

Úvod

Vyhlazování je statistická technika pro rekonstrukci reálné funkce na základě pozorovaných nebo naměřených dat. Cílem vyhlazování je nalezení takového odhadu neznámé funkce, aby byly odfiltrovány náhodné výkyvy a bylo možné lépe poznat strukturu dat. K tomuto úkolu lze přistoupit dvěma způsoby – parametricky a neparametricky:

- *Parametrické odhady* jsou založeny na předpokladu, že neznámá funkce patří do třídy funkcí závislých na parametrech, a cílem je odhadnout tyto parametry.
- *Neparametrické odhady* nepředepisují datům „Prokrustovo lože“ parametrizace, ale nechávají „hovořit samotná data“.

V tomto učebním textu se zaměříme na neparametrické odhady, a to zejména na jádrové odhady, které patří mezi efektivní neparametrické odhady. Budeme se zabývat jádrovými odhady regresní funkce, hustoty, distribuční funkce a také odhadem dvourozměrné hustoty. Všechny jádrové odhady závisí na jádře, které má roli vahové funkce, a na vyhlazovacím parametru, který řídí hladkost odhadu.

Budeme zabývat následujícími otázkami:

- Jaké jsou statistické vlastnosti jádrových odhadů.
- Jaký vliv má tvar jádra na odhad.
- Jaký vliv má šířka vyhlazovacího okna na odhad.
- Jak lze tuto šířku stanovit v praxi.

Volba vhodného vyhlazovacího parametru je zásadním problémem ve všech typech jádrových odhadů a tomuto problému budeme věnovat značnou pozornost.

Všechny uvedené metody jsou implementovány v Matlabu, příslušný toolbox je dostupný na adrese:

```
https://www.math.muni.cz/veda-a-vyzkum/vyvijeny-software/  
274-matlab-toolbox.html
```

Na konci textu (kapitola 7) jsou uvedeny soubory dat pro samostatnou práci studentů. Tyto soubory již byly zpracovány v příslušných kapitolách a studenti si tak mohou ověřit správnost svých výsledků.

Definice základních statistických pojmů a jejich vlastností lze najít např. v elektronických skriptech Pravděpodobnost a statistika I autorů M. Forbelské a J. Kolářka (jsou dostupná na Elportálu Informačního systému).

Na tomto místě bych ráda poděkovala Mgr. Kamile Hasilové, Ph.D., za pomoc při sazbě tohoto textu a za příspěvek ke kapitole 5 a kapitolám o reálných datech.

Obsah

1	Jádrové funkce a jejich vlastnosti	6
1	Základní pojmy a definice	6
1.1	Jádra s minimálním rozptylem	7
1.2	Optimální jádra	8
2	Jádrové odhady regresní funkce	11
1	Motivace	11
2	Základní typy neparametrických odhadů	12
3	Statistické vlastnosti jádrových odhadů	16
4	Volba jádra	21
5	Volba vyhlazovacího parametru	22
5.1	Metoda křížového ověřování	22
6	Automatická procedura	24
7	Aplikace na reálná data	25
3	Jádrové odhady hustoty	30
1	Motivace	30
2	Základní typy neparametrických odhadů	30
3	Statistické vlastnosti jádrových odhadů hustoty	32
3.1	Odhad derivace hustoty	36
4	Volba jádra	37
5	Volba vyhlazovacího parametru	38
5.1	Metoda referenční hustoty	38
5.2	Metoda maximálního vyhlazení	39
5.3	Metoda křížového ověřování	41
5.4	Iterační metoda	43
6	Automatická procedura	45
7	Aplikace na reálná data	49
4	Jádrové odhady distribuční funkce	53
1	Motivace	53
2	Základní typy neparametrických odhadů	54
3	Statistické vlastnosti odhadu	55
4	Volba jádra	59
5	Volba vyhlazovacího parametru	59
5.1	Metody křížového ověřování	59
5.2	Princip maximálního vyhlazení	59
5.3	Plug-in metoda	60
6	Aplikace na reálná data	61

5	Jádrové odhady dvourozměrných hustot	66
1	Motivace	66
2	Základní typy odhadů	67
3	Statistické vlastnosti jádrových odhadů hustoty	68
4	Volba jádra	71
5	Volba vyhlazovacího parametru	71
	5.1 Metoda referenční hustoty	71
	5.2 Metoda křížového ověřování	72
6	Aplikace na reálná data	73
6	Návody ke cvičením a odpovědi na otázky	78
7	Datové soubory	85
8	Přílohy	99

Seznam použitého značení

h	vyhlazovací parametr
\mathbf{H}	matice vyhlazovacích parametrů
$m(\cdot)$	regresní funkce
$f(\cdot)$	hustota pravděpodobnosti
$F(\cdot)$	distribuční funkce
\hat{h}	odhad vyhlazovacího parametru h
$\hat{\sigma}$	odhad směrodatné odchylky σ
\hat{m}	odhad regresní funkce m
\hat{f}	odhad hustoty f
\hat{F}	odhad distribuční funkce F
\int	značí integrál $\int_{-\infty}^{\infty}$, pokud není uvedeno jinak
$K(\cdot)$	jádrová funkce (jádro)
$V(K)$	$V(K) = \int K^2(x) dx$
$V(g)$	$V(g) = \int g^2(x) dx$
$\beta_k(K)$	$\beta_k(K) = \int x^k K(x) dx$
$f * g$	konvoluce funkcí f a g , $(f * g)(x) = \int f(t)g(x - t) dt$
$W(\cdot)$	integrál z jádra, $W(x) = \int_{-\infty}^x K(t) dt$
$C^k[0, 1]$	prostor funkcí, které mají spojitě derivace až do řádu k včetně na intervalu $[0, 1]$
NW	Nadarayovy-Watsonovy odhady
PC	Priestleyovy-Chaovy odhady
LL	lokálně lineární odhady
GM	Gasserovy-Müllerovy odhady

MSE	střední kvadratická chyba (mean square error)
MISE	střední integrální kvadratická chyba (mean integrated square error)
AMISE	asymptotická střední integrální kvadratická chyba (asymptotic mean integrated square error)
AIV	asymptotický tvar integrálu rozptylu (asymptotic integrated variance)
AISB	asymptotický tvar integrálu druhé mocniny vychýlení (asymptotic integrated square bias)
AMSE	střední průměrná kvadratická chyba (average mean square error)
CV	metoda křížového ověřování
REF	metoda referenční hustoty
MS	metoda maximálního vyhlazení
PI	plug-in metoda
IT	iterační metoda

Symbolika O, o

Symboly O a o se často používají pro vyjádření chyb matematických výrazů. (Nazývají se také řádové vztahy.)

Nechť f je funkce definovaná v okolí bodu a . Symbol

$$f(x) = O(g(x)) \quad \text{pro } x \rightarrow a$$

značí, že

$$\limsup_{x \rightarrow a} \frac{|f(x)|}{g(x)} < \infty.$$

Podobně symbol

$$f(x) = o(g(x)) \quad \text{pro } x \rightarrow a$$

značí, že

$$\lim_{x \rightarrow a} \frac{|f(x)|}{g(x)} = 0.$$

Dodatek „pro $x \rightarrow a$ “ se často vynechává, pokud je jasné, o které a se jedná. Je to zejména v případech $a = 0$ či $a = \infty$. Často používaným výraz je $O(h^k)$, resp. $o(h^k)$, kde $g(h) = h^k$, přičemž zpravidla $h \rightarrow 0$.

Pro počítání s výrazy obsahujícími symboly O a o platí následující pravidla:

$$\begin{aligned} O(g(x)) + O(g(x)) &= O(g(x)) \\ o(g(x)) + o(g(x)) &= o(g(x)) \\ O(g(x)) + o(g(x)) &= O(g(x)) \\ O(g(x)) \cdot O(h(x)) &= O(g(x) \cdot h(x)) \\ o(g(x)) \cdot o(h(x)) &= o(g(x) \cdot h(x)) \\ O(g(x)) \cdot o(h(x)) &= o(g(x) \cdot h(x)) \\ o(g(x)) &= O(g(x)) \end{aligned}$$

Pozor! Tyto rovnice nejsou symetrické, platí jen zleva doprava. Např. poslední rovnice značí, že je-li funkce $f(x) = o(g(x))$, pak je $f(x) = O(g(x))$. Opačně to ovšem neplatí.

Kapitola 1

Jádrové funkce a jejich vlastnosti

1 Základní pojmy a definice

V úvodu bylo uvedeno, že všechny jádrové odhady závisí na jádrové funkci (jádře), a proto se v této kapitole budeme zabývat jádrovými funkcemi. Nyní uvedeme definici jádra a jeho vlastnosti.

Definice 1.1. Nechť ν, k jsou nezáporná celá čísla, $0 \leq \nu < k$, nechť K je reálná funkce s těmito vlastnostmi

1. K splňuje Lipschitzovu podmínku na intervalu $[-1, 1]$, tj. $|K(x) - K(y)| \leq L|x - y|$ pro $\forall x, y \in [-1, 1], L > 0$,
2. nosič(K) = $[-1, 1]$, tj. $K = 0$ vně intervalu $[-1, 1]$,
3. K splňuje momentové podmínky

$$\int_{-1}^1 x^j K(x) dx = \begin{cases} 0 & 0 \leq j < k, j \neq \nu, \\ (-1)^\nu \nu! & j = \nu \end{cases} \quad (1.1)$$

a $\int_{-1}^1 x^k K(x) dx \neq 0$, tuto hodnotu označíme $\beta_k(K)$.

Taková funkce K se nazývá *jádro řádu k* a třída všech těchto funkcí se označuje $S_{\nu k}$.

Definice 1.2. Označme $V(K) = \int_{-1}^1 K^2(x) dx$, $K \in S_{\nu k}$ pro jádro $K \in S_{\nu k}$ a definujme veličinu

$$\delta_{\nu k} = \left(\frac{V(K)}{\beta_k^2(K)} \right)^{\frac{1}{2k+1}}.$$

Tuto veličinu budeme nazývat *kanonický faktor*.

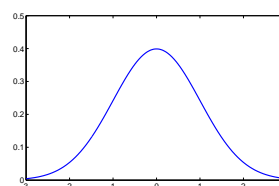
Příklad 1.1. V tabulce 1.1 jsou uvedeny příklady několika jader společně s jejich grafy. Funkce $I_{[-1,1]}(x)$ je indikátorová funkce intervalu $[-1, 1]$, tj.

$$I_{[-1,1]}(x) = \begin{cases} 1 & \text{pro } x \in [-1, 1], \\ 0 & \text{jinak.} \end{cases}$$

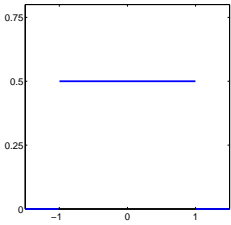
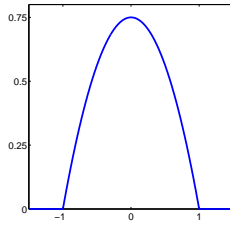
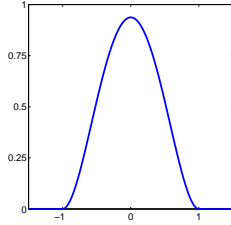
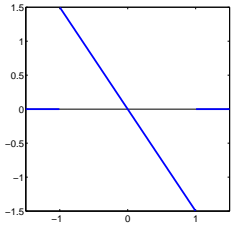
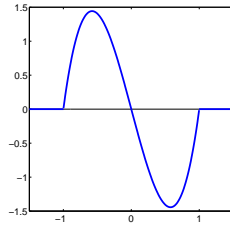
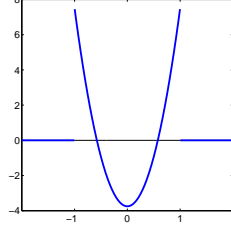
Poznámka 1.1. V praxi se také používá *Gaussovo jádro*

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}.$$

Jde sice o jádro řádu 2, avšak nemá vlastnost 2 z definice 1.1.



Tabulka 1.1: Jádra

S_{02}		
$K(x) = \frac{1}{2}I_{[-1,1]}(x)$ obdélkové jádro 	$K(x) = \frac{3}{4}(1-x^2)I_{[-1,1]}(x)$ Epanečnikovo jádro 	$K(x) = \frac{15}{16}(1-x^2)^2 I_{[-1,1]}(x)$ kvartické jádro 
S_{13}		S_{24}
$K(x) = -\frac{3}{2}xI_{[-1,1]}(x)$ 	$K(x) = -\frac{15}{4}x(1-x^2)I_{[-1,1]}(x)$ 	$K(x) = -\frac{15}{4}(1-3x^2)I_{[-1,1]}(x)$ 

Nyní uvedeme dva typy jader – jádra s minimálním rozptylem a optimální jádra.

1.1 Jádra s minimálním rozptylem

Předpokládejme, že $K \in S_{\nu k}$, $0 \leq \nu \leq k-2$, ν a k jsou obě sudá nebo lichá¹. Uvažujme funkcionál $V(K) = \int_{-1}^1 K^2(x) dx$ a zabývejme se problémem najít takové jádro $K \in S_{\nu k}$, pro které tento funkcionál nabývá minimální hodnoty, tj. řešíme variační úlohu

$$\min V(K) \quad \text{za předpokladu } K \in S_{\nu k}.$$

Řešení této úlohy se nazývají *jádra s minimálním rozptylem*, což jsou polynomy stupně $k-2$ na intervalu $[-1, 1]$. Tyto polynomy jsou sudé funkce pro k sudé a liché funkce pro k liché a mají $k-2$ různých kořenů v intervalu $(-1, 1)$. Obecný vztah pro jádra s minimálním rozptylem lze nalézt v [3].

Příklad 1.2. Jádra s minimálním rozptylem:

$$S_{02}: \quad K(x) = \frac{1}{2}I_{[-1,1]}(x)$$

$$S_{13}: \quad K(x) = -\frac{3}{2}xI_{[-1,1]}(x)$$

$$S_{24}: \quad K(x) = -\frac{15}{4}(1-3x^2)I_{[-1,1]}(x)$$

Poznámka 1.2. Jádra s minimálním rozptylem mají skoky v koncových bodech intervalu $[-1, 1]$, což negativně ovlivňuje hladkost výsledného odhadu.

¹Obvykle se používá pojem *parita*, tedy ν a k mají stejnou paritu.

Tabulka 1.2: Jádra s minimálním rozptylem (MR) a optimální jádra (OPT)

ν	k	typ	vzorec (na intervalu $[-1, 1]$)	$V(K)$	$\beta(K)$	$T(K)$
0	2	OPT	$\frac{3}{4}(1 - x^2)$	0,600	0,2000	0,3491
		MR	$\frac{1}{2}$	0,500	0,3333	0,3701
0	4	OPT	$\frac{15}{32}(x^2 - 1)(7x^2 - 3)$	1,250	-0,0476	0,6199
		MR	$\frac{3}{8}(3 - 5x^2)$	1,125	-0,0857	0,6432
1	3	OPT	$\frac{15}{4}x(x^2 - 1)$	2,143	-0,4286	0,7477
		MR	$-\frac{3}{2}x$	1,500	-0,6000	0,8137
1	5	OPT	$-\frac{105}{32}x(x^2 - 1)(9x^2 - 5)$	11,93	0,1515	2,1675
		MR	$\frac{15}{8}x(7x^2 - 5)$	9,375	0,2381	2,3278
2	4	OPT	$-\frac{105}{16}(x^2 - 1)(5x^2 - 1)$	35,00	1,3333	6,6846
		MR	$\frac{15}{4}(1 - 3x^2)$	22,50	1,7143	7,2622
2	6	OPT	$\frac{315}{64}(x^2 - 1)(77x^4 - 58x^2 + 5)$	381,6	-0,6293	27,162
		MR	$\frac{105}{32}(-45x^4 + 42x^2 - 5)$	275,6	-0,9091	29,503

1.2 Optimální jádra

Při vyšetřování statistických vlastností se setkáváme s následujícím funkcioálem

$$T(K) = \left(\underbrace{\left| \int_{-1}^1 x^k K(x) dx \right|}_{\beta_k(K)} \right)^{2\nu+1} \left(\underbrace{\int_{-1}^1 K^2(x) dx}_{V(K)} \right)^{k-\nu} \frac{1}{2^{k+1}},$$

který lze zkráceně psát $T(K) = (|\beta_k(K)|^{2\nu+1} V(K)^{(k-\nu)})^{2/(2k+1)}$. Jádra, pro která tento funkcioál nabývá minimální hodnoty, se nazývají *optimální jádra*. Jde o polynomy stupně k , které mají $k - 2$ různých kořenů v intervalu $(-1, 1)$ a body $-1, 1$ jsou rovněž kořeny těchto polynomů. Obecný vzorec pro tvar optimálních jader lze nalézt např. v [3].

Příklad 1.3. Optimální jádra:

$$\begin{aligned} S_{02}: \quad K_{opt,0,2}(x) &= \frac{3}{4}(1 - x^2)I_{[-1,1]}(x) \\ S_{13}: \quad K_{opt,1,3}(x) &= \frac{15}{4}x(x^2 - 1)I_{[-1,1]}(x) \\ S_{24}: \quad K_{opt,2,4}(x) &= -\frac{105}{16}(x^2 - 1)(5x^2 - 1)I_{[-1,1]}(x) \end{aligned}$$

Jádra $K_{opt,1,k}$ a $K_{opt,2,k}$ se používají pro odhad první a druhé derivace hustoty (viz kapitola 3.1) a z toho důvodu pro ně zavedeme dodatečné označení $K^{(1)}$, respektive $K^{(2)}$.

Přehled vybraných jader s minimálním rozptylem a optimálních jader je uveden v tabulce 1.2

Poznámka 1.3. Pro jádra s minimálním rozptylem a optimální jádra platí následující tvrzení, jehož důkaz je v [3].

Nechť $K \in S_{\nu+1,k+1}$ je jádro s minimálním rozptylem a $K_{opt,\nu,k} \in S_{\nu,k}$ je optimální jádro. Pak platí

$$K'_{opt,\nu,k}(x) = K(x), \quad x \in [-1, 1]$$

Jako příklad lze uvést jádro $K_{opt,0,2}(x) = \frac{3}{4}(1 - x^2) \in S_{02}$ a $K'_{opt,0,2}(x) = K_{1,3}(x) = -\frac{3}{2}x \in S_{13}$.

Poznámka 1.4. Optimální jádra jsou spojité funkce na celé reálné ose, což znamená, že jsou „hladší“ než jádra s minimálním rozptylem. Odhadovaná funkce „zdědí“ hladkost jádra a to znamená, že hladší jádra produkují hladší křivku. Nejčastěji používaným jádrem je Epanečnikovo jádro.

Shrnutí
<p>Funkcionály závislé na jádře K</p> $\beta_k(K) = \int_{-1}^1 x^k K(x) dx \quad V(K) = \int_{-1}^1 K^2(x) dx$ $T(K) = (\beta_k(K) ^{2\nu+1} V(K)^{(k-\nu)})^{\frac{2}{2k+1}} \quad \delta_{\nu k} = \left(\frac{V(K)}{\beta_k^2(K)} \right)^{\frac{1}{2k+1}}$
<p>Jádra s minimálním rozptylem minimalizují funkcionál $V(K)$. Optimální jádra minimalizují funkcionál $T(K)$.</p>
<p>Podrobnější popis jader a vzorců s nimi souvisejících lze nalézt v toolboxu Matlabu.</p>

Výstupy z výukové jednotky

Student

- zná základní třídy jádrových funkcí, jejich vlastností a metody jejich konstrukce

Dodatek

Nechť $K \in S_{\nu k}$, $\nu \in \mathbb{N}$, pro $\gamma > 0$ položíme

$$K_\gamma(\cdot) = \frac{1}{\gamma^{\nu+1}} K\left(\frac{\cdot}{\gamma}\right).$$

Jádra K a K_γ nazýváme *ekvivalentní*.

Nechť jsou dány funkce f a g , pro které platí

$$\int f^2(x) dx < \infty \quad \text{a} \quad \int g^2(x) dx < \infty.$$

Konvoluci $f * g$ definujeme vztahem

$$(f * g)(x) = \int f(t)g(x-t) dt.$$

Cvičení

1. Dokažte, že funkcionál $T(K)$ je invariantní vzhledem k transformaci $K_\gamma(\cdot)$, tj.

$$T(K) = T(K_\gamma).$$

2. Vypočítejte kanonické faktory $\delta_{\nu k}$ a hodnoty funkcionálu $T(K)$ pro jádra z tabulky 1.2.

3. Ukažte, že pro konvoluci $f * g$ platí

- $f * g = g * f$,

- $f * (g * h) = (f * g) * h$,
- $f * (g + h) = f * g + f * h$.

4. Nechť $K \in S_{02}$. Ukažte, že platí následující vztah

$$\int (K * K)(x) x^j dx = \begin{cases} 1 & j = 0, \\ 0 & j = 1, \\ 2\beta_2(K) & j = 2. \end{cases}$$

5. Jak je třeba zvolit konstantu A , aby funkce

$$K(x) = \begin{cases} A \cos \frac{\pi}{2} x & x \in [-1, 1], \\ 0 & \text{jinak,} \end{cases}$$

byla jádrem třídy S_{02} ? Dále dopočítejte hodnoty $\beta(K)$ a $V(K)$ pro toto jádro.

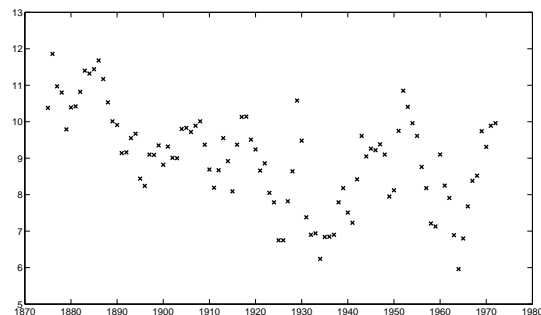
Kapitola 2

Jádrové odhady regresní funkce

1 Motivace

Uvažujme datový soubor, který obsahuje měření úrovně hladiny Huronského jezera. Huronské jezero je druhé největší jezero v systému pěti velkých jezer v Severní Americe. Jezerem prochází státní hranice mezi Kanadou a USA.¹

Měření byla prováděna ročně, v letech 1875 až 1972, a výsledky měření jsou zobrazeny na obrázku 2.1. Naším cílem je najít funkci popisující úroveň hladiny v uvedených letech. Vidíme,



Obrázek 2.1: Úroveň hladiny Huronského jezera

že pouhý pohled na tento dvourozměrný bodový diagram obvykle nestačí k tomu, abychom určili tento funkční vztah.

