Neparametrické modelování hustot a distribucí.

K danému náhodnému výběru $\{x_1, x_2, ..., x_n\}$ pevného rozsahu n hledáme neznámou hustotu ve tvaru: $f_a(x) = \frac{1}{nh} \sum_{i=1}^{n} \kappa(\frac{x-x_i}{h})$ kde:

$h; h > 0$ je parametr měřítka – pokud na něčem závislý, pak pouze nanejvýš na rozsahu výběru n (někdy též i na pozorovaných hodnotách), vyhlazovací parametr, scaling parameter, … a

$\kappa(x)$ je jádro aproximace – nějaká hustota, splňující následující podmínky:

1. $\kappa(x) \geq 0; \forall x \in R_1; \int_{-\infty}^{+\infty} \kappa(x)dx = 1$, triviální podmínky toho, že za jádro budeme považovat jen hustoty (v obecné teorii tomu tak nemusí být, často je opouštěna podmínka nezápornosti).

2. $\int_{-\infty}^{+\infty} x\kappa(x)dx = 0$; jádro je centrované, jaderná hustota má nulovou střední hodnotu

3. $\int_{-\infty}^{+\infty} x^2\kappa(x)dx = 1$; jádro je normované, jaderná hustota má jednotkový rozptyl (samozřejmě za platnosti podmínky 2.).

Základní vlastnosti odhadu $f_a(x)$

$\int_{-\infty}^{+\infty} f_a(x)dx = \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{+\infty} \kappa(\frac{x-x_i}{h})dx = (A)$ po substituci $\frac{x-x_i}{h} = y \Rightarrow x = hy + x_i \Rightarrow dx = hdy$

dostaneme $(A) = \frac{1}{nh} \sum_{i=1}^{n} h\int_{-\infty}^{+\infty} \kappa(y)dy = \frac{1}{nh} \sum_{i=1}^{n} h = 1$. Tedy zvolený typ aproximace je opět hustotou.

$\mu_a = E_{f_a}\{x\} = \int_{-\infty}^{+\infty} x f_a(x)dx = \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{+\infty} x\kappa(\frac{x-x_i}{h})dx = (B)$ po substituci

$\frac{x-x_i}{h} = y \Rightarrow x = hy + x_i \Rightarrow dx = hdy$; $(B) = \frac{1}{nh} \sum_{i=1}^{n} h\int_{-\infty}^{+\infty} (hy + x_i)\kappa(y)dy =$

$= \frac{1}{n} \sum_{i=1}^{n} h\int_{-\infty}^{+\infty} y\kappa(y)dy + \frac{1}{n} \sum_{i=1}^{n} x_i \int_{-\infty}^{+\infty} \kappa(y)dy = \frac{1}{n} \sum_{i=1}^{n} x_i$

=0 podle podmínky 2.

= 1 podle podmínky 1.

Tedy střední hodnota počítaná podle aproximační hustoty je rovna výběrovému průměru.

$$\sigma_a^2\{x\} = E\int_{-\infty}^{+\infty}(x-\mu_a)^2 f_a(x)dx = \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}(x-\mu_a)^2\kappa\left(\frac{x\text{-}x_i}{h}\right)dx = (C)$$ po substituci

$$\frac{x\text{-}x_i}{h} = y \Rightarrow x = hy + x_i \Rightarrow dx = hdy;\quad (C) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}(hy + x_i - \mu_a)^2\kappa(y)dy =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}\left[(hy)^2 + (x_i-\mu_a)^2 + 2hy(x_i-\mu_a)\right]\kappa(y)dy = \frac{1}{n}\sum_{i=1}^{n}\left[\int_{-\infty}^{+\infty}(hy)^2\kappa(y)dy + \int_{-\infty}^{+\infty}(x_i-\mu_a)^2\kappa(y)dy\right] +$$

$$+\frac{1}{n}\sum_{i=1}^{n}\left[\int_{-\infty}^{+\infty}2hy(x_i-\mu_a)^2\kappa(y)dy\right] = h^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_a)^2$$

$=h^2$ podle podmínky 3.

$=0$ podle podmínky 2.

$=(x_i-\mu_a)^2$ podle podmínky 1.

Pokud označíme $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_a)^2$, výběrový rozptyl, dostaneme:

$\sigma_a^2\{x\} = h^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_a)^2 = h^2 + \left(1-\frac{1}{n}\right)s_n^2$. Pokud bychom požadovali $\sigma_a^2\{x\} = s_n^2$, dostaneme volbu $h = \sqrt{\frac{s_n^2}{n}}$. **(1)**

„Nestrannost“ odhadu $f_a(x)$

Podmínky náhodného výběru nám dávají shodnou hustotu $f(x)$, nám nedostupnou, pro všechna pozorování tohoto náhodného výběru. Dále podle podmínek kladených na náhodný výběr (iid) platí pro sdruženou hustotu $f(x_1,...,x_n) = \prod_{i=1}^{n} f(x_i)$. Pak:

$$E_{x_1,x_2,...,x_n}\{f_a(x)\} = \int_{-\infty}^{+\infty}.....\int_{-\infty}^{+\infty}f_a(x)\left[\prod_{i=1}^{n}f(x_i)\right]dx_1.....dx_n =$$

$$= \int_{-\infty}^{+\infty}.....\int_{-\infty}^{+\infty}\left[\frac{1}{nh}\sum_{j=1}^{n}\kappa\left(\frac{x\text{-}x_j}{h}\right)\right]\left[\prod_{i=1}^{n}f(x_i)\right]dx_1....dx_n = \frac{1}{nh}\sum_{j=1}^{n}\int_{-\infty}^{+\infty}.....\int_{-\infty}^{+\infty}\kappa\left(\frac{x\text{-}x_j}{h}\right)\left[\prod_{i=1}^{n}f(x_i)\right]dx_1....dx_n =$$

$$= \frac{1}{nh}\sum_{j=1}^{n}\int_{-\infty}^{+\infty}\kappa\left(\frac{x\text{-}x_j}{h}\right)f(x_j)dx_j = \frac{1}{h}\int_{-\infty}^{+\infty}\kappa\left(\frac{x\text{-}z}{h}\right)f(z)dz = (D)$$

po substituci: $\frac{x-z}{h} = y; z = x-hy; dz = -hdy$ dostaneme $(D) = \int_{-\infty}^{+\infty}\kappa(y)f(x-hy)dy$.

