

Volba vyhlazovací matice pro jádrové odhady vícerozměrných hustot

Ivana Horová, Jan Kolářek, Kamila Vopatová



Oddělení aplikované matematiky
Přírodovědecká fakulta
Masarykova univerzita



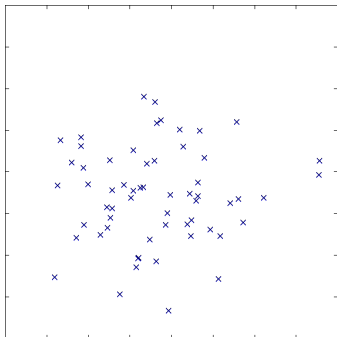
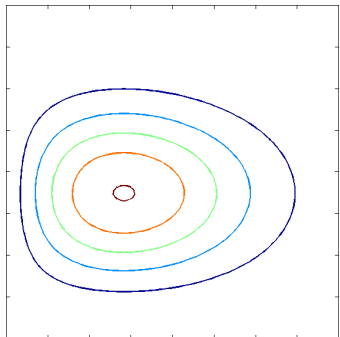
Obsah

- 1 Úvod
- 2 Metody pro odhad H – Metody založené na MISE
- 3 Metody pro odhad H – Metody založené na AMISE
- 4 Jiné metody pro odhad H
- 5 Simulovaná data
- 6 Reálná data
- 7 Reference

Obsah

- 1 Úvod
- 2 Metody pro odhad H – Metody založené na MISE
- 3 Metody pro odhad H – Metody založené na AMISE
- 4 Jiné metody pro odhad H
- 5 Simulovaná data
- 6 Reálná data
- 7 Reference

Motivace



Jádrový odhad vícerozměrné hustoty

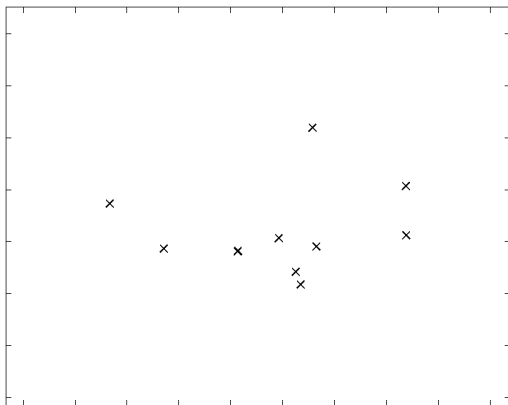
Pro d -rozměrný náhodný výběr $\mathbf{X}_1, \dots, \mathbf{X}_n$ z rozdělení s hustotou f definujeme jádrový odhad hustoty

$$\hat{f}(\mathbf{x}, H) = \frac{1}{n} \sum_{i=1}^n K_H(\mathbf{x} - \mathbf{X}_i) = \frac{1}{n} |H|^{-1/2} \sum_{i=1}^n K(H^{-1/2}(\mathbf{x} - \mathbf{X}_i)),$$

- K je d -rozměrná jádrová funkce, pro kterou platí $\int_{\mathbb{R}^d} K(\mathbf{x}) \, d\mathbf{x} = 1$
- H je matice vyhlazovacích parametrů z množiny \mathcal{F} symetrických pozitivně definitních matic typu $d \times d$
- $\mathbf{x} = (x_1, \dots, x_d)^T \in \mathbb{R}^d$

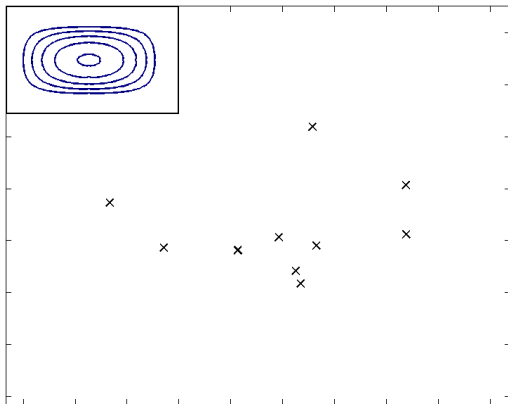
Data – Jádro – Hustota

Pro daná data



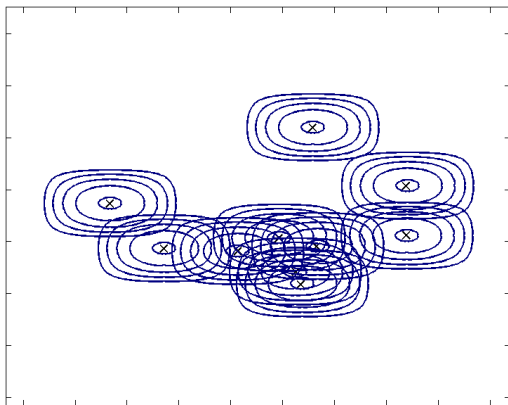
Data – Jádru – Hustota

vybereme jádro, např. Epanečnikovo



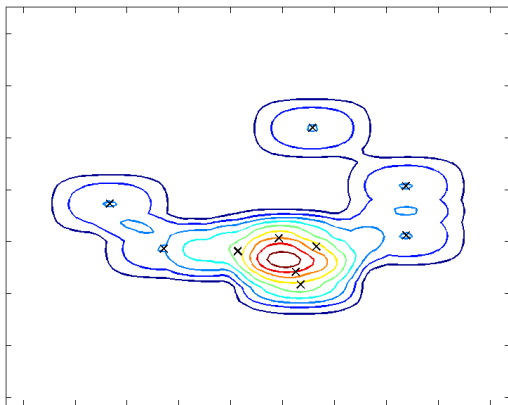
Data – Jádro – Hustota

vyčíslíme v každém bodě



Data – Jádro – Hustota

a získáme rekonstruovanou hustotu.

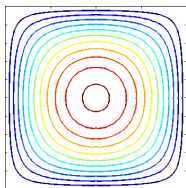


Jádrový odhad vícerozměrné hustoty

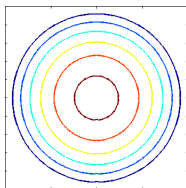
Typ jádra – z jednorozměrného jádra lze vytvořit dvě různá vícerozměrná jádra:

- součinnové jádro
- sféricky symetrické jádro

$$K^P(\mathbf{x}) = \prod_{i=1}^d K(x_i)$$



$$K^S(\mathbf{x}) = \frac{K((\mathbf{x}^T \mathbf{x})^{1/2})}{\int K((\mathbf{x}^T \mathbf{x})^{1/2}) d\mathbf{x}}$$



Výběr mnohorozměrného jádra

Vybrat si součinné jádro K^P nebo sféricky symetrické K^S ?
Měřítkem optimality je funkcionál C_d [Wand & Jones, 1995]

$$C_d(K) = (V(K)^4 \beta_2(K)^{2d})^{1/(d+4)} \rightarrow \min$$

Řešením této úlohy jsou *optimální jádra*.

Mezi součinnými jádry nejlépe vychází *Epanečnikovo* součinné jádro.