Statistická úloha, kterou se budeme zabývat, spočívá v proložení vhodné křivky těmito body tak, aby byly odfiltrovány náhodné výkyvy a bylo možné lépe poznat strukturu dat. Tuto křivku nazýváme regresní křivkou.

Formalizujme nyní tuto úlohu: Uvažujme standardní regresní model

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \dots, n, \quad (2.1)$$

kde m je neznámá regresní funkce, $x_i, i = 1, \dots, n$, jsou body plánu, $Y_i, i = 1, \dots, n$, jsou hodnoty závisle proměnné a $\varepsilon_i, i = 1, \dots, n$, jsou chyby měření, o nichž se předpokládá, že jsou nezávislémi, identicky rozdělenými náhodnými veličinami splňujícími podmínky

$$E\varepsilon_i = 0, \quad \text{var } \varepsilon_i = \sigma^2, \quad i = 1, \dots, n. \quad (2.2)$$

Poznámka 2.1. Jsou-li body plánu uspořádaná nenáhodná čísla, mluvíme o regresním modelu s *pevným plánem*. V případě, že body plánu X_1, \dots, X_n jsou náhodné veličiny se stejnou hustotou

¹„Great Lakes from space“ od SeaWiFS Project, NASA/Goddard Space Flight Center, and ORBIMAGE. – http://visibleearth.nasa.gov/view_rec.php?id=793. Licencováno pod Public domain via Wikimedia Commons.

f , jedná se o regresní model s *náhodným plánem* (podrobněji např. [14]). Budeme se dále zabývat modelem s *pevným plánem*.

Bez újmy na obecnosti budeme v dalším předpokládat, že pro body $x_i, i = 1, \dots, n$, platí

$$0 \leq x_1 \leq x_2 \leq \dots \leq x_n \leq 1.$$

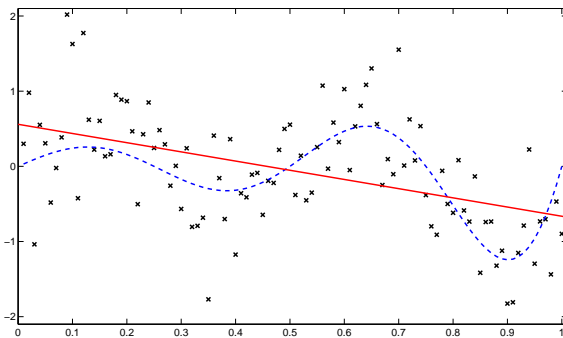
Cílem regresní analýzy je nalézt vhodnou aproximaci \hat{m} neznámé funkce m . Tento proces odhadu regresní funkce se obvykle nazývá *vyhlazování*. K tomuto úkolu lze přistoupit dvěma způsoby – parametricky a neparametricky. Příkladem parametrického odhadu regresní funkce je regresní přímka vyjadřující lineární závislost. Naopak u neparametrického přístupu nepředpokládáme, že funkce má nějaký předepsaný tvar, pouze předpokládáme jistou hladkost odhadované funkce (tj. dostatečný počet spojitých derivací).

V první polovině dvacátého století byla věnována pozornost zejména parametrickým metodám. V posledních letech však zaznamenaly značný rozvoj neparametrické metody. Tento vývoj souvisí s rostoucími požadavky na zpracování dat, ať už jde o rozsah souborů, rozmanitost těchto dat apod. Čistě parametrický přístup nevyhovuje vždy potřebám flexibility a nebývalý rozmach výpočetní techniky vytvořil dobré předpoklady pro rozvoj neparametrických metod. I přes tento vývoj si oba způsoby zachovávají své výhody a nijak si nekonkurují. Někdy je vhodné užít neparametrické metody a pak na výsledný odhad použít parametrickou metodu.

Příklad 2.1. Obr. 2.2 ilustruje na simulovaných datech nevhodnost aplikace parametrického přístupu. V tomto případě byla data generována podle vztahu

$$Y_i = \frac{\sin 4\pi x_i}{(1 + \cos 0,6\pi x_i)^2} + \varepsilon_i,$$

kde body $x_i = i/100, i = 1, \dots, 100$, a chyby $\varepsilon_i, i = 1, \dots, 100$, mají normální rozdělení $N(0; 0,25)$. (Data jsou v tabulce 7.1.)



Obrázek 2.2: Simulovaná data (\times) s regresní přímkou (červená, plná) a původní funkcí (modrá, čárkovaná)

Předpokládejme, že hledaná křivka je přímka a známou metodou nejmenších čtverců určíme rovnici této přímky. Obr. 2.2 znázorňuje přesnou funkci, generovaná data a výslednou přímku. Je zřejmé, že náš předpoklad, že hledaná funkce je přímka, není správný.

2 Základní typy neparametrických odhadů

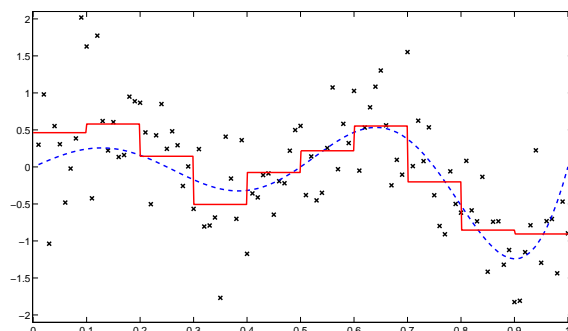
Pokud jde o historii neparametrických metod, připomeňme, že v r. 1857 saský ekonom Engel analyzoval data týkající se nákladů na domácnost a pro vyjádření závislosti použil schodovitou (tj. po částech konstantní funkci), kterou dnes nazýváme *regresogram*. Regresogram užívá stejné základní myšlenky jako histogram pro odhad hustoty. Tato myšlenka spočívá v rozdělení

množiny hodnot proměnné X na intervaly $B_j, j = 1, \dots, J$, a za odhad v bodě $x \in B_j$ se vezme průměr hodnot Y na tomto subintervalu, tj.

$$\hat{m}(x) = \frac{\sum_{i=1}^n Y_i I_{[B_j]}(x_i)}{\sum_{i=1}^n I_{[B_j]}(x_i)},$$

kde $I_{[B_j]}$ je indikátorová funkce subintervalu B_j .

Výsledek aplikace regresogramu na simulovaná data z příkladu 2.1 je znázorněn na obr. 2.3. Vidíme, že tento odhad „vhodně“ vystihuje tvar funkce, ale výsledný odhad je příliš hrubý.

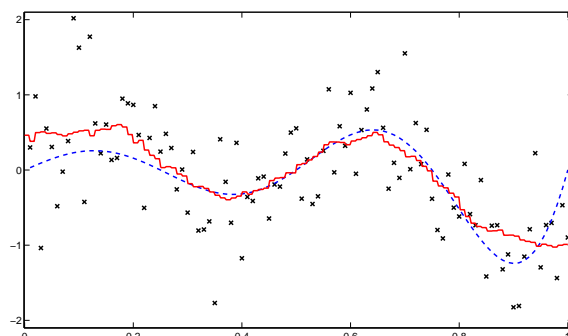


Obrázek 2.3: Regresogram (červená, plná) pro simulovaná data z příkladu 2.1 s původní funkcí (modrá, čárkovaná)

Přirozeným zobecněním regresogramu je *metoda klouzavých průměrů*. Tato metoda používá lokálních průměrů hodnot Y , ale odhad v bodě x je založen na centrovaném okolí bodu x : $[x - h, x + h]$, $h > 0$, přesněji

$$\hat{m}(x, h) = \frac{\sum_{i=1}^n Y_i I_{[x-h, x+h]}(x_i)}{\sum_{i=1}^n I_{[x-h, x+h]}(x_i)}. \quad (2.3)$$

Obr. 2.4 ilustruje aplikaci této metody na simulovaných datech příkladu 2.1. Uvedené metody



Obrázek 2.4: Klouzavý průměr (červená, plná) pro simulovaná data z příkladu 2.1 s původní funkcí (modrá, čárkovaná)

patří mezi nejjednodušší neparametrické vyhlazovací metody. Jádrové odhady lze považovat za zobecnění těchto metod.

Připomeňme zde základní myšlenku vyhlazování tak, jak ji formuloval R. Eubank v r. 1988:

Jestliže předpokládáme, že m je hladká funkce, pak pozorování v bodech x_i blízko bodu x obsahují informace o hodnotě m v bodě x . Bylo by tedy vhodné užít lokálních průměrů dat blízko bodu x , abychom získali odhad $m(x)$.

Obecně lze jádrové odhady regresní funkce m v bodě x definovat takto

$$\hat{m}(x, h) = \sum_{i=1}^n W_i(x, h) Y_i, \quad (2.4)$$

kde funkce $W_i(x, h)$, $i = 1, \dots, n$, se nazývají *váhy*, nezávisí na hodnotách Y_i , ale závisí na kladném čísle h , které se nazývá *vyhlazovací parametr* (nebo také šířka vyhlazovacího okna). Speciální, velmi užitečný typ vah, závisí na jádrové funkci K .

Nechť $K \in S_{0k}$, k je sudé číslo, položme $K_h(\cdot) = \frac{1}{h} K(\frac{\cdot}{h})$. Mezi nejznámější typy jádrových odhadů regresní funkce patří ([8]):

1. Nadarajovy-Watsonovy odhady (1964)

$$\hat{m}_{NW}(x, h) = \frac{\sum_{i=1}^n K_h(x - x_i) Y_i}{\sum_{i=1}^n K_h(x - x_i)},$$

2. Priestleyovy-Chaovy odhady (1972)

$$\hat{m}_{PC}(x, h) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) Y_i,$$

3. lokálně lineární odhady (Stone 1977, Cleveland 1979)

$$\hat{m}_{LL}(x, h) = \frac{1}{n} \sum_{i=1}^n \frac{\{\hat{s}_2(x, h) - \hat{s}_1(x, h)(x - x_i)\} K_h(x - x_i) Y_i}{\hat{s}_2(x, h) \hat{s}_0(x, h) - \hat{s}_1(x, h)^2},$$

kde

$$\hat{s}_r(x) = \frac{1}{n} \sum_{i=1}^n (x - x_i)^r K_h(x - x_i), \quad r = 0, 1, 2,$$

4. Gasserovy-Müllerovy odhady (1979)

$$\hat{m}_{GM}(x, h) = \sum_{i=1}^n Y_i \int_{s_{i-1}}^{s_i} K_h(x - t) dt,$$

kde $s_0 = 0$, $s_i = (x_i + x_{i+1})/2$, $s_n = 1$. Tento odhad je konvolučním typem odhadu.

Úmluva. Uvedené jádrové odhady budeme zapisovat ve tvaru

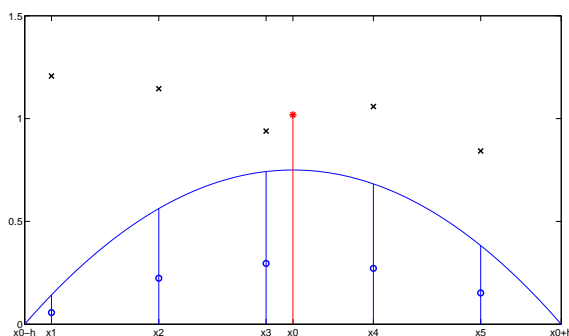
$$\hat{m}_j(x, h) = \sum_{i=1}^n W_i^{(A)}(x, h) Y_i,$$

kde index A značí příslušný typ odhadu NW , PC , LL , GM s danou váhovou funkcí.

V mnoha aplikacích je užitečný zejména Nadarayův-Watsonův odhad \hat{m}_{NW} . Popíšeme nyní jeho konstrukci a budeme ilustrovat vliv vyhlazovacího parametru na kvalitu odhadu. Pro daný bod x_0 , $h < x_0 < 1 - h$, jsou váhy Nadarayova-Watsonova odhadu dány vztahem

$$W_i(x_0, h) = \frac{K_h(x_0 - x_i)}{\sum_{j=1}^n K_h(x_0 - x_j)}, \quad \sum_{j=1}^n W_j(x_0, h) = 1.$$

Obrázek 2.5 ilustruje konstrukci odhadu v bodě x_0 , který je založen na pěti pozorováních $(x_1, Y_1), \dots, (x_5, Y_5)$ (černé křížky). Parabola reprezentuje Epanečnikovo jádro K_h a kroužky znázorňují hodnoty vah $W_i(x_0, h) = K_h(x_0 - x_i) / \sum_{i=1}^5 K_h(x_0 - x_i)$ pro $i = 1, \dots, 5$. Výsledný odhad regresní funkce \hat{m} v bodě x_0 je označen hvězdičkou.



Obrázek 2.5: Ilustrace Nadarayova-Watsonova odhadu v bodě x_0

OTÁZKA. Popište konstrukci Nadarayova-Watsonova odhadu, použijeme-li obdélníkové jádro místo Epanečnikova jádra. Vypočtete váhy $W_i(x_0, h)$ pro odhad s obdélníkovým jádrem. (Odpověď viz kapitola 6.)

Jádrový odhad není definován pro $\sum_{i=1}^n K_h(x - x_i) = 0$. Jestliže nastane případ „0/0“, pak klademe $\hat{m}_{NW}(x, h) = 0$. Omezíme se nyní na odhady funkce m v bodech plánu x_i , $i = 1, \dots, n$.

Pro malé hodnoty h je výraz $\left| \frac{x_i - x_j}{h} \right| > 1$ pro $x_i \neq x_j$, a tedy hodnota jádra v těchto bodech je rovna nule, a pro bod x_i dostáváme odhad

$$\hat{m}_{NW}(x_i, h) \rightarrow \frac{K(0)Y_i}{K(0)} = Y_i.$$

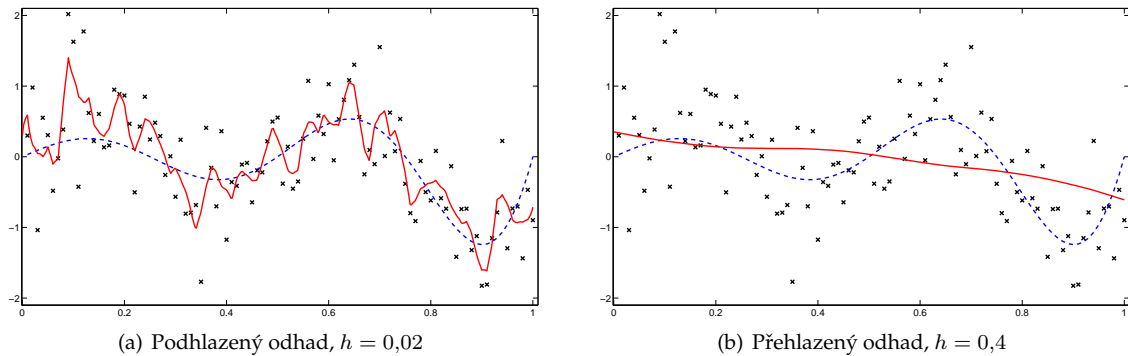
To znamená, že při malé šířce vyhlazovacího okna ($h \rightarrow 0$) odhad reprodukuje data (viz obr. 2.6(a)).

Podobně pro velké hodnoty h je výraz $\left| \frac{x_i - x_j}{h} \right| \approx 0$, tedy pro všechny body plánu je hodnota jádrové funkce $K\left(\frac{x_i - x_j}{h}\right) \approx K(0)$ a dostaneme tak průměr dat

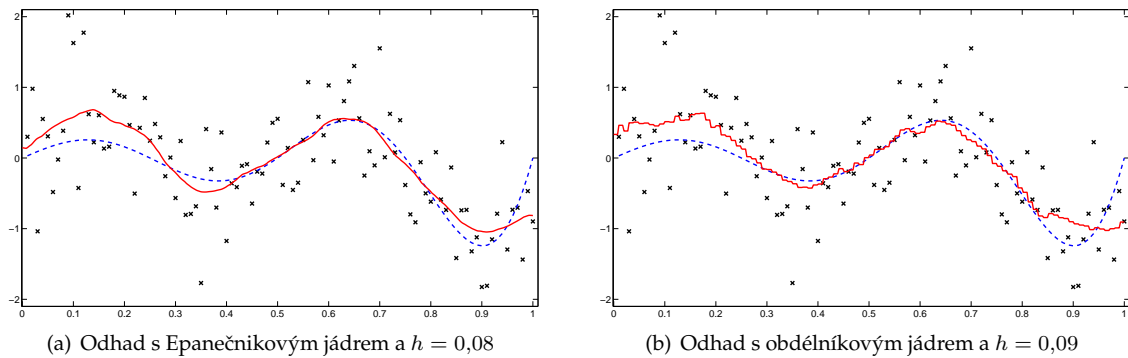
$$\hat{m}_{NW}(x_i, h) \rightarrow \frac{\sum_{j=1}^n K(0)Y_j}{\sum_{j=1}^n K(0)} = \frac{K(0) \sum_{j=1}^n Y_j}{nK(0)} = \frac{1}{n} \sum_{j=1}^n Y_j.$$

Tedy velká šířka okna ($h \rightarrow \infty$) vede k přehlazení, a to k průměru dat (viz obr. 2.6(b)).

Na obrázku 2.7(a) je znázorněn odhad s Epanečnikovým jádrem. Tento odhad se nejvíce blíží skutečné regresní funkci. Pokud jde o volbu vyhlazovacího parametru, je třeba si uvědomit, že konečné rozhodnutí o odhadované křivce je částečně subjektivní, neboť i asymptoticky optimální odhady obsahují poměrně značné „množství šumu“ a to nechává prostor pro subjektivní posouzení.



Obrázek 2.6: Podhlazený a přehladený odhad (červená, plná) regresní funkce (modrá, čárkovaná) z příkladu 2.1



Obrázek 2.7: Odhady (červená, plná) regresní funkce (modrá, čárkovaná) z příkladu 2.1 s Epanečnickým a obdélníkovým jádrem

Poznámka 2.2. *Intervaly spolehlivosti* pro hodnotu regresní funkce m v bodě x jsou užitečné v mnoha aplikacích. Bodový interval spolehlivosti udává interval, v němž s pravděpodobností $1 - \alpha$ leží hodnota funkce m v bodě x . Jsou definovány takto

$$\left[\hat{m}(x, h) - u_{1-\frac{\alpha}{2}} \sqrt{\frac{V(K)\hat{\sigma}^2(x)}{nh}}, \hat{m}(x, h) + u_{1-\frac{\alpha}{2}} \sqrt{\frac{V(K)\hat{\sigma}^2(x)}{nh}} \right],$$

kde $u_{1-\frac{\alpha}{2}}$ je $(1 - \alpha/2)$ -kvantil standardního normálního rozdělení a odhad rozptylu v bodě x je dán vztahem

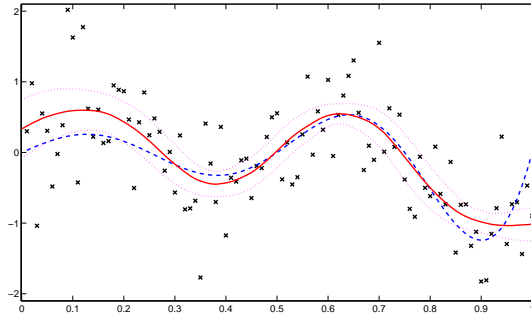
$$\hat{\sigma}^2(x) = \sum_{i=1}^n W_i(x, h) (Y_i - \hat{m}(x, h))^2.$$

Ukázka intervalu spolehlivosti pro $\alpha = 0,05$ je na obrázku 2.8.

3 Statistické vlastnosti jádrových odhadů

Kvalitu jádrového odhadu lze lokálně popsat pomocí střední kvadratické chyby (MSE) odhadu \hat{m} v bodě x , která je obecně dána vztahem

$$\text{MSE } \hat{m}(x, h) = E(\hat{m}(x, h) - m(x))^2.$$



Obrázek 2.8: Interval spolehlivosti pro data z příkladu 2.1 při $\alpha = 0,05$ (růžová, tečkovaná) se zobrazeným odhadem regresní funkce (červená, plná) a původní funkcí (modrá, čárkovaná)

Upravíme tento vztah

$$\begin{aligned} \text{MSE } \hat{m}(x, h) &= E\hat{m}^2(x, h) - 2m(x)E\hat{m}(x, h) + m^2(x) \\ &= \underbrace{(E\hat{m}(x, h) - m(x))^2}_{\text{bias}^2} + \underbrace{E\hat{m}^2(x, h) - (E\hat{m}(x, h))^2}_{\text{var}}, \end{aligned} \quad (2.5)$$

což znamená, že střední kvadratická chyba může být vyjádřena jako součet *rozptylu odhadu* $\text{var } \hat{m}(x, h)$ a čtverce *vychýlení* $\text{bias}^2 \hat{m}(x, h)$. Tento rozklad *rozptyl-vychýlení* usnadňuje analýzu vlastností odhadu.

Všechny uvedené jádrové odhady regresní funkce \hat{m}_{NW} , \hat{m}_{PC} , \hat{m}_{LL} , \hat{m}_{GM} jsou asymptoticky ekvivalentní (viz např. [8, 14]). Z tohoto důvodu budeme dále podrobněji zabývat *Priestleyovými-Chaovými odhady*, které budeme psát bez uvedení označení *PC*, tedy: $\hat{m}(x, h)$ a $W_i(x, h)$.

Připomeňme, že pro Priestleyovy-Chaovy odhady je váhová funkce tvaru

$$W_i(x, h) = \frac{1}{n} K_h(x - x_i) = \frac{1}{nh} K\left(\frac{x - x_i}{h}\right).$$

Pro další výpočty budeme předpokládat:

- (i) Jádrová funkce K je sudou funkcí na intervalu $[-1, 1]$, $K \in S_{02}$,
- (ii) vyhlazovací parametr $h = h(n)$ je nenáhodnou posloupností kladných čísel splňující $h \rightarrow 0$ a $nh \rightarrow \infty$ pro $n \rightarrow \infty$,
- (iii) bod x , v němž počítáme odhad, splňuje nerovnost $h < x < 1 - h$ pro všechna $n \geq n_0$, kde n_0 je pevné,
- (iv) $m \in C^2[0, 1]$,
- (v) $x_i = \frac{i}{n}$, $i = 1, \dots, n$.

