# POROVNÁNÍ METOD ODHADU VYHLAZOVACÍHO PARAMETRU PŘI JÁDROVÝCH ODHADECH HUSTOTY

**JAN ORAVA**

orava@mail.muni.cz
Ústav matematiky a statistiky, MU Brno

## SUMMARY

**Jádrový odhad pravděpodobnostní hustoty patří do skupiny neparametrických odhadů, tedy nepotřebujeme znát žádné apriorní informace o neznámé hustotě. Kvalita výsledného odhadu závisí zejména na vhodné volbě vyhlazovacího parametru. Příspěvek srovnává účinnost šesti různých metod odhadu vyhlazovacího parametru.**

## JÁDROVÝ ODHAD HUSTOTY

Jádrový odhad hustoty funkce $f$ označíme $\hat{f}(\cdot, h)$, $h$ značí vyhlazovací parametr a funkci $\mathsf{K}$ nazveme jádrem.

**1 JÁDROVÝ ODHAD HUSTOTY**

$$\hat{f}(x,h) = \frac{1}{nh}\sum_{i=1}^{n} \mathsf{K}\left(\frac{x - X_i}{h}\right).$$

**2 OBECNÁ DEFINICE JÁDRA**

Nechť je dána reálná funkce $\mathsf{K}$, která splňuje podmínky

1. $\mathsf{K} \in Lip[-1,1]$,
2. $supp(\mathsf{K}) = [-1,1]$,
3. $\nu$, $k$, $0 \le \nu < k$ jsou celá nezáporná čísla stejné parity a

$$\int_{-1}^{1} x^j \mathsf{K}(x)\,\mathrm{d}x = \begin{cases} 0, & 0 \le j < k,\ j \neq \nu, \\ (-1)^\nu \nu!, & j = \nu, \\ \mu_k \neq 0, & j = k. \end{cases}$$

nazýváme jádrem řádu $(\nu, k)$. Třídu takových funkcí označíme $S_{\nu,k}$.

## VYHLAZOVACÍ PARAMETR

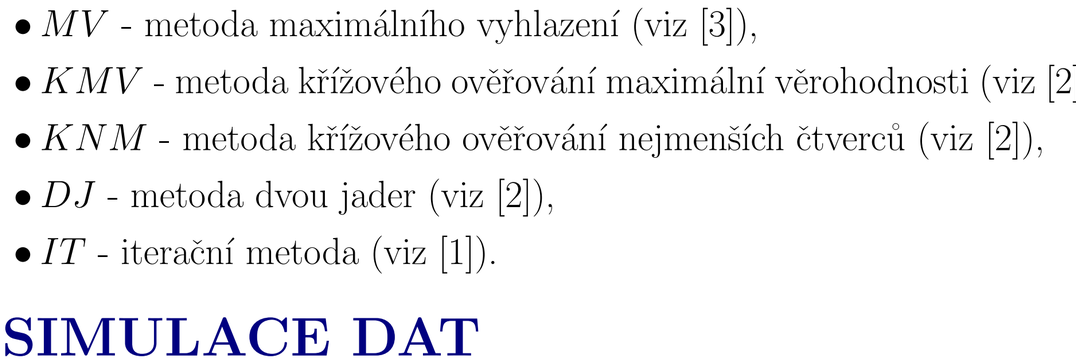
Existuje několik metod volby vyhlazovacího parametru s různou výpočetní náročností a s různou přesností výsledného odhadu. V simulační studii porovnáme přesnost šesti metod odhadu:

- $REF$ - metoda referenční hustoty (viz [3]),
- $MV$ - metoda maximálního vyhlazení (viz [3]),
- $KMV$ - metoda křížového ověřování maximální věrohodnosti (viz [2]),
- $KNM$ - metoda křížového ověřování nejmenších čtverců (viz [2]),
- $DJ$ - metoda dvou jader (viz [2]),
- $IT$ - iterační metoda (viz [1]).

## SIMULACE DAT

Pro simulaci zvolíme hustoty čtyř různých rozdělení:

- $f_1 \sim N(0,1)$, tj. $f_1(x) = \frac{1}{\sqrt{2\pi}} \exp -\frac{x^2}{2}$,
- $f_2 \sim 0.5 \cdot N(-2, 0.6^2) + 0.5 \cdot N(1, 1.3^2)$, tj. $f_2(x) = \frac{0{,}5}{0{,}6\sqrt{2\pi}} \exp -\frac{(x+2)^2}{2\cdot 0{,}6^2} + \frac{0{,}5}{1{,}3\sqrt{2\pi}} \exp -\frac{(x-1)^2}{2\cdot 1{,}3^2}$,
- $f_3 \sim \beta(2,8)$, tj. $f_3(x) = \frac{1}{\int_0^1 t(1-t)^7\,\mathrm{d}t} x(1-x)^7$,
- $f_4 \sim lnN(0, 0.5^2)$, tj. $f_4(x) = \frac{1}{0{,}5x\sqrt{2\pi}} \exp -\frac{\ln(x)^2}{2\cdot 0{,}5^2}$.




