Metody odhadu šířky vyhlazovacích parametrů Priestley-Chaova odhadu podmíněné hustoty

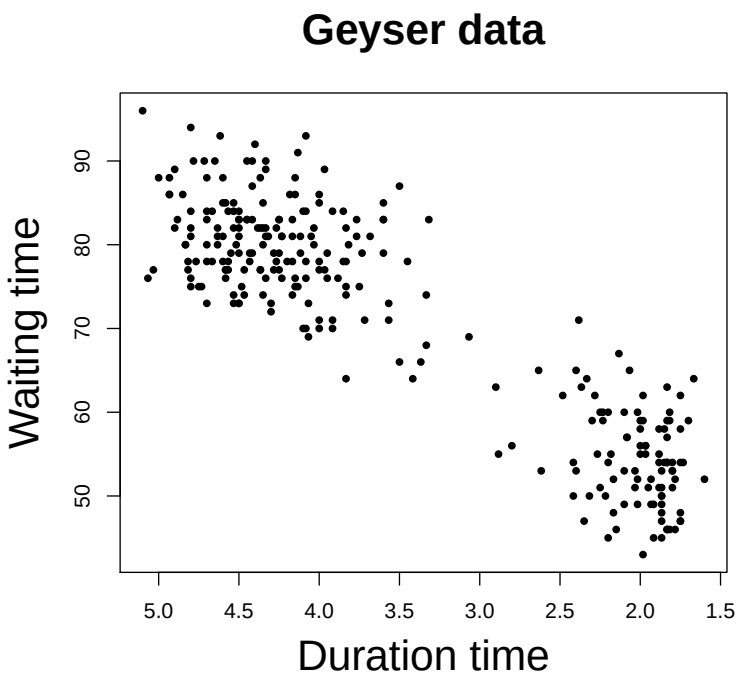
Kateřina Konečná

Ústav matematiky a statistiky, Masarykova univerzita, Brno
Ústav matematiky a deskriptivní geometrie, Vysoké učení technické v Brně, Brno

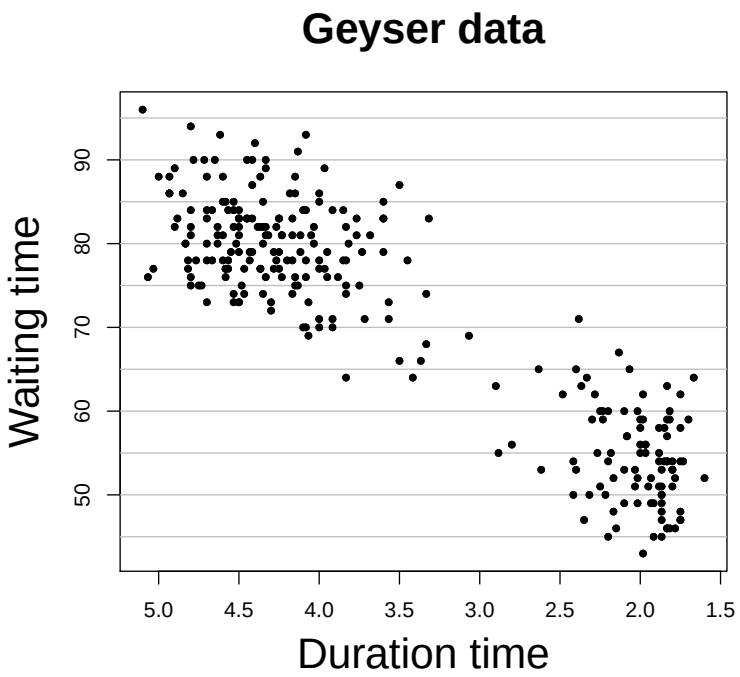
ROBUST 2018
rekreační zařízení Rybník
21. – 26. ledna 2018

