c = empmeasure(a,x,s) c = 0.3333 (5 pontos em 15) >> d = distemp(f,a,x,s) d = 0.2758 >> d = distemp(g,a,x,s) d = 0.1300 Escolhe g

Teorema

Sejam duas estimativas f_n e g_n de $X_i,\ldots,X_n$ i.i.d. de $f(x)$ com área unitária. Para a estimativa de Scheffé, temos:

$$\int\left|f_n^*-f\right|\le 3\min\left(\int\left|f_n-f\right|,\int\left|g_n-f\right|\right)+4\max_{A\in\mathcal{A}}\left|\int_A f-\mu_n(A)\right|$$

com $\mathcal{A}=\{\{f_n>g_n\},\{g_n>f_n\}\}$ •

Suponhamos k fdp's candidatas, f_{ni}, e que escolhemos a que vence para o maior número de $k(k\text{-}1)/2$ pares de comparações. Chamemos-lhe a fdp vencedora do torneio Scheffé.

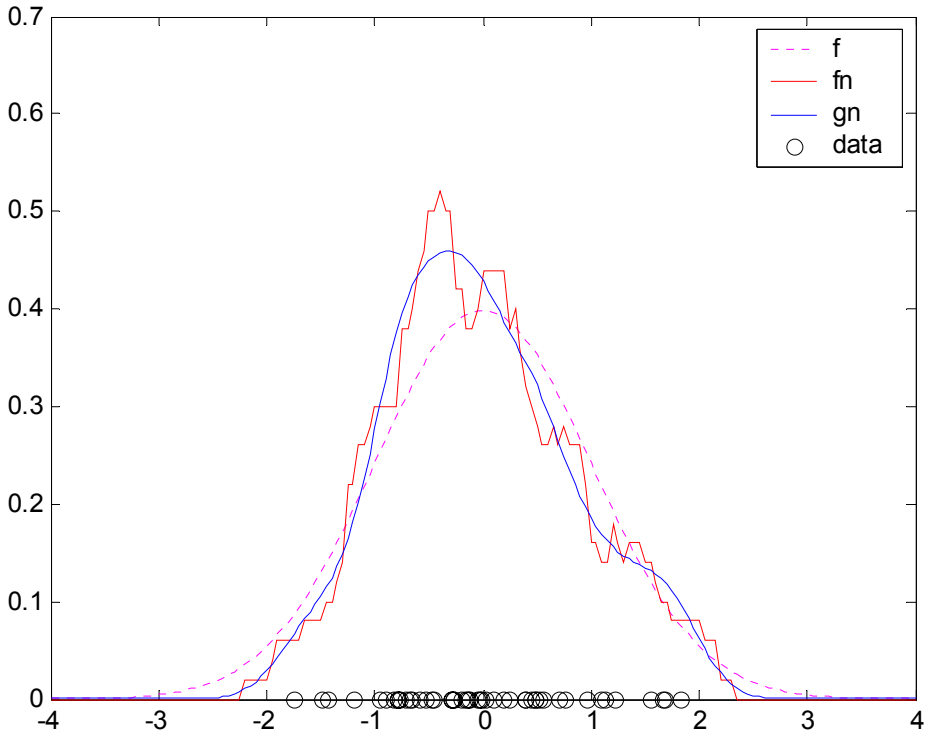
Teorema

Para a vencedora do torneio Scheffé temos: $\int\left|f_n^*-f\right|\le 9\min_i\left(\int\left|f_{ni}-f\right|\right)+16\Delta$,

onde D é o supremo para todos os conjuntos de Scheffé de $\left|\int_A f-\mu_n(A)\right|$