Shrnuto: $E_{x_1,x_2,...,x_n}\{f_a(x)\} = \int_{-\infty}^{+\infty}\kappa(y)f(x-hy)dy$.

Potom $\lim_{h\to 0_+} E_{x_1,x_2,\ldots,x_n}\{f_a(x)\} = \lim_{h\to 0_+}\int_{-\infty}^{+\infty}\kappa(y)f(x-hy)dy = (E)$ a pokud jsou splněny předpoklady záměny limity a integrálu (zde v nevlastních mezích) a funkce $f(x)$ je spojitá, platí: $(E) = \int_{-\infty}^{+\infty}\kappa(y)\left(\lim_{h\to 0_+} f(x-hy)\right)dy = \int_{-\infty}^{+\infty}\kappa(y)\left(\lim_{h\to 0_+} f(x-\lim_{h\to 0_+}(hy))\right)dy = f(x)$.

Důkaz tohoto tvrzení za daleko obecnějších předpokladů je v: L. Devroye and L. Gyorfi: Nonparametric Density Estimation: The L1 View, str. 6-11. Pokud tedy je $h = h(n) \underset{n\to+\infty}{\to} 0$ je odhad hustoty $f_a(x)$ asymptoticky nestranný.

V L. Devroye and L. Gyorfi: Nonparametric Density Estimation: The L1 View, str.12-19 je dokázáno pro $J_n = \int_{-\infty}^{+\infty}\left|f_a(x) - f(x)\right|dx$: Následující podmínky jsou ekvivalentní:

1. $J_n \underset{n\to+\infty}{\to} 0$ v pravděpodobnosti pro některou hustotu $f(x)$.
2. $J_n \underset{n\to+\infty}{\to} 0$ v pravděpodobnosti pro každou hustotu $f(x)$.
3. $J_n \underset{n\to+\infty}{\to} 0$ skoro všude pro každou hustotu $f(x)$.
4. $J_n \underset{n\to+\infty}{\to} 0$ „exponenciálně“[1] pro každou hustotu $f(x)$. $J_n \underset{n\to+\infty}{\to} 0$ „exponenciálně“ znamená: $\forall\varepsilon>0;\exists r>0, n_0$ takové, že $n \ge n_0 \Rightarrow P(J_n \ge \varepsilon) \le e^{-rn}$. To ale také znamená konvergenci řady $\sum_{n=1}^{+\infty} P(J_n \ge \varepsilon)$ neboť má konvergentní majorantu e^{-rn}.

Tam uvedený důkaz potřebuje jen předpoklad 1. na jádro $\kappa(x)$.

Diskuse „nestrannosti“ jádrového odhadu hustoty

Platí: $E_{x_1,x_2,\ldots,x_n}\{f_a(x)\} = \int_{-\infty}^{+\infty}\kappa(y)f(x-hy)dy$, odtud:

$E_{x_1,x_2,\ldots,x_n}\{f_a(x)\}$ je hustotou

- Nezápornost je zřejmá
- $\int_{-\infty}^{+\infty} E_{x_1,x_2,\ldots,x_n}\{f_a(x)\}dx = \int_{-\infty-\infty}^{+\infty+\infty}\kappa(y)f(x-hy)dydx = 1$

Dále, použijeme-li ve vztahu $E_{x_1,x_2,\ldots,x_n}\{f_a(x)\} = \int_{-\infty}^{+\infty}\kappa(y)f(x-hy)dy$ substituci $z = hy \Rightarrow dy = \frac{dz}{h}$, dostaneme: $E_{x_1,x_2,\ldots,x_n}\{f_a(x)\} = \int_{-\infty}^{+\infty}\kappa(y)f(x-hy)dy = \int_{-\infty}^{+\infty}\frac{1}{h}\kappa(\frac{z}{h})f(x-z)dz$.

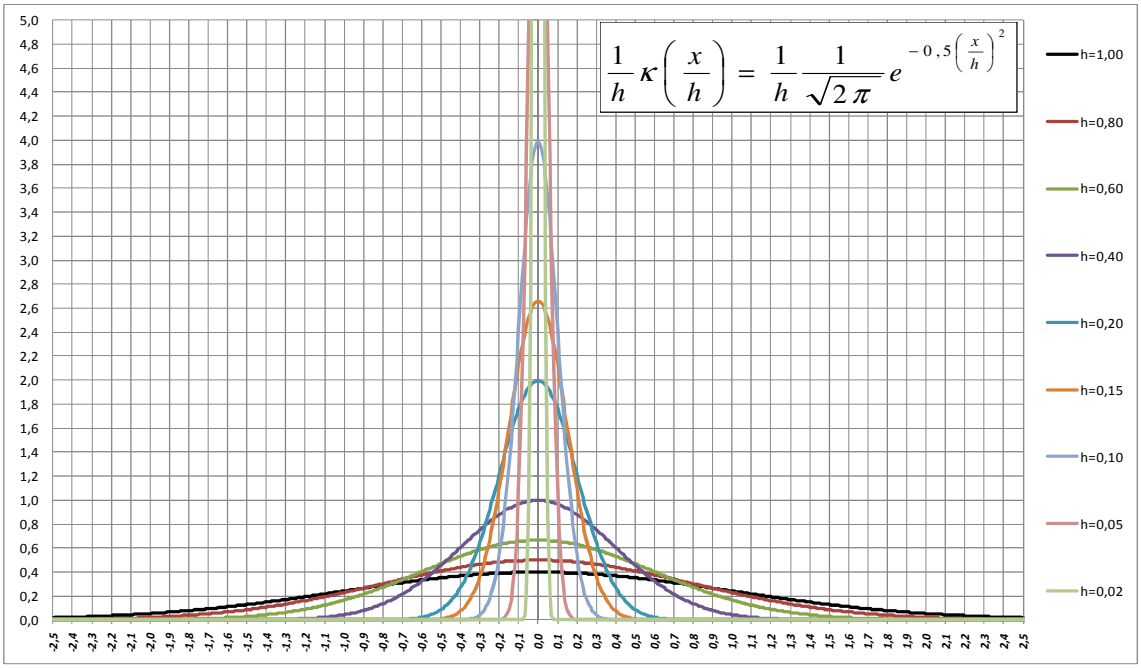
---

[1] $J_n \underset{n\to+\infty}{\to} 0$ „exponenciálně“ znamená: $\forall\varepsilon>0;\exists r>0, n_0$ takové, že $n \ge n_0 \Rightarrow P(J_n \ge \varepsilon) \le e^{-rn}$. To ale také znamená, že řada $\sum_{n=1}^{+\infty} P(J_n \ge \varepsilon)$ je konvergentní neboť má konvergentní majorantu e^{-rn}.

