Jádrové odhady podmíněné hustoty

Kateřina Konečná, Ivanka Horová a Jan Koláček

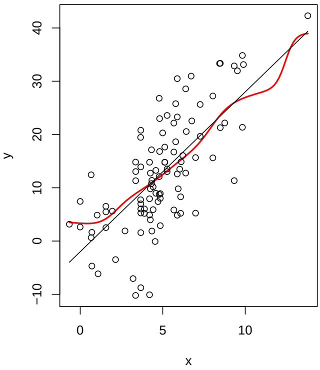
Ústav matematiky a statistiky
Masarykova univerzita
Brno

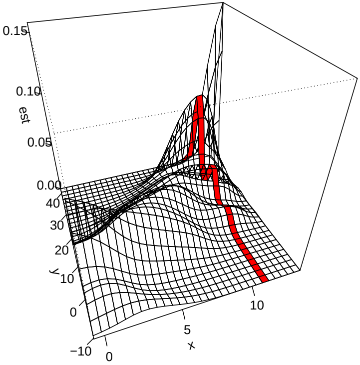
workshopy
Finanční matematika v praxi III
Matematické modely a aplikace
Podlesí
3. – 6. září 2013

evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
AMathNet síť pro transfer znalostí v aplikované matematice
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

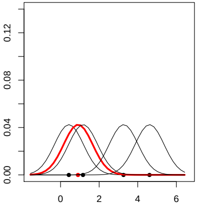
Obsah

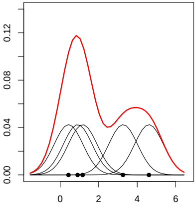
Motivace



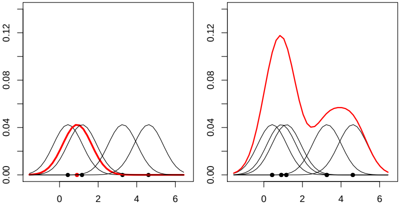


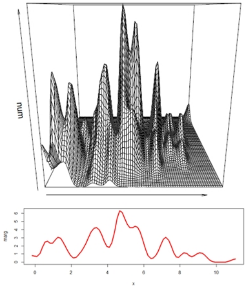
Konstrukce jádrových odhadů podmíněné hustoty

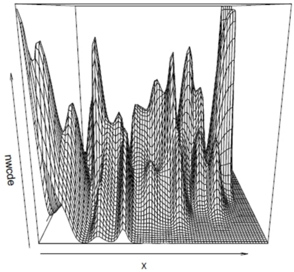




Konstrukce jádrových odhadů podmíněné hustoty







Jádro

Definition

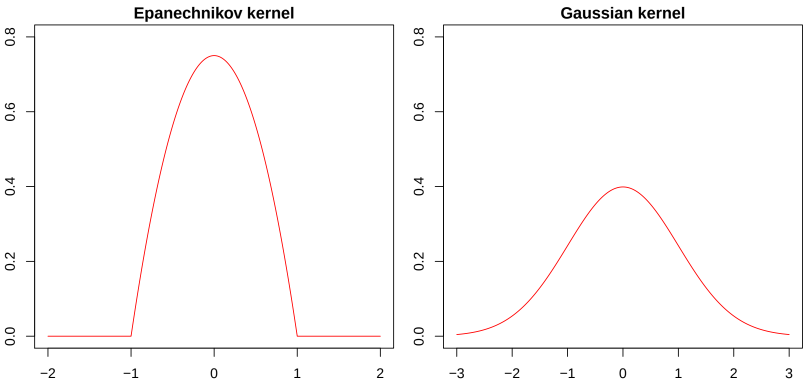
Reálná funkce K splňující

1. $K \in Lip[-1,1]$, tj. $|K(x) - K(y)| \leq L|x - y|$, $\forall x, y \in [-1,1]$, $L > 0$,
2. $supp(K) = [-1,1]$,
3. momentové podmínky:

$$\int_{-1}^{1} x^j K(x)\,\mathrm{d}x = \begin{cases} 1 & j = 0, \\ 0 & j = 1, \\ \beta_2 \neq 0 & j = 2 \end{cases}$$

se nazývá jádro řádu 2.