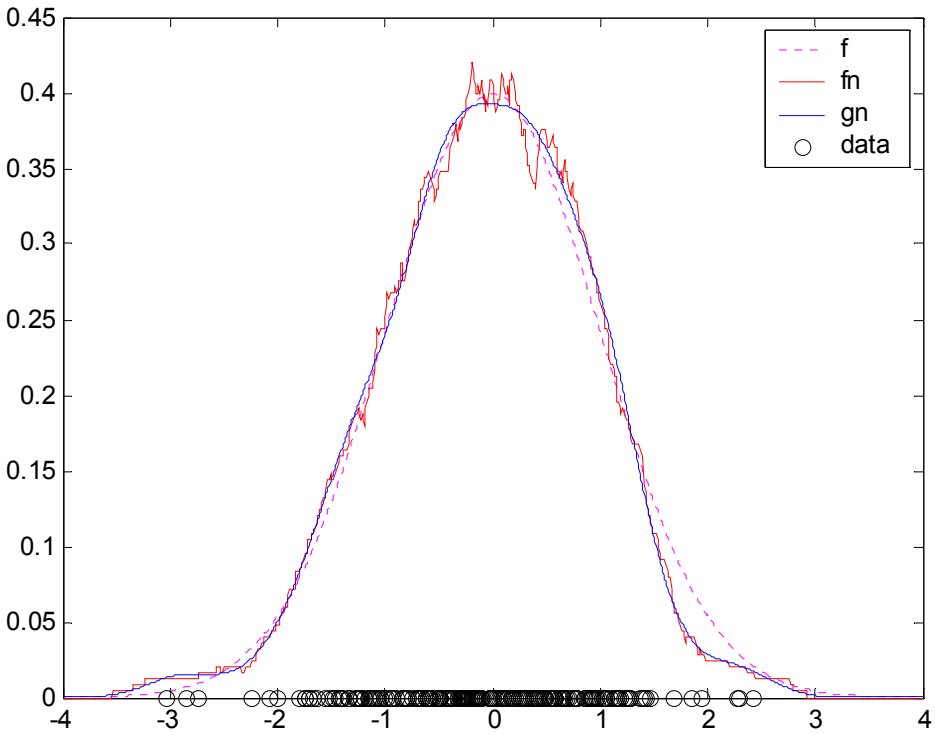
•

Exemplo 1



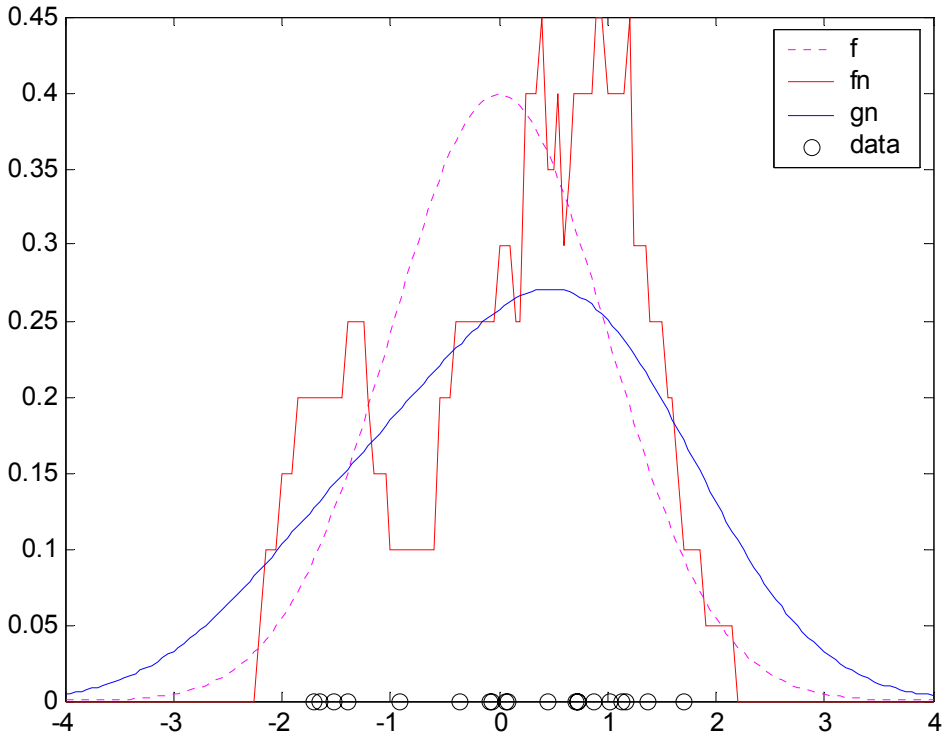

```
>> expscheffe(50,0.3,0.05);   %h1=1; h2=0.3
Distancia L1 de f a fn:          0.2005
Distancia L1 de f a gn:          0.1816
Variaçao total em {fn>gn} de fn face `a dist. empirica:      0.2330
Variaçao total em {fn>gn} de gn face `a dist. empirica:      0.1853
A estimativa Scheffe e´ gn
Variaçao total em {fn>gn} de f face `a dist. empirica:       0.1747
Variaçao total em {gn>fn} de f face `a dist. empirica:       0.2019
Majorante:      1.3524
```