To je ale konvoluce dvou hustot $\frac{1}{h}\kappa(\frac{x}{h})$ a $f(x)$.

Proto: ***hustota*** $E_{x_1,x_2,\ldots,x_n}\{f_a(x)\}$ ***je hustotou náhodné proměnné, která je tvořena součtem původní (pozorované) náhodné proměnné a „na ní“ nezávislého šumu*** ε_h ***s hustotou*** $\frac{1}{h}\kappa(\frac{x}{h})$.

Máme-li „šum“ ε_1 s hustotou $\kappa(x)$, šum s hustotou $\frac{1}{h}\kappa(\frac{x}{h})$ bude $\varepsilon_h = h\varepsilon_1$. Příklad hustot takového „šumu“ je na následujícím obrázku:



„Vydatnost“ odhadu $f_a(x) = \frac{1}{nh}\sum_{i=1}^{n}\kappa(\frac{x-x_i}{h})$

$$E_{x_1,x_2,\ldots,x_n}\{f_a^2(x)\} = \int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty} f_a^2(x)\left[\prod_{i=1}^{n} f(x_i)\right]dx_1\ldots dx_n =$$

$$= \int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty}\left[\frac{1}{nh}\sum_{j=1}^{n}\kappa(\frac{x-x_j}{h})\right]^2\left[\prod_{i=1}^{n} f(x_i)\right]dx_1\ldots dx_n =$$

$$= \frac{1}{n^2h^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty}\kappa(\frac{x-x_j}{h})\kappa(\frac{x-x_k}{h})\left[\prod_{i=1}^{n} f(x_i)\right]dx_1\ldots dx_n =$$

$$= \frac{1}{n^2h^2}\left[\sum_{i=1}^{n}\int_{-\infty}^{+\infty}\kappa^2\left(\frac{x\text{-}x_k}{h}\right)f(x_k)dx_k + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\kappa\left(\frac{x\text{-}x_k}{h}\right)\kappa\left(\frac{x\text{-}x_j}{h}\right)f(x_k)f(x_j)dx_k dx_j\right] =$$

[2]

$$= \frac{1}{nh^2}\int_{-\infty}^{+\infty}\kappa^2\left(\frac{x\text{-}x_k}{h}\right)f(x_k)dx_k + \frac{n-1}{nh^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\kappa\left(\frac{x\text{-}x_k}{h}\right)\kappa\left(\frac{x\text{-}x_j}{h}\right)f(x_k)f(x_j)dx_k dx_j =$$

$$= \frac{1}{nh}\int_{-\infty}^{+\infty}\kappa^2(y)f(x-hy)dy + \frac{n-1}{n}\left(\int_{-\infty}^{+\infty}\kappa(y)f(x-hy)dy\right)^2.$$

Protože: $\sigma^2_{x_1,x_2,\ldots,x_n}\{f_a(x)\} = E_{x_1,x_2,\ldots,x_n}\{f_a^2(x)\} - \left(E_{x_1,x_2,\ldots,x_n}\{f_a(x)\}\right)^2$ a $E_{x_1,x_2,\ldots,x_n}\{f_a(x)\} = \int_{-\infty}^{+\infty}\kappa(y)f(x-hy)dy$, dostaneme pro rozptyl odhadu:

$\sigma^2_{x_1,x_2,\ldots,x_n}\{f_a(x)\} = \frac{1}{nh}\int_{-\infty}^{+\infty}\kappa^2(y)f(x-hy)dy - \frac{1}{n}\left(\int_{-\infty}^{+\infty}\kappa(y)f(x-hy)dy\right)^2$. Odtud bezprostředně (a z asymptotické nestrannosti) plyne potřeba $n\to+\infty \Rightarrow [(h\to 0)\wedge(nh\to+\infty)]$ pro existenci rozptylu a asymptotickou vydatnost (za předpokladu $h = h(n)$).

Shrnuto: $\sigma^2_{x_1,x_2,\ldots,x_n}\{f_a(x)\} = \frac{1}{nh}\int_{-\infty}^{+\infty}\kappa^2(y)f(x-hy)dy - \frac{1}{n}\left(\int_{-\infty}^{+\infty}\kappa(y)f(x-hy)dy\right)^2$.

Proto, aby byl takový odhad asymptoticky nestranný a asymptoticky vydatný, stačí:

$$n\to+\infty \Rightarrow (h\to 0)$$
$$n\to+\infty \Rightarrow (nh\to+\infty)$$

Příklady jádrových funkcí

„Obdélníkové jádro“:

$$\kappa(x) = \frac{1}{2a} \Leftrightarrow -a \le x \le +a; \quad a>0; \text{ jinak } \kappa(x) = 0 \qquad \textbf{(K1)}$$

1. Podmínka je zřejmá
2. Podmínka je též zřejmá, hustota je symetrická .

---

[2] $\sum_{j=n+1}^{n}\ldots = 0$

3. Podmínka $\int_{-a}^{+a} x^2 \frac{1}{2a}\,dx = \frac{1}{2a}\int_{-a}^{+a} x^2\,dx = \frac{1}{2a}\left[\frac{x^3}{3}\right]_{-a}^{+a} = \frac{1}{2a}\left[\frac{a^3}{3}+\frac{a^3}{3}\right] = \frac{a^2}{3} = 1 \Rightarrow a = \sqrt{3}$