Jádro


Epanechnikov kernel
0.0
0.2
0.4
0.6
0.8
−2
−1
0
1
2
Gaussian kernel
0.0
0.2
0.4
0.6
0.8
−3
−2
−1
0
1
2
3

Podmíněná hustota

Nechť $X \in \mathbb{R}$ je náhodná veličina a $h(x)$ její hustota, $Y \in \mathbb{R}$ náhodná veličina a $f(x, y)$ sdružená hustota náhodného vektoru (X, Y). Podmíněná hustota $Y|(X = x)$ je definována

$$f(y|x) = \frac{f(x, y)}{h(x)}.$$

Jádrový odhad marginální hustoty:

$$\widehat{h}(x) = \frac{1}{nh_x} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_x}\right) \tag{1}$$

Jádrový odhad sdružené hustoty:

$$\widehat{f}(x, y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_x}\right) K\left(\frac{y - Y_i}{h_y}\right) \tag{2}$$

Nadaraya-Watsonův odhad podmíněné hustoty

Jádrový odhad podmíněné hustoty náhodné veličiny $Y|(X = x)$ je definován

$$\hat{f}_{NW}(y|x) = \frac{1}{h_y} \sum_{i=1}^{n} w_i(x) K\left(\frac{y - Y_i}{h_y}\right), \tag{3}$$

kde

$$w_i(x) = \frac{K\left(\frac{x - X_i}{h_x}\right)}{\sum\limits_{j=1}^{n} K\left(\frac{x - X_j}{h_x}\right)} \tag{4}$$

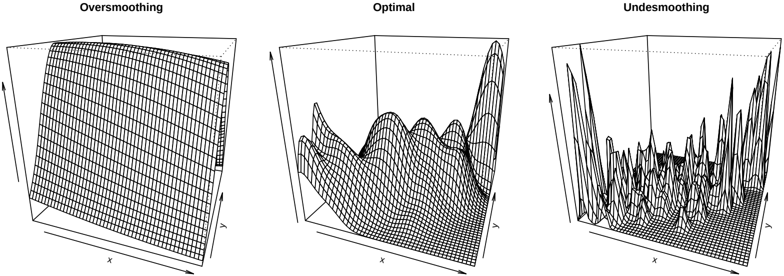
je váhová funkce.
Odhad se nazývá Nadaraya-Watsonův odhad podmíněné hustoty.

Míra kvality odhadu $\widehat{f}(y|x)$

- vyhlazovací matice $\mathbf{H} = \begin{pmatrix} h_x & 0 \\ 0 & h_y \end{pmatrix}$
- vyhlazovací parametry h_x, h_y určují vyhlazení ve směru x a y

Míra kvality odhadu $\widehat{f}(y|x)$

- vyhlazovací matice $\mathbf{H} = \begin{pmatrix} h_x & 0 \\ 0 & h_y \end{pmatrix}$
- vyhlazovací parametry h_x, h_y určují vyhlazení ve směru x a y



Míra kvality odhadu $\widehat{f}(y|x)$

- MISE: **M**ean **I**ntegrated **S**quare **E**rror

$$\mathrm{MISE}\left\{\widehat{f}(y|x)\right\} = \iint \mathrm{E}\left\{\widehat{f}(y|x) - f(y|x)\right\}^2 h(x)\,\mathrm{d}x\,\mathrm{d}y$$

- AMISE: **A**symptotic **M**ean **I**ntegrated **S**quare **E**rror

$$\mathrm{AMISE}\{\hat{f}_{NW}(y|x)\} = \frac{c_1}{nh_x h_y} - \frac{c_2}{nh_x} + c_3 h_x^4 + c_4 h_y^4 + c_5 h_x^2 h_y^2$$