Je zřejmé, že pro $n \rightarrow \infty$ platí (jedná se o přibližný výpočet integrálu – viz Dodatek na str. 29)²

$$E\hat{m}(x, h) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) EY_i = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) m(x_i) = \int_0^1 \frac{1}{h} K\left(\frac{x-t}{h}\right) m(t) dt + O(n^{-1}). \quad (2.6)$$

Nechť $u = \frac{x-t}{h}$, odtud $dt = -h du$, a tedy s využitím Taylorova rozvoje

$$\begin{aligned} E\hat{m}(x, h) &= \int_{-(1-x)/h}^{x/h} K(u) m(x - hu) du + O(n^{-1}) \\ &= \int_{-(1-x)/h}^{x/h} K(u) \left[m(x) - uhm'(x) + \frac{u^2 h^2}{2} m''(x) + o(h^2) \right] du + O(n^{-1}). \end{aligned} \quad (2.7)$$

²Symbolika O a o viz úvodní část na straně 5.

Podle výše uvedených předpokladů platí $h < x < 1 - h$, tedy $x/h \rightarrow \infty$ a $-(1-x)/h \rightarrow -\infty$ pro $h \rightarrow 0$. Odtud, s využitím faktu, že nosičem funkce K je interval $[-1, 1]$, plyne

$$E\hat{m}(x, h) = m(x) \underbrace{\int_{-1}^1 K(u) du}_{=1} - hm'(x) \underbrace{\int_{-1}^1 uK(u) du}_{=0} + \frac{h^2}{2} m''(x) \underbrace{\int_{-1}^1 u^2 K(u) du}_{=\beta_2(K)} + o(h^2) + O(n^{-1}).$$

Celkem dostaneme

$$\begin{aligned} \text{bias } \hat{m}(x, h) &= \frac{h^2}{2} \beta_2(K) m''(x) + o(h^2) + O(n^{-1}) \\ &\approx \frac{h^2}{2} \beta_2(K) m''(x). \end{aligned}$$

Podobně pro rozptyl platí

$$\begin{aligned} \text{var } \hat{m}(x, h) &= E(\hat{m}(x, h) - E\hat{m}(x, h))^2 \\ &= E\left(\frac{1}{n} \sum_{i=1}^n K_h(x - x_i) Y_i - \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) m(x_i)\right)^2 \\ &= \frac{1}{n^2} E\left(\sum_{i=1}^n K_h(x - x_i) \underbrace{(Y_i - m(x_i))}_{\varepsilon_i}\right)^2, \end{aligned}$$

z vlastností (2.2) plyne $EK_h(x_i)K_h(x_j)\varepsilon_i\varepsilon_j = 0$ pro $i \neq j$, tedy

$$\begin{aligned} &= \frac{1}{n^2} \sum_{i=1}^n K_h^2(x - x_i) E\varepsilon_i^2 \\ &= \frac{\sigma^2}{n^2 h^2} \sum_{i=1}^n K^2\left(\frac{x - x_i}{h}\right) = \frac{\sigma^2}{nh} \left(\int_0^1 \frac{1}{h} K^2\left(\frac{x-t}{h}\right) dt + O(n^{-1})\right), \end{aligned}$$

Zde jsme opět použili přibližného výpočtu integrálu. Opět s využitím substituce $u = \frac{x-t}{h}$ a vztahu $O(n^{-1}) = o((nh)^{-1})$ můžeme pro $n \rightarrow \infty$ psát

$$\text{var } \hat{m}(x, h) = \frac{\sigma^2}{nh} \int_{-1}^1 K^2(u) du + o((nh)^{-1}) = \frac{\sigma^2}{nh} \int_{-1}^1 K^2(u) du + o((nh)^{-1}) \approx \frac{\sigma^2}{nh} V(K).$$

Tímto jsme dokázali následující větu o tvaru střední kvadratické chyby.

Věta 2.1. *Nechť jsou splněny předpoklady (i) – (iii), pak střední kvadratická chyba nabývá tvaru*

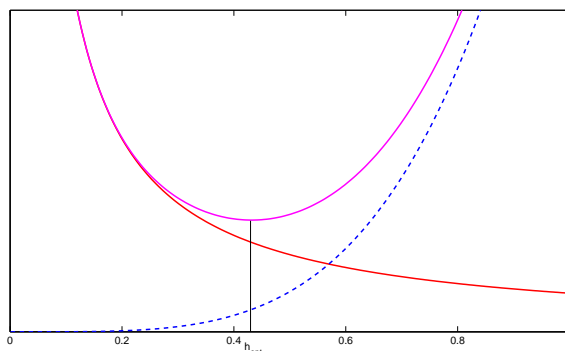
$$\text{MSE } \hat{m}(x, h) = \underbrace{\frac{\sigma^2}{nh} V(K)}_{\text{var}} + \underbrace{h^4 \beta_2^2(K) \frac{1}{4} (m''(x))^2}_{\text{bias}^2} + o(h^4 + (nh)^{-1}). \quad (2.8)$$

Chyba MSE dává pouze lokální pohled na chybu odhadu, proto se častěji používá globální tvar chyby – AMISE – asymptotická střední integrální kvadratická chyba. AMISE je součástí střední integrální kvadratické chyby (MISE) a vztah mezi chybami MSE, MISE a AMISE je následující

$$\text{MISE } \hat{m}(\cdot, h) = \int_0^1 \text{MSE } \hat{m}(x, h) dx = \text{AMISE } \hat{m}(\cdot, h) + o(h^4 + (nh)^{-1}).$$

AMISE je tvaru

$$\text{AMISE } \hat{m}(\cdot, h) = \underbrace{\frac{\sigma^2}{nh} V(K)}_{\text{AIV}(h)} + \underbrace{\frac{h^4}{4} \beta_2^2(K) V(m'')}_{\text{AISB}(h)}, \quad (2.9)$$



Obrázek 2.9: AMISE (růžová, plná) jako součet rozptylu AIV (červená, plná) a vychýlení AISB (modrá, čárkovaná)

kde $V(m'') = \int_0^1 (m''(x))^2 dx$ a AIV značí asymptotický tvar rozptylu (*asymptotic integrated variance*) a AISB asymptotický tvar druhé mocniny vychýlení (*asymptotic integrated square bias*).

Na obrázku 2.9 je znázorněn průběh AIV(h) a AISB(h) a také výsledné chyby AMISE $\widehat{m}(\cdot, h)$. Je vidět, že rozptyl AIV(h) nabývá velkých hodnot pro h malé, ale AISB(h) klesá. Pro velké h je tomu naopak. Volba vyhlazovacího parametru je zřejmě klíčovým problémem jádrového vyhlazování.

Naším cílem je minimalizovat AMISE $\widehat{m}(\cdot, h)$, tzn. najít takovou hodnotu vyhlazovacího parametru h , pro kterou asymptotická střední integrální kvadratická chyba nabývá minimální hodnoty, a tedy odhad bude nejlepší ve smyslu AMISE. Užijeme metody matematické analýzy a spočítáme derivaci

$$\frac{d\text{AMISE } \widehat{m}(\cdot, h)}{dh} = -\frac{\sigma^2 V(K)}{nh^2} + h^3 \beta_2^2(K) V(m''),$$

položíme ji rovnu nule a vyjádříme h

$$h_{opt,0,2}^5 = \frac{\sigma^2 V(K)}{n \beta_2^2(K) V(m'')}. \quad (2.10)$$

Poznámka 2.3. Tento výpočet vede k nalezení minima AMISE, protože platí

$$\frac{d^2 \text{AMISE}}{dh^2} > 0.$$

Vztah (2.10) má pouze teoretický charakter, protože hodnota $h_{opt,0,2}$ závisí na neznámých veličinách σ^2 a $m''(x)$, a tedy není užitečná pro praktické účely. Abychom odhadli optimální hodnotu vyhlazovacího parametru, musíme použít metody, které jsou založeny na datech (*data-driven methods*). Nejznámější z těchto metod bude uvedena v dalším odstavci.

Vztah (2.10) pro optimální šířku vyhlazovacího okna ukazuje, že řád konvergence optimální šířky vyhlazovacího okna $h_{opt,0,2}$ závisí na řádu jádra K , tedy pro jádra řádu 2 je $O(n^{-1/5})$. Dosaďme-li (2.10) do vztahu (2.9) pro AMISE, dostaneme

$$\begin{aligned} \text{AMISE}(h_{opt,0,2}) &= \frac{\sigma^2}{n} V(K) \left(\frac{\sigma^2 V(K)}{n \beta_2^2(K) V(m'')} \right)^{-1/5} + \frac{1}{4} \left(\frac{\sigma^2 V(K)}{n \beta_2^2(K) V(m'')} \right)^{4/5} \beta_2^2(K) V(m'') \\ &= (\sigma^2)^{4/5} (V(K))^{4/5} n^{-4/5} (\beta_2^2(K))^{1/5} (V(m''))^{1/5} \\ &\quad + \frac{1}{4} (\sigma^2)^{4/5} (V(K))^{4/5} n^{-4/5} (\beta_2^2(K))^{1/5} (V(m''))^{1/5} \\ &= \frac{5}{4} (\sigma^2 V(K))^{4/5} (\beta_2^2(K) V(m''))^{1/5} n^{-4/5}, \end{aligned} \quad (2.11)$$

tj. $\text{AMISE } \widehat{m}(\cdot, h_{opt,0,2}) = O(n^{-1/5})$.

Poznámka 2.4. Jestliže jádro K náleží do třídy S_{0k} , pak AMISE je tvaru

$$\text{AMISE } \hat{m}(\cdot, h) = \frac{\sigma^2}{nh} V(K) + \frac{1}{(k!)^2} h^{2k} \beta_k^2(K) V(m^{(k)}) \quad (2.12)$$

a pro optimální vyhlazovací parametr h platí

$$h_{opt,0,k}^{2k+1} = \frac{\sigma^2 V(K) (k!)^2}{2kn \beta_k^2(K) V(m^{(k)})}, \quad (2.13)$$

kde $V(m^{(k)}) = \int_0^1 (m^{(k)}(x))^2 dx$, podrobněji např. [7].

Nyní uvedeme důležité lemma, které ukazuje zajímavou vlastnost vyhlazovacího parametru.

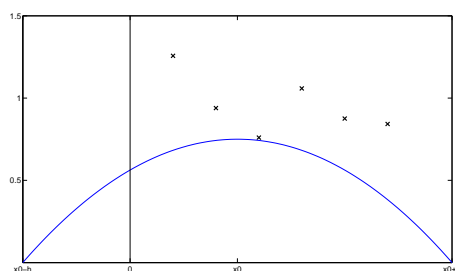
Lemma 2.1. Pro $h_{opt,0,k}$ platí

$$\text{AIV}(h_{opt,0,k}) = 2k \text{AISB}(h_{opt,0,k}).$$

Důkaz. Viz cvičení. □

Lze ukázat, že pro jádra $K \in S_{0k}$ je AMISE $\hat{m}(\cdot, h) = O(n^{-\frac{2k}{2k+1}})$. To znamená, že s rostoucím k se zvyšuje asymptotická rychlost konvergence. Ale není zcela jasné, zda tato zvýšená rychlost konvergence vede již k zlepšení pro konečné rozsahy výběrů, neboť ostatní veličiny se rovněž mění s k . Nevýhodou jader vyšších řádů je fakt, že pro tato jádra je optimální šířka okna větší, což může mít negativní dopad na hraniční efekty [9]. Na druhé straně, chování jádrových odhadů s jádry vyšších řádů je méně citlivé na volbu šířky okna, není-li určena zcela optimálně, neboť křivka AMISE $\hat{m}(\cdot, h)$ je plošší.

Poznámka 2.5. Vyšetřování kvality odhadu obvykle probíhá za předpokladu, že pracujeme s vnitřními body intervalu $[0, 1]$. V hraničních oblastech, tj. v intervalech $[0, h) \cup (1-h, 1]$, je kvalita odhadu ovlivněna negativně skutečností, že jádro K zde nespĺňuje momentové podmínky (1.1). To je způsobeno tím, že blízko krajních bodů nosič jádra K zasahuje do oblasti, kde nejsou žádná data, což zhoršuje odhad – viz obr. 2.10. Hraniční efekty jsou také patrné na obrázcích 2.7(a)



Obrázek 2.10: Hraniční efekt

a 2.14, zejména u pravého okraje intervalu. Problém okrajových efektů lze překonat např. použitím hraničních jader (viz [9]) nebo reflexní metodou (viz [3]).

Ukázkový příklad 2.2. Uvažujme simulovaná data generovaná regresní funkcí $m(x) = \sin^2 \pi x$ na intervalu $x \in [0, 1]$ s chybami $\varepsilon_i \sim N(0; 0,25^2)$. Vypočítejme hodnotu optimálního vyhlazovacího parametru pro odhad s jádrem řádu 2.

Podle vztahu (2.10) potřebujeme spočítat výraz $V(m'')$.

$$\begin{aligned}
 m(x) &= \sin^2 \pi x \\
 m'(x) &= 2 \sin \pi x \cos \pi x \pi = \pi \sin 2\pi x \\
 m''(x) &= \pi \cos 2\pi x 2\pi = 2\pi^2 \cos 2\pi x \\
 V(m'') &= \int_0^1 [m''(x)]^2 dx = \int_0^1 [2\pi^2 \cos 2\pi x]^2 dx \\
 &= 4\pi^4 \int_0^1 \cos^2 2\pi x dx = 4\pi^4 \int_0^1 \frac{1 + \cos 4\pi x}{2} dx \\
 &= 2\pi^4.
 \end{aligned}$$

Výpočet $h_{opt,0,2}$ pro

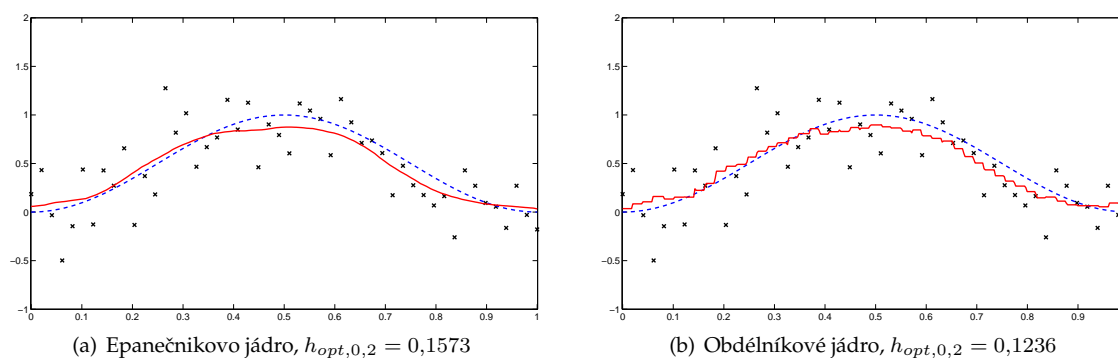
- Epanečnikovo jádro: $V(K) = 3/5, \beta_2(K) = 1/5,$

$$h_{opt,0,2}^5 = \frac{\sigma^2 3/5}{n(1/5)^2 2\pi^4} = \frac{15\sigma^2}{2\pi^4 n},$$

- obdélníkové jádro: $V(K) = 1/2, \beta_2(K) = 1/3,$

$$h_{opt,0,2}^5 = \frac{\sigma^2 1/2}{n(1/3)^2 2\pi^4} = \frac{9\sigma^2}{4\pi^4 n}.$$

Odhady s optimálním vyhlazovacím parametrem pro soubor o velikosti 50 hodnot jsou na obrázku 2.11. (Data jsou v tabulce 7.2.) Vidíme, že odhad s „hladším“ Epanečnikovým jádrem generuje „hladší“ křivku.



Obrázek 2.11: Odhad regresní funkce z ukázkového příkladu 2.2, odhad (červená, plná) a původní funkce (modrá, čárkovaná)

4 Volba jádra

Volba jádra není z asymptotického hlediska podstatná, jak je zřejmé z faktu (2.11). Je vhodné zvolit optimální jádro, které minimalizuje funkcionál $T(K)$, neboť tato jádra jsou spojitá na \mathbb{R} a odhadovaná regresní funkce tak „zdedí“ hladkost jádra. Vhodná jsou jádra třídy S_{02} a S_{04} , lze je vybrat z tabulek 8.1 pro S_{0k} .

5 Volba vyhlazovacího parametru

5.1 Metoda křížového ověřování

Jednou z nejrozšířenějších a nejpoužívanějších metod pro určení optimální hodnoty parametru h je metoda křížového ověřování (*cross-validation method*). Tato metoda je založena na odhadu regresní funkce (2.4), v němž vynecháme i -té pozorování:

$$\hat{m}_{-i}(x_i, h) = \sum_{\substack{\ell=1 \\ \ell \neq i}}^n W_\ell(x_i, h) Y_\ell, \quad i = 1, \dots, n.$$

Funkce křížového ověřování je definována takto

$$CV(h) = \frac{1}{n} \sum_{i=1}^n (\hat{m}_{-i}(x_i, h) - Y_i)^2 \quad (2.14)$$

a odhadem optimální hodnoty vyhlazovacího parametru je bod, v němž nastává minimum této funkce, tj.

$$\hat{h}_{opt,0,k} = h_{CV} = \arg \min_{h \in H_n} CV(h).$$

Hledáme tedy minimum na intervalu $H_n = [a_k n^{-1/(2k+1)}, b_k n^{-1/(2k+1)}]$, jehož tvar plyne ze vztahu (2.13), přičemž a_k, b_k jsou konstanty ($0 < a_k < b_k < \infty$), které ovšem neznáme. A proto pro ekvidistantní body plánu byl na základě zkušeností doporučen interval $[\frac{1}{n}, 2]$.

Poznámka 2.6. Někdy se místo chyby MISE používá průměrná střední kvadratická chyba AMSE (*average mean square error*)

$$AMSE \hat{m}(\cdot, h) = \frac{1}{n} E \sum_{i=1}^n (\hat{m}(x_i, h) - m(x_i))^2$$

Využívá se zejména v případech, kdy není vhodné použít numerické integrování související s chybou, která se vyskytuje v MISE.

Věta 2.2. Pro střední hodnotu funkce $CV(h)$ platí

$$E CV(h) = \underbrace{\frac{1}{n} E \sum_{i=1}^n (\hat{m}(x_i, h) - m(x_i))^2}_{AMSE} + \sigma^2.$$

Důkaz. Funkci křížového ověřování lze rozepsat

$$\begin{aligned} CV(h) &= \frac{1}{n} \sum_{i=1}^n (\hat{m}_{-i}(x_i, h) - m(x_i) - \underbrace{(Y_i - m(x_i))}_{\varepsilon_i})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (\hat{m}_{-i}(x_i, h) - m(x_i))^2 - 2 \frac{1}{n} \sum_{i=1}^n (\hat{m}_{-i}(x_i, h) - m(x_i)) \varepsilon_i + \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \\ &= \frac{1}{n} \sum_{i=1}^n (\hat{m}_{-i}(x_i, h) - m(x_i))^2 - \frac{2}{n} \sum_{i=1}^n \varepsilon_i \left(\sum_{\substack{\ell=1 \\ \ell \neq i}}^n W_\ell(x_i, h) Y_\ell - m(x_i) \right) + \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2. \end{aligned}$$

Střední hodnota $E CV(h)$ je rovna součtu tří veličin. Předpokládejme, že $\hat{m}_{-i}(x_i, h) \approx \hat{m}(x_i, h)$, pak první ze sčítanců je roven přímo AMSE $\hat{m}(\cdot, h)$:

$$\frac{1}{n} E \sum_{i=1}^n (\hat{m}_{-i}(x_i, h) - m(x_i))^2 = AMSE \hat{m}(\cdot, h).$$

Dále víme, že $Y_\ell = m(x_\ell) + \varepsilon_\ell$, a tedy pro druhý sčítanec platí:

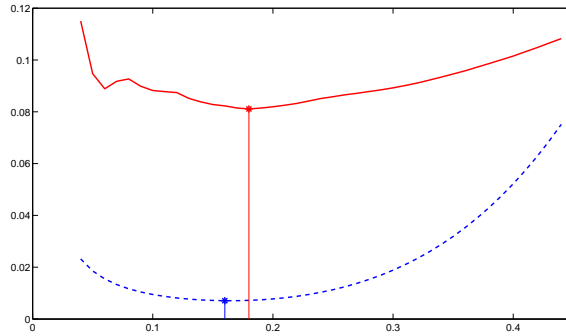
$$\begin{aligned}
& -\frac{2}{n} E \sum_{i=1}^n \varepsilon_i \left(\sum_{\substack{\ell=1 \\ \ell \neq i}}^n W_\ell(x_i, h) Y_\ell - m(x_i) \right) \\
&= -\frac{2}{n} E \sum_{i=1}^n \varepsilon_i \left(\sum_{\substack{\ell=1 \\ \ell \neq i}}^n W_\ell(x_i, h) (m(x_\ell) + \varepsilon_\ell) - m(x_i) \right) \\
&= -\frac{2}{n} E \left[\sum_{i=1}^n \sum_{\substack{\ell=1 \\ \ell \neq i}}^n \varepsilon_i W_\ell(x_i, h) m(x_\ell) + \sum_{i=1}^n \sum_{\substack{\ell=1 \\ \ell \neq i}}^n \varepsilon_i W_\ell(x_i, h) \varepsilon_\ell - \sum_{i=1}^n \varepsilon_i m(x_i) \right] \\
&= -\frac{2}{n} \left[\sum_{i=1}^n \sum_{\substack{\ell=1 \\ \ell \neq i}}^n W_\ell(x_i, h) m(x_\ell) \underbrace{E \varepsilon_i}_{=0} + \sum_{i=1}^n \sum_{\substack{\ell=1 \\ \ell \neq i}}^n W_\ell(x_i, h) \underbrace{E \varepsilon_i \varepsilon_\ell}_{=0} - \sum_{i=1}^n m(x_i) \underbrace{E \varepsilon_i}_{=0} \right] \\
&= 0,
\end{aligned}$$

Stejně jako pro druhý sčítanec, i pro třetí sčítanec využijeme vlastnosti (2.2):

$$\frac{1}{n} E \sum_{i=1}^n \varepsilon_i^2 = \sigma^2.$$

□

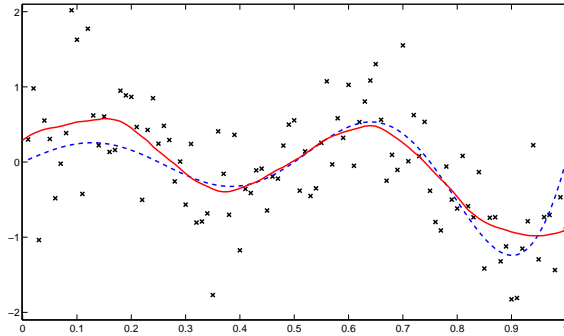
Tento výsledek znamená, že minimalizace $E CV(h)$ odpovídá minimalizaci AMSE. Jestliže tedy předpokládáme, že minimum $CV(h)$ je blízko minima $E CV(h)$, pak tato minimalizace dává dobrou volbu vyhlazovacího parametru – viz ilustrace na obr. 2.12.