„Epanechnikovo jádro“:

$$\kappa(x) = b\left(1-\left(\frac{x}{a}\right)^2\right) \Leftrightarrow -a \le x \le +a; \quad a,b>0;\ \text{jinak}\ \kappa(x)=0 \qquad \textbf{(K2)}$$

1. Podmínka: nezápornost je zřejmá, $\int_{-a}^{+a} b\left(1-\left(\frac{x}{a}\right)^2\right)dx = b\int_{-a}^{+a}\left(1-\left(\frac{x}{a}\right)^2\right)dx =$

   $= b\left[\left(1-\frac{x^3}{3a^2}\right)\right]_{-a}^{+a} = \frac{4}{3}ab = 1 \Rightarrow b = \frac{3}{4a}.$
2. Podmínka: je zřejmá, hustota je symetrická.
3. Podmínka:

   $\int_{-a}^{+a} bx^2\left(1-\left(\frac{x}{a}\right)^2\right)dx = b\int_{-a}^{+a}\left(x^2-\frac{x^4}{a^2}\right)dx = b\left[\frac{x^3}{3}-\frac{x^5}{5a^2}\right]_{-a}^{+a} = b\left(\frac{2a^3}{3}-\frac{2a^5}{5a^2}\right) = \frac{4a^3b}{15} = 1$

Řešení obou podmínek $b = \frac{3}{4a}$ a $\frac{4a^3b}{15} = 1$ dá $a = \sqrt{5};\ \ b = \frac{3}{4\sqrt{5}}$.

„Trojúhelníkové jádro“:

$$\kappa(x) = b\left(1-\left|\frac{x}{a}\right|\right) \Leftrightarrow -a \le x \le +a; \quad a,b>0;\ \text{jinak}\ \kappa(x)=0 \qquad \textbf{(K3)}$$

1. Podmínka: nezápornost je zřejmá,

   $2\int_{0}^{+a} b\left(1-\left(\frac{x}{a}\right)\right)dx = 2b\left[x-\frac{x^2}{2a}\right]_0^a = 2b\left(a-\frac{a^2}{2a}\right) = ab = 1$
2. Podmínka: je zřejmá, hustota je symetrická
3. Podmínka: $2\int_{0}^{+a} bx^2\left(1-\left(\frac{x}{a}\right)\right)dx = 2b\left[\frac{x^3}{3}-\frac{x^4}{4a}\right]_0^a = 2b\left(\frac{a^3}{3}-\frac{a^4}{4a}\right) = \frac{a^3b}{6} = 1$

Řešení obou podmínek $ab = 1$ a $\frac{a^3b}{6} = 1$ dá $a = \sqrt[3]{6};\ \ b = \frac{1}{\sqrt[3]{6}}$.

„Mocninné jádro“:

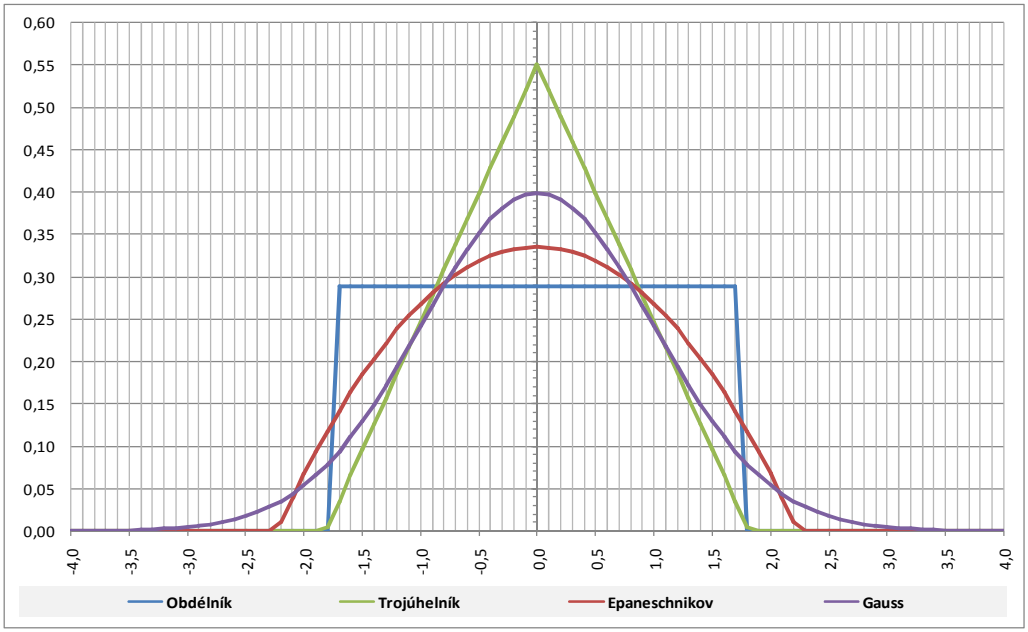
$$\kappa(x) = b\left(1-\left|\frac{x}{a}\right|^{\alpha}\right) \Leftrightarrow -a \le x \le +a;\ \alpha>0;\ \ a,b>0;\ \text{jinak}\ \kappa(x)=0 \qquad \textbf{(K4)}$$

Podmínky a hodnoty konstant a,b se odvodí obdobným způsobem jako u „trojúhelníkového jádra“

„Gaussovo jádro“:

$$\kappa(x) = \frac{1}{\sqrt{2\pi}} e^{-0,5x^2} \qquad \textbf{(K5)}$$

1. Podmínka je zřejmá, jde o normální, centrovanou a normovanou hustotu.
2. Podmínka je opět zřejmá, hustota je symetrická kolem nuly.
3. Podmínka je také zřejmá, hustota je centrovaná a normovaná.