Matrice vyhlazovacích parametrů H

- nejdůležitější složka při jádrovém vyhlazování
- má vliv na orientaci jádra a jeho „šířku“
- 3 třídy vyhlazovacích matic
 - \mathcal{F} : obecná třída symetrických pozitivně definitních matic s $\frac{1}{2}d(d+1)$ nezávislými prvky,
 - \mathcal{D} : diagonální matice tvaru $H = \text{diag}(h_1^2, \dots, h_d^2)$,
 - \mathcal{S} : třída nejjednodušších matic typu $H = h^2 \cdot I_d$

Pro třídy \mathcal{D} a \mathcal{S} lze psát odhad hustoty v jednodušším tvaru

$$\mathcal{D}: \hat{f}(\mathbf{x}, H) = \frac{1}{nh_1 \cdots h_d} \sum_{i=1}^n K\left(\frac{x_1 - X_{1i}}{h_1}, \dots, \frac{x_d - X_{di}}{h_d}\right)$$

$$\mathcal{S}: \hat{f}(\mathbf{x}, H) = \frac{1}{nh^d} \sum_{i=1}^n K((\mathbf{x} - \mathbf{X}_i)/h)$$

Obsah

- 1 Úvod
- 2 Metody pro odhad H – Metody založené na MISE**
- 3 Metody pro odhad H – Metody založené na AMISE
- 4 Jiné metody pro odhad H
- 5 Simulovaná data
- 6 Reálná data
- 7 Reference

MISE

Kvalitu odhadu \hat{f} můžeme měřit např. pomocí střední integrální kvadratické chyby (MISE) [Wand & Jones, 1995]

$$\begin{aligned} MISE \hat{f}(\cdot, H) &= E \int (\hat{f}(\mathbf{x}, H) - f(\mathbf{x}))^2 d\mathbf{x} \\ &= \int \text{var } \hat{f}(\mathbf{x}, H) d\mathbf{x} + \int (\text{bias } \hat{f}(\mathbf{x}, H))^2 d\mathbf{x}, \end{aligned}$$

kde

$$\begin{aligned} \int \text{var } \hat{f}(\mathbf{x}, H) d\mathbf{x} &= \frac{1}{n} \left[\int K_H^2(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\mathbf{y} - \left(\int K_H(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\mathbf{y} \right)^2 \right] \\ \int (\text{bias } \hat{f}(\mathbf{x}, H))^2 d\mathbf{x} &= \int \left(\int K_H(\mathbf{x} - \mathbf{y}) f(\mathbf{y}, H) d\mathbf{y} - \hat{f}(\mathbf{x}, H) \right)^2 d\mathbf{x}. \end{aligned}$$

LSCV

Nevychýlená metoda křížového ověřování odhaduje integrální kvadratickou chybu (ISE) [Sain et al, 1994]

$$\begin{aligned} ISE\hat{f}(\cdot, H) &= \int \left(\hat{f}(\mathbf{x}, H) - f(\mathbf{x}) \right)^2 d\mathbf{x} \\ &= \int [\hat{f}(\mathbf{x}, H)]^2 d\mathbf{x} - 2 \int \hat{f}(\mathbf{x}, H) \cdot f(\mathbf{x}) d\mathbf{x} + \int f^2(\mathbf{x}) d\mathbf{x} \end{aligned}$$

$$LSCV(H) = \int [\hat{f}(\mathbf{x}, H)]^2 d\mathbf{x} - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}(\mathbf{X}_i, H),$$

kde $\hat{f}_{-i}(\mathbf{x}, H) = \frac{1}{(n-1)} \sum_{\substack{j=1 \\ j \neq i}}^n K_H(\mathbf{x} - \mathbf{X}_j)$

$$LSCV(H) = n^{-2} \sum_{i,j=1}^n (K_H * K_H - 2K_H)(\mathbf{X}_i - \mathbf{X}_j) + 2n^{-1} K_H(0)$$

Obsah

- 1 Úvod
- 2 Metody pro odhad H – Metody založené na MISE
- 3 Metody pro odhad H – Metody založené na AMISE**
- 4 Jiné metody pro odhad H
- 5 Simulovaná data
- 6 Reálná data
- 7 Reference

Od MISE k AMISE

$$\begin{aligned} MISE \hat{f}(\cdot, H) &= E \int (\hat{f}(\mathbf{x}, H) - f(\mathbf{x}))^2 d\mathbf{x} \\ &= \int \text{var} \hat{f}(\mathbf{x}, H) d\mathbf{x} + \int (\text{bias} \hat{f}(\mathbf{x}, H))^2 d\mathbf{x}, \end{aligned}$$

→

$$\begin{aligned} AMISE \hat{f}(\cdot, H) &= \frac{1}{n} \underbrace{\left[\int K_H^2(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\mathbf{y} \right]}_{AI\text{Var}} \\ &+ \underbrace{\int \left(\int K_H(\mathbf{x} - \mathbf{y}) f(\mathbf{y}, H) d\mathbf{y} - \hat{f}(\mathbf{x}, H) \right)^2 d\mathbf{x}}_{AI\text{Bias}^2} \end{aligned}$$

AMISE

Předpoklady pro přechod od MISE k asymptotickému tvaru střední integrální kvadratické chyby (AMISE):

- K je ohraničená jádrová funkce s kompaktním nosičem, pro niž platí

$$\int K(\mathbf{x}) \, d\mathbf{x} = 1, \quad \int \mathbf{x}K(\mathbf{x}) \, d\mathbf{x} = 0, \quad \int \mathbf{x}\mathbf{x}^T K(\mathbf{x}) \, d\mathbf{x} = \beta_2(K)I_d,$$

kde $\beta_2(K) = \int x_i^2 K(x) \, dx$ nezávisí na i a I_d je jednotková matice řádu d .

- $H = H_n$ je posloupnost vyhlazovacích matic takových, že $n^{-1}|H|^{-1/2}$ a všechny prvky H se blíží k nule pro $n \rightarrow \infty$.
- Všechny prvky matice \mathcal{D}_f^2 druhých derivací hustoty f jsou po částech spojitě a integrovatelné se čtvercem.

AMISE

Asymptotický tvar střední integrální kvadratické chyby [Wand & Jones, 1995]

$$AMISE(H) = AIVar + AIBias^2$$

Vychýlení (*Bias*) lze s užitím vícerozměrné Taylorovy věty rozepsat

$$\begin{aligned} (K_H * f)(\mathbf{x}) - f(\mathbf{x}) &= \int K_H(\mathbf{x} - \mathbf{y})f(\mathbf{y}) d\mathbf{y} - f(\mathbf{x}) \\ &= \int K(\mathbf{z})f(\mathbf{x} - H^{1/2}\mathbf{z}) d\mathbf{z} - f(\mathbf{x}) \\ &= f(\mathbf{x}) + \frac{1}{2}\beta_2(K) \operatorname{tr}[HD_f^2(\mathbf{x})] + o(\operatorname{tr} H) - f(\mathbf{x}) \end{aligned}$$

$$Bias \sim \frac{1}{2}\beta_2(K) \operatorname{tr}[HD_f^2(\mathbf{x})]$$

AMISE

Podobně můžeme rozepsat i rozptyl (Var)

$$\begin{aligned} \text{var } \hat{f}(\mathbf{x}, H) &= \frac{1}{n} \left[\int K_H^2(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\mathbf{y} - \left(\int K_H(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) \right)^2 \right] \\ &= \frac{1}{n} |H|^{-1/2} \int K^2(\mathbf{z}) f(\mathbf{x} - H^{1/2}\mathbf{z}) d\mathbf{z} \\ &= \frac{1}{n} |H|^{-1/2} V(K) f(\mathbf{x}) + o(n^{-1} |H|^{-1/2}) \end{aligned}$$

$$\text{Var} \sim \frac{1}{n} |H|^{-1/2} V(K) f(\mathbf{x})$$

Tedy

$$AMISE(H) = \frac{1}{n} |H|^{-1/2} V(K) + \frac{1}{2} \beta_2^2(K) \int \text{tr}^2 [H D_f^2(\mathbf{x})] d\mathbf{x}$$