Obrázek 2.12: Porovnání minima AMSE (modrá, čárkovaná) a minima funkce křížového ověřování CV (červená, plná) pro simulovaná data z ukázkového příkladu

Příklad 2.3. Použijeme metodu křížového ověřování pro nalezení vyhlazovacího parametru pro data z příkladu 2.1. Při použití Epanečnikova jádra získáme vyhlazovací parametr $h_{CV} = 0,1158$. Na obrázku 2.13 je zobrazen odhad regresní funkce s tímto parametrem.

Kromě metody křížového ověřování se také pro odhad optimálního vyhlazovacího parametru používají metody založené na ASE (average square error), metody plug-in, metody odvozené z Fourierovy transformace a bootstrapové metody (podrobněji např. [3, 6]).



Obrázek 2.13: Simulovaná data (\times) s jádrovým odhadem regresní funkce ($h_{CV} = 0,1158$) (červená, plná) a původní funkcí (modrá, čárkovaná)

6 Automatická procedura

Z dříve uvedených odhadů chyb je zřejmé, že kvalita jádrového odhadu závisí na šířce okna h , na jádře K a na řádu jádra k , což je číslo, které odpovídá předpokládanému počtu derivací v odhadovaném modelu. Je zřejmé, že všechny tyto tři veličiny se vzájemně ovlivňují, a proto je třeba zabývat se jejich volbou současně.

Pro simultánní volbu jádra, optimálního vyhlazovacího parametru a řádu jádra byla navržena automatická procedura (viz [3]), která odhadne všechny parametry tak, aby byla minimalizována AMISE. Procedura byla původně odvozena pro odhad hustoty pravděpodobnosti ([4]), ale lze ji aplikovat i pro odhad regresní funkce. Uvedeme zde její zjednodušenou verzi.

Podle vztahů (2.12) a (2.13) je AMISE tvaru

$$\text{AMISE } \hat{m}(\cdot, h) = \frac{\sigma^2}{nh} V(K) + \frac{1}{(k!)^2} h^{2k} \beta_k^2(K) V(m^{(k)})$$

a $h_{opt,0,k}$ tvaru

$$h_{opt,0,k}^{2k+1} = \frac{\sigma^2 \delta_{0k}^{2k+1} (k!)^2}{2kn V(m^{(k)})}$$

Ze vztahu pro $h_{opt,0,k}$ vypočteme $V(m^{(k)})$ a dosadíme do vztahu pro AMISE. Dále použijeme vztahy

$$\delta_{0k} = \left(\frac{V(K)}{\beta_k^2(K)} \right)^{\frac{1}{2k+1}} \quad T(K) = (\beta_k(K) V^k(K))^{\frac{2}{2k+1}} \quad T(K) = \delta_{0k}^{2k} \beta_k^2(K)$$

a dostaneme vyjádření

$$\text{AMISE } \hat{m}(\cdot, h_{opt,0,k}) = \frac{\sigma^2 (2k+1) \delta_{0k}}{2nkh_{opt,0,k}} T(K),$$

ve kterém jsou parametry K , h a k separovány, což umožňuje vybrat tyto parametry simultánně. Právě tento vztah je základem automatické procedury.

Položme

$$L(k) = \frac{(2k+1) \delta_{0k}}{2nkh_{opt,0,k}} T(K).$$

a množinu vhodných řádů k označme

$$I(k_0) = \left\{ 2j, j = 0, \dots, \left[\frac{k_0}{2} \right] \right\},$$

kde $[z]$ značí celou část čísla z . Procedura pak probíhá v pěti krocích:

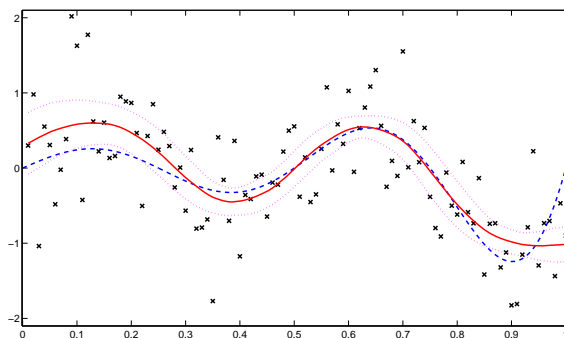
1. Pro $k \in I(k_0)$ najděte optimální jádro $K_{opt,0,k} \in S_{0k}$, které je dáno tabulkou 8.1 a vypočtěte kanonický faktor δ_{0k} .
2. Pro $k \in I(k_0)$ a $K_{opt,0,k} \in S_{0k}$ najděte optimální vyhlazovací parametr $\hat{h}_{opt,0,k}$.
3. Pro $k \in I(k_0)$ vypočtěte hodnotu výběrového kritéria $L(k)$ s využitím hodnot získaných v krocích 1 a 2.
4. Vypočtěte optimální hodnotu řádu \hat{k} , které minimalizuje funkcionál $L(k)$.
5. Použijte parametry z předchozích kroků k získání optimálního jádrového odhadu regresní funkce, tj.

$$\hat{m}(x, \hat{h}_{opt,0,\hat{k}}) = \sum_{i=1}^n W_i(x, \hat{h}_{opt,0,\hat{k}}) Y_i.$$

Příklad 2.4. Aplikace procedury na data z příkladu 2.1. Maximální řád jádra zvolme $k_0 = 8$, tedy množina možných řádů jader je $I(8) = \{0, 2, 4, 6, 8\}$. Pro tyto řády spočítejme hodnoty z kroků 1–3, v kroku 2 jsme použili metodu křížového ověřování pro nalezení optimálního vyhlazovacího parametru $\hat{h}_{opt,0,k}$.

k	$K_{opt,0,k}$	δ_{0k}	$\hat{h}_{opt,0,k}$	$L(k)$
2	$-\frac{3}{4}(x^2 - 1)$	1,7188	0,1158	0,0648
4	$\frac{15}{32}(x^2 - 1)(7x^2 - 3)$	2,0165	0,2446	0,0575
6	$-\frac{105}{256}(x^2 - 1)(33x^4 - 30x^2 + 5)$	2,0834	0,3575	0,0574
8	$\frac{315}{4096}(x^2 - 1)(715x^6 - 1001x^4 + 385x^2 - 35)$	2,1021	0,4543	0,0592

Z tabulky vidíme, že optimální řád jádra je $\hat{k} = 6$. Výsledný odhad je uveden na obrázku 2.14.



Obrázek 2.14: Simulovaná data (\times) s jádrovým odhadem regresní funkce při použití procedury (červená, plná) a skutečnou funkcí (modrá, čárkovaná) společně s 95% intervalem spolehlivosti (růžová, tečkovaná)

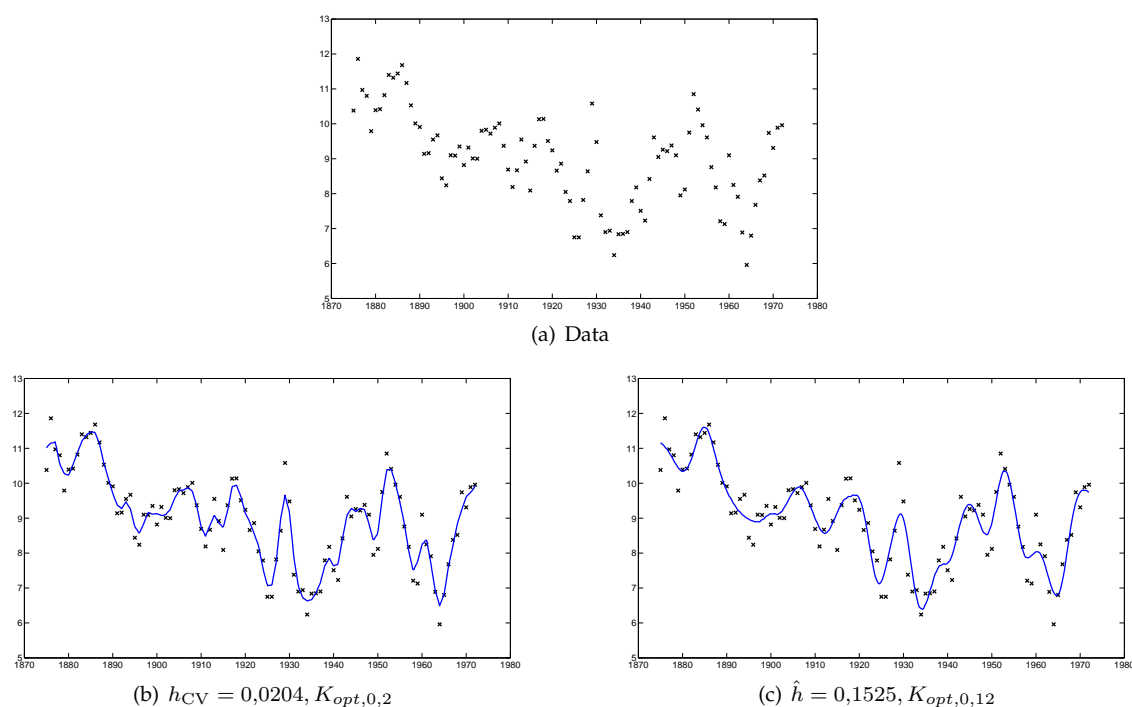
7 Aplikace na reálná data

Nyní se vrátíme k motivačnímu příkladu. Datový soubor obsahuje měření úrovně hladiny vody v Huronském jezeře. Měření byla prováděna ročně, v letech 1875 až 1972, tedy naměřené hodnoty jsou ekvidistantní. Data jsou shrnuta v tabulce 7.8 a na obrázku 2.15(a).

Vyhlazovací parametry jsme odhadli pomocí metody křížového ověřování (za použití Epanečnikova jádra) a také pomocí automatické procedury. Hodnoty vyhlazovacích parametrů jsou následující:

$$h_{CV} = 0,0204 \quad (K_{opt,0,2}), \quad \hat{h} = 0,1525 \quad (K_{opt,0,12}).$$

Výsledná křivka $\hat{m}(x)$ je odhadem funkce $m(x)$ popisující průběh úrovně hladiny v letech 1875 až 1972. Odhady na obrázku 2.15 ukazují, že metoda křížového ověřování spíše podhlazuje.



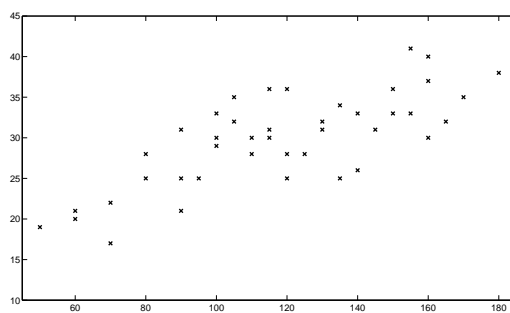
Obrázek 2.15: Úroveň hladiny Huronského jezera – data a odhadnuté regresní funkce, na ose x jsou roky a na ose y je hladina vody ve stopách (snížená o 570 stop – viz poznámka u dat na str. 90)

Druhým datovým souborem jsou měření axiální délky krystalů ledu. Měření byla prováděna v místnosti s konstantní teplotou $-5\text{ }^{\circ}\text{C}$ v časových intervalech 50 až 180 vteřin po přinesení krystalu do místnosti. V tomto případě nejde o ekvidistanční data, protože hodnoty se liší o pět či deset vteřin. Data jsou v tabulce 7.9 a na obrázku 2.16(a). Chceme odhadnout funkci, která vyjadřuje závislost axiální délky krystalů na čase.

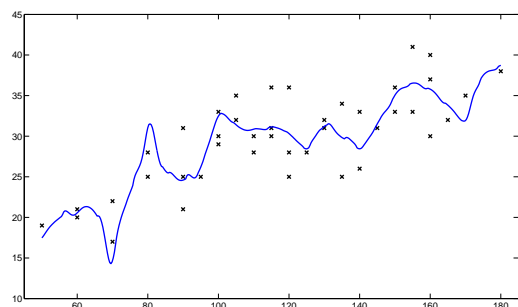
Odhady vyhlazovacích parametrů podle metody křížového ověřování a pomocí automatické procedury:

$$h_{CV} = 0,1865 \quad (K_{opt,0,2}), \quad \hat{h} = 0,8826 \quad (K_{opt,0,10}).$$

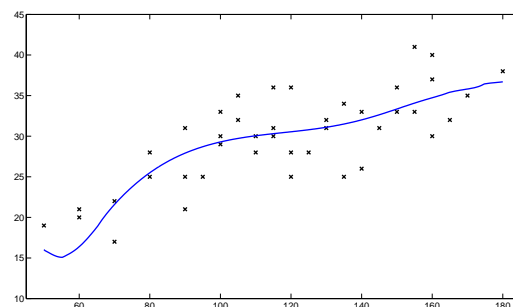
Výsledné odhady regresní funkce jsou na obrázku 2.16. Je vidět, že metoda křížového ověřování dává spíše podhlazený odhad. Na druhou stranu, odhad pomocí procedury se může zdát již přehlázený.



(a) Data



(b) $h_{CV} = 0,1865, K_{opt,0,2}$



(c) $\hat{h} = 0,8826, K_{opt,0,10}$

Obrázek 2.16: Axiální délka krystalů ledu – data a odhadnuté regresní funkce, na ose x je vynesena čas ve vteřinách a na ose y délka krystalu v mikrometrech

Shrnutí
<p>Odhad regresní funkce $m(x)$ v bodě x je tvaru</p> $\hat{m}(x, h) = \sum_{i=1}^n W_i(x, h) Y_i.$ <p>$W_i(x, h)$ závisí na K a h.</p>
<p>Vychýlení (bias) a rozptyl (var) odhadu s jádrem řádu 2 jsou</p> $\text{bias } \hat{m}(x, h) \approx \frac{h^2}{2} \beta_2(K) m''(x), \quad \text{var } \hat{m}(x, h) \approx \frac{\sigma^2}{nh} V(K).$
<p>Asymptotická střední kvadratická chyba jádrového odhadu regresní funkce s jádrem řádu 2 je</p> $\text{AMISE } \hat{m}(\cdot, h) = \frac{\sigma^2}{nh} V(K) + \frac{1}{4} h^4 \beta_2^2(K) V(m'').$
<p>Optimální vyhlazovací parametr vzhledem k AMISE pro $k = 2$ je tvaru</p> $h_{\text{AMISE}}^5 = \frac{\sigma^2 V(K)}{n \beta_2^2(K) V(m'')}$ <p>s řádem konvergence $n^{-1/5}$.</p>
<p>Metoda křížového ověřování pro odhad optimálního vyhlazovacího parametru</p> $\text{CV}(h) = \frac{1}{n} \sum_{i=1}^n (\hat{m}_{-i}(x_i, h) - Y_i)^2 \Rightarrow h_{\text{CV}} = \arg \min \text{CV}(h).$
<p>Automatická procedura pro simultánní volbu optimálního jádra, jeho řádu a vyhlazovacího parametru je dostupná v toolboxu Matlabu.</p>

Výstupy z výukové jednotky

Student

- zná základní typy jádrových odhadů regresní funkce a jejich derivací
- je schopen analyzovat statistické vlastnosti odhadů
- má přehled o metodách pro volbu vyhlazovacího parametru
- se seznámil s automatickou procedurou pro simultánní volbu vyhlazovacího parametru, jádra a jeho řádu
- je schopen analyzovat daný soubor dat a aplikovat uvedenou proceduru na tento soubor
- je schopen použít příslušný toolbox v Matlabu a zkonstruovat odhad regresní funkce

Dodatek

Výpočet integrálu, $t_i = \frac{i}{n}$, $i = 1, \dots, n$,

$$\int_0^1 G(t) dt = \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} G(t) dt$$

s využitím Taylorova rozvoje funkce $G(t)$

$$\begin{aligned} &= \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} (G(t_{i+1}) + (t - t_{i+1})G'(t_{i+1}) + o(1)) dt \\ &= \sum_{i=0}^{n-1} G(t_{i+1}) \underbrace{(t_{i+1} - t_i)}_{=\frac{1}{n}} + \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} (t - t_{i+1})G'(t_{i+1}) dt + o(1) \\ &= \frac{1}{n} \sum_{i=1}^n G(t_i) + \sum_{i=0}^{n-1} \frac{(t_{i+1} - t_i)^2}{2} G'(t_{i+1}) + o(1) \\ &= \frac{1}{n} \sum_{i=1}^n G(t_i) + \frac{1}{2n^2} \sum_{i=0}^{n-1} G'(t_{i+1}) + o(1). \end{aligned}$$

Za předpokladu $|G'(t_{i+1})| \leq M$, pak platí

$$\int_0^1 G(t) dt = \frac{1}{n} \sum_{i=1}^n G(t_i) + O(n^{-1}).$$

Cvičení

1. Odvoďte tvar Priestleyova-Chaova odhadu regresní funkce v bodě x pro obdélníkové jádro.
2. Pro odhad \hat{m}_{NW} dokažte, že „množství“ vyhlazení pomocí jádra K s vyhlazovacím parametrem h je stejné, jako „množství“ vyhlazení jádrem K_δ s parametrem $h^* = h/\delta$, tj.

$$\hat{m}(x, h, K) = \hat{m}(x, h^*, K_\delta).$$

3. Uvažujte funkci $m(x) = \sin^2 \pi x$ a regresní model $Y_i = m(x) + \varepsilon_i$, $i = 1, \dots, 100$, kde $\varepsilon_i \sim N(0; 0,25)$. Vypočtete hodnotu $h_{opt,0,4}$ pro odhad s jádrem $K(x) = \frac{15}{32}(x^2 - 1)(7x^2 - 3)$. Pomůcka: $V(m^{(4)}) = 32\pi^8$.
4. Vypočtete h_{CV} pro data z ukázkového příkladu 2.2 a porovnejte s h_{opt} .
5. Dokažte vztah (2.11) pro obecné k , tj. že platí

$$AMISE(h_{opt,0,k}) = n^{-2k/(2k+1)} \left(\sigma^2 V(K) \right)^{2k/(2k+1)} \left(\beta_k^2(K) (k!)^{-2} V(m^{(k)}) \right)^{1/(2k+1)} \frac{2k+1}{(2k)^{\frac{2k}{2k+1}}}.$$

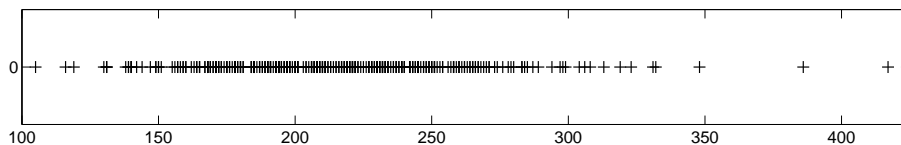
6. Dokažte vztah z lemmatu 2.1.
7. Aplikujte metodu křížového ověřování a automatickou proceduru na simulovaná data z ukázkového příkladu 2.2.

Kapitola 3

Jádrové odhady hustoty

1 Motivace

Velká část pozorování a měření prováděných v biologii i jiných vědách má výsledek vyjádřený reálným číslem. Tato čísla jsou hodnotami nějaké reálné náhodné veličiny. Jako příklad můžeme vzít měření obsahu cholesterolu v krevní plazmě pacientů. Jde o soubor 371 měření – hodnoty jsou na obrázku 3.1 – která byla provedena u pacientů, kteří si stěžovali na bolest na hrudi. Data pocházejí z rozsáhlé studie, v níž autoři zkoumali vliv lipidů a jiných látek na nemoci srdce [10, 11].



Obrázek 3.1: Měření cholesterolu v krevní plazmě 371 pacientů

Chceme zjistit, jak často se v populaci vyskytuje dané množství cholesterolu v krvi? Případně, jaká je pravděpodobnost, že pacient bude mít zvýšený obsah cholesterolu? Rozdělení pravděpodobnosti je popsáno reálnou funkcí, která se nazývá hustota pravděpodobnosti a značí se f . Hustota pravděpodobnosti je základním pojmem ve statistice [1, 2]. Funkce $f(x)$ hustotou spojitě náhodné veličiny, jestliže platí

- $f(x)$ je nezápornou funkcí pro všechna x ,
- $\int f(x) dx = 1$.

Odhadem hustoty rozumíme rekonstrukci hustoty z množiny naměřených dat. Tato rekonstrukce může poskytnout důležité informace o dané množině dat. Předpokládejme, že máme k dispozici nezávislé náhodné proměnné X_1, \dots, X_n , které mají tutéž spojitou hustotu f . Můžeme předpokládat, že neznámá hustota patří do třídy hustot, které závisejí na nějakých parametrech. Pro odhad hledané hustoty je tedy třeba odhadnout tyto parametry. Tento přístup se nazývá parametrický.

My se zaměříme na neparametrický přístup, ve kterém se předpokládá pouze jistá hladkost odhadované hustoty (tj. dostatečný počet spojitých derivací).

2 Základní typy neparametrických odhadů

Nejstarším neparametrickým odhadem hustoty je *histogram* [12, 11, 14]. Histogram zobrazuje relativní četnosti třídicích intervalů jako plochy obdélníků sestrojených nad těmito intervaly. Pak

definujeme odhad hustoty četnosti

$$\hat{f}(x, h) = \frac{1}{nh} (\text{počet } X_i \text{ ve stejném intervalu jako } x),$$

kde h značí šířku třídicích intervalů (obvykle se volí stejná šířka pro všechny intervaly).