Obsah

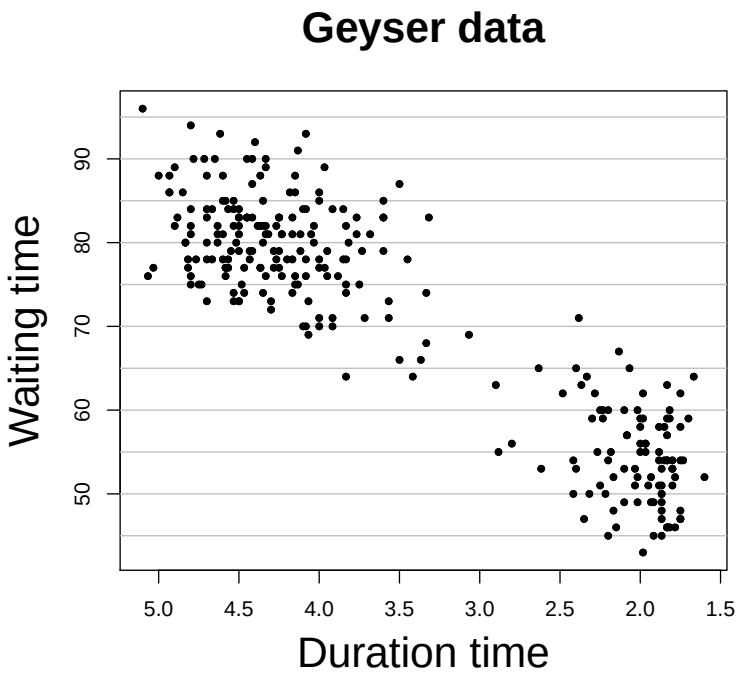
Motivace

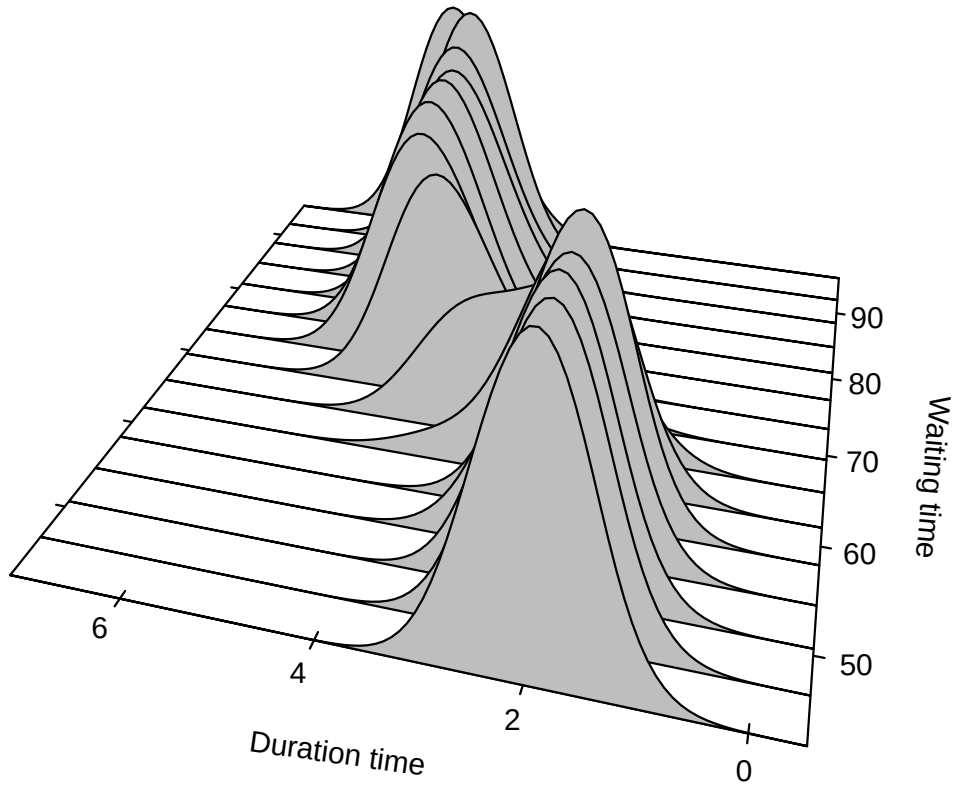

Geyser data
Waiting time
90
80
70
60
50
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
Duration time

Motivace


Geyser data
Waiting time
90
80
70
60
50
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
Duration time

Motivace


Geyser data
Waiting time
90
80
70
60
50
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
Duration time


Waiting time
90
80
70
60
50
6
4
2
0
Duration time

Cíle práce

- Priestley-Chaův odhad podmíněné hustoty
- statistické vlastnosti odhadu
- optimální šířky vyhlazovacích parametrů
- metody pro odhad šířky vyhlazovacích parametrů
- simulační studie

Jádrová funkce

Definice

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.