AMISE

Dále ještě upravíme tvar vychýlení

$$\int \text{tr}^2 [HD_f^2(\mathbf{x})] d\mathbf{x} = (\text{vech } H)^T \Psi_{\mathcal{F}} \text{vech } H.$$

Operace vech (z angl. vector half):

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \Rightarrow \text{vech } A = \begin{pmatrix} a \\ c \\ d \end{pmatrix}$$

Pro $d = 2$ je matice $\Psi_{\mathcal{F}}$ tvaru

$$\begin{pmatrix} \psi_{4,0} & 2\psi_{3,1} & \psi_{2,2} \\ 2\psi_{3,1} & 4\psi_{2,2} & 2\psi_{1,3} \\ \psi_{2,2} & 2\psi_{1,3} & \psi_{0,4} \end{pmatrix}$$

kde

$$\psi_{k,l} = \int \frac{\partial^{k+l} f(\mathbf{x})}{\partial x_i^k \partial x_j^l} f(\mathbf{x}) d\mathbf{x},$$

AMISE

$$AMISE(H) = \frac{1}{n} |H|^{-1/2} V(K) + \frac{1}{2} \beta_2^2(K) (\text{vech } H)^T \Psi_{\mathcal{F}} \text{vech } H$$

I přes tato zjednodušení nelze vyjádřit optimální H vzhledem k AMISE a musí se počítat numericky. [Wand & Jones, 1995]

\mathcal{D} : Je-li matice H diagonální, pak lze psát

$H = \text{diag}(h_1^2, \dots, h_d^2) = \text{diag}(\mathbf{h}^2)$ a AMISE můžeme psát ve tvaru

$$AMISE(H) = \frac{V(K)}{nh_1 \dots h_d} + \frac{1}{4} \beta_2^2(K) (\mathbf{h}^2)^T \Psi_{\mathcal{D}} (\mathbf{h}^2)$$

kde $\Psi_{\mathcal{D}}$ obsahuje prvky $\psi_{2e_i+2e_j}$, e_i je jednotkový vektor s 1 na i -tém místě.

AMISE

\mathcal{S} : V případě $H = h^2 \cdot I_d$ dostaneme

$$AMISE(H) = \frac{V(K)}{nh^d} + \underbrace{\frac{1}{4}\beta_2^2(K)h^4 \int \left[\sum_{i=1}^d \frac{\partial^2 f(\mathbf{x})}{\partial x_i^2} \right]^2 d\mathbf{x}}_{I(D^2)}$$

Jen v tomto případě lze vyjádřit optimální hodnotu vyhlazovacího parametru h

$$h_{AMISE} = \left[d \frac{V(K)}{n\beta_2^2(K)I(D^2)} \right]^{1/(d+4)}$$

AMISE – 2D

Zaměříme se na odhady dvourozměrné hustoty s diagonální vyhlazovací maticí H . Pak asymptotický tvar střední integrální kvadratické chyby (AMISE) lze jednoduše vyjádřit takto:

$$AMISE(H) = \frac{V(K)}{nh_1h_2} + \frac{1}{4}\beta_2(K)^2(h_1^4\psi_{4,0} + 2h_1^2h_2^2\psi_{2,2} + h_2^4\psi_{0,4}),$$

kde

- $V(K) = \iint K^2(x_1, x_2) dx_1 dx_2 < \infty$
- $\beta_2(K) = \iint x_i^2 K(x_1, x_2) dx_1 dx_2, i = 1, 2$
- $\psi_{k,l} = \iint \left(\frac{\partial^4 f(x_1, x_2)}{\partial x_1^k \partial x_2^l} \right) f(x_1, x_2) dx_1 dx_2, \quad k, l = 0, 2, 4, \quad k + l = 4$

Lemma

Označme $H_{AMISE} = \arg \min_{H \in \mathcal{D}} AMISE(H)$. Pak pro vyhlazovací matici H_{AMISE} s prvky

$$h_{1,AMISE} = \left[\frac{\psi_{0,4}^{3/4} V(K)}{\beta_2(K)^2 \psi_{4,0}^{3/4} (\psi_{2,2} + \psi_{0,4}^{1/2} \psi_{4,0}^{1/2}) n} \right]^{1/6},$$

$$h_{2,AMISE} = \left[\frac{\psi_{4,0}^{3/4} V(K)}{\beta_2(K)^2 \psi_{0,4}^{3/4} (\psi_{2,2} + \psi_{0,4}^{1/2} \psi_{4,0}^{1/2}) n} \right]^{1/6},$$

platí

$$\iint \text{var } \hat{f}(\mathbf{x}, H_{AMISE}) dx_1 dx_2 = 2 \iint \left(\text{bias } \hat{f}(\mathbf{x}, H_{AMISE}) \right)^2 dx_1 dx_2.$$

Odhad AMISE

$$\widehat{AMISE}(H) = \iint \widehat{\text{var}} \hat{f}(\mathbf{x}, H) dx_1 dx_2 + \iint \left(\widehat{\text{bias}} \hat{f}(\mathbf{x}, H) \right)^2 dx_1 dx_2,$$

kde

$$\begin{aligned} \iint \widehat{\text{var}} \hat{f}(\mathbf{x}, H) dx_1 dx_2 &= \frac{1}{n} \iint |H|^{-1/2} V(K) \hat{f}(\mathbf{x}, H) dx_1 dx_2, \\ \iint \left(\widehat{\text{bias}} \hat{f}(\mathbf{x}, H) \right)^2 dx_1 dx_2 &= \iint \left(\iint K_H(\mathbf{x} - \mathbf{y}) \hat{f}(\mathbf{y}, H) dy_1 dy_2 \right. \\ &\quad \left. - \hat{f}(\mathbf{x}, H) \right)^2 dx_1 dx_2. \end{aligned}$$

\widehat{AMISE} – optimální H

Úpravou předešlých rovnic získáme

$$\iint \widehat{\text{var}} \hat{f}(\mathbf{x}; H) dx_1 dx_2 = \frac{V(K)}{nh_1 h_2},$$

$$\begin{aligned} & \iint (\widehat{\text{bias}} \hat{f}(\mathbf{x}; H))^2 dx_1 dx_2 \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (K_H * K_H * K_H * K_H \\ & \quad - 2K_H * K_H * K_H + K_H * K_H)(\mathbf{X}_i - \mathbf{X}_j), \end{aligned}$$

kde * značí operaci konvoluce.