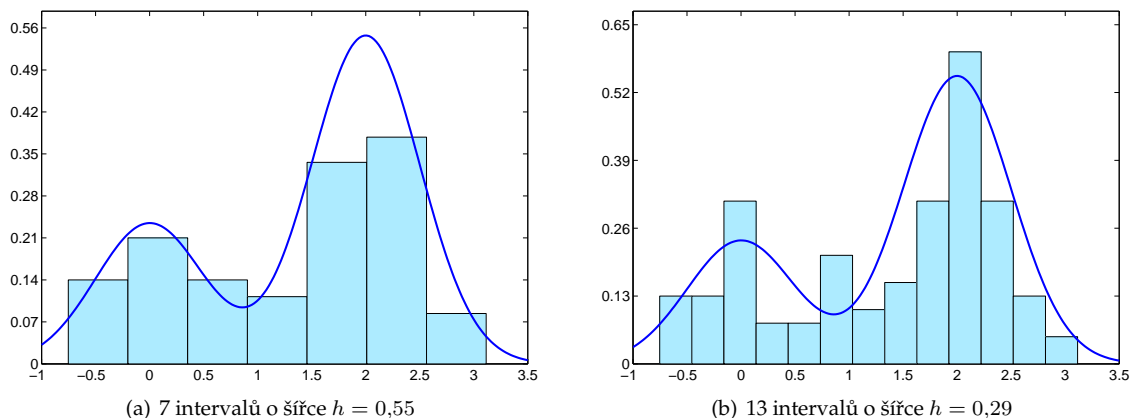
Nevýhody histogramu:

- Histogram je citlivý na počet tříd a jejich šířku.
- Histogram je schodovitá funkce, ale přitom předpokládáme, že neznámá hustota je spojitá.

Příklad 3.1. Mějme dán datový soubor generovaných ze směsi dvou normálních rozdělení $N(0; 0,25)$ a $N(2; 0,25)$ s hustotou

$$f(x) = 0,3 \frac{1}{\sqrt{0,5\pi}} e^{-\frac{x^2}{0,5}} + 0,7 \frac{1}{\sqrt{0,5\pi}} e^{-\frac{(x-2)^2}{0,5}},$$

který má rozsah $n = 100$. (Data jsou v tabulce 7.3.) Z obrázku 3.2 je patrné, že histogram nevystihuje korektně hustotu pravděpodobnosti dat.



Obrázek 3.2: Histogramy s různými počty třídicích intervalů

Výše uvedené problémy lze odstranit použitím jádrových odhadů. Jádrový odhad hustoty f v bodě $x \in \mathbb{R}$ je definovaný vztahem [14].

$$\hat{f}(x, h) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i), \quad (3.1)$$

$K \in S_{0k}$ a h je vyhlazovací parametr nebo také šířka vyhlazovacího okna.

Jádrový odhad hustoty závisí na třech parametrech: jádře, které hraje roli vahové funkce, vyhlazovacím parametru, který řídí hladkost odhadu, a na řádu jádra, který odpovídá předpokládanému počtu derivací neznámé hustoty.

Popíšeme konstrukci jádrového odhadu:

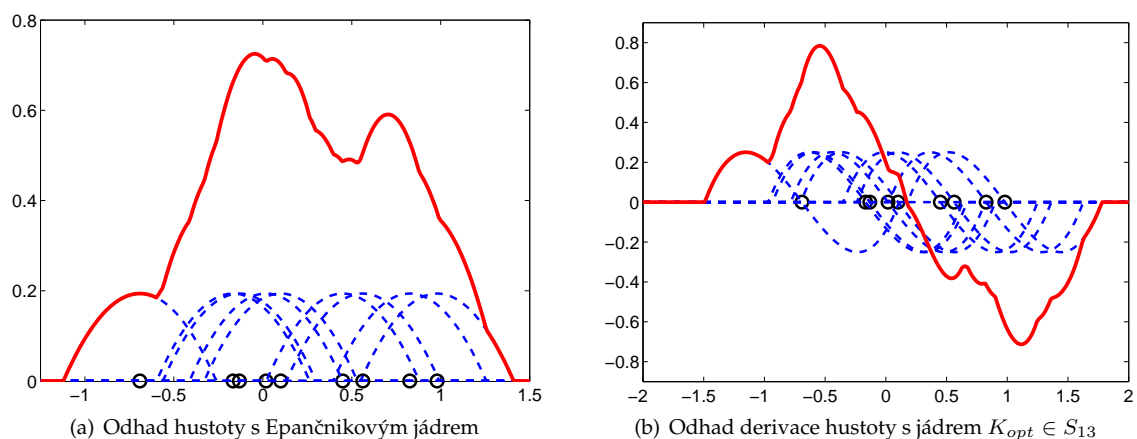
V každém bodě X_i sestrojíme jádro K_h a odhad v bodě x je průměr n hodnot jader v tomto bodě. Na obrázku 3.3(a) jsou čárkovaně zobrazena jednotlivá Epančnikova jádra v bodech X_i (kroužky) a plnou čarou je pak zobrazen odhad hustoty.

OTÁZKA. Popište konstrukci odhadu s obdélníkovým jádrem. Jak bude tento odhad vypadat?

Nyní uvedeme ještě vztah pro jádrový odhad ν -té derivace hustoty. Budeme předpokládat, že $0 \leq \nu \leq k - 2$ a k, ν jsou stejné parity. Pak

$$\hat{f}^{(\nu)}(x, h) = \frac{1}{nh^{\nu+1}} \sum_{i=1}^n K^{(\nu)}\left(\frac{x - X_i}{h}\right), \quad K^{(\nu)} \in S_{\nu k}. \quad (3.2)$$

Konstrukce jádrového odhadu derivace je stejná jako konstrukce odhadu hustoty. V každém bodě X_i sestrojíme jádro K_h ze třídy $S_{\nu k}^1$ a odhad v bodě x je průměrem hodnot jader v tomto bodě. Na obrázku 3.3(b) je zobrazen odhad první derivace hustoty pro soubor o devíti pozorování a bylo zde použito jádro $K(x) = \frac{15}{4}x(x^2 - 1)$.



Obrázek 3.3: Konstrukce jádrového odhadu hustoty a její derivace

3 Statistické vlastnosti jádrových odhadů hustoty

Stejně jako u jádrových odhadů regresní funkce lze kvalitu jádrového odhadu hustoty popsat lokálně pomocí střední kvadratické chyby.

Věta 3.1. *Střední kvadratická chyba je tvaru*

$$\begin{aligned} \text{MSE } \hat{f}(x, h) &= E(\hat{f}(x, h) - f(x))^2 \\ &= \underbrace{\frac{1}{n}((K_h^2 * f)(x) - (K_h * f)^2(x))}_{\text{var}} + \underbrace{((K_h * f)(x) - f(x))^2}_{\text{bias}}. \end{aligned}$$

Důkaz. Spočítejme střední hodnotu odhadu $\hat{f}(x, h)$

$$E\hat{f}(x, h) = E\frac{1}{n} \sum_{i=1}^n K_h(x - X_i) = EK_h(x - X) = \int K_h(x - y)f(y) dy = (K_h * f)(x).$$

Vychýlení (bias) pak bude mít tvar

$$\text{bias } \hat{f}(x, h) = E\hat{f}(x, h) - f(x) = (K_h * f)(x) - f(x).$$

¹Jádra viz tabulku 1.2.

Dále upravíme vztah pro rozptyl

$$\begin{aligned}
\text{var } \hat{f}(x, h) &= \text{var } \frac{1}{n} \sum_{i=1}^n K_h(x - X_i) = \frac{1}{n^2} \text{var } \sum_{i=1}^n K_h(x - X_i) = \frac{1}{n} \text{var } K_h(x - X) \\
&= \frac{1}{n} EK_h^2(x - X) - \frac{1}{n} (EK_h(x - X))^2 \\
&= \frac{1}{n} \int K_h^2(x - y) f(y) dy - \frac{1}{n} ((K_h * f)(x))^2 \\
&= \frac{1}{n} ((K_h^2 * f)(x) - (K_h * f)^2(x)).
\end{aligned}$$

□

Důsledek. *Střední integrální kvadratická chyba nabývá tvaru*

$$\begin{aligned}
\text{MISE } \hat{f}(\cdot, h) &= \int \text{MSE } \hat{f}(\cdot, h) dx \\
&= \frac{1}{n} \left(\int (K_h^2 * f)(x) dx - \int (K_h * f)^2(x) dx \right) + \int ((K_h * f)(x) - f(x))^2 dx.
\end{aligned}$$

Podobně jako u odhadu regresní funkce můžeme použít globální pohled na kvalitu odhadu, a to pomocí střední integrální kvadratické chyby (MISE) a jejího asymptotického tvaru (AMISE).

Věta 3.2. *Nechť funkce f má spojitě derivace až do řádu k_0 (tj. $f \in C^{k_0}$) pro $0 < k \leq k_0$, $K \in S_{0k}$ a $\int (f^{(k)}(x))^2 dx < \infty$, dále předpokládejme $h \rightarrow 0$ a $nh \rightarrow \infty$ pro $n \rightarrow \infty$. Pak platí*

$$\text{MISE } \hat{f}(\cdot, h) = \frac{V(K)}{nh} + \frac{1}{(k!)^2} h^{2k} \beta_k^2(K) V(f^{(k)}) + o(h^{2k} + (nh)^{-1}),$$

kde $V(f^{(k)}) = \int (f^{(k)}(x))^2 dx$.

Důkaz. Nejprve vypočteme střední hodnotu

$$\begin{aligned}
E\hat{f}(x, h) &= (K_h * f)(x) = \int K_h(x - y) f(y) dy = \int \frac{1}{h} K\left(\frac{x - y}{h}\right) f(y) dy \\
&= \int K(z) f(x - hz) dz
\end{aligned}$$

dále použijeme Taylorův rozvoj: $f(x - hz) = f(x) - f'(x)hz + \dots + \frac{(-1)^k}{k!} f^{(k)}(x) h^k z^k + o(h^k)$

$$\begin{aligned}
&= \int K(z) [f(x) - f'(x)hz + \dots + \frac{(-1)^k}{k!} f^{(k)}(x) h^k z^k + o(h^k)] dz \\
&= f(x) \underbrace{\int K(z) dz}_{=1} - f'(x)h \underbrace{\int zK(z) dz}_{=0} + \dots + \frac{(-1)^k}{k!} f^{(k)}(x) h^k \underbrace{\int z^k K(z) dz}_{=\beta_k(K)} + o(h^k) \\
&= f(x) + \frac{(-1)^k f^{(k)}(x)}{k!} h^k \beta_k(K) + o(h^k).
\end{aligned}$$

Tedy vychýlení odhadu je tvaru

$$\text{bias } \hat{f}(\cdot, h) = E\hat{f}(x, h) - f(x) = \underbrace{\frac{(-1)^k f^{(k)}(x)}{k!} h^k \beta_k(K)}_{=o(1)} + o(h^k),$$

a tedy

$$E\hat{f}(x, h) = f(x) + o(1).$$

Nyní dokážeme vztah pro rozptyl. Víme, že

$$\text{var } \hat{f}(x, h) = \frac{1}{n} \left((K_h^2 * f)(x) - \underbrace{(K_h * f)^2(x)}_{(f(x)+o(1))^2} \right)$$

a dále počítáme

$$\begin{aligned} &= \frac{1}{n} \int \frac{1}{h^2} K^2\left(\frac{x-y}{h}\right) f(y) dy - \frac{1}{n} (f(x) + o(1))^2 \\ &= \frac{1}{nh} \int K^2(z) \underbrace{f(x-hz)}_{=f(x)+o(1)} dz - \frac{1}{n} (f(x) + o(1))^2 \\ &= \frac{1}{nh} \int K^2(z) (f(x) + o(1)) dz - \frac{1}{n} (f(x) + o(1))^2 \\ &= \frac{f(x)}{nh} \int K^2(z) dz + \underbrace{\frac{o(1)}{nh} \int K^2(z) dz}_{o((nh)^{-1})} - \frac{1}{n} (f(x) + o(1))^2 \\ &= \frac{f(x)}{nh} \int K^2(z) dz + o((nh)^{-1}). \end{aligned}$$

Tedy

$$\begin{aligned} \text{MSE } \hat{f}(x, h) &= \frac{f(x)}{nh} \int K^2(z) dz + o((nh)^{-1}) + \left(\frac{(-1)^k f^{(k)}(x)}{k!} h^k \beta_k(K) + o(h^k) \right)^2 \\ &= \frac{f(x)}{nh} \int K^2(z) dz + o((nh)^{-1}) + f^{(k)}(x) \frac{h^{2k}}{(k!)^2} \beta_k^2(K) \\ &\quad + \underbrace{2 \frac{(-1)^k f^{(k)}(x)}{k!} h^k \beta_k(K) (h^k)}_{=o(h^{2k})} + o(h^{2k}) \end{aligned}$$

a pak využijeme faktu, že $\int f(x) dx = 1$

$$\begin{aligned} \text{MISE } \hat{f}(\cdot, h) &= \int \text{MSE } \hat{f}(x, h) dx \\ &= \frac{V(K)}{nh} + \frac{1}{(k!)^2} h^{2k} \beta_k^2(K) V(f^{(k)}) + o(h^{2k} + (nh)^{-1}). \end{aligned}$$

□

Důsledek. Necht $h \rightarrow 0$, $nh \rightarrow \infty$ pro $n \rightarrow \infty$, pak \hat{f} je konzistentním odhadem f , tj. $E\hat{f} \rightarrow f$ a $\text{var } \hat{f} \rightarrow 0$.

Stejně jako u odhadu regresní funkce má význam asymptotická integrální střední kvadratická chyba AMISE $\hat{f}(\cdot, h)$:

$$\text{MISE } \hat{f}(\cdot, h) = \text{AMISE } \hat{f}(\cdot, h) + o(h^{2k} + (nh)^{-1}),$$

kde AMISE je tvaru

$$\text{AMISE } \hat{f}(\cdot, h) = \frac{V(K)}{nh} + \frac{1}{(k!)^2} h^{2k} \beta_k^2(K) V(f^{(k)}). \quad (3.3)$$

V dalších částech textu budeme využívat označení jednotlivých částí chyby AMISE, která je součtem asymptotického tvaru integrálu rozptylu AIV (*asymptotic integrated variance*) a asymptotického tvaru integrálu druhé mocniny vychýlení AISB (*asymptotic integrated squared bias*):

$$\text{AIV}(h) = \frac{V(K)}{nh}, \quad \text{AISB}(h) = \frac{1}{(k!)^2} h^{2k} \beta_k^2(K) V(f^{(k)}),$$

tedy

$$\text{AMISE } \hat{f}(\cdot, h) = \text{AIV}(h) + \text{AISB}(h).$$

Užitím vztahů $T(K) = (V(K)^k \beta_k(K))^{2/(2k+1)}$ a $\delta_{0k}^{2k+1} = \frac{V(K)}{\beta_k^2(K)}$ pro $K \in S_{0k}$ lze AMISE zapsat ve tvaru

$$\text{AMISE } \hat{f}(\cdot, h) = T(K) \left(\frac{\delta_{0k}}{nh} + \frac{h^{2k} V(f^{(k)})}{\delta_{0k}^{2k} (k!)^2} \right). \quad (3.4)$$

Důkaz viz cvičení.

Odtud je zřejmé, že vyhlazovací parametr, pro nějž AMISE nabývá minimální hodnoty, je dán vztahem

$$h_{opt,0,k}^{2k+1} = \frac{\delta_{0k}^{2k+1} (k!)^2}{2knV(f^{(k)})}, \quad (3.5)$$

tj. $h_{opt,0,k} = O(n^{-1/(2k+1)})$.

Vypočteme hodnotu AMISE při dosažení optimálního parametru $h_{opt,0,k}$:

$$\text{AMISE } \hat{f}(\cdot, h_{opt,0,k}) = T(K) V(f^{(k)})^{1/(2k+1)} n^{-2k/(2k+1)} \frac{2k+1}{(2k(k!)^2)^{1/(2k+1)}}, \quad (3.6)$$

tj. $\text{AMISE } \hat{f}(\cdot, h_{opt,0,k}) = O(n^{-2k/(2k+1)})$.

I v tomto případě, podobně jako u odhadu regresní funkce, platí vztah mezi asymptotickým rozptylem $\text{AIV}(h)$ a vychýlením $\text{AISB}(h)$:

$$\text{AIV}(h_{opt,0,k}) = 2k \text{AISB}(h_{opt,0,k}). \quad (3.7)$$

Nyní uvedeme zajímavou vlastnost vyhlazovacího parametru.

Poznámka 3.1. Nechť $K \in S_{02}$. Pak optimální hodnota vyhlazovacího parametru je

$$h_{opt,0,2}^5 = \frac{\delta_{02}^5}{nV(f'')}.$$

Počítejme derivace $\text{AMISE } \hat{f}(\cdot, h)$ dané rovnicí (3.3) pro $k = 2$

$$\begin{aligned} \frac{d^2 \text{AMISE } \hat{f}(\cdot, h)}{dh^2} &= \frac{2V(K)}{nh^3} + 3h^2 \beta_2^2(K) V(f'') \\ \frac{d^3 \text{AMISE } \hat{f}(\cdot, h)}{dh^3} &= \frac{-6V(K)}{nh^4} + 6h \beta_2^2(K) V(f''). \end{aligned}$$

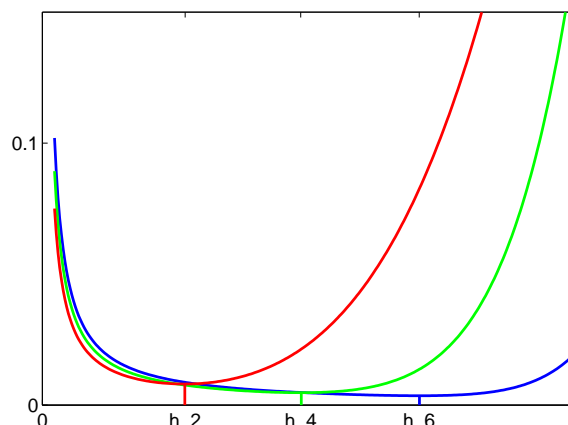
Řešením rovnice $d^3 \text{AMISE } \hat{f}(\cdot, h) / dh^3 = 0$ je

$$h^5 = \frac{V(K)}{n\beta_2^2(K)V(f'')} = \frac{\delta_{02}^5}{nV(f'')} = h_{opt,0,2}^5,$$

tj. $h_{opt,0,2}$ také realizuje minimum $d^2 \text{AMISE } \hat{f}(\cdot, h) / dh^2$.

Lze ukázat, že

$$\left. \frac{d^2 \text{AMISE}(\hat{f}(\cdot, h))}{dh^2} \right|_{h=h_{opt,0,k}} = O(n^{-\frac{2k-2}{2k+1}})$$



Obrázek 3.4: AMISE $\hat{f}(\cdot, h)$ pro jádra vyšších řádů s vyznačenými minimálními hodnotami pro jádra řádu 2, 4, 6

a to znamená, že pro jádra vyšších řádů je minimum AMISE $\hat{f}(\cdot, h)$ plošší a tedy volba h blízká optimální hodnotě $h_{opt,0,k}$ nevede k velkému růstu AMISE $\hat{f}(\cdot, h)$. Na obrázku 3.4 jsou zobrazeny body minima funkce AMISE $\hat{f}(\cdot, h)$ pro hustotu normálního rozdělení $N(0; 1)$ se sto prvky.

Vztah pro optimální hodnotu vyhlazovacího parametru poskytuje informaci, že asymptoticky je $h = O(n^{-1/(2k+1)})$. Ale vztah má pouze teoretický charakter, protože optimální parametr závisí na neznámé hustotě f . Je zde tedy opět problém s volbou tohoto parametru. Metodám pro odhad vyhlazovacího parametru je věnován odstavec 5.

Poznámka 3.2. Z předchozích úvah je zřejmé, že množina přípustných hodnot vyhlazovacích parametrů je dána vztahem

$$H_n = [a_k n^{-1/(2k+1)}, b_k n^{-1/(2k+1)}],$$

kde a_k, b_k jsou konstanty, $0 < a_k < b_k < \infty$.

Ukázkový příklad 3.2. Máme k dispozici data, která pocházejí z rozdělení s hustotou $f(x) = \frac{35}{32}(1-x^2)^3$ pro $x \in [-1, 1]$. Vypočítejme hodnotu optimálního vyhlazovacího parametru pro odhad s jádrem řádu 2.

Podle vztahu (3.5) potřebujeme spočítat výraz $V(f'')$.

$$f(x) = \frac{35}{32}(1-x^2)^3$$

$$V(f'') = \int_{-1}^1 [f''(x)]^2 dx = 35$$

Výpočet $h_{opt,0,2}$ pro Epanečnikovo jádro: $V(K) = 3/5, \beta_2(K) = 1/5$, tedy $\delta_{02}^5 = \frac{V(K)}{\beta_2^2(K)} = 15$

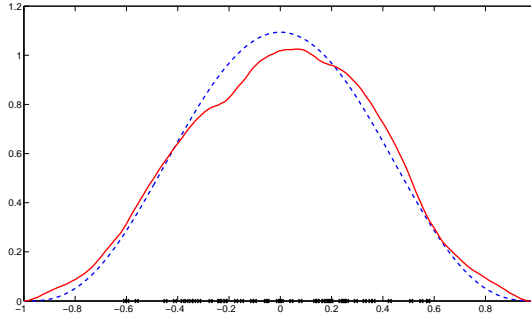
$$h_{opt,0,2}^5 = \frac{15(2!)^2}{2 \cdot 2 \cdot n \cdot 35} = \frac{3}{7n}.$$

Tedy pro soubor 50 hodnot bude $h_{opt,0,2} = 0,8441 \cdot 50^{-1/5} = 0,3860$. Odhad s optimálním vyhlazovacím parametrem pro tento datový soubor (viz tabulku 7.4) je na obrázku 3.5.