- optimální šířka vyhlazovacích parametrů:

$$(h_x, h_y) = \operatorname*{arg\,min}_{(h_x,h_y)\in H} \mathrm{AMISE}$$

Metody volby vyhlazovacích parametrů

- metody kombinující metody pro odhad parametrů regrese a hustoty
  - pro hustotu: cross-validace, metoda referenční hustoty
  - pro regresi: cross-validace, metoda penalizačních funkcí
- metoda referenční hustoty
  - předpoklad normálního rozložení podmíněné hustoty

$$f(y|x) = \frac{1}{\sqrt{2\pi}(p+qx)} \mathrm{e}^{-\frac{1}{2}\left(\frac{y-a-bx}{p+qx}\right)^2}.$$

  - předpoklad normální nebo rovnoměrně spojité marginální hustoty
- cross-validace
  - $CV(h_x, h_y) = \frac{1}{n}\sum_{i=1}^{n}\int \hat{f}_{-i}(y|X_i)^2\,\mathrm{d}y - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(Y_i|X_i)$
- metoda penalizačních funkcí
- bootstrap

Iterační metoda

- vyhlazovací parametry dány minimalizací

$$\mathrm{AMISE}\left\{\widehat{f}(y|x)\right\} = \mathrm{AIV}\left\{\widehat{f}(y|x)\right\} + \mathrm{AISB}\left\{\widehat{f}(y|x)\right\}$$

- namísto vztahu

$$\mathrm{AIV}\left\{\widehat{f}(y|x)\right\} = 2\mathrm{AISB}\left\{\widehat{f}(y|x)\right\} \tag{5}$$

závisejícího na neznámé hustotě budeme uvažovat

$$\widehat{\mathrm{AIV}}\left\{\widehat{f}(y|x)\right\} = 2\widehat{\mathrm{ISB}}\left\{\widehat{f}(y|x)\right\} \tag{6}$$

Iterační metoda

$$\widehat{\mathrm{AIV}}\left\{\widehat{f}(y|x)\right\} = 2\widehat{\mathrm{ISB}}\left\{\hat{f}(y|x)\right\},\ \text{kde}$$

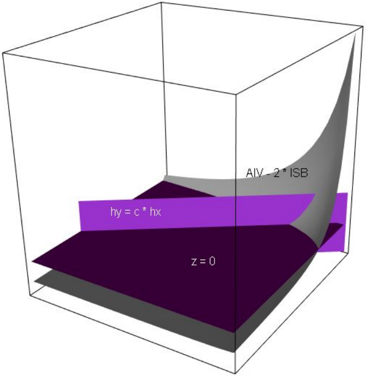
$$\widehat{\mathrm{AIV}}\left\{\widehat{f}(y|x)\right\} = \frac{c_1}{nh_x h_y},\ c_1 = \int R^2(K)\,\mathrm{d}x \text{ a } R(K) = \int K^2(x)\,\mathrm{d}x$$

$$\widehat{\mathrm{ISB}}\left\{\hat{f}(y|x)\right\} = \iint \left\{ \frac{1}{h_y} \frac{\sum_i (K*K)\left(\frac{x-X_i}{h_x}\right)(K*K)\left(\frac{y-Y_i}{h_y}\right)}{\sum_i (K*K)\left(\frac{x-X_i}{h_x}\right)} - \frac{1}{h_y} \frac{\sum_i K\left(\frac{x-X_i}{h_x}\right) K\left(\frac{y-Y_i}{h_y}\right)}{\sum_i K\left(\frac{x-X_i}{h_x}\right)} \right\}^2 h(x)\,\mathrm{d}x\,\mathrm{d}y$$