Typy odhadů

- jádrový odhad podmíněné hustoty

$$\hat{f}\left(y|x\right) = \sum_{i=1}^{n} w_i(x) K_{h_y}\left(y - Y_i\right),$$

Typy odhadů

- jádrový odhad podmíněné hustoty

$$\hat{f}\left(y|x\right) = \sum_{i=1}^{n} w_i(x) K_{h_y}\left(y - Y_i\right),$$

  - Nadaraya-Watsonovy váhy

$$w_i(x) = \frac{K_{h_x}(x - X_i)}{\sum_{j=1}^{n} K_{h_x}\left(x - X_j\right)}$$

  - lokálně lineární váhy

$$w_i(x) = \frac{(\hat{s}_2(x) - \hat{s}_1(x)(x - X_i))K_{h_x}(x - X_i)}{n\left(\hat{s}_0(x) \cdot \hat{s}_2(x) - \hat{s}_1^2(x)\right)}$$

  - Priestley-Chaovy váhy

$$w_i(x) = \frac{1}{n} K_{h_x}\left(x - X_i\right)$$

Statistické vlastnosti odhadu

Věta

Nechť $Y_1, Y_2, \ldots, Y_n$ jsou pozorované hodnoty v bodech plánu $x_i = \frac{i}{n}$, $i = 1, 2, \ldots, n$ a $\delta = x_{i+1} - x_i$, $i = 1, 2, \ldots, n-1$. Nechť je $f(y|x)$ nejméně 2krát spojitě diferencovatelná a $K(x)$ jádrová funkce splňující momentové podmínky. Pro $h_x \to 0$, $h_y \to 0$ a $nh_x h_y \to \infty$ pro $n \to \infty$ jsou asymptotické vychýlení (AB) a asymptotický rozptyl (AV) rovny

$$\mathrm{AB}\left\{\hat{f}_{PC}(y|x)\right\} = \frac{1}{2}h_x^2\beta_2(K)\frac{\partial^2 f(y|x)}{\partial x^2} + \frac{1}{2}h_y^2\beta_2(K)\frac{\partial^2 f(y|x)}{\partial y^2}, \qquad (1)$$

$$\mathrm{AV}\left\{\hat{f}_{PC}(y|x)\right\} = \frac{\delta}{h_x h_y}R^2(K)f(y|x), \qquad (2)$$

kde $R(K) = \int K^2(u)\,\mathrm{d}u$.

Míry kvality odhadu

- Asymptotická střední kvadratická chyba (lokální chyba odhadu):

$$\mathrm{AMSE}\left\{\widehat{f}_{PC}(y|x)\right\} = \mathrm{ASB}\left\{\widehat{f}_{PC}(y|x)\right\} + \mathrm{AV}\left\{\widehat{f}_{PC}(y|x)\right\}$$

- Asymptotická střední integrální kvadratická chyba (globální chyba odhadu):

$$\mathrm{AMISE}\left\{\hat{f}_{PC}(\cdot|\cdot)\right\} = \iint \mathrm{AMSE}\left\{\widehat{f}_{PC}(y|x)\right\}\,\mathrm{d}x\,\mathrm{d}y.$$

- Optimální šířky vyhlazovacích parametrů:

$$\begin{aligned} h_x^* &= \delta^{1/6} c_1^{1/6} \left(4\left(\frac{c_2^5}{c_3}\right)^{1/4} + 2c_4\left(\frac{c_2}{c_3}\right)^{3/4}\right)^{-1/6} \\ h_y^* &= \left(\frac{c_2}{c_3}\right)^{1/4} h_x^*. \end{aligned} \tag{3}$$

Metoda křížového ověřování

- Hlavní myšlenka: minimalizace integrální kvadratické chyby (ISE).
- Cross-validační funkce $CV(h_x, h_y)$ je vhodným odhadem ISE:

$$CV(h_x, h_y) = \delta^2 \sum_i \sum_{j \neq i} h_x h_y K_{h_x\sqrt{2}}(x_i - x_j) K_{h_y\sqrt{2}}(Y_i - Y_j) - 2\delta \sum_i \hat{f}_{-i,PC}(Y_i|x_i),$$

kde $\hat{f}_{-i,PC}(y|x)$ je odhad v bodě (x_i, Y_i) s využitím bodů $\{(x_j, Y_j), j \neq i\}$.

- Odhady vyhlazovacích parametrů:

$$(\hat{h}_x, \hat{h}_y)_{CV} = \arg\min_{(h_x,h_y)} CV(h_x, h_y).$$

Metoda referenční hustoty

- Motivace: předpoklad rozdělení pravděpodobnosti pro náhodné veličiny X a $Y|(X = x)$
- Hlavní myšlenka:

$$Y|(X = x) \sim N\left(c + dx, p^2\right),$$

pro $d \neq 0$, $p > 0$.