3.1 Odhad derivace hustoty

Pojednáme nyní stručně o statistických vlastnostech jádrových odhadů derivace hustoty. Připomeňme, že jádrový odhad derivace hustoty je dán vztahem (3.2), tj.

$$\hat{f}^{(\nu)}(x, h) = \frac{1}{nh^{\nu+1}} \sum_{i=1}^n K^{(\nu)}\left(\frac{x-X_i}{h}\right), \quad K^{(\nu)} \in S_{\nu k}.$$



Obrázek 3.5: Odhad hustoty z ukázkového příkladu 3.2, odhad (červená, plná) a původní funkce (modrá, čárkovaná) při použití Epanečnikova jádra a $h_{opt,0,2} = 0,3860$

Předpokládejme nyní, že platí $0 \leq \nu \leq k - 2$, ν a k mají stejnou paritu, $\lim_{n \rightarrow \infty} h = 0$, $\lim_{n \rightarrow \infty} nh^{2\nu+1} = \infty$, $f \in C^{k_0}$ ($k \leq k_0$) a $V(f^{(k)}) = \int (f^{(k)}(x))^2 dx < \infty$. Pak lze ukázat, že asymptotická střední kvadratická chyba AMISE $\hat{f}^{(\nu)}(\cdot, h)$ je tvaru

$$\text{AMISE } \hat{f}^{(\nu)}(\cdot, h) = \frac{V(K^{(\nu)})}{nh^{2\nu+1}} + \frac{1}{(k!)^2} h^{2(k-\nu)} \beta_k^2(K^{(\nu)}) V(f^{(k)}).$$

Důkaz je založen na použití vhodného Taylorova rozvoje hustoty f , podobně jako u důkazu tvaru AMISE u odhadu hustoty [3].

Optimální hodnota vyhlazovacího parametru je dána vztahem

$$h_{opt,\nu,k}^{2k+1} = \frac{\delta_{\nu k}^{2k+1} (2\nu + 1)(k!)^2}{2n(k - \nu)V(f^{(k)})}, \quad \text{kde } \delta_{\nu k}^{2k+1} = \frac{V(K^{(\nu)})}{\beta_k^2(K^{(\nu)})}. \quad (3.8)$$

Tento vzorec umožňuje výpočet optimálního vyhlazovacího parametru pro $\hat{f}^{(\nu)}$ pomocí $h_{opt,0,k}$ a $h_{opt,1,k}$. Předpokládejme nejdříve, že ν a k jsou sudá čísla. Pak

$$\frac{h_{opt,\nu,k}}{h_{opt,0,k}} = \left(\frac{(2\nu + 1)k}{k - \nu} \right)^{1/(2k+1)} \frac{\delta_{\nu k}}{\delta_{0k}}. \quad (3.9)$$

Pro ν a k lichá platí

$$\frac{h_{opt,\nu,k}}{h_{opt,1,k}} = \left(\frac{(2\nu + 1)(k - 1)}{3(k - \nu)} \right)^{1/(2k+1)} \frac{\delta_{\nu k}}{\delta_{1k}}. \quad (3.10)$$

Speciálně pro $\nu = 2, k = 4$ dostáváme velmi užitečný vztah

$$h_{opt,2,4} = 10^{1/9} \frac{\delta_{24}}{\delta_{04}} h_{opt,0,4}, \quad (3.11)$$

přičemž

$$K_{opt,0,4}(x) = \frac{15}{32}(x^2 - 1)(7x^2 - 3), \quad \delta_{04} = 2,0165,$$

$$K^{(2)}(x) = K_{opt,2,4}(x) = \frac{105}{16}(1 - x^2)(5x^2 - 1), \quad \delta_{24} = 1,3925.$$

4 Volba jádra

Volba jádra není z asymptotického hlediska podstatná, jak je zřejmé z faktu (3.6). Je vhodné zvolit optimální jádro, které minimalizuje funkcionál $T(K)$, neboť tato jádra jsou spojitá na \mathbb{R} a hladkost jádra také „zdědí“ odhadovaná hustota.

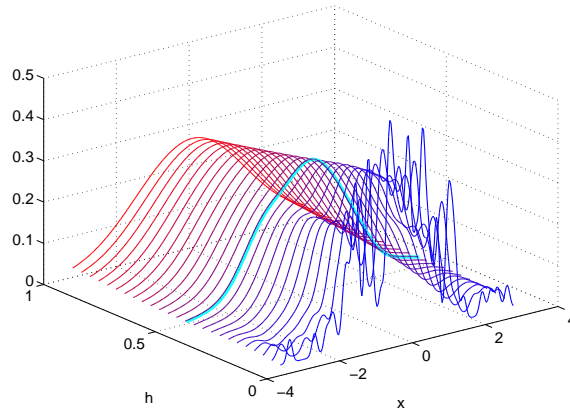
Poznámka 3.3. Při výpočtech se používá nejčastěji Epanečnikovo jádro.

5 Volba vyhlazovacího parametru

Volba vyhlazovacího parametru pro jádrový odhad hustoty je, stejně jako u regrese, zásadním problémem.

I když tomuto problému byla a je věnována značná pozornost, doposud neexistuje univerzální přístup k řešení tohoto problému. Nejjednodušší metoda je „okometrická“. Je účelné „nakreslit“ několik křivek s různými vyhlazovacími parametry dříve než uijeme nějakou automatickou proceduru.

Je třeba zdůraznit, že z hlediska analýzy všechny volby vyhlazovacího parametru vedou k užitečnému odhadu hustoty. Velká šířka okna charakterizuje globální strukturu hustoty a naopak malá šířka odhaluje lokální strukturu, která může nebo nemusí být přítomná v přesné hustotě. Tuto myšlenku ilustruje obrázek 3.6, na němž jsou zobrazeny odhady pro simulovaná data ($f(x) \sim N(0; 1)$, $n = 100$) s hodnotami vyhlazovacího parametru z intervalu $[0,05, 1]$. Jednotlivé odhady hustoty příslušející těmto hodnotám jsou znázorněny tenkými čarami. Silná křivka znázorňuje odhad s optimální hodnotou $h = 0,4217$. Třída těchto odhadů ukazuje široký rozsah vyhlazení od podhlazení (modré křivky) až k přehlazení (červené křivky).



Obrázek 3.6: Volba vyhlazovacího parametru

V dalších úvahách bude užitečná následující definice:

Definice 3.1. Nechť \hat{h} je odhad $h_{opt,0,k}$. Řekneme, že \hat{h} konverguje k $h_{opt,0,k}$ s relativní rychlostí $n^{-\alpha}$, jestliže

$$\frac{\hat{h} - h_{opt,0,k}}{h_{opt,0,k}} = O(n^{-\alpha}).$$

5.1 Metoda referenční hustoty

Nejčastěji se pro odhad neznámé veličiny $V(f^{(k)})$ (viz rovnice (3.3)) používá parametrické třídy hustot. Jednou z možností je použít standardní normální hustotu f s rozptylem σ^2 , tj. předpokládáme, že

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}.$$

V tomto případě je odhad optimálního vyhlazovacího parametru tvaru pro $K \in S_{0k}$

$$h_{REF} = \left(\frac{2^{2k}(k!)^3 \sqrt{\pi}}{(2k)!k} \right)^{\frac{1}{2k+1}} \delta_{0k} \sigma n^{-\frac{1}{2k+1}}, \quad (3.12)$$

Je třeba ještě odhadnout směrodatnou odchylku σ . To lze dvěma způsoby:

$$\hat{\sigma}_{SD} = \left(\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \right)^{1/2}, \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad (3.13)$$

$$\hat{\sigma}_{IQR} = \frac{X_{[3n/4]} - X_{[n/4]}}{\Phi^{-1}(\frac{3}{4}) - \Phi^{-1}(\frac{1}{4})}, \quad (3.14)$$

kde Φ^{-1} je standardní normální kvantilová funkce a číslo $X_{[3n/4]}$, respektive $X_{[n/4]}$, je horní, respektive dolní výběrový kvartil. Je vhodné volit $\min\{\hat{\sigma}_{SD}, \hat{\sigma}_{IQR}\}$.

Poznámka 3.4. Pokud za jádro K zvolíme Gaussovo jádro ($k = 2$)

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

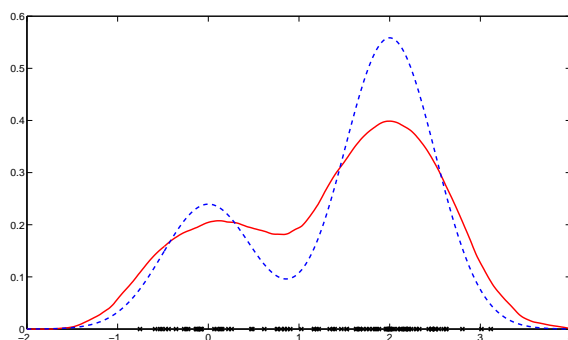
pak dostaneme jednoduchý vztah [11, 12]

$$h_{REF} = \left(\frac{4}{3n} \right)^{1/5} \sigma. \quad (3.15)$$

Příklad 3.3. Použijme odhad vyhlazovacího parametru pro data z příkladu 3.1 metodou referenční hustoty. Pro Epanečnikovo jádro, které je řádu $k = 2$, se vztah (3.12) zjednoduší na tvar

$$h_{REF} = \left(\frac{8\sqrt{\pi}}{3} \right)^{1/5} \delta_{02} \hat{\sigma} n^{-1/5}.$$

Dále odhadneme směrodatnou odchylku: $\hat{\sigma}_{SD} = 1,0325$, $\hat{\sigma}_{IQR} = 1,3344$, tedy $\hat{\sigma} = 1,0325$. Po dosažení počtu prvků $n = 100$ a parametru $\delta_{02} = 1,7188$ získáme hodnotu vyhlazovacího parametru pro odhad hustoty $h_{REF} = 0,9639$. Na obrázku 3.7 je vykreslen odhad hustoty s tímto parametrem.



Obrázek 3.7: Odhad hustoty s $h_{REF} = 0,9639$, odhad (červená, plná), původní funkce (modrá, čárkovaná)

OTÁZKA. Jak hodnotu bude mít odhad vyhlazovacího parametru pomocí metody referenční hustoty pro hustotu z ukázkového příkladu 3.2?

5.2 Metoda maximálního vyhlazení

Princip maximálního vyhlazení (*maximal smoothing*) – MS (nebo přehlazení) znamená, že vybereme největší stupeň přehlazení kompatibilní s odhadovanou hustotou. Získáme tak horní hranici pro odhad optimální šířky vyhlazovacího okna. Tato hodnota pak může sloužit jako počáteční aproximace pro některé z dalších metod. Princip spočívá v tom, že hledáme hustotu, pro kterou $V(f^{(k)})$ nabývá minimální hodnoty, a tedy vztah pro $h_{opt,0,k}$ nabývá maximální hodnoty.

Věta 3.3 (Terrell 1990). Mezi všemi hustotami f s nosičem $[-1, 1]$ má hustota rozdělení $Beta(k+2, k+2)$

$$g_k(x) = \begin{cases} \frac{(2k+3)!}{((k+1)!)^2 2^{2k+3}} (1-x^2)^{k+1} & |x| \leq 1, \\ 0 & \text{jinak,} \end{cases} \quad (3.16)$$

nejmenší hodnotu integrálu $\int_{-1}^1 (f^{(k)}(x))^2 dx$.

Lze ukázat, že platí

1. $\sigma_k^2 = \int_{-1}^1 x^2 g_k(x) dx = \frac{1}{2k+5}$.
2. Pro $r > 0$, $\int (r g^{(k)}(rx))^2 dx = r^{2k+1} \int (g^{(k)}(x))^2 dx$ pro každou hustotu, pro kterou integrál existuje.
3. Jestliže hustota g má rozptyl σ_g^2 , pak hustota $\frac{\sigma_g}{\sigma} g\left(\frac{\sigma_g}{\sigma} x\right)$ má rozptyl σ^2 .

Jestliže f je neznámá hustota s rozptylem σ^2 a g_k je hustota rozdělení $Beta(k+2, k+2)$, pro kterou je $V(g^{(k)})$ minimální, pak využitím faktu 1-3 lze ukázat, že

$$h_{opt,0,k} \leq \delta_{0k} \left(\frac{(k!)^2}{2nk} \right)^{\frac{1}{2k+1}} \frac{\sigma}{\sigma_k} (V(g_k^{(k)}))^{\frac{-1}{2k+1}}.$$

Hodnotu σ lze odhadnout pomocí dříve uvedených vztahů a $\sigma_k = \frac{1}{2k+5}$.

Hodnotu $V(g_k^{(k)})$ lze vypočítat pomocí speciálních ortogonálních polynomů [5]:

$$V(g_k^{(k)}) = \int_{-1}^1 (g_k^{(k)}(x))^2 dx = \frac{(2k+3)!(2k+2)!}{2^{2k+2}(2k+1)(2k+5)(k+1)!^2}. \quad (3.17)$$

Použijeme-li poslední vyjádření (3.17), dostaneme horní hranici pro vyhlazovací parametry

$$\hat{h}_{opt,0,k} \leq h_{MS} = \hat{\sigma} n^{-1/(2k+1)} b_k,$$

přičemž

$$b_k = \sqrt{2k+5} \left(\frac{2^{2k+2} V(K) (2k+1)(2k+5)(k+1)^2 (k!)^2}{\beta_k^2(K) (2k+3)! (2k+2)!} \right)^{\frac{1}{2k+1}}.$$

Tabulka 3.1: Hodnoty b_k pro optimální jádro $K_{opt,0,k} \in S_{0k}$

k	2	4	6	8	10
b_k	2,5324	3,3175	3,9003	4,3949	4,8349

Příklad 3.4. Určeme hodnotu h_{MS} pro odhad hustoty s jádrem řádu $k=2$, tj. $K \in S_{02}$. Podle věty 3.3 je hustota g_2 tvaru

$$g_2(x) = \frac{35}{32} (1-x^2)^3, \quad x \in [-1, 1],$$

a dále z vlastností funkce g_2 a ze vztahu (3.17) plyne

$$\sigma_2^2 = \int_{-1}^1 x^2 g_2(x) dx = \frac{1}{9} \quad \text{a} \quad \int_{-1}^1 (g_2''(x))^2 dx = 35.$$

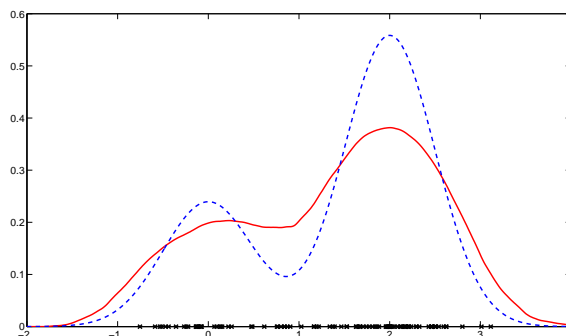
Pak pro $K_{opt,0,2}$

$$h_{MS} = \hat{\sigma} n^{-1/5} \left(\frac{V(K)}{\beta_2^2(K)} \cdot \frac{243}{35} \right)^{1/5}.$$

Příklad 3.5. Pro data z příkladu 3.1 bude vyhlazovací parametr určený metodou maximálního vyhlazení s Epanečnickovým jádrem ($k = 2$, $V(K) = 3/5$, $\beta_2(K) = 1/5$) roven

$$h_{\text{MS}} = \hat{\sigma} n^{-1/5} \left(\frac{3/5}{1/25} \cdot \frac{243}{35} \right)^{1/5} = 1,0409,$$

protože $\hat{\sigma} = 1,0325$ a $n = 100$. Výsledný odhad je vidět na obr. 3.8.



Obrázek 3.8: Odhad hustoty s $h_{\text{MS}} = 1,0409$, odhad (červená, plná), původní funkce (modrá, čárkovaná)

Poznámka 3.5. Hodnota h_{MS} může sloužit jako horní hranice pro množinu vyhlazovacích parametrů volených podle jiné metody, např. metody křížového ověřování. Tedy $H_n = [h_\ell, h_{\text{MS}}]$, kde h_ℓ je nejmenší vzdálenost mezi po sobě jdoucími body $X_i, i = 1, \dots, n$.

OTÁZKA. Jak hodnotu bude mít odhad vyhlazovacího parametru pomocí metody maximálního vyhlazení pro hustotu z ukázkového příkladu 3.2?

5.3 Metoda křížového ověřování

Metoda křížového ověřování patří mezi nejužívanější metody pro odhad vyhlazovacího parametru. Myšlenka této metody je založena na minimalizaci MISE, jak je zřejmé z následující úvahy:

$$\begin{aligned} \text{MISE } \hat{f}(\cdot, h) &= E \int (\hat{f}(x, h) - f(x))^2 dx \\ &= E \int \hat{f}^2(x, h) dx - 2E \int \hat{f}(x, h) f(x) dx + \int f^2(x) dx. \\ \text{MISE } \hat{f}(\cdot, h) - \int f^2(x) dx &= E \left(\int \hat{f}^2(x, h) dx - 2 \int \hat{f}(x, h) f(x) dx \right). \end{aligned}$$

Definujme funkci křížového ověřování

$$\text{CV}(h) = \int \hat{f}^2(x, h) dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i, h), \quad (3.18)$$

kde $\hat{f}_{-i}(X_i, h)$ je odhad v bodě X_i bez použití tohoto bodu.

Věta 3.4. Platí

$$E \text{CV}(h) = \text{MISE } \hat{f}(\cdot, h) - \int f^2(x) dx,$$

tj. $\text{CV}(h)$ je nevychýleným odhadem

$$E \left(\int \hat{f}^2(x, h) dx - 2 \int \hat{f}(x, h) f(x) dx \right).$$

Důkaz. Spočítejme střední hodnotu funkce $CV(h)$, tj.

$$E CV(h) = E \int \hat{f}^2(x, h) dx - \frac{2}{n} E \sum_{i=1}^n \hat{f}_{-i}(X_i, h).$$

Dále upravme druhý člen tohoto vyjádření:

$$\begin{aligned} E \frac{1}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i, h) &= E \frac{1}{n} \sum_{i=1}^n \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n K_h(X_i - X_j) \\ &= E \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n K_h(X_i - X_j) = E K_h(X_1 - X_2) \\ &= \iint \underbrace{K_h(x-y)}_{E \hat{f}(x, h)} f(y) f(x) dx dy = E \int \hat{f}(x, h) f(x) dx. \end{aligned}$$

Odtud

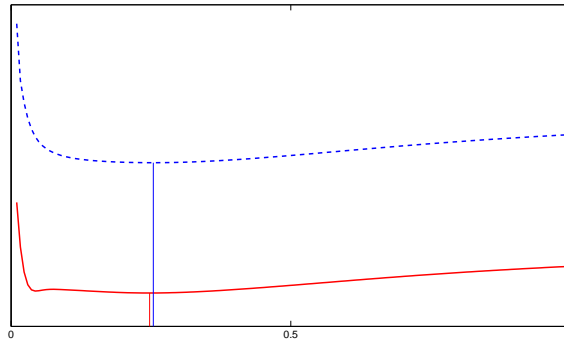
$$E CV(h) = E \int \hat{f}^2(x, h) dx - 2E \int \hat{f}(x, h) f(x) dx.$$

□

Odhad $h_{opt,0,k}$ je dán vztahem

$$h_{CV} = \arg \min_{h \in H_n} CV(h).$$

Odtud plyne, že $CV(h) + \int f^2(x) dx$ je pro každé h nevychýleným odhadem MISE $\hat{f}(\cdot, h)$. Protože $\int f^2(x) dx$ nezávisí na h , minimalizace $E CV(h)$ odpovídá minimalizaci MISE. Jestliže předpokládáme, že $\min CV(h) \sim \min E CV(h)$, dostaneme dobrou aproximaci optimální hodnoty h .

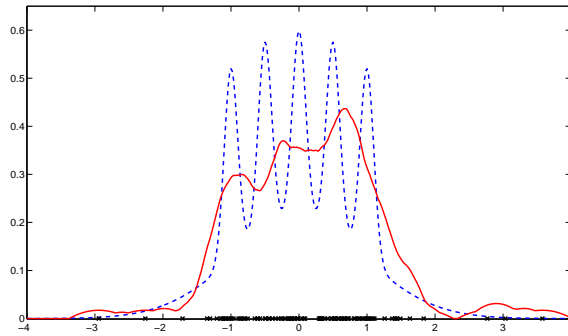


Obrázek 3.9: Porovnání minima MISE (modrá, čárkovaná) a minima funkce křížového ověřování CV (červená, plná) pro simulovaná data z příkladu 3.1

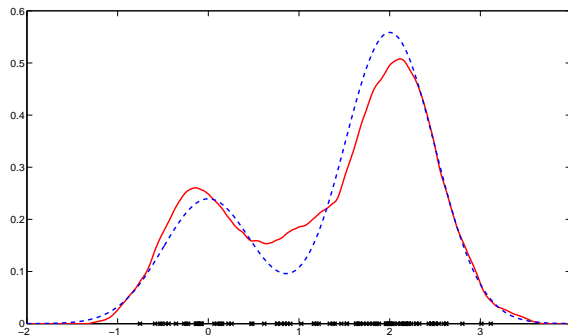
OTÁZKA. Jak hodnotu bude mít odhad vyhlazovacího parametru pomocí metody křížového ověřování pro hustotu z ukázkového příkladu 3.2?

Poznámka 3.6. Předpokládejme, že $k = 2$. Pak vychýlení odhadu může být velké, jestliže $(f'')^2$ nabývá velkých hodnot, tj. křivost hustoty je velká. Při vyhlazování se tato objevuje ve „vrcholech“, kde je vychýlení záporné, nebo v „údolích“, kde je vychýlení kladné. Odhad má tendenci „vyhladit“ tyto jevy, jak je patrné z obrázku 3.10.

Příklad 3.6. Jádrový odhad hustoty dat z příkladu 3.1 je zobrazen na obr. 3.11. Pro rekonstrukci bylo použito Epanečnikovo jádro a vyhlazovací parametr určený metodou křížového ověřování $h_{CV} = 0,5628$.