AMISE – optimální H

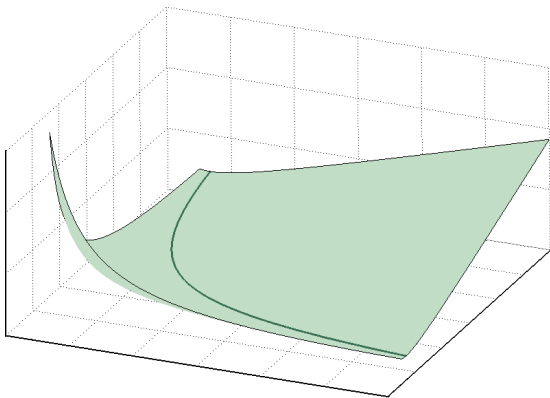
Označme

$$g(h_1, h_2) = \sum_{i,j=1}^n K_H * K_H * K_H * K_H - 2K_H * K_H * K_H + K_H * K_H)(\mathbf{X}_i - \mathbf{X}_j)$$

Myšlenka této metody je založena na Lemmatu, tedy hledáme taková \hat{h}_1 a \hat{h}_2 , aby platilo

$$\begin{aligned} IVar &= 2 \cdot IBias^2 \\ \frac{V(K)}{n\hat{h}_1\hat{h}_2} &= 2 \cdot \frac{1}{n^2}g(\hat{h}_1, \hat{h}_2) \\ nV(K) &= 2\hat{h}_1\hat{h}_2g(\hat{h}_1, \hat{h}_2) \end{aligned}$$

\widehat{AMISE} – optimální H



Optimální $H \rightarrow M1$

Jak najít další vztah mezi \hat{h}_1 a \hat{h}_2 ?

- ① Scottovo pravidlo: $\hat{h}_i = \hat{\sigma}_i n^{-1/6}$, ($i = 1, 2$) [Scott, 1992]

$$\hat{h}_2 = \hat{c} h_1, \quad \hat{c} = \frac{\hat{\sigma}_2}{\hat{\sigma}_1}$$

$$M1 \begin{cases} 2\hat{h}_1 \hat{h}_2 g(\hat{h}_1, \hat{h}_2) = nV(K) \\ \hat{h}_2 = \hat{c} \hat{h}_1 \end{cases}$$

K odhadu $\hat{\sigma}_i$ můžeme použít výběrovou směrodatnou odchylku

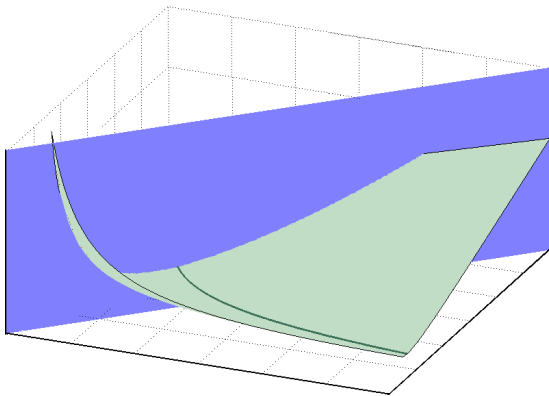
$$\hat{\sigma}_i^2 = \frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \bar{X}_i)^2, \quad i = 1, 2,$$

nebo robustnější odhad

$$\hat{\sigma}_{i,IQR} \approx \frac{X_{i[3n/4]} - X_{i[n/4]}}{1.349}, \quad i = 1, 2.$$

Doporučuje se použít menší z odhadů: $\min(\hat{\sigma}_i, \hat{\sigma}_{i,IQR})$, $i = 1, 2$.

Řešení metody M1



Optimální $H \rightarrow M2$

- 2 Podle vztahu pro optimálními hodnoty h_1 a h_2 (vzhledem k AMISE) platí

$$\frac{h_2}{h_1} = \left(\frac{\psi_{40}}{\psi_{04}} \right)^{1/4} \quad \Leftrightarrow \quad h_2^4 \psi_{04} = h_1^4 \psi_{40},$$

$$\psi_{04} = \int \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 \mathbf{d}\mathbf{x} \approx n^{-2} \sum_{i,j=1}^n \left(\frac{\partial^2 K_H}{\partial x_2^2} * \frac{\partial^2 K_H}{\partial x_2^2} \right) (\mathbf{X}_i - \mathbf{X}_j)$$

$$\psi_{40} = \int \left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 \mathbf{d}\mathbf{x} \approx n^{-2} \sum_{i,j=1}^n \left(\frac{\partial^2 K_H}{\partial x_1^2} * \frac{\partial^2 K_H}{\partial x_1^2} \right) (\mathbf{X}_i - \mathbf{X}_j)$$

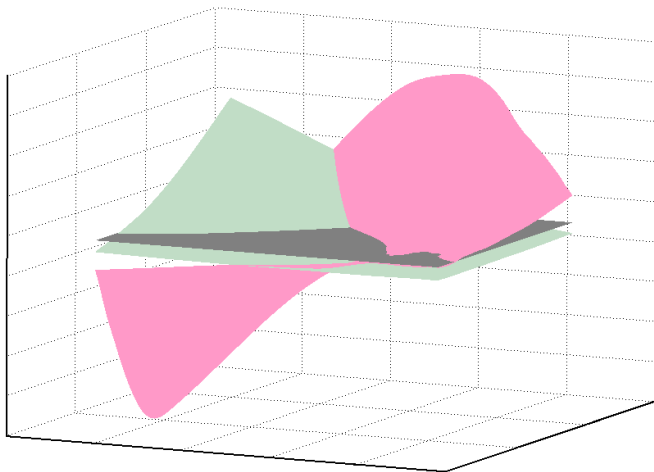
Optimální $H \rightarrow M2$

Pak ze vztahu $h_2^4 \psi_{04} = h_1^4 \psi_{40}$ plyne

$$h_2^4 n^{-2} \sum_{i,j=1}^n \left(\frac{\partial^2 K_H}{\partial x_2^2} * \frac{\partial^2 K_H}{\partial x_2^2} \right) (\mathbf{X}_i - \mathbf{X}_j) = h_1^4 n^{-2} \sum_{i,j=1}^n \left(\frac{\partial^2 K_H}{\partial x_1^2} * \frac{\partial^2 K_H}{\partial x_1^2} \right) (\mathbf{X}_i - \mathbf{X}_j)$$

$$M2 \left\{ \begin{array}{l} 2\hat{h}_1 \hat{h}_2 g(\hat{h}_1, \hat{h}_2) = nV(K) \\ \hat{h}_2^4 n^{-2} \sum_{i,j=1}^n \left(\frac{\partial^2 K_{\hat{H}}}{\partial x_2^2} * \frac{\partial^2 K_{\hat{H}}}{\partial x_2^2} \right) (\mathbf{X}_i - \mathbf{X}_j) = \\ = \hat{h}_1^4 n^{-2} \sum_{i,j=1}^n \left(\frac{\partial^2 K_{\hat{H}}}{\partial x_1^2} * \frac{\partial^2 K_{\hat{H}}}{\partial x_1^2} \right) (\mathbf{X}_i - \mathbf{X}_j) \end{array} \right.$$

Řešení metody M2



BCV a Plug-in metoda

Cílem těchto metod je odhadnout funkci

$$\psi_{k,l} = \int \frac{\partial^{k+l} f(\mathbf{x})}{\partial x_1^k \partial x_2^l} \cdot f(\mathbf{x}) \, d\mathbf{x}, \quad k, l = 0, 2, 4, \quad k + l = 4.$$

- Vychýlená metoda křížového ověřování (BCV) – 2 typy,
- plug-in metoda.