Exemplo 2



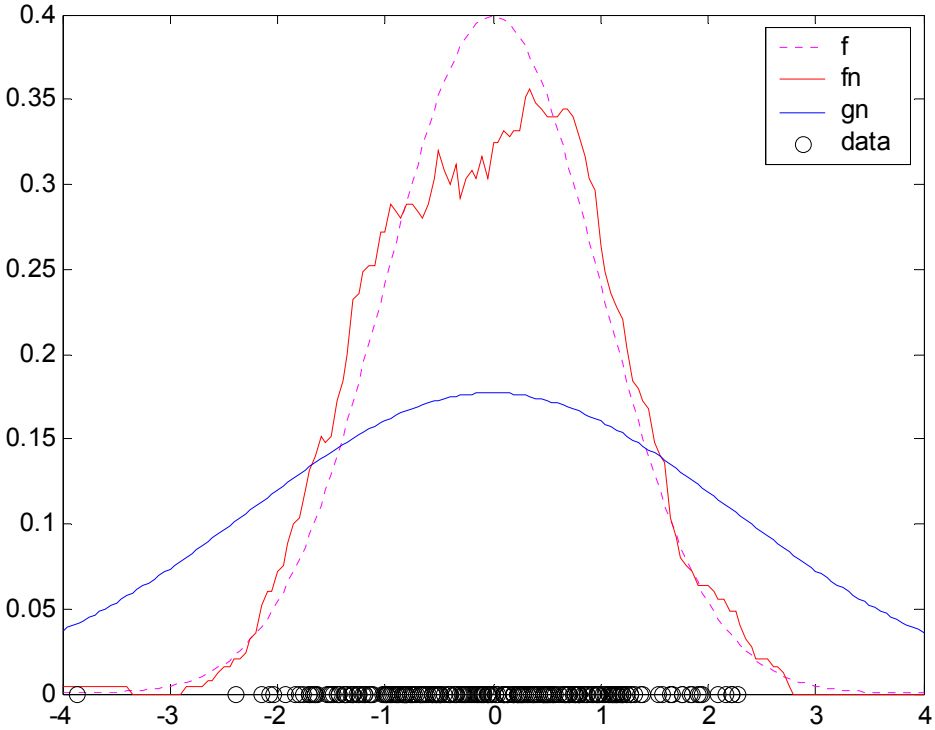

```
>> expscheffe(250,0.3,0.01); %h1=1; h2=0.3
Distancia L1 de f a fn:          0.0743
Distancia L1 de f a gn:          0.0639
Variaçao total em {fn>gn} de fn face `a dist. empirica:      0.0466
Variaçao total em {fn>gn} de gn face `a dist. empirica:      0.0276
A estimativa Scheffe e´ gn
Variaçao total em {fn>gn} de f face `a dist. empirica:       0.0279
Variaçao total em {gn>fn} de f face `a dist. empirica:       0.0244
Majorante:      0.3034
```

Exemplo 3




```
>> expscheffe(20,1,1,0.05);
Distancia L1 de f a fn:          0.4662
Distancia L1 de f a gn:          0.4075
Variaçao total em {fn>gn} de fn face `a dist. empirica:      0.0425
Variaçao total em {fn>gn} de gn face `a dist. empirica:      0.2590
A estimativa Scheffe e´ fn
Variaçao total em {fn>gn} de f face `a dist. empirica:       0.2186
Variaçao total em {gn>fn} de f face `a dist. empirica:       0.1042
Majorante:      2.0968
```

Exemplo 4




```
>> expscheffe(250,1,2,0.05);
Distancia L1 de f a fn:          0.1542
Distancia L1 de f a gn:          0.6704
Variaçao total em {fn>gn} de fn face `a dist. empirica:      0.0100
Variaçao total em {fn>gn} de gn face `a dist. empirica:      0.3615
A estimativa Scheffe e´ fn
Variaçao total em {fn>gn} de f face `a dist. empirica:       0.0059
Variaçao total em {gn>fn} de f face `a dist. empirica:       0.0034
Majorante:      0.4861
```

Exemplo 5

Se f_n e g_n são fdp's normais que estimam f, é possível mostrar que:

$$\mathrm{E}\left[\int \left|f_n^* - f\right| - 3\min\left(\int |f_n - f|, \int |g_n - f|\right)\right] \leq \frac{8}{\sqrt{n}}$$

Desigualdades deste tipo existem para outras situações de f_n e g_n; o factor "3" é inescapável.

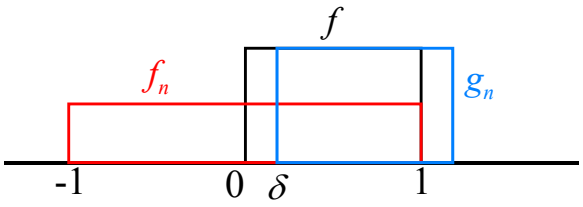
A estimativa de Scheffé parece ser a única que oferece garantias de majorante do erro.