- Numerický výpočet (2D lichoběžníkové pravidlo) konstant c_1, c_2, c_3, $c_4 \rightarrow$ dosazením do (3) dostáváme $\left(\hat{h}_x, \hat{h}_y\right)_{REF}$

Simulační studie

zvolený model, generování simulovaných hodnot

$\downarrow$

odhad šířek vyhlazovacích parametrů $\hat{h}_x$, $\hat{h}_y$, metody: CV, REF

$\downarrow$

$$\widehat{\mathrm{ISE}}\left\{\hat{f}_{PC}(\cdot|\cdot)\right\} = \frac{\Delta}{n}\sum_{j=1}^{N}\sum_{i=1}^{n}\left(\hat{f}_{PC}(y_j|x_i) - f(y_j|x_i)\right)^2,$$

kde $\mathbf{y} = (y_1, \ldots, y_N)$ - vektor ekvidistantních hodnot na nosiči rozdělení pravděpodobnosti náhodné veličiny Y, $\Delta = |y_{j+1} - y_j|$, $j = 1, \ldots, N-1$

$\downarrow$

opakování 200krát $\rightarrow$ porovnání:

Simulační studie

zvolený model, generování simulovaných hodnot

↓

odhad šířek vyhlazovacích parametrů $\hat{h}_x$, $\hat{h}_y$, metody: CV, REF

↓

$$\widehat{\mathrm{ISE}}\left\{\hat{f}_{PC}(\cdot|\cdot)\right\} = \frac{\Delta}{n}\sum_{j=1}^{N}\sum_{i=1}^{n}\left(\hat{f}_{PC}(y_j|x_i) - f(y_j|x_i)\right)^2,$$

kde $\mathbf{y} = (y_1, \ldots, y_N)$ - vektor ekvidistantních hodnot na nosiči rozdělení pravděpodobnosti náhodné veličiny Y, $\Delta = |y_{j+1} - y_j|$, $j = 1, \ldots, N-1$

↓

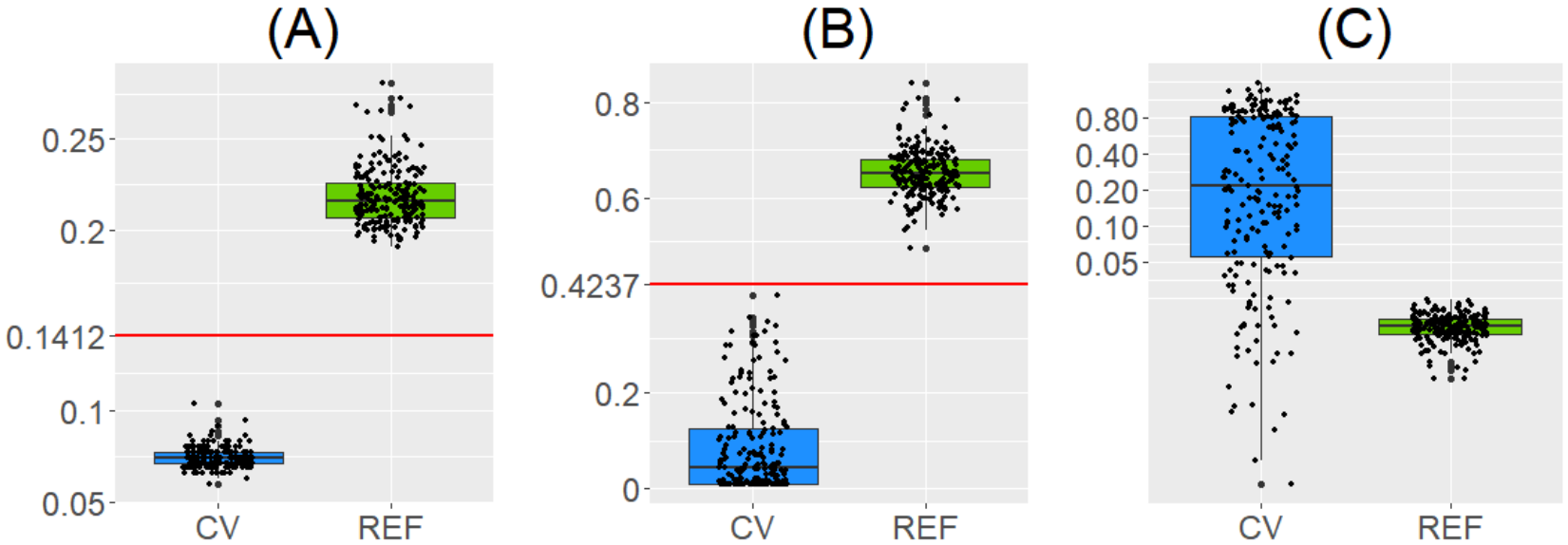
opakování 200krát → porovnání:
(A) $\hat{h}_x$, (B) $\hat{h}_y$, (C) $\widehat{\mathrm{ISE}}$, (D) výpočetní čas