BCV

1

$$\psi_{k,\ell} \approx \int \frac{\partial^{k+\ell} \hat{f}(\mathbf{x}, H)}{\partial x_1^k \partial x_2^\ell} \cdot \hat{f}(\mathbf{x}, H) \, d\mathbf{x} - \frac{1}{n} \int \frac{\partial^{k+\ell} K_H(\mathbf{x})}{\partial x_1^k \partial x_2^\ell} \cdot K_H(\mathbf{x}) \, d\mathbf{x}$$

$$\hat{\psi}_{k,\ell} = n^{-2} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \left(\frac{\partial^{k+\ell} K_H}{\partial x_1^k \partial x_2^\ell} * K_H \right) (\mathbf{X}_i - \mathbf{X}_j)$$

2

$$\psi_{k,\ell} = E \left(\frac{\partial^{k+\ell} f(\mathbf{x})}{\partial x_1^k \partial x_2^\ell} \right) \approx \frac{1}{n} \sum_{i=1}^n \frac{\partial^{k+\ell} \hat{f}_{-i}(\mathbf{X}_i, H)}{\partial x_1^k \partial x_2^\ell}$$

$$\tilde{\psi}_{k,\ell} = n^{-1} (n-1)^{-1} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \frac{\partial^{k+\ell} K_H(\mathbf{X}_i - \mathbf{X}_j)}{\partial x_1^k \partial x_2^\ell}$$

kde $\hat{f}_{-i}(\mathbf{x}, H) = (n-1)^{-1} \sum_{j=1, j \neq i}^n K_H(\mathbf{X}_i - \mathbf{X}_j)$. [Duong & Hazelton, 2004]

Plug-in metoda

$$\psi_{k,\ell} = E\left(\frac{\partial^{k+\ell} f(\mathbf{x})}{\partial x_1^k \partial x_2^\ell}\right) \approx \frac{1}{n} \sum_{i=1}^n \frac{\partial^{k+\ell} \hat{f}(\mathbf{X}_i, \mathbf{G})}{\partial x_1^k \partial x_2^\ell}$$

$$\hat{\psi}_{k,\ell}(\mathbf{G}) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \frac{\partial^{k+\ell} K_{\mathbf{G}}(\mathbf{X}_i - \mathbf{X}_j)}{\partial x_1^k \partial x_2^\ell},$$

kde předpokládáme $\mathbf{G} = g^2 \mathbf{I}$

$$g_{AMSE} = \left[\frac{-2 \frac{\partial^{k+\ell} K(0)}{\partial x_1^k \partial x_2^\ell}}{\beta_2(K) \{ \psi_{k+2,\ell} + \psi_{k,\ell+2} \} n} \right]^{\frac{1}{4+k+\ell}}$$

[Wand & Jones, 1994]

Obsah

- 1 Úvod
- 2 Metody pro odhad H – Metody založené na MISE
- 3 Metody pro odhad H – Metody založené na AMISE
- 4 Jiné metody pro odhad H**
- 5 Simulovaná data
- 6 Reálná data
- 7 Reference

PLCV

Metoda křížového ověřování pomocí pseudověrohodnostní funkce
[Cao, 1993]

$$L(H) = \prod_{i=1}^n \hat{f}_{-i}(\mathbf{x}_i, H) \longrightarrow \max$$

$$\ell(H) = \ln L(H)$$

$$= -n \ln(n-1) - \frac{n}{2} \ln \det(H) + \sum_{i=1}^n \ln \left[\sum_{\substack{j=1 \\ j \neq i}}^n K(H^{-1/2}(\mathbf{x}_i - \mathbf{x}_j)) \right]$$

Metoda referenční hustoty

Za předpokladu, že f je d -rozměrná normální hustota, pak optimální hodnota matice vyhlazovacích parametrů je

$$H_{ref} = \left(\frac{4}{d+2} \right)^{2/(d+4)} \widehat{\Sigma} n^{-2/(d+4)},$$

$\widehat{\Sigma}$ je odhad kovarianční matice. [Wand & Jones, 1995]

Metoda maximálního vyhlazení

H_{MS} je řešením rovnice

$$H \cdot H^T = \left(\frac{(d+8)^{\frac{d+6}{2}} \pi^{d/2} V(K)}{16n \Gamma(\frac{d+8}{2})(d+2)} \right)^{\frac{2}{d+4}} \cdot \widehat{\Sigma},$$

pro $d = 2$

$$H \cdot H^T = \left(\frac{625\pi V(K)}{96n} \right)^{\frac{1}{3}} \cdot \widehat{\Sigma},$$

přičemž $\widehat{\Sigma}$ značí odhad kovarianční matice.

Pro $H = \text{diag}\{h_1, h_2\}$

$$h_i = \left(\frac{625\pi V(K)}{96n} \right)^{\frac{1}{6}} \cdot \hat{\sigma}_i, \quad i = 1, 2$$

[Sain et al, 1994]

Obsah

- 1 Úvod
- 2 Metody pro odhad H – Metody založené na MISE
- 3 Metody pro odhad H – Metody založené na AMISE
- 4 Jiné metody pro odhad H
- 5 Simulovaná data**
- 6 Reálná data
- 7 Reference

Simulace

Aplikace na simulovaných datech – porovnání metod M1 a M2 s metodou LSCV.

- Epanečnikovo součinné jádro

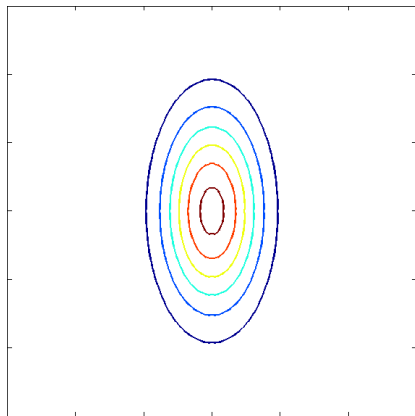
$$K(x_1, x_2) = \frac{9}{16}(1 - x_1^2)(1 - x_2^2), \quad x_1, x_2 \in [-1, 1],$$

- n – počet pozorování v náhodném výběru,
- R – počet opakování.