Obrázek 3.10: Zahlazení vrcholů a údolí při odhadu hustoty směsi normálních rozdělání, odhad (červená, plná), původní funkce (modrá, čárkovaná)



Obrázek 3.11: Odhad hustoty s $h_{CV} = 0,5628$, odhad (červená, plná), původní funkce (modrá, čárkovaná)

5.4 Iterační metoda

V tomto odstavci pojednáme o další metodě pro odhad vyhlazovacího parametru. Metodu pouze popíšeme a nebudeme se zabývat statistickými vlastnostmi získané aproximace. Podrobnou analýzu lze najít v monografii [3]. Tato metoda je založena na vztahu (3.7):

$$AIV(h_{opt,0,k}) = 2k AISB(h_{opt,0,k}).$$

Přepíšeme tuto rovnici

$$\frac{V(K)}{nh} - 2kh^{2k} \frac{\beta_k^2(K)}{(k!)^2} V(f^{(k)}) = 0. \quad (3.19)$$

Problém nalézt $h_{opt,0,k}$, pro které AMISE $\hat{f}(\cdot, h)$ nabývá minimální hodnoty, je tedy ekvivalentní řešení této rovnice. Zde se ovšem vyskytuje stejný problém – neznáme hodnotu $V(f^{(k)})$, a proto budeme počítat s odhady rozptylu a vychýlení. Tyto odhady uvažujeme ve tvaru

$$\widehat{\text{var}}\hat{f}(x, h) = \frac{1}{nh} \int K^2(y) \hat{f}(x - hy) dy$$

$$\widehat{\text{bias}}\hat{f}(x, h) = (K_h * \hat{f})(x, h) - \hat{f}(x, h) = \int \hat{f}(x - hy, h) K(y) dy - \hat{f}(x, h).$$

Odtud plyne

$$\widehat{AIV}(h) = \frac{V(K)}{nh} \quad (3.20)$$

a

$$\begin{aligned}\overline{\text{AISB}}(h) &= \int \left(\int \hat{f}(x - hy, h) K(y) dy - \hat{f}(x, h) \right)^2 dx \\ &= \int \left(\int K(y) \sum_{i=1}^n \frac{1}{nh} K\left(\frac{x - hy - X_i}{h}\right) dy - \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \right)^2 dx.\end{aligned}$$

Výraz lze upravit pomocí konvolucí a dostaneme

$$\begin{aligned}&= \frac{1}{n^2 h} \sum_{i,j=1}^n (K * K * K * K - 2K * K * K + K * K) \left(\frac{X_i - X_j}{h} \right) \\ &= \frac{1}{n^2 h} \sum_{i,j=1}^n \Lambda \left(\frac{X_i - X_j}{h} \right).\end{aligned}$$

Funkce $\Lambda(z) = (K * K * K * K - 2K * K * K + K * K)(z)$ má tyto vlastnosti

$$\begin{aligned}\int z^j \Lambda(z) dz &= 0, \quad j = 0, 1, \dots, 2k - 1, \\ \int z^{2k} \Lambda(z) dz &= \binom{2k}{k} \beta_k^2, \\ \Lambda_h(z) &= \frac{1}{h} \Lambda\left(\frac{z}{h}\right).\end{aligned}\tag{3.21}$$

$\overline{\text{AISB}}\hat{f}(\cdot, h)$ je vychýleným odhadem AISB, a proto budeme uvažovat

$$\widehat{\text{AISB}}(h) = \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^n \Lambda_h(X_i - X_j).\tag{3.22}$$

Místo rovnice (3.19) řešíme rovnici

$$\frac{V(K)}{nh} - \frac{2k}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^n \Lambda_h(X_i - X_j) = 0.$$

Rovnici lze také zapsat ve tvaru:

$$h - \frac{nV(K)}{2k \sum_{\substack{i,j=1 \\ i \neq j}}^n \Lambda_h(X_i - X_j)} = 0.\tag{3.23}$$

Uvedenou rovnici lze řešit Newtonovou metodou, neboť derivaci funkce lze snadno spočítat užitím konvolucí. Řešení této rovnice označíme h_{IT} .

Řešení rovnice (3.23) lze považovat za vhodnou aproximaci $h_{\text{opt},0,k}$. Tato skutečnost je dokázána v následující větě [3].

Věta 3.5. *Nechť*

$$P(h) = \frac{V(K)}{nh} - 2kh^{2k} \frac{\beta_k^2(K)}{(k!)^2} V(f^{(k)})$$

a

$$\hat{P}(h) = \frac{V(K)}{nh} - \frac{2k}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^n \Lambda_h(X_i - X_j).$$

Pak platí

$$E\hat{P}(h) = P(h) + o(h^{2k+1}),$$

$$\text{var } \hat{P}(h) = \frac{8k^2}{n^2h} V(\Lambda)V(f) + o(n^{-2}h^{-1}).$$

Poznámka 3.7. Rychlost konvergence pro metodu křížového ověřování:

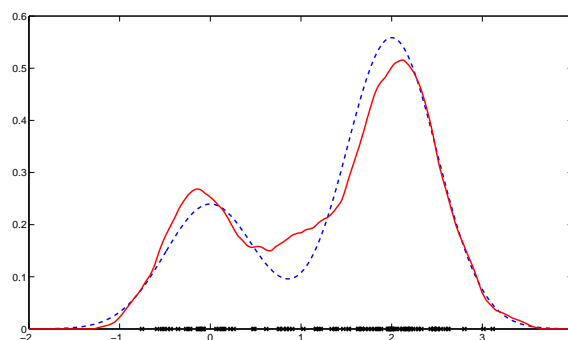
$$\frac{\hat{h}_{CV} - h_{opt,0,2}}{h_{opt,0,2}} = O(n^{-1/10}).$$

Rychlost konvergence pro iterační metodu

$$\frac{\hat{h}_{IT} - h_{opt,0,2}}{h_{opt,0,2}} = O(n^{-1/10})$$

Řády rychlosti konvergence pro CV metodu a iterační metodu jsou stejné, ale výhodou iterační metody je podstatně menší výpočetní náročnost.

Příklad 3.7. Jádrový odhad hustoty dat z příkladu 3.1 s Epanečnikovým jádrem a vyhlazovacím parametrem určeným iterační metodou je uveden na obr. 3.12.



Obrázek 3.12: Odhad hustoty s $h_{IT} = 0,5314$, odhad (červená, plná), původní funkce (modrá, čárkovaná)

6 Automatická procedura

Obdobně jako v případě regresní funkce můžeme nalézt podobnou formuli pro AMISE $\hat{f}(\cdot, h)$, ve které budou jednotlivé parametry K, h, k separovány, což nám umožní navrhnout proceduru pro simultánní volbu těchto parametrů.

Vyjdeme ze vztahu

$$\text{AMISE } \hat{f}(\cdot, h_{opt,0,k}) = T(K) \left(\frac{\delta_{0k}}{nh_{opt,0,k}} + \frac{h_{opt,0,k}^{2k} V(f^{(k)})}{\delta_{0k}^{2k} (k!)^2} \right).$$

Ze vztahu pro $h_{opt,0,k}$ vypočteme $V(f^{(k)})$

$$V(f^{(k)}) = \frac{\delta_{0k}^{2k+1} (k!)^2}{2knh_{opt,0,k}^{2k+1}}$$

a tuto hodnotu dosadíme do předchozího vztahu:

$$\text{AMISE } \hat{f}(\cdot, h_{opt,0,k}) = T(K) \frac{(2k+1)\delta_{0k}}{2knh_{opt,0,k}}$$

Tento vztah je základem automatické procedury, označme jej $L(k)$ ve shodě s označením u regresní funkce. Podobně množinu vhodných řádů k označme

$$I(k_0) = \left\{ 2j, j = 0, \dots, \left\lfloor \frac{k_0}{2} \right\rfloor \right\}.$$

Procedura

1. Pro $k \in I(k_0)$ najděte optimální jádro $K_{opt,0,k} \in S_{0k}$, které je dáno tabulkou 8.1, k němu příslušný kanonický faktor δ_{0k} .
2. Pro $k \in I(k_0)$ a $K_{opt,0,k} \in S_{0k}$ najděte optimální vyhlazovací parametr $\hat{h}_{opt,0,k}$.
3. Pro $k \in I(k_0)$ vypočítejte hodnotu výběrového kritéria $L(k)$ s využitím hodnot získaných v krocích 1 a 2.
4. Vypočítejte optimální hodnotu řádu \hat{k} , které minimalizuje funkcional $L(k)$.
5. Použijte parametry z předchozích kroků ke konstrukci optimálního jádrového odhadu hustoty, tj.

$$\hat{f}(x, \hat{h}_{opt,0,\hat{k}}) = \frac{1}{n\hat{h}_{opt,0,\hat{k}}} \sum_{i=1}^n K\left(\frac{x - X_i}{\hat{h}_{opt,0,\hat{k}}}\right).$$

Příklad 3.8. Aplikace procedury na data z příkladu 3.1. Maximální řád jádra zvolme $k_0 = 8$, tedy množina možných řádů jader je $I(8) = \{0, 2, 4, 6, 8\}$. Pro tyto řády spočítejme hodnoty z kroků 1–3.

V toolboxu Matlabu, který je doprovodným materiálem těchto skript, je jako implicitní metoda pro odhad vyhlazovacího parametru automatickou procedurou použita iterační metoda (podrobněji např. [3]). Proto při výpočtu optimálních parametrů použijeme mezivýpočty z tohoto toolboxu.

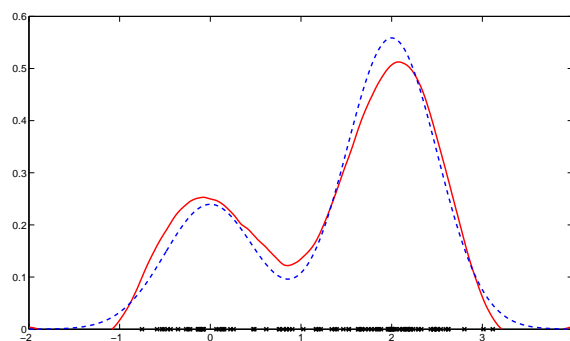
k	$K_{opt,0,k}$	δ_{0k}	$\hat{h}_{opt,0,k}$	$L(K)$
2	$-\frac{3}{4}(x^2 - 1)$	1,7188	0,5314	0,0141
4	$\frac{15}{32}(x^2 - 1)(7x^2 - 3)$	2,0165	1,0734	0,0131
6	$-\frac{105}{256}(x^2 - 1)(33x^4 - 30x^2 + 5)$	2,0834	1,6460	0,0125
8	$\frac{315}{4096}(x^2 - 1)(715x^6 - 1001x^4 + 385x^2 - 35)$	2,1021	2,1367	0,0126

Z tabulky vidíme, že optimální řád jádra je $\hat{k} = 6$. Výsledný odhad je uveden na obrázku 3.13.

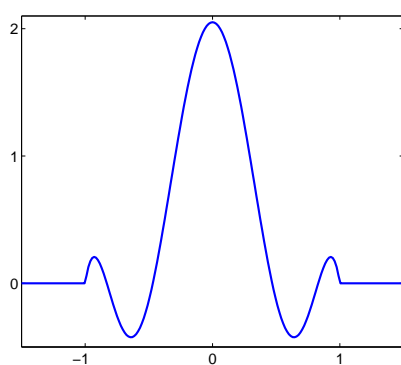
Při bližším pohledu na odhadnutou hustotu je patrný vliv použití optimálního jádra vyššího řádu. Jádra vyšších řádů mohou nabývat záporných hodnot a tím ovlivnit výslednou odhadnutou funkci – viz obrázek 3.14. V takovém případě je vhodné použít jinou metodu pro nalezení vyhlazovacího parametru, případně použít jiné jádro. Lze doporučit jádra třídy S_{02} , např. kvartické jádro: $K(x) = \frac{15}{16}(1 - x^2)^2 I_{[-1,1]}(x)$, nebo jádro triweight: $K(x) = \frac{35}{32}(1 - x^2)^3 I_{[-1,1]}(x)$.

Shrneme-li na závěr vypočítané hodnoty vyhlazovacích parametrů pro simulovaná data z příkladu 3.1, můžeme vizuálně porovnat jednotlivé odhady – viz obrázek 3.15.

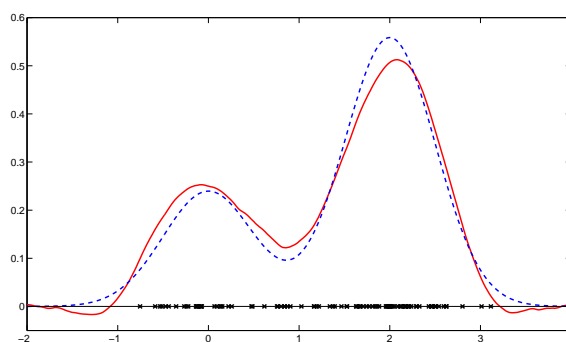
$$\begin{aligned} h_{opt,0,2} &= 0,5122, & h_{REF} &= 0,9639, \\ h_{MS} &= 1,0409, & h_{CV} &= 0,5628, \\ h_{IT} &= 0,5314, & \hat{h} &= 1,6460 \ (K_{opt,0,6}). \end{aligned}$$



Obrázek 3.13: Simulovaná data s jádrovým odhadem hustoty při použití procedury, odhad (červená, plná), původní funkce (modrá, čárkovaná)

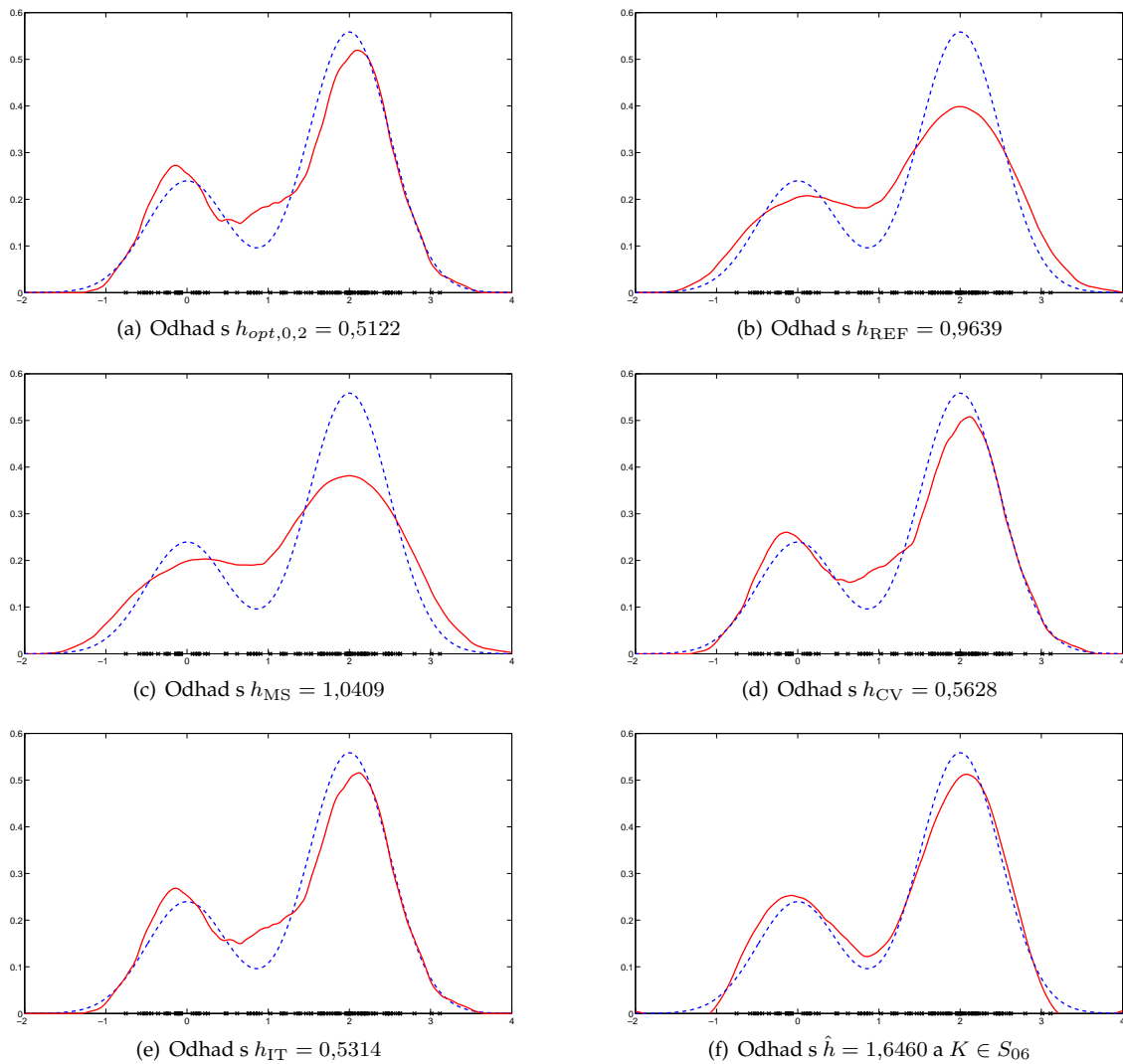


(a) Jádro $K_{opt,0,6}$

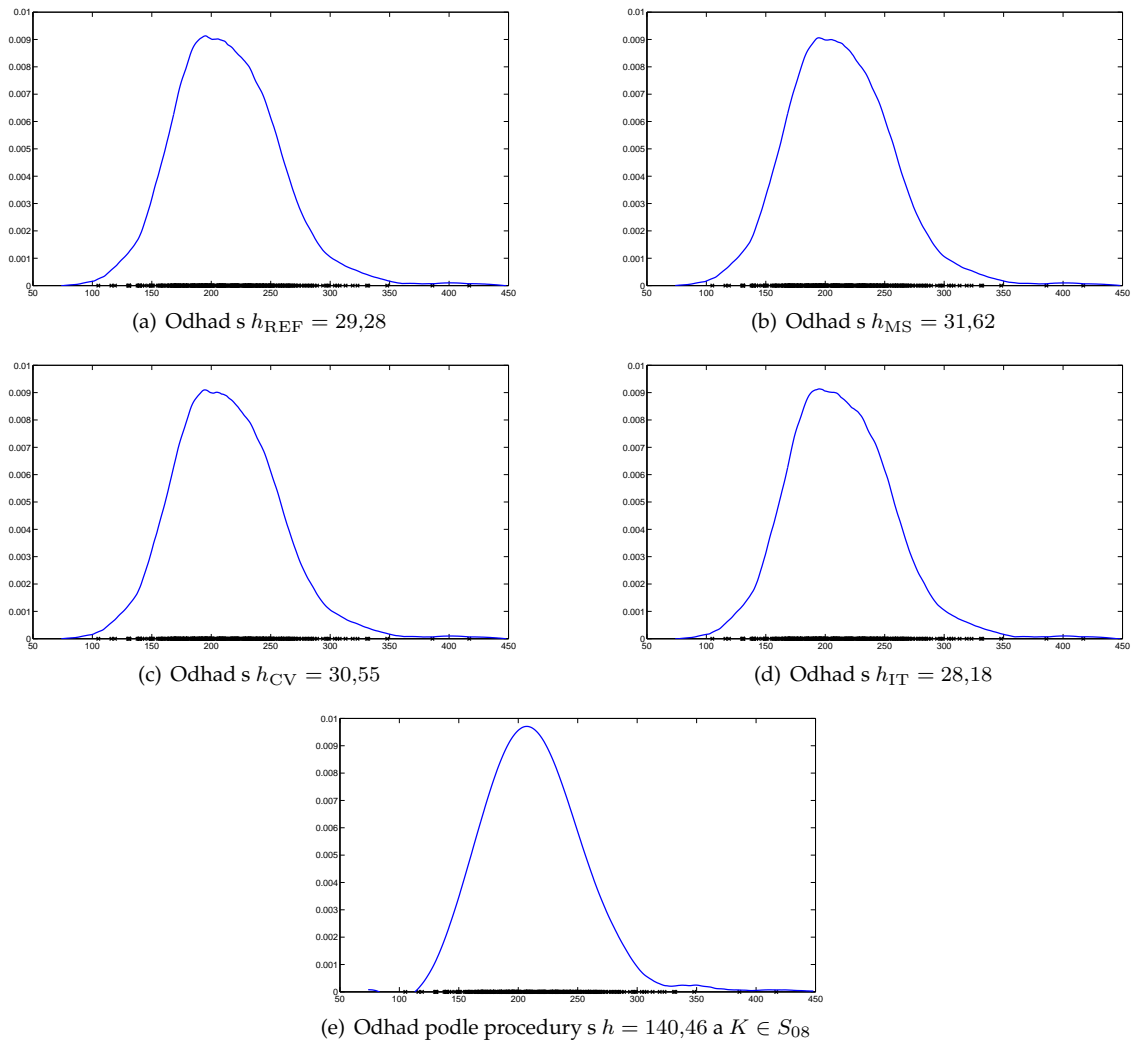


(b) Odhad hustoty

Obrázek 3.14: Jádro třídy S_{06} a k němu příslušný odhad hustoty při použití procedury (červená, plná) a původní funkcí (modrá, čárkovaná)



Obrázek 3.15: Srovnání odhadů pro data z příkladu 3.1



Obrázek 3.16: Grafy odhadnutých hustot pro obsah cholesterolu

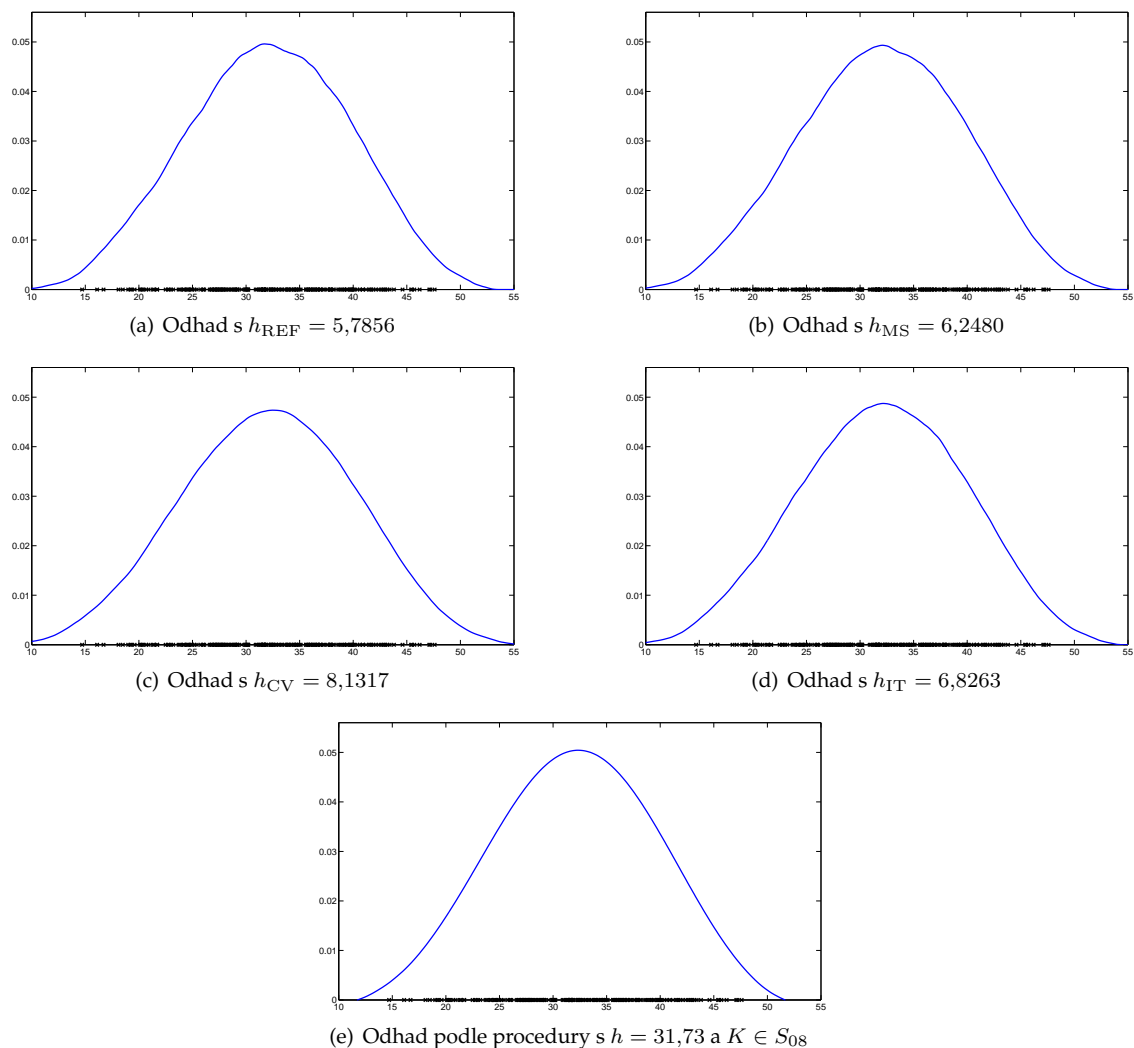
7 Aplikace na reálná data

Vraťme se k úvodnímu příkladu, který obsahuje měření množství cholesterolu v krevní plazmě u 371 pacientů. Data jsou shrnuta v tabulce 7.11. Skupina pacientů obsahovala dvě podskupiny, a to pacienty, u nichž bylo potvrzeno zúžení tepen (320), a pacienty zdravé (51). Máme-li k dispozici tuto informaci, mohli bychom očekávat, že odhadovaná hustoty může být bimodální. Avšak na druhou stranu, pokud při rekonstrukci hustoty neodhalíme tuto strukturu, neznamená to, že tam není, může být skrytá.