Simulace 1

- model: $x_i = \frac{i}{n},\ \varepsilon_i \sim N(0,1),\ Y_i = 3x_i - 2 + \varepsilon_i,\ i = 1, \ldots, 100$

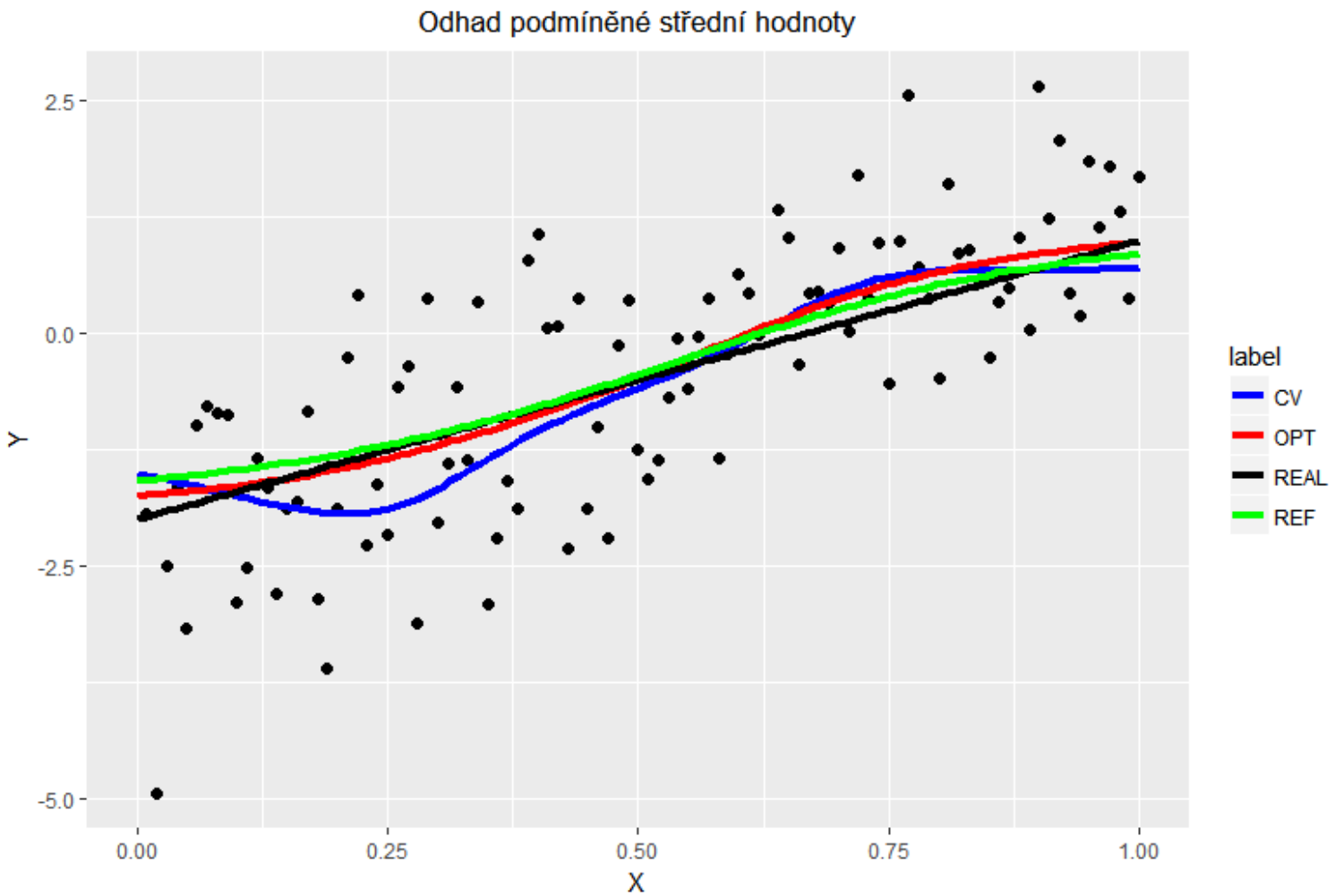
Simulace 1

- model: $x_i = \frac{i}{n},\ \varepsilon_i \sim N(0,1),\ Y_i = 3x_i - 2 + \varepsilon_i,\ i = 1, \ldots, 100$