2.2 Outros critérios de escolha

Escolha de Máxima Verosimilhança

O problema com este critério é que mesmo em situações idealizadas não minimiza L_1.

Exemplo:
Seja:
f: uniforme em [0,1]
f_n: uniforme em [-1,1]
g_n: uniforme em $[\delta,1+\delta]$



Dado $X_i,\ldots,X_n$ i.i.d. de $f(x)$, a escolha de máxima verosimilhança (m.v.) consiste em escolher:

$$\phi_n(x) \overset{\text{def}}{=} \arg\max\left(\prod_{i=1}^{n} f_n(x_i), \prod_{i=1}^{n} g_n(x_i)\right)$$

Seja N o número de pontos que cai em $[0,\delta[$, uma v.a. binomial (n,δ). Então g_n é escolhido sse $N=0$.
Dado que as distâncias L_1 são $\int|f_n - f| = 1$ e $\int|g_n - f| = 2\delta$, a melhor escolha para $\delta < ½$ é g_n. Ora, vejamos a diferença entre a distância L_1 média de ϕ_n face a f e a distância de g_n face a f (o mesmo é dizer a diferença entre o erro médio da escolha m.v. e o erro da escolha óptima):

$$\begin{aligned}\mathrm{E}\left[\int|\phi_n - f|\right] - \int|g_n - f| &= \left(P(N=0)\times 2\delta + P(N>0)\times 1\right) - 2\delta = \\ &P(N>0) + (1 - P(N>0)2\delta - 2\delta = (1-2\delta)P(N>0) = \\ &(1-2\delta)\left(1-(1-\delta)^n\right) \underset{n\to\infty}{\to} 1-2\delta\end{aligned}$$

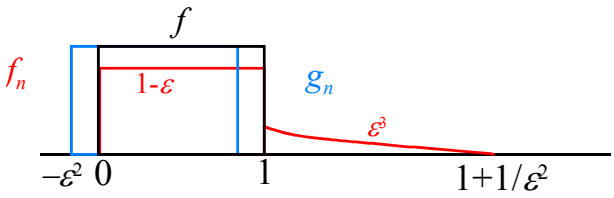
Isto é, à medida que δ decresce, sendo o erro da escolha g_n desprezável, obtém-se com $n \to \infty$ uma escolha de m.v. com erro catastrófico!

Escolha baseada na distância L_2

Mesmo em casos simples de funções de quadrado integrável a escolha basead na distância L_2 pode levar a decisões incorrectas.

Exemplo

Seja:
f: uniforme em [0,1]
f_n: 1-ε [0,1] e ε^3 em [1,1+1/ε^2]
g_n: uniforme em [$-\varepsilon^2$,1- ε^2]



Temos:

$$\int |f_n - f| = 2\varepsilon; \qquad \int |g_n - f| = 2\varepsilon^2$$

Logo a escolha acertada é claramente g_n.

Contudo:

$$\int (f_n - f)^2 = \varepsilon^2 + \varepsilon^4 < 2\varepsilon^2 = \int (g_n - f)^2$$

3 Estimação Usando Kernel

A variação total de uma fdp relativamente a uma estimativa discreta como μ_n é 1. Precisamos de suavizar μ_n por forma a diminuir a variação total. Fazemos isso recorrendo à convolução com um kernel.

3.1 Aproximação funcional por convolução

Convolução: $f * g(x) = \int f(y)g(x-y)dy = \int g(y)f(x-y)dy$

Se f e g são as fdp's de X e Y, f*g é a fdp de X+Y.
Se g é concentrada em torno de 0, f*g aproxima-se de f.

Teorema
Seja K uma qualquer função integrável em $\Re^d$ ($\int |K| < \infty$) e f uma fdp em $\Re^d$. Designando $K_h(x) = (1/h)^d K(x/h), x \in \Re^d \ h > 0$, temos:

$$\lim_{h \to 0} \int \left| f * K_h - f \int K \right| = 0$$

•

3.2 Estimativa de kernel

Kernel: função K tal que $\int |K| < \infty$ e $\int K = 1$.