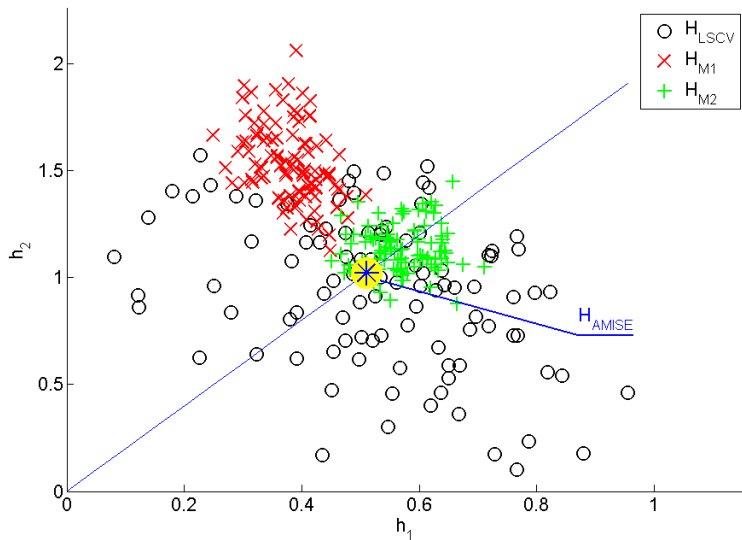
Normální hustota I

$$n = 100, R = 100$$

$$\mathbf{X} \sim N_2(0, 0; \frac{1}{4}, 1, 0)$$

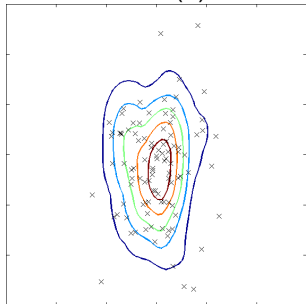


Normální hustota I – H

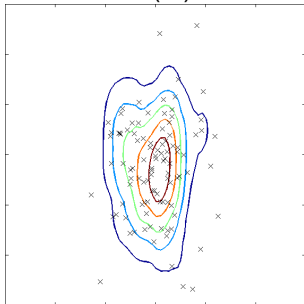


Normální hustota I – rekonstrukce

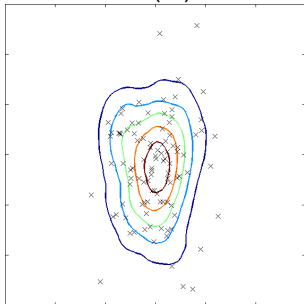
LSCV (○)



M1 (×)



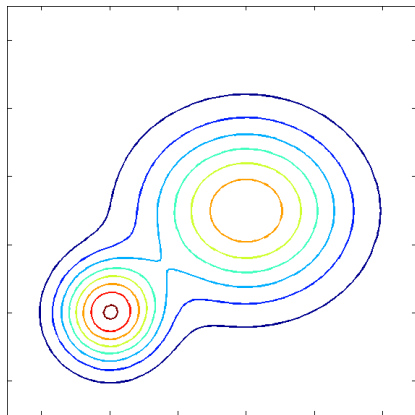
M2 (+)

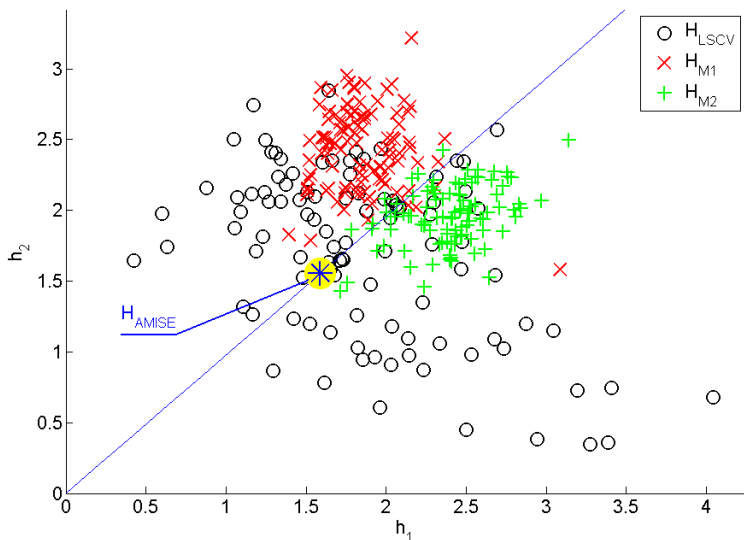


Normální hustota II

$n = 100, R = 100$

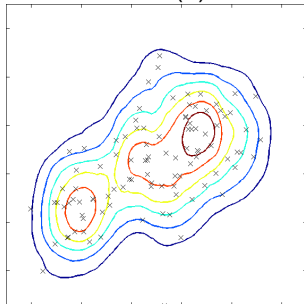
$$\mathbf{X} \sim 0.25N_2(0, 0; 1, 1, 0) \\ + 0.75N_2(4, 3; 4, 3, 0)$$



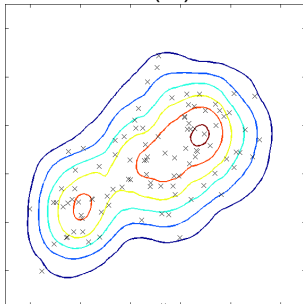
Normální hustota II – H 

Normální hustota II – rekonstrukce

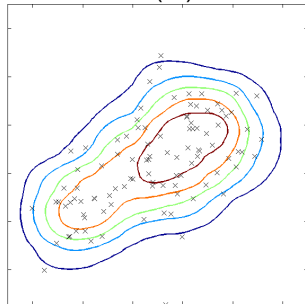
LSCV (○)



M1 (×)



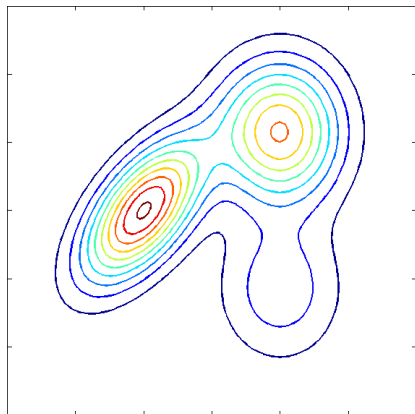
M2 (+)

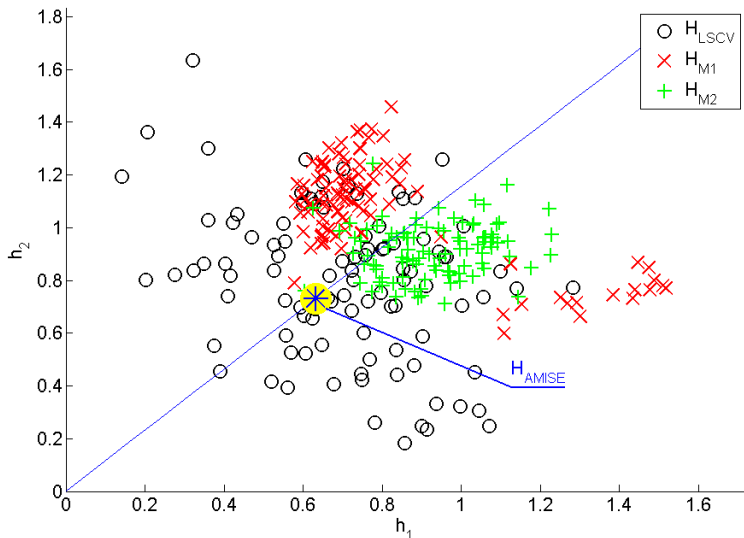


Normální hustota III

$$n = 100, R = 100$$

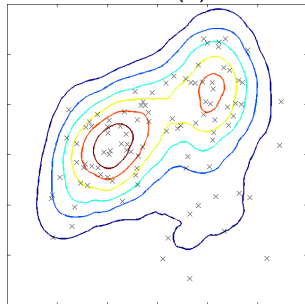
$$\begin{aligned} \mathbf{X} \sim & \frac{3}{7} N_2\left(-1, 0; \frac{9}{25}, \frac{49}{100}, \frac{63}{250}\right) \\ & + \frac{3}{7} N_2\left(1, \frac{2}{\sqrt{3}}; \frac{9}{25}, \frac{49}{100}, 0\right) \\ & + \frac{1}{7} N_2\left(1, -\frac{2}{\sqrt{3}}; \frac{9}{25}, \frac{49}{100}, 0\right) \end{aligned}$$