Iterační metoda

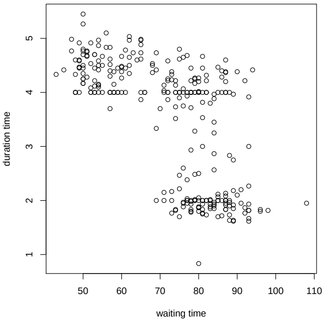
- systém dvou nelineárních rovnic:

$$\widehat{\mathrm{AIV}}(\hat{h}_x, \hat{h}_y) = 2\widehat{\mathrm{ISB}}(\hat{h}_x, \hat{h}_y)$$
$$\hat{h}_y = \hat{c}\hat{h}_x$$

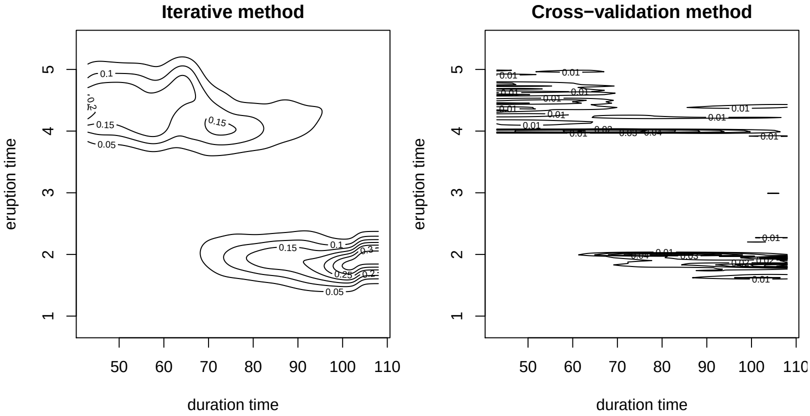


Reálná data: gejzír Old Faithful

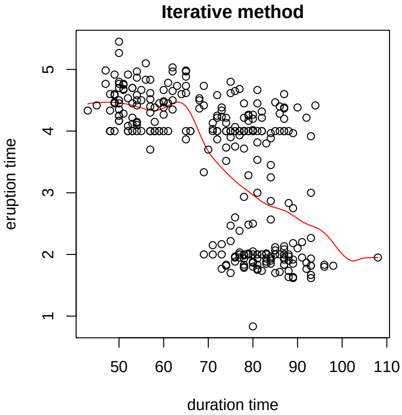
- 299 pozorování erupcí gejzíru Old Faithful, Yellowstonský národní park, Wyoming, USA, 1989 – 1990
- 2 proměnné:
    - nezávislá proměnná: doba čekání do následující erupce [min]
    - závislá proměnná: doba erupce [min]

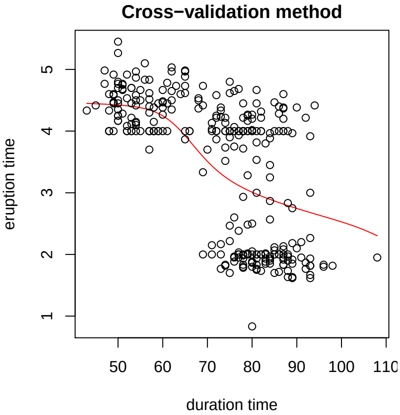


Gejzír Old Faithful



Old Faithful geyser





Literatura

- BASHTANNYK, D. M., HYNDMANN, R. J. *Bandwidth Selection for Kernel Conditional Density Estimation.* Computational Statistics & Data Analysis, 2001, Volume 36, Issue 3, Pages 279–298
- DE GOOIJER, J. G., ZEROM, D. *On Conditional Density Estimation.* Statistica Neerlandica, Netherlands Society for Statistics and Operations Research, 2003, Volume 57, Issue 2, Pages 159–176
- HANSEN, B. E. *Nonparametric Conditional Density Estimation.* 2004. [online][cit. leden 2012]. Dostupné z: `http://www.ssc.wisc.edu/~bhansen/papers/ncde.pdf`
- HYNDMAN, R. J., BASHTANNYK, D. M., GRUNWALD, G. K. *Estimating and Visualising Conditional Densities.* Journal of Computational and Graphical Statistics, 1996, Volume 5, Number 4, Pages 315–336

Děkuji za pozornost.