(D) výpočetní čas [s]	CV	REF
průměr	118.102	0.005
medián	119.115	0
sm. odchylka	4.062	0.008
IQR	1.163	0.01

Simulace 1

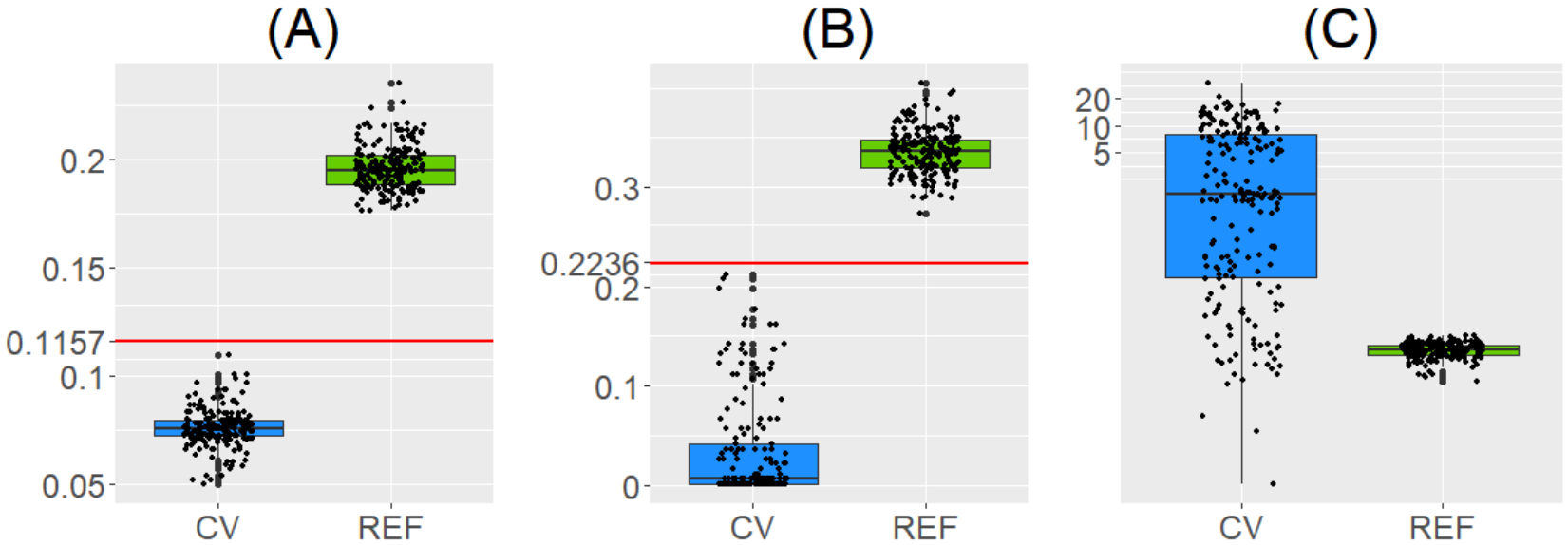

Odhad podmíněné střední hodnoty
2.5
0.0
-2.5
-5.0
Y
0.00
0.25
0.50
0.75
1.00
X
label
CV
OPT
REAL
REF

Simulace 2

- model: $x_i = \frac{i}{n},\ \varepsilon_i \sim N(0, 0.5^2),\ Y_i = \mathrm{e}^{x_i} + \varepsilon_i,\ i = 1, \ldots, 100$