Odpovídající modely distribučních funkcí

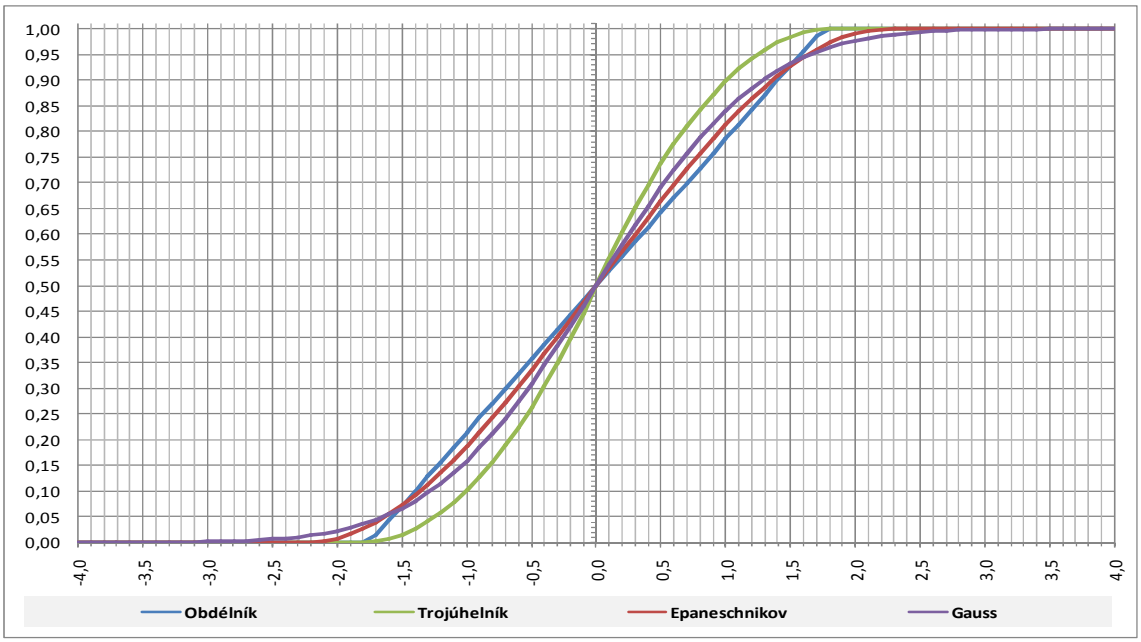
K danému tvaru jádrové funkce přiřadíme: $K(x) = \int_{-\infty}^{x} \kappa(z)dz$. Pak pro aproximaci distribuční funkce dostaneme: $F_a(x) = \frac{1}{n} \sum_{i=1}^{n} K(\frac{x\text{-}x_i}{h})$.

Např.: pro „obdélníkové jádro“: $\kappa(x) = \frac{1}{2a} \Leftrightarrow -a \le x \le +a; \quad a > 0;$ jinak $\kappa(x) = 0$ je odpovídající „distribuční jádro“:

$$K(x) = \frac{1}{a}x + 1 \Leftrightarrow -a \le x \le a;\ K(x) = 0 \Leftrightarrow x < -a;\ K(x) = 1 \Leftrightarrow x > a$$

Po dosazení $a = \sqrt{3}$ máme:

$$K(x) = \frac{1}{\sqrt{2}}x + 1 \Leftrightarrow -\sqrt{2} \le x \le \sqrt{2};\ K(x) = 0 \Leftrightarrow x < -\sqrt{2};\ K(x) = 1 \Leftrightarrow x > \sqrt{2}$$



Volba parametru h

Jedna možnost byla uvedena výše: $h = \sqrt{\frac{s_n^2}{n}}$ z podmínky aby neparametrická hustota měla stejný rozptyl jako výběrový. Další přístupy vycházejí z některých kritérií:

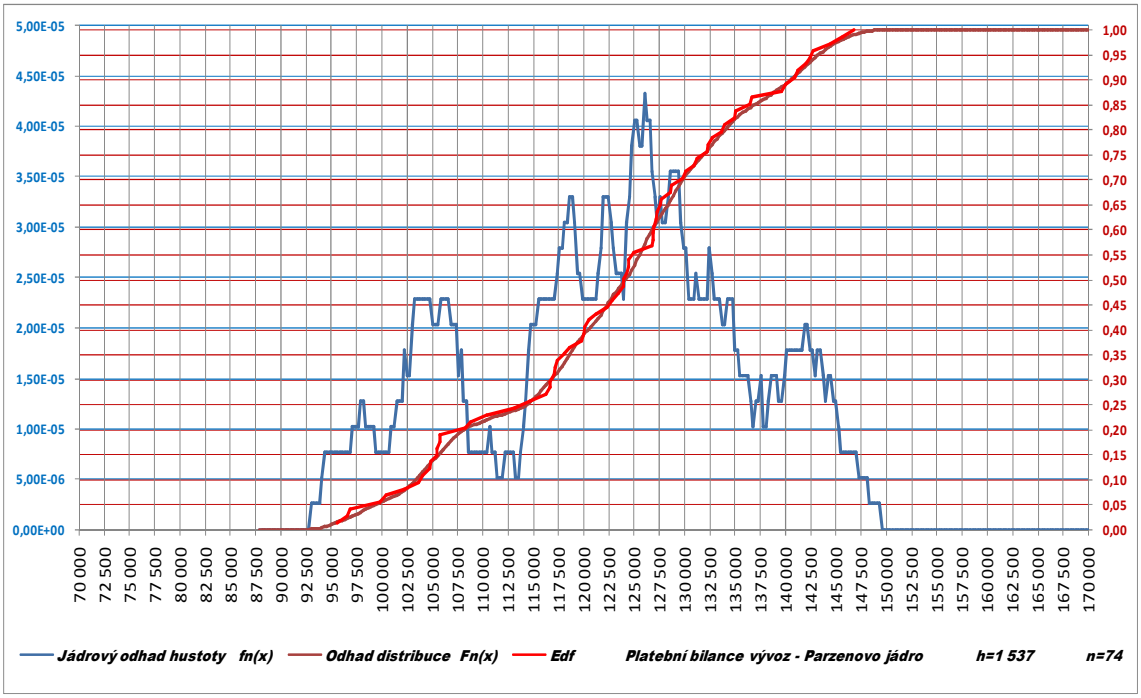
$E\left\{\left|f_{a,h,n}(x) - f(x)\right|\right\} \xrightarrow[h]{} min \quad \forall x$; $E\left\{\sup_x \left|f_{a,h,n}(x) - f(x)\right|\right\} \xrightarrow[h]{} min(inf)$;