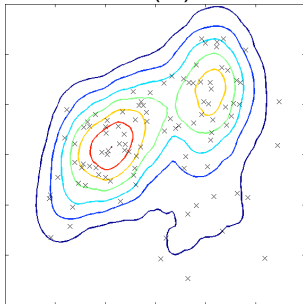
Normální hustota III – H 

Normální hustota III – rekonstrukce

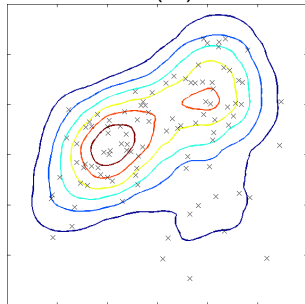
LSCV (o)



M1 (x)



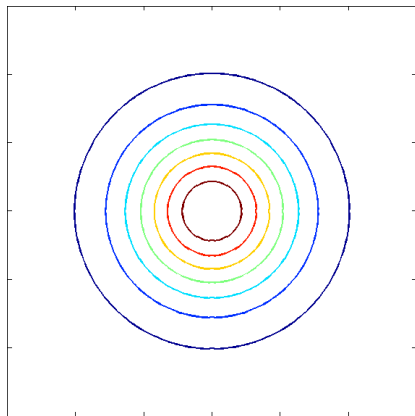
M2 (+)

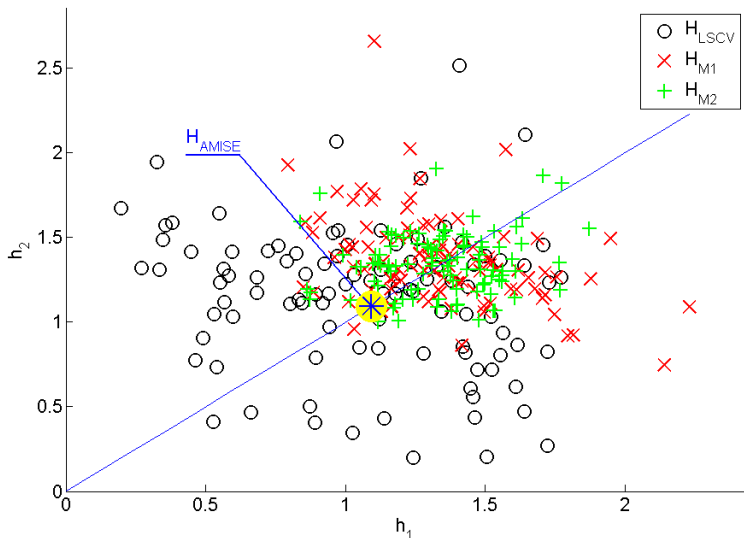


Studentovo rozdělení

$$n = 60, R = 100$$

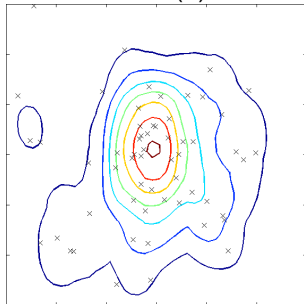
$$\mathbf{X} \sim t(5) \cdot t(5)$$



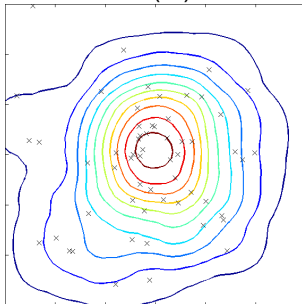
Studentovo rozdělení – H 

Studentovo rozdělení – rekonstrukce

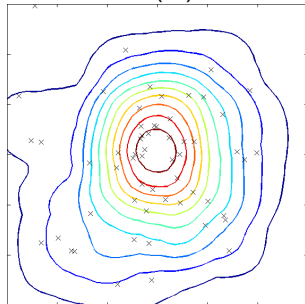
LSCV (o)



M1 (x)



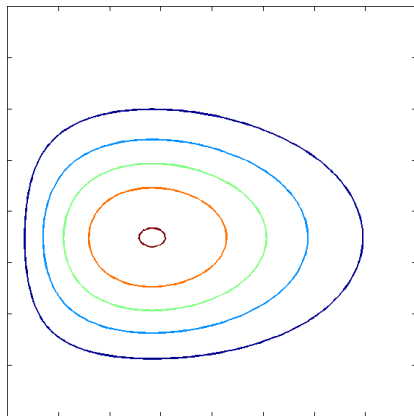
M2 (+)

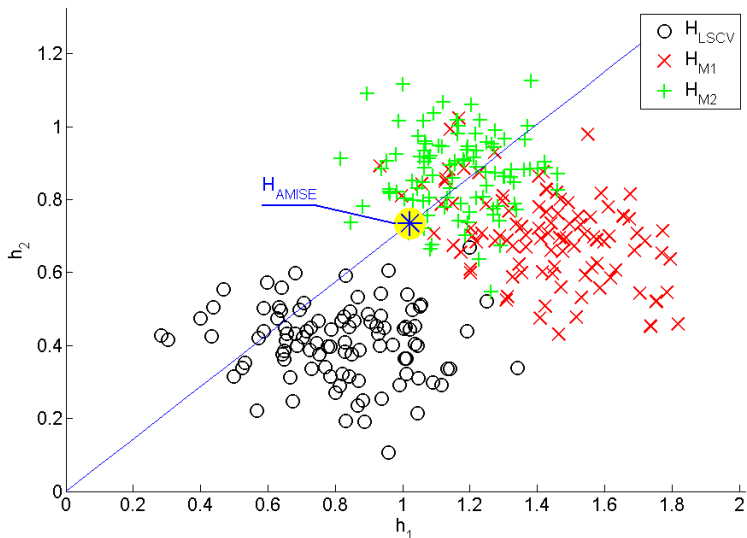


Weibullovo rozdělení

$n = 60, R = 100$

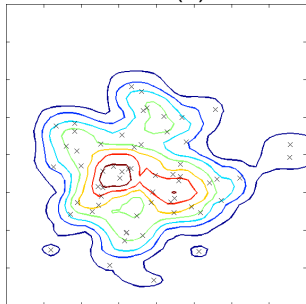
$$\mathbf{X} \sim Wb(2, 2) \cdot Wb(2, 3)$$



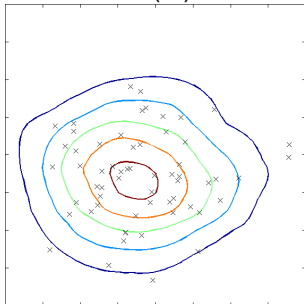
Weibullovo rozdělení – H 

Weibullovo rozdělení – rekonstrukce

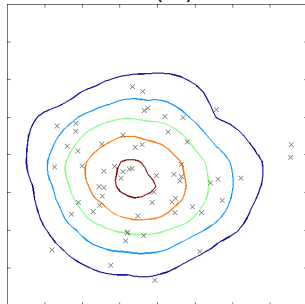
LSCV (o)



M1 (x)



M2 (+)