Para escolher f_n tal que $\int |f_n - f|$ seja pequeno fazemos uso do facto de que $\int |f * K_h - f|$ é pequeno quando h é pequeno e aproximamos f*K_h por μ_n*K_h.
A *estimativa de kernel* é:

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

"Variable kernel estimate": $h = h(x)$
"Data-dependent kernel estimate: $h = h(X_i, \ldots, X_n)$

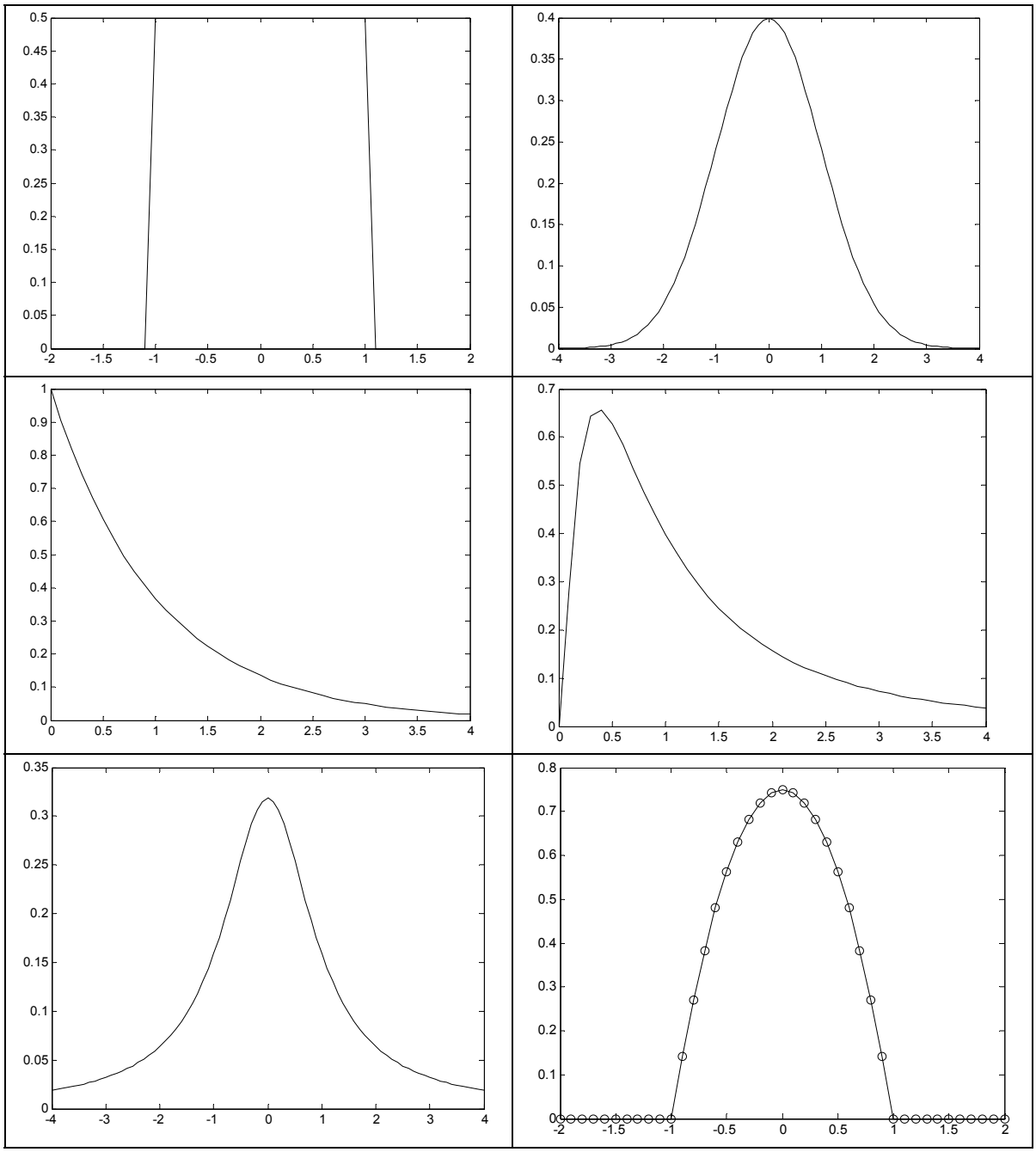
Consistência das estimativas de kernel:

Teorema
Seja K um kernel fixo e h dependente apenas de n. Se $h \to 0$ e $nh^d \to \infty$ quando $n \to \infty$, então $\mathrm{E}[|f_n - f|] \to 0$.

•

```
FTYPE
   case 0: uniforme(-1,1);
   case 1: normal(0,1);
   case 2: exp(1);
```

case 3: lognormal(0,1);
case 4: cauchy(0);



KERNELTYPE
0 % uniform
1 % gaussian
2 % Epanechnikov
3 % Cauchy