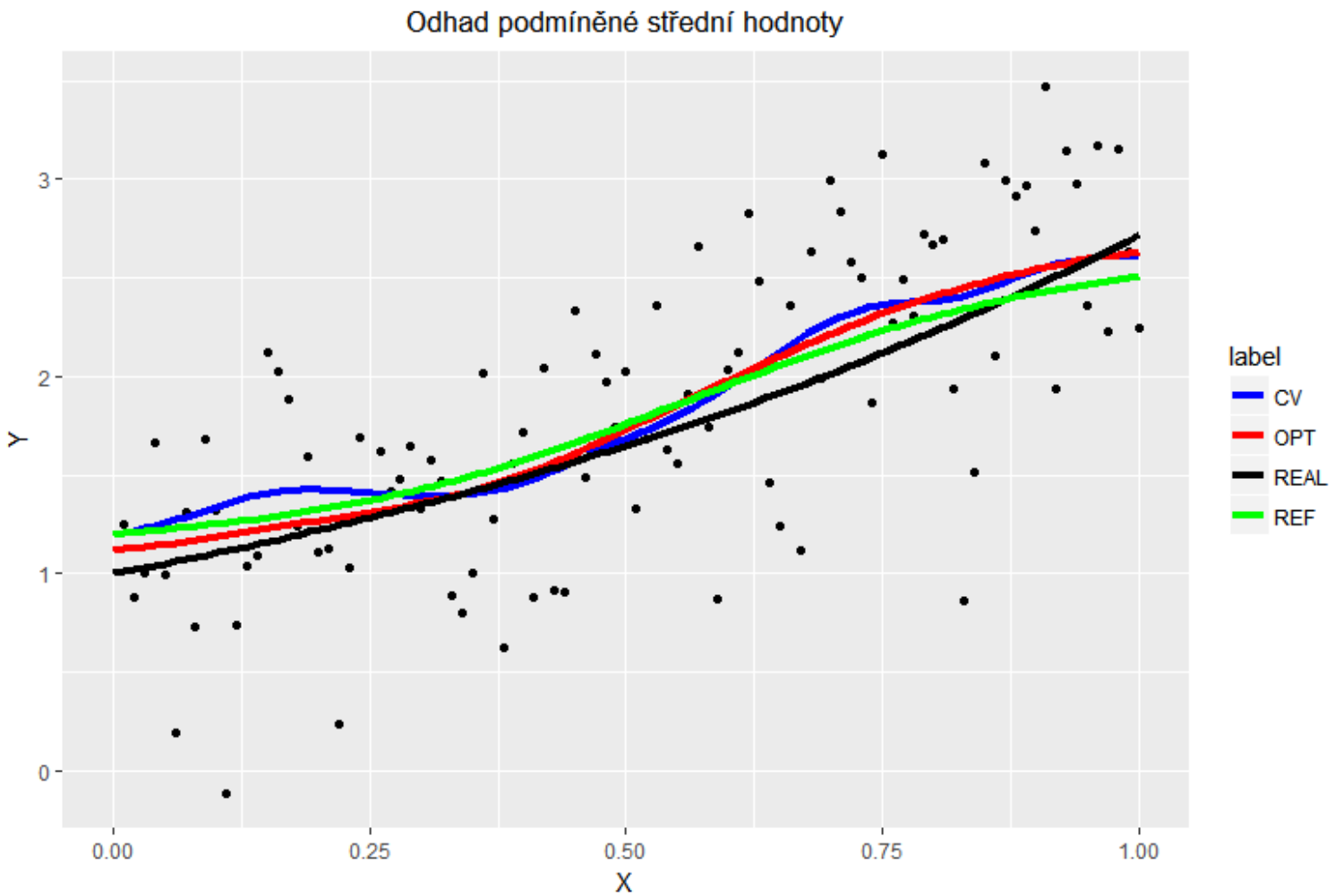
Simulace 2

- model: $x_i = \frac{i}{n},\ \varepsilon_i \sim N(0, 0.5^2),\ Y_i = e^{x_i} + \varepsilon_i,\ i = 1, \ldots, 100$



(D) výpočetní čas [s]	CV	REF
průměr	110.775	0.005
medián	110.79	0
sm. odchylka	0.980	0.008
IQR	0.685	0.01

Simulace 2


Odhad podmíněné střední hodnoty
Y
X
0
1
2
3
0.00
0.25
0.50
0.75
1.00
label
CV
OPT
REAL
REF