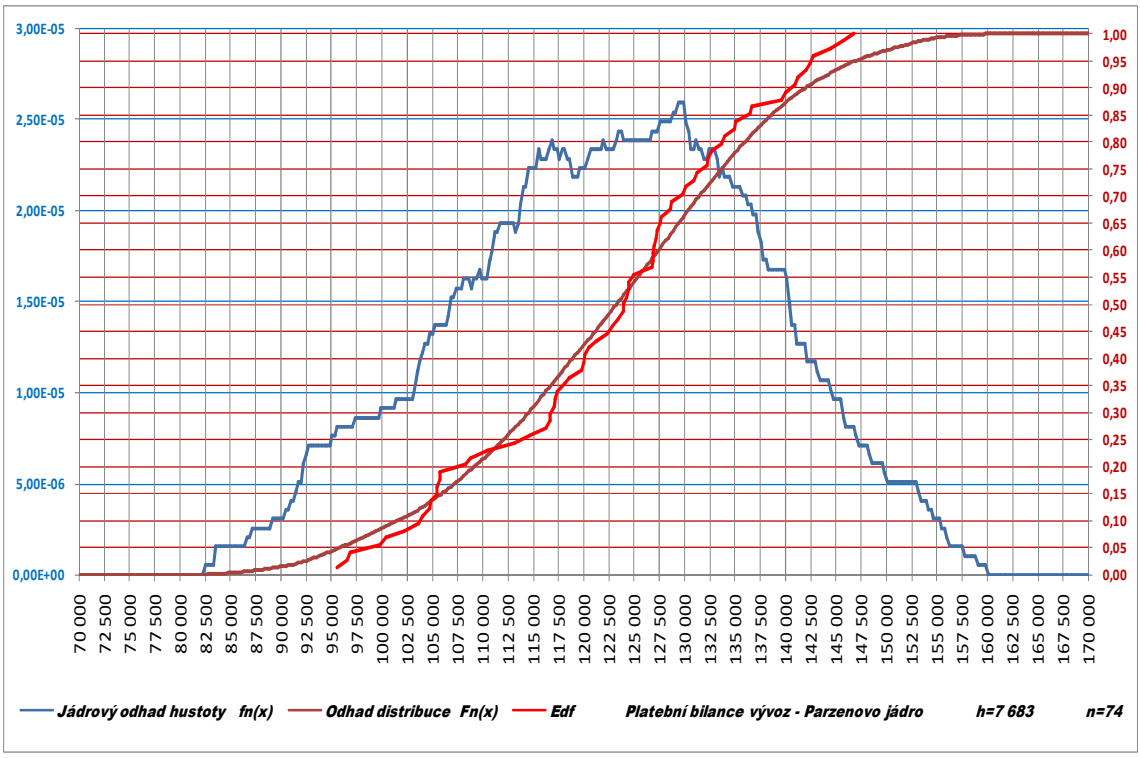
$\sup_f \left|E\left\{f_{a,h,n}(x)\right\} - f(x)\right| \xrightarrow[h]{} min(inf) \quad \forall x$; … . Jejich přehled lze nalézt např. v:

L. Devroye and L. Gyorfi Nonparametric Density Estimation: The L1 View. John Wiley, New York, 1985. Nebo v:

Luc Devroye and Gábor Lugosi: Variable Kernel Estimates: on the Impossibility of Tuning the Parameters. (School of Computer Science, McGill University, Montreal, Canada H3A2A 7, E-mail: luc@cs.mcgill.ca; Department of Economics, Pompeu Fabra University, Ramon Trias Fargas, 25-27, 08005 Barcelona, Spain, E-mail: gabor.lugosi@econ.upf.ed)

Příklady


5,00E-05
4,50E-05
4,00E-05
3,50E-05
3,00E-05
2,50E-05
2,00E-05
1,50E-05
1,00E-05
5,00E-06
0,00E+00
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
70 000
72 500
75 000
77 500
80 000
82 500
85 000
87 500
90 000
92 500
95 000
97 500
100 000
102 500
105 000
107 500
110 000
112 500
115 000
117 500
120 000
122 500
125 000
127 500
130 000
132 500
135 000
137 500
140 000
142 500
145 000
147 500
150 000
152 500
155 000
157 500
160 000
162 500
165 000
167 500
170 000
Jádrový odhad hustoty fn(x)
Odhad distribuce Fn(x)
Edf
Platební bilance vývoz - Parzenovo jádro
h=1 537
n=74


3,00E-05
2,50E-05
2,00E-05
1,50E-05
1,00E-05
5,00E-06
0,00E+00
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
70 000
72 500
75 000
77 500
80 000
82 500
85 000
87 500
90 000
92 500
95 000
97 500
100 000
102 500
105 000
107 500
110 000
112 500
115 000
117 500
120 000
122 500
125 000
127 500
130 000
132 500
135 000
137 500
140 000
142 500
145 000
147 500
150 000
152 500
155 000
157 500
160 000
162 500
165 000
167 500
170 000
Jádrový odhad hustoty fn(x)
Odhad distribuce Fn(x)
Edf
Platební bilance vývoz - Parzenovo jádro
h=7 683
n=74

Jádrové modely dvou proměnných

$$f_a(x,y)=\frac{1}{nh_xh_y}\sum_{i=1}^{n}\kappa_x\left(\frac{x-x_i}{h_x}\right)\kappa_y\left(\frac{y-y_i}{h_y}\right)$$

$$F_a(x,y)=\frac{1}{n}\sum_{i=1}^{n}K_x\left(\frac{x-x_i}{h_x}\right)K_y\left(\frac{y-y_i}{h_y}\right)$$

Je evidentní, že obě aproximace jsou „konzistentní“ vůči operaci přechodu k marginálním rozdělením.

Tj.: $f_a(x)=\int_{-\infty}^{+\infty}f_a(x,y)dy=\frac{1}{nh_xh_y}\sum_{i=1}^{n}\kappa_x\left(\frac{x-x_i}{h_x}\right)\int_{-\infty}^{+\infty}\kappa_y\left(\frac{y-y_i}{h_y}\right)dy=\frac{1}{nh_x}\sum_{i=1}^{n}\kappa_x\left(\frac{x-x_i}{h_x}\right)$

Pro další výpočty je užitečný následující triviální vztah: $\int_{-\infty}^{+\infty}x\frac{1}{h_x}\kappa_x\left(\frac{x-x_i}{h_x}\right)dx=x_i$.