Průměr relativních chyb AMISE

Relativní chyba AMISE

$$\frac{|AMISE(H_{metoda}) - AMISE(H_{AMISE})|}{AMISE(H_{AMISE})}$$

<i>data</i>	LSCV	M1	M2
Norm I	0.5046	0.4486	0.0841
Norm II	0.5392	0.7227	0.7200
Norm III	0.4548	0.9232	0.6392
Student	0.5407	0.4052	0.3063
Weibull	0.8424	0.2987	0.1619

Průměry IAE

Integrální absolutní chyba (IAE)

$$IAE(H) = \int |\hat{f}(\mathbf{x}, H) - f(\mathbf{x})| d\mathbf{x}$$

<i>data</i>	LSCV	M1	M2
Norm I	0.3213	0.3016	0.2687
Norm II	0.3619	0.3468	0.3313
Norm III	0.4041	0.3828	0.3700
Student	0.4098	0.3378	0.3313
Weibull	0.3895	0.2942	0.2920

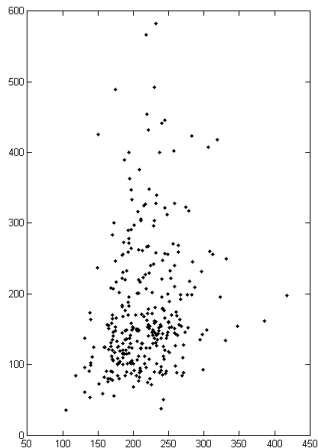
Obsah

- 1 Úvod
- 2 Metody pro odhad H – Metody založené na MISE
- 3 Metody pro odhad H – Metody založené na AMISE
- 4 Jiné metody pro odhad H
- 5 Simulovaná data
- 6 Reálná data**
- 7 Reference

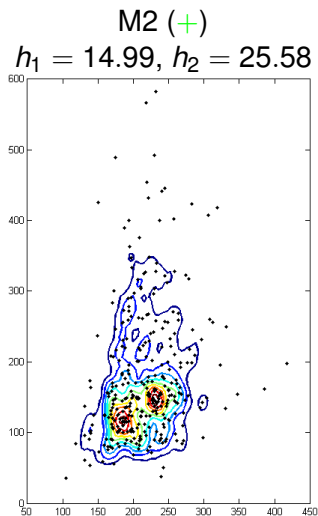
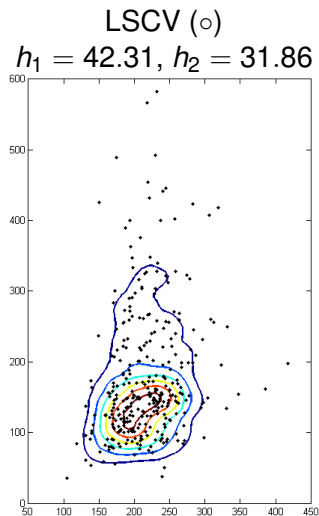
Lipidy v krvi

Koncentrace lipidů v krevní plazmě pacientů.

- 320 pacientů, u kterých bylo zjištěno zúžení cév,
- X_1 – cholesterol [mg/100 ml],
- X_2 – triglycerid [mg/100 ml].

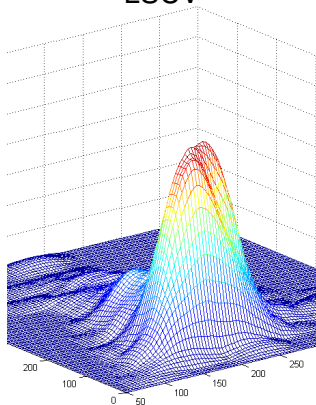


Lipidy v krvi – rekonstrukce

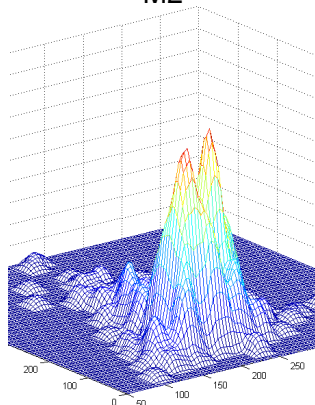


Lipidy v krvi – rekonstrukce

LSCV



M2



Obsah

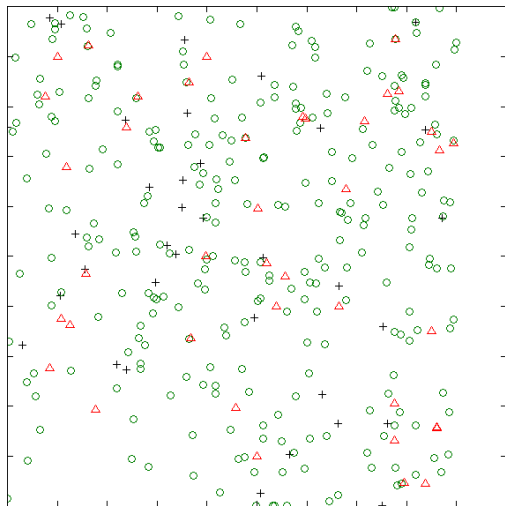
- 1 Úvod
- 2 Metody pro odhad H – Metody založené na MISE
- 3 Metody pro odhad H – Metody založené na AMISE
- 4 Jiné metody pro odhad H
- 5 Simulovaná data
- 6 Reálná data
- 7 Reference**

Reference

- ▷ Cao R., Cuevas A., González Manteiga W. **A comparative study of several smoothing methods on density estimation** *Comput. Statist. Data Anal.* 17, 1994.
- ▷ Duong T., Hazelton M.L. **Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation.** *J. Multivariate Anal.* 93, 2005.
- ▷ Duong T., Hazelton M.L. **Cross-validation Bandwidth matrices for Multivariate Kernel Density Estimation.** *Scand. J. Statist.* 31, 2004.
- ▷ Horová I. *et al.* **Bandwidth choice for kernel density estimates.** *Proceedings IASC.* Yokohama: IASC, 2008.
- ▷ Sain S.R., Baggerly K.A. and Scott D.W. **Cross-Validation of Multivariate Densities.** *J. Amer. Statist. Assoc.* 89, 807–817, 1994.
- ▷ Scott D.W. **Multivariate Density Estimation: Theory, Practice, and Visualization** New York: John Wiley & Sons, 1992.
- ▷ Wand M.P., Jones M.C. **Multivariate Plug-in Bandwidth Selection.** *Comput. Statist.* 9, 1994.
- ▷ Wand, M.P. and Jones, M.C. **Kernel Smoothing.** London: Chapman and Hall, 1995.

Pouštní rostliny

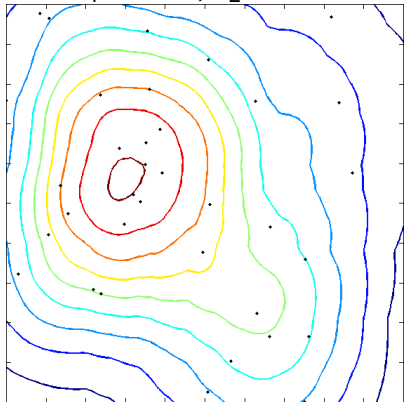
Měření dlouhověkosti u rostlin vybrané lokality Mohavské pouště.



Pouštní rostliny – dead

LSCV (○)

$h_1 = 6.01, h_2 = 7.38$



M2 (+)

$h_1 = 7.36, h_2 = 7.96$