Dosažené výsledky práce

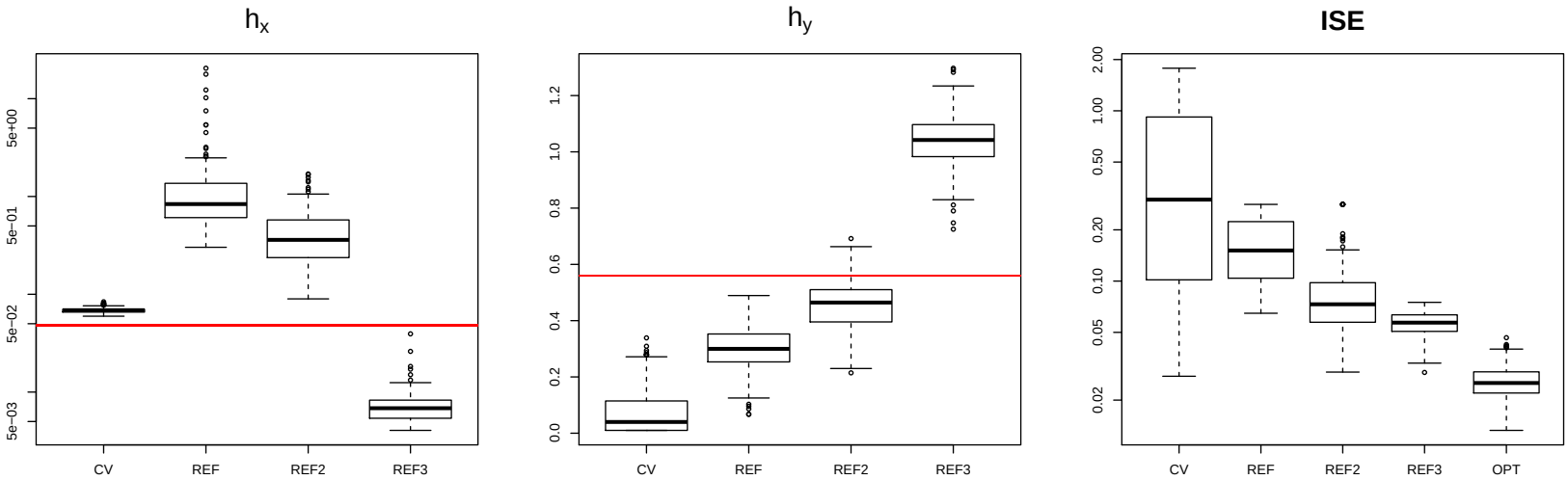
Priestley-Chaův odhad podmíněné hustoty:

- statistické vlastnosti odhadu (střední hodnota, rozptyl)
- míry kvality odhadu (lokální, globální)
- optimální šířky vyhlazovacích parametrů $\hat{h}_x^*$, $\hat{h}_y^*$
- metody pro odhad šířky vyhlazovacích parametrů:
  - metoda křížového ověřování
  - metoda referenční hustoty
  - leave-one-out metoda maximální věrohodnosti

Budoucí práce

- Co dělat v případě dat s nelineární podmíněnou střední hodnotou?

$$x_i = \frac{i}{n},\ \varepsilon_i \sim N(0,1),\ \ Y_i = \sin\left(3\pi x_i^2\right) + \varepsilon_i,\ \ i = 1, \ldots, 100$$



- Další metody pro odhad šířky vyhlazovacích parametrů:
  - odvození vztahů pro metodu referenční hustoty
  - iterační metoda

Reference

- BASHTANNYK, D. M., HYNDMAN, R. J., Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, vol. 36, no. 3, 2001: pp. 279–298, ISSN 0167-9473.
- KONEČNÁ, K., Priestley-chao estimator of conditional density, In *Mathematics, Information Technologies and Applied Sciences 2017, post-conference proceedings of extended versions of selected papers*, Brno: University of Defence, 2017, ISBN 978-80-7582-026-6, pp. 151–163.
- PRIESTLEY, M. B., CHAO, M. T., Non-parametric function fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 3, 1972: pp. 385–392, ISSN 00359246.
- ROSENBLATT, M., Conditional probability density and regression estimators. *Multivariate analysis II*, vol. 25, 1969: p. 31.
- WAND, M. P., JONES, M. C., *Kernel smoothing*. Crc Press, 1994, ISBN 9780412552700.

Optimální šířky vyhlazovacích parametrů

$$\mathrm{AMISE}\left\{\hat{f}_{PC}\left(\cdot|\cdot\right)\right\} = \frac{\delta}{h_x h_y} c_1 + c_2 h_x^4 + c_3 h_y^4 + c_4 h_x^2 h_y^2, \text{ kde}$$

$$c_1 = \int R^2(K)\,\mathrm{d}x,$$

$$c_2 = \frac{1}{4}\beta_2^2(K) \iint \left(\frac{\partial^2 f(y|x)}{\partial x^2}\right)^2 \mathrm{d}x\,\mathrm{d}y,$$

$$c_3 = \frac{1}{4}\beta_2^2(K) \iint \left(\frac{\partial^2 f(y|x)}{\partial y^2}\right)^2 \mathrm{d}x\,\mathrm{d}y,$$

$$c_4 = \frac{1}{2}\beta_2^2(K) \iint \frac{\partial^2 f(y|x)}{\partial x^2} \frac{\partial^2 f(y|x)}{\partial y^2}\,\mathrm{d}x\,\mathrm{d}y.$$