K tomu aby platil, stačí předpoklady: $\int_{-\infty}^{+\infty}\kappa(x)dx=1,\ \int_{-\infty}^{+\infty}x\kappa(x)dx=0$.

Míry vztahu obou proměnných:

Korelace

$$E\{xy\}=\frac{1}{nh_xh_y}\sum_{i=1}^{n}\int_{-\infty-\infty}^{+\infty+\infty}xy\,\kappa_x\left(\frac{x-x_i}{h_x}\right)\kappa_y\left(\frac{y-y_i}{h_y}\right)dxdy=\frac{1}{n}\sum_{i=1}^{n}x_iy_i\ \Rightarrow$$

$$Cov(x,y)=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{x}\bar{y}\ \Rightarrow$$

$$Corr(x,y)=\frac{\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{x}\bar{y}}{\sqrt{h_x^2+\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{h_y^2+\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$, a při volbě $h_x=\sqrt{\frac{s_x^2}{n}}$, $h_y=\sqrt{\frac{s_y^2}{n}}$

dostaneme $Corr(x,y)=\frac{\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{x}\bar{y}}{\sqrt{s_x^2}\sqrt{s_y^2}}$.

Regrese

Podmíněné rozdělení:

$$f_a(x,y)=\frac{1}{nh_xh_y}\sum_{i=1}^{n}\kappa_x\left(\frac{x-x_i}{h_x}\right)\kappa_y\left(\frac{y-y_i}{h_y}\right)\Rightarrow$$

$$f_a(y/x)=\frac{\frac{1}{nh_xh_y}\sum_{i=1}^{n}\kappa_x\left(\frac{x-x_i}{h_x}\right)\kappa_y\left(\frac{y-y_i}{h_y}\right)}{\frac{1}{nh_x}\sum_{i=1}^{n}\kappa_x\left(\frac{x-x_i}{h_x}\right)}=\frac{\frac{1}{h_y}\sum_{i=1}^{n}\kappa_x\left(\frac{x-x_i}{h_x}\right)\kappa_y\left(\frac{y-y_i}{h_y}\right)}{\sum_{i=1}^{n}\kappa_x\left(\frac{x-x_i}{h_x}\right)}$$ a odtud

vlastní regrese:

$$r(x)=E\{y/x\}=\int_{-\infty}^{+\infty}yf_a(y/x)dy=\frac{\frac{1}{h_y}\sum_{i=1}^{n}\kappa_x\left(\frac{x-x_i}{h_x}\right)\int_{-\infty}^{+\infty}y\kappa_y\left(\frac{y-y_i}{h_y}\right)dy}{\sum_{i=1}^{n}\kappa_x\left(\frac{x-x_i}{h_x}\right)}=\frac{\sum_{i=1}^{n}y_i\kappa_x\left(\frac{x-x_i}{h_x}\right)}{\sum_{i=1}^{n}\kappa_x\left(\frac{x-x_i}{h_x}\right)}$$

Skedastická funkce

$sked\ (x)=E\left\{\left(y-E\{y/x\}\right)^2/x\right\}=E\left\{y^2/x\right\}-\left(E\{y/x\}\right)^2$. Pak

$$E\left\{y^2/x\right\}=\frac{\sum_{i=1}^{n}\kappa_x\left(\frac{x-x_i}{h_x}\right)\int_{-\infty}^{+\infty}y^2\frac{1}{h_y}\kappa_y\left(\frac{y-y_i}{h_y}\right)dy}{\sum_{i=1}^{n}\kappa_x\left(\frac{x-x_i}{h_x}\right)}=\frac{\sum_{i=1}^{n}\kappa_x\left(\frac{x-x_i}{h_x}\right)\left(h_y^2+y_i^2\right)}{\sum_{i=1}^{n}\kappa_x\left(\frac{x-x_i}{h_x}\right)}=(F),\ \text{protože}$$

$$\int_{-\infty}^{+\infty}y^2\frac{1}{h_y}\kappa_y\left(\frac{y-y_i}{h_y}\right)dy=\int_{-\infty}^{+\infty}\left(h_yz+y_i\right)^2\kappa_y(z)dz=\int_{-\infty}^{+\infty}\left(h_y^2z^2+2h_yzy_i+y_i^2\right)\kappa_y(z)dz=h_y^2+y_i^2,$$

potom: $(F)=h_y^2+\frac{\sum_{i=1}^{n}y_i^2\kappa_x\left(\frac{x-x_i}{h_x}\right)}{\sum_{i=1}^{n}\kappa_x\left(\frac{x-x_i}{h_x}\right)}=h_y^2+\sum_{i=1}^{n}y_i^2\frac{\kappa_x\left(\frac{x-x_i}{h_x}\right)}{\sum_{j=1}^{n}\kappa_x\left(\frac{x-x_j}{h_x}\right)}$.

Shrnuto:

$$sked(x)=h_y^2+\sum_{i=1}^{n}y_i^2\frac{\kappa_x\left(\frac{x-x_i}{h_x}\right)}{\sum_{j=1}^{n}\kappa_x\left(\frac{x-x_j}{h_x}\right)}-\left(\sum_{i=1}^{n}y_i\frac{\kappa_x\left(\frac{x-x_i}{h_x}\right)}{\sum_{j=1}^{n}\kappa_x\left(\frac{x-x_j}{h_x}\right)}\right)^2=h_y^2+\sum_{i=1}^{n}y_i^2\frac{\kappa_x\left(\frac{x-x_i}{h_x}\right)}{\sum_{j=1}^{n}\kappa_x\left(\frac{x-x_j}{h_x}\right)}-r^2(x)$$