Pro každou hustotu provedeme 200 simulací náhodného výběru velikosti $n = 50$. Úmyslně volíme soubor menšího rozsahu. V případě velkého $n$ budou dávat metody lepší výsledky, ovšem ne vždy je v praxi možné získat dostatečný počet pozorování.

[1] V případě metody dvou jader byly použity jádra řádu k=2/l=4, k=4/l=6, k=6/l=8 a k=2/l=8.

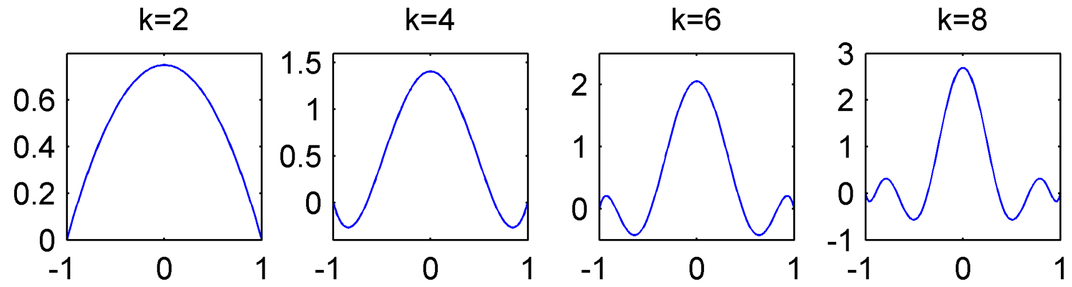
## HODNOTÍCÍ FUNKCE

Pro simulaci definujeme hodnotící funkci ve tvaru

$$F_{XY} = \sum_{\frac{k}{2}=1}^{4} \sum_{i=1}^{4} \frac{|\overline{h}_{XY,i,k} - h_{AMISE,i,k}|}{h_{AMISE,i,k}} \sigma(\overline{h}_{XY,i,k}),$$

kde $XY$ označuje použitou metodu odhadu vyhlazovacího parametru, $i$ je pořadí simulované hustoty a $k$ určuje řád jádra použitého při odhadu hustoty[1] (viz tabulka). Dále $h_{AMISE,i,k}$ označuje příslušnou $AMISE$-optimální hodnotu vyhlazovacího parametru, $\overline{h}_{XY,i,k}$ označuje aritmetický průměr odhadů vyhlazovacího parametru metodou $XY$ pro simulovaná data příslušející $i$-té hustotě a $\sigma(\overline{h}_{XY,i,k})$ značí směrodatnou odchylku těchto odhadů.

| $k$ | $\mathsf{K}_{opt}$ (grafy použitých jader viz níže) |
|---|---|
| 2 | $-\frac{3}{4}(x^2-1)$ |
| 4 | $\frac{15}{32}(x^2-1)(7x^2-3)$ |
| 6 | $-\frac{105}{256}(x^2-1)(33x^4-30x^2+5)$ |
| 8 | $\frac{315}{4096}(x^2-1)(715x^6-1001x^4+385x^2-35)$ |




## ZÁVĚR

První řádek níže uvedené tabulky shrnuje hodnoty hodnotící funkce pro jednotlivé metody odhadu vyhlazovacího parametru. Další významný faktor, který použijeme pro srovnání metod odhadu je časová výpočetní náročnost. Druhý řádek tabulky ukazuje relativní dobu výpočtu v porovnání s metodou referenční hustoty.

| $XY$ | $REF$ | $MV$ | $KMV$ | $KNM$ | $DJ$ | $IT$ |
|---|---|---|---|---|---|---|
| $F$ | 4,2663 | 5,5442 | 10,5582 | 4,6573 | 1,2063 | 6,0079 |
| $\frac{\bar{t}_{XY}}{\bar{t}_{REF}}$ | 1,0 | 0,9 | 1137,9 | 2131,0 | 1623,1 | 162,6 |

Dle hodnotící funkce dosáhla na simulovaných datech nejlepšího výsledku metoda dvou jader. V praxi nejpoužívanější metoda nejmenších čtverců dosáhla průměrného výsledku. Celkem dobře dopadla také iterační metoda, která dává poměrně uspokojivé výsledky a rychlost výpočtu je o řád rychlejší, než je tomu u předchozích metod. Metoda referenční hustoty a metoda maximálního vyhlazení najdou využití zejména v případě, kdy potřebujeme rychlý předběžný odhad.

Podrobnější výsledky simulační studie naleznete v [2].

**Poděkování.** Děkuji touto cestou prof. RNDr. Ivance Horová, CSc. za cenné rady a připomínky k této práci.

**Literatura.**

[1] Horová I. a Zelinka J. (2007). *Contributions to the bandwidth choice for kernel density estimates.* Computational Statistics, Springer-Verlag, 32–47.

[2] Orava J. (2008). *Volba vyhlazovacího parametru při jádrových odhadech hustoty.* Diplomová práce, PŘF MU, Brno

[3] Řezáč M. (2007). *Jádrové odhady hustoty.* Disertační práce, PŘF MU, Brno.