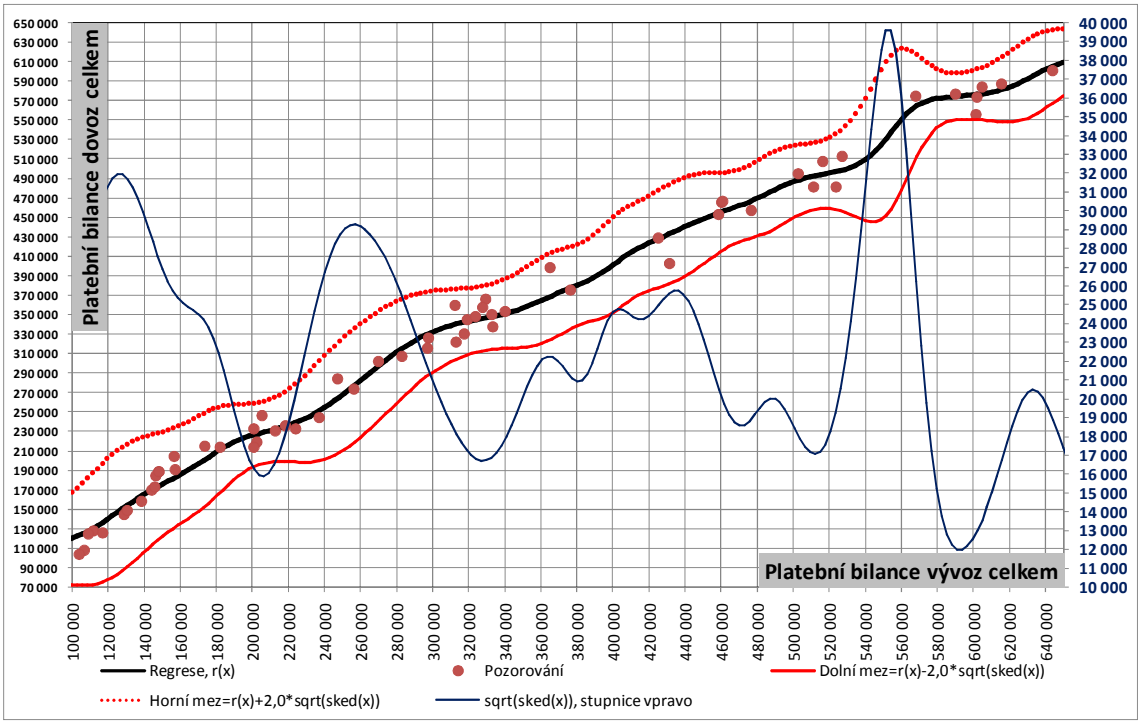
V obou, jak regresní, tak skedastická funkci je důležitá řada funkcí $w_i(x)=\frac{\kappa_x\left(\frac{x-x_i}{h_x}\right)}{\sum_{j=1}^{n}\kappa_x\left(\frac{x-x_j}{h_x}\right)}=$

vah. S tímto označením dostaneme:

$r(x)=\sum_{i=1}^{n}y_iw_i(x)$, kde $\forall x\in defX;\ \ w_i(x)\geq 0;\ \ \sum_{i=1}^{n}w_i(x)=1$

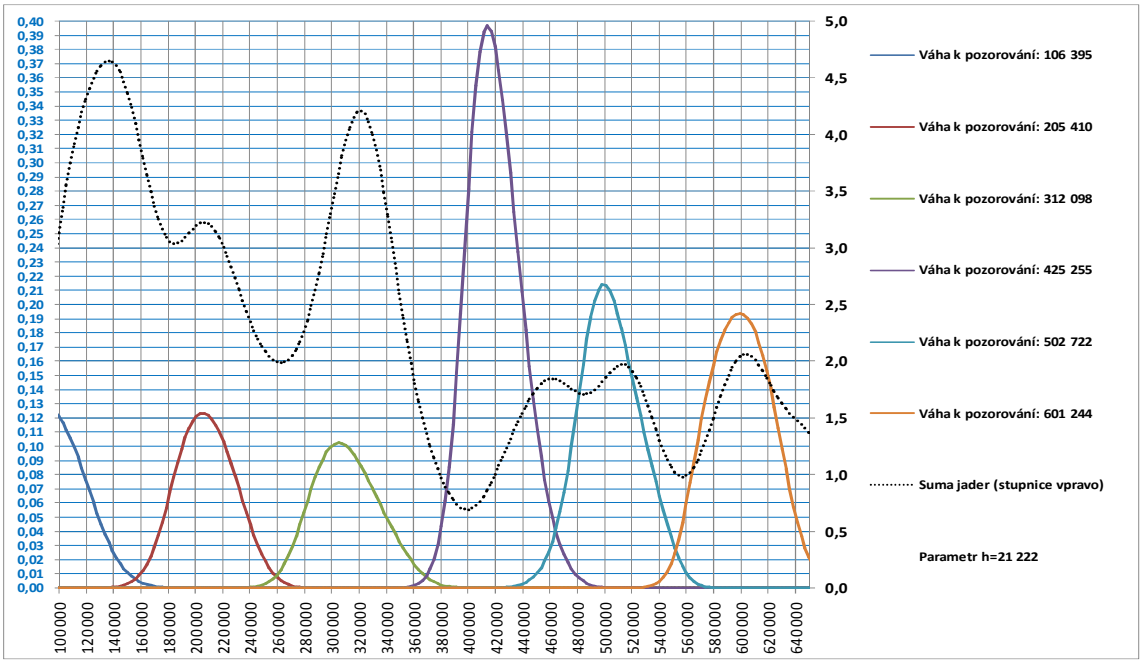
$sked(x)=h_y^2+\sum_{i=1}^{n}y_i^2w_i(x)-r^2(x).$

Příklad regrese a skedastické funkce



Gaussovo jádro h_x=21 222, h_y=19 166, n=65, $h = \sqrt{\frac{s_n^2}{n}}$. Data z časových řad nebyla stacionarizována.

Příklad váhových funkcí, odpovídajících výše zobrazené regresi



Doporučená a zdrojová literatura:

L. Devroye and L. Gyorfi	Nonparametric Density Estimation: The L1 View	John Wiley, New York, 1985.
A. Berlinet and L. Devroye	A Comparison of Kernel Density Estimates	Working paper, School of Computer Science, McGill University, Montreal, Canada, H3A 2A7.
Alfréd Rényi	Teorie pravděpodobnosti	ACADEMIA, Praha 1972.