S použitím všech uvedených metod pro odhad optimálního vyhlazovacího parametru jsme vypočítali tyto hodnoty:

$$h_{REF} = 29,28, \quad h_{MS} = 31,62, \quad h_{CV} = 30,55, \quad h_{IT} = 28,18.$$

Výsledné odhady s Epanečnickovým jádrem jsou zobrazeny na obrázku 3.16, kde je také odhad hustoty při použití automatické procedury, u níž vypočítáme odhad optimálního vyhlazovacího parametru $\hat{h} = 140,46$ a jádra K_{08} .



Obrázek 3.17: Grafy odhadnutých hustot pro délku krunýře

Druhý datový soubor obsahuje morfologická měření padesáti exemplářů od obojího pohlaví a obou barevných forem (oranžová a modrá) krabů rodu *Leptograpsus*.² Pro odhad hustoty nám postačuje jeden druh měření, vybrali jsme délku podél středové osy krunýře, která byla měřena v milimetrech. Data jsou uvedena v tabulce 7.10.

Užitím výše uvedených metod pro odhad vyhlazovacího parametru jsme (při použití Epanečnikova jádra) dostali následující hodnoty:

$$h_{\text{REF}} = 5,7856, \quad h_{\text{MS}} = 6,2480, \quad h_{\text{CV}} = 8,1317, \quad h_{\text{IT}} = 6,8263.$$

U automatické procedury je v toolboxu implicitně nastavena iterační metoda pro odhad vyhlazovacího parametru, proto jsme tuto volbu ponechali i zde, ať má čtenář možnost porovnání při vlastních výpočtech. Při použití procedury vyjde vyhlazovací parametr roven $\hat{h} = 31,7329$ s optimálním jádrem $K_{\text{opt},0,8}$. Výsledné odhady hustoty na jsou zachyceny na obrázku 3.17.

²Celý datový soubor je dostupný v programu R.

Shrnutí
<p>Odhad hustoty pravděpodobnosti f v bodě x je tvaru</p> $\hat{f}(x, h) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right).$
<p>Asymptotická střední kvadratická chyba jádrového odhadu hustoty pravděpodobnosti s jádrem řádu k je součtem asymptotického tvaru rozptylu (AIV) a druhé mocniny vychýlení (AISB)</p> $\text{AMISE } \hat{f}(\cdot, h) = \underbrace{\frac{V(K)}{nh}}_{\text{AIV}(h)} + \underbrace{\frac{1}{(k!)^2} h^{2k} \beta_k^2(K) V(f^{(k)})}_{\text{AISB}(h)}.$
<p>Optimální vyhlazovací parametr vzhledem k AMISE pro odhad hustoty pravděpodobnosti s jádrem řádu k je tvaru</p> $h_{opt,0,k}^{2k+1} = \frac{\delta_{0k}^{2k+1} (k!)^2}{2knV(f^{(k)})},$ <p>tj. $h_{opt,0,k} = O(n^{-1/(2k+1)})$, $\text{AMISE } \hat{f}(\cdot, h_{opt,0,k}) = O(n^{-2k/(2k+1)})$.</p>
<p>Metody pro odhad optimální hodnoty vyhlazovacího parametru h</p> <ul style="list-style-type: none"> metoda referenční hustoty $h_{\text{REF}} = \left(\frac{2^{2k} k!^3 \sqrt{\pi}}{(2k)!k} \right)^{\frac{1}{2k+1}} \delta_{0k} \sigma n^{-\frac{1}{2k+1}},$ <ul style="list-style-type: none"> metoda maximálního vyhlazení $h_{\text{MS}} = \hat{\sigma} n^{-1/(2k+1)} b_k,$ <ul style="list-style-type: none"> metoda křížového ověřování $h_{\text{CV}} = \arg \min_{h \in H_n} \text{CV}(h) = \int \hat{f}^2(x, h) dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i, h),$ <ul style="list-style-type: none"> iterační metoda $h_{\text{IT}} = \text{řešení rovnice: } h - \frac{nV(K)}{2k \sum_{\substack{i,j=1 \\ i \neq j}}^n \Lambda_h(X_i - X_j)} = 0.$
<p>Automatická procedura pro simultánní volbu optimálního jádra, vyhlazovacího parametru a řádu jádra je dostupná v toolboxu Matlabu.</p>

Výstupy z výukové jednotky

Student

- zná tvar jádrových odhadů hustoty pravděpodobnosti
- je schopen analyzovat statistické vlastnosti těchto odhadů

- se seznámil s metodami pro volbu vyhlazovacího parametru
- porozuměl automatické proceduře pro simultánní volbu parametrů vyhlazování
- zvládne použití toolboxu v Matlabu a dokáže pro daný soubor dat zkonstruovat jádrový odhad hustoty a jejích derivací

Cvičení

1. Dokažte vztah (3.4) pro tvar chyby AMISE.
2. Uvažujte náhodný výběr, který pochází z rozdělení s hustotou $f(x) = 20x(1-x)^3$ pro $x \in [0, 1]$ a který obsahuje 50 prvků. Vypočítejte hodnotu $h_{opt,0,4}$. Je výsledná hodnota správná?
3. Vypočítejte hodnotu optimálního vyhlazovacího parametru pro odhad s jádrem řádu 2 pro soubor dat s hustotou $K(x) = \frac{15}{16}(1-x^2)^2$ pro $x \in [-1, 1]$.
Pomůcka: $V(f'') = 22,5$.
4. Dokažte:
 - a) $\sigma_k^2 = \int_{-1}^1 x^2 g_k(x) dx = \frac{1}{2k+5}$, pro hustotu $g_k(x)$ Beta($k+2, k+2$) rozdělení (rovnice (3.16)).
 - b) Jestliže hustota g má rozptyl σ_g^2 , pak hustota $\frac{\sigma_g}{\sigma} g\left(\frac{\sigma_g}{\sigma} x\right)$ má rozptyl σ^2 .
5. Dokažte vztah (3.15).
6. Spočítejte h_{MS} pro
 - Epanečnikovo jádro $K(x) = \frac{3}{4}(1-x^2)$,
 - kvartické jádro $K(x) = \frac{15}{16}(1-x^2)^2$.
7. Aplikujte metody pro odhad vyhlazovacího parametru a automatickou proceduru na simulovaná data z ukázkového příkladu 3.2.

Kapitola 4

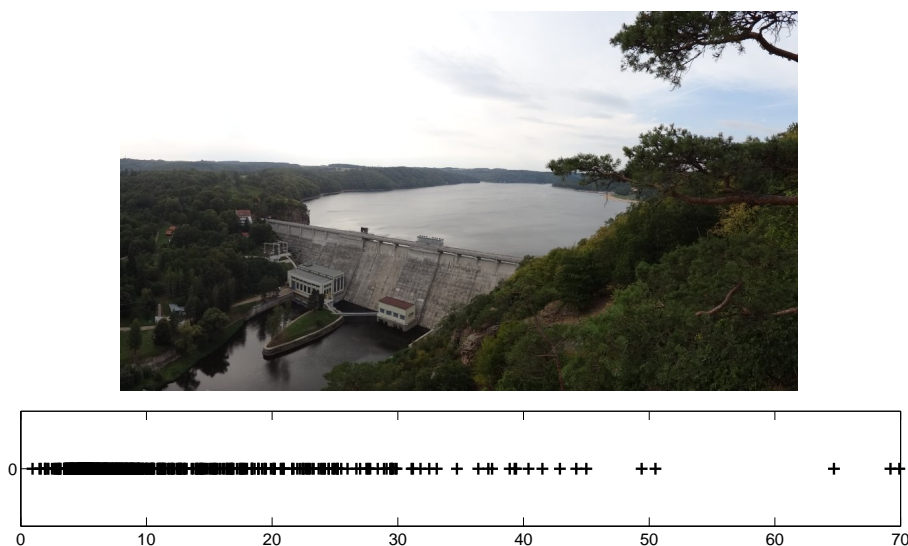
Jádrové odhady distribuční funkce

1 Motivace

Distribuční funkce $F(x)$ popisuje rozložení pravděpodobnosti náhodné veličiny X (budeme předpokládat spojitost náhodné veličiny). Stejně jako při rekonstrukci hustoty z množiny pozorovaných dat lze distribuční funkci odhadnout parametrickými nebo neparametrickými metodami. Zaměříme se výhradně na neparametrické metody, kdy předpokládáme pouze jistou hladkost odhadované distribuční funkce.

Distribuční funkce má použití v analýze přežití, v teorii spolehlivosti, klasifikaci diagnostických testů (ROC křivky) aj. Funkce $\bar{F}(x) = 1 - F(x)$ se nazývá funkce přežití a vyjadřuje pravděpodobnost, že daná událost (úmrtí, selhání, porucha) nenastane dříve než po nějakém čase x . S touto funkcí se můžeme setkat i v hydrologii, kde se nazývá křivkou zabezpečení.

Křivka zabezpečení je jedním z nejdůležitějších indikátorů toku řeky během daného časového období. Tato křivka vyjadřuje míru, s jakou je zajištěn vybraný průtok. Máme k dispozici průměrné měsíční průtoky řeky Dyje v profilu Vranov¹ (pod přehradou Vranov, plocha povodí asi 2 220 km²) v období 1931–1990. Zajímá nás stabilita měsíčních průměrných průtoků.



Obrázek 4.1: Průměrné měsíční průtoky řeky Dyje

Máme-li tedy náhodnou veličinu X , která je popsána svou distribuční funkcí $F(x)$ a k ní

¹Data byla poskytnuta ČHMÚ, pobočka Brno.

příslušnou hustotou $f(x)$, pak platí

- $0 \leq F(x) \leq 1$,
- $F(x)$ je zprava spojitá,
- $\lim_{x \rightarrow -\infty} F(x) = 0, \lim_{x \rightarrow \infty} F(x) = 1$,
- $F(x) = \int_{-\infty}^x f(t) dt, F'(x) = f(x)$.

Nejužívanějším neparametrickým odhadem distribuční funkce je empirická distribuční funkce F_n . Ovšem F_n je schodovitá funkce i v případě, že F je spojitá. Nadaraya (1964) navrhl „hladkou“ alternativu k F , a to jádrový odhad \hat{F} , který se získá integrací známého jádrového odhadu hustoty (3.1).

2 Základní typy neparametrických odhadů

Nechť X_1, \dots, X_n jsou nezávislé náhodné proměnné, které mají tutéž spojitou hustotu f a distribuční funkci F . Nejjednodušší neparametrický odhad distribuční funkce F je *empirická distribuční funkce* \hat{F}_n definovaná v bodě x vztahem

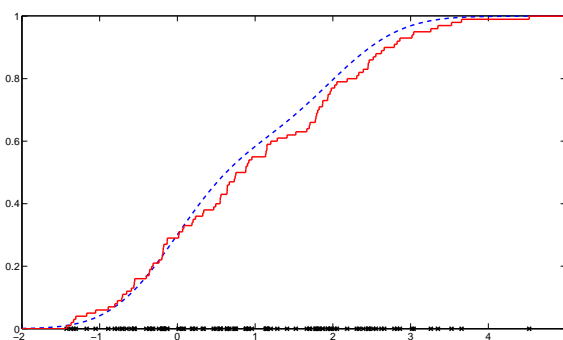
$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(X_i).$$

Tento odhad má sice dobré statistické vlastnosti, ale je to schodovitá funkce (viz obr. 4.2, a proto se budeme zabývat postupy, které umožní zkonstruovat „hladký“ odhad distribuční funkce F .

Příklad 4.1. Mějme dán náhodný výběr o velikosti $n = 100$ ze směsi dvou normálních hustot $N(0; 4/9)$ a $N(2; 1)$ s hustotou (viz kapitola 3 o hustotě) (Data jsou v tabulce 7.5.)

$$f(x) = 0,5 \frac{3}{2\sqrt{2\pi}} e^{-\frac{9x^2}{8}} + 0,5 \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2}}.$$

Z obrázku 4.2 je patrné, že schodovitá funkce nevystihuje plně charakter distribuční funkce.



Obrázek 4.2: Empirická distribuční funkce F_n (červená, plná) a skutečná distribuční funkce F (modrá, čárkovaná) pro data z příkladu 4.1

Nejznámější postup, jak odvodit neparametrický odhad distribuční funkce, spočívá v integraci jádrového odhadu hustoty, t.j.

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

Užijeme-li substituce $y = (t - X_i)/h$, dostaneme

$$\widehat{F}(x, h) = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{\frac{x-X_i}{h}} K(y) dy = \frac{1}{n} \sum_{i=1}^n W\left(\frac{x-X_i}{h}\right).$$

To znamená, že odhad F v bodě $x \in \mathbb{R}$ je definován takto

$$\widehat{F}(x, h) = \frac{1}{n} \sum_{i=1}^n W\left(\frac{x-X_i}{h}\right), \quad W(x) = \int_{-\infty}^x K(t) dt. \quad (4.1)$$

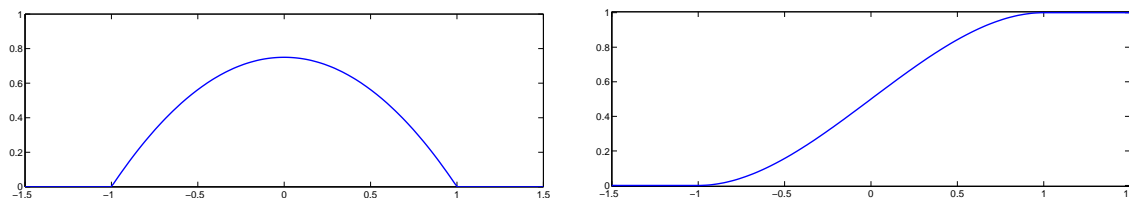
Zde předpokládáme, že $K \in S_{02}$, $K(x) \geq 0$ pro $x \in [-1, 1]$. Níže jsou uvedeny základní vlastnosti funkce W :

1. $W(x) = 0$ pro $x \in (-\infty, -1]$ a $W(x) = 1$ pro $x \in [1, \infty)$,
2. $\int_{-1}^1 W^2(x) dx \leq \int_{-1}^1 W(x) dx = 1$,
3. $\int_{-1}^1 W(x)K(x) dx = \frac{1}{2}$,
4. $\int_{-1}^1 xW(x)K(x) dx = \frac{1}{2} \left(1 - \int_{-1}^1 W^2(x) dx\right)$.

OTÁZKA. Lze použít obdélníkové jádro? Jaký bude tvar a vlastnosti funkce $W(x)$?

Příklad 4.2. Použijeme-li Epanečnikovo jádro $K(x) = \frac{3}{4}(1-x^2)I_{[-1,1]}(x)$, pak funkce W je tvaru

$$W(x) = \begin{cases} 0 & x \leq -1, \\ \frac{1}{4}(-x^3 + 3x + 2) & |x| < 1, \\ 1 & x \geq 1. \end{cases}$$



Obrázek 4.3: Epanečnikovo jádro K (vlevo) a k němu příslušná funkce W (vpravo)

Pro data z příkladu 4.1 je jádrový odhad distribuční funkce prezentován na obrázku 4.4.

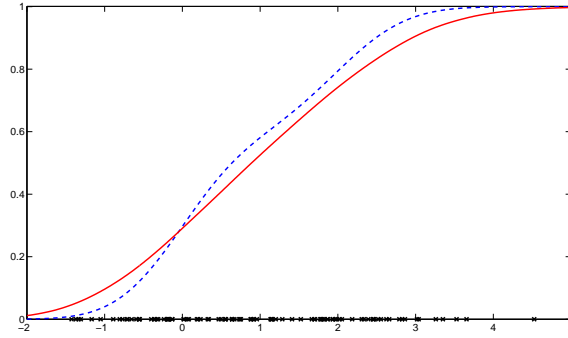
3 Statistické vlastnosti odhadu

Kvalitu jádrového odhadu lze lokálně popsat pomocí střední kvadratické chyby MSE:

$$\begin{aligned} \text{MSE } \widehat{F}(x, h) &= E(\widehat{F}(x, h) - F(x))^2 \\ &= \underbrace{(E\widehat{F}(x, h) - F(x))^2}_{\text{bias}^2} + \underbrace{E(\widehat{F}(x, h))^2 - (E\widehat{F}(x, h))^2}_{\text{var}}. \end{aligned}$$

Spočítejme nejdříve hodnotu $E\widehat{F}(x, h)$ v bodě $x \in \mathbb{R}$:

$$\begin{aligned} E\widehat{F}(x, h) &= \int W\left(\frac{x-y}{h}\right) f(y) dy \\ &= h \int_{-\infty}^1 W(t) f(x-ht) dt + h \int_1^{\infty} W(t) f(x-ht) dt. \end{aligned}$$



Obrázek 4.4: Jádrový odhad distribuční funkce s parametrem $h = 1,5$, odhad (červená, plná), původní funkce (modrá, čárkovaná)

Předpokládejme dále, že $F \in C^2$. Označme první integrál I_1 a druhý I_2 . Integrál I_1 počítáme metodou per partes a využijeme vlastnosti funkce $W(t)$

$$\begin{aligned}
 I_1 &= h \int_{-\infty}^1 W(t)f(x - ht) dt \\
 &= \left| \begin{array}{l} u = W(t) \quad u' = W'(t) = K(t) \\ v' = f(x - ht)h \quad v = -F(x - ht) \end{array} \right| \\
 &= [-W(t)F(x - ht)]_{-1}^1 + \int_{-1}^1 F(x - ht)W'(t) dt \\
 &= -F(x - h) + \int_{-1}^1 K(t)F(x - ht) dt. \tag{4.2}
 \end{aligned}$$

Dále použijeme Taylorův rozvoj $F(x - ht) = F(x) - htF'(x) + \frac{h^2t^2}{2}F''(x) + o(h^2)$, tedy

$$= -F(x - h) + F(x) + \frac{1}{2}F''(x)h^2\beta_2(K) + o(h^2).$$

Počítejme nyní integrál I_2 :

$$I_2 = h \int_1^{\infty} W(t)f(x - ht) dt,$$

uvažujeme-li substituce $x - ht = z$, dostaneme

$$= - \int_{x-h}^{-\infty} f(z) dz = \int_{-\infty}^{x-h} f(z) dz = F(x - h). \tag{4.3}$$

Vychýlení odhadu je tedy tvaru

$$\text{bias } \widehat{F}(x, h) = \frac{1}{2}F''(x)h^2\beta_2(K) + o(h^2).$$

Poznámka 4.1. Vztahy (4.2) a (4.3) dávají zajímavý vztah pro vychýlení

$$E\widehat{F}(x, h) - F(x) = -F(x - h) + \int_{-1}^1 K(t)F(x - ht) dt + F(x - h) - F(x).$$

Odtud plyne

$$E\widehat{F}(x, h) = \int_{-1}^1 K(t)F(x - ht) dt.$$

A dále (z Taylorova vzorce)

$$\begin{aligned} E\widehat{F}(x, h) &= \int_{-1}^1 K(t) \left(F(x) - htF'(x) + \frac{1}{2}h^2t^2F''(x) + o(h^2) \right) dt \\ &= F(x) + \underbrace{\frac{1}{2}h^2\beta_2(K)F''(x) + o(h^2)}_{=o(h)} \\ &= F(x) + o(h). \end{aligned}$$

Nyní dokážeme tvar rozptylu.

$$\text{var } \widehat{F}(x, h) = \frac{1}{n} \left(EW^2 \left(\frac{x-X}{h} \right) - E^2W \left(\frac{x-X}{h} \right) \right).$$

Zde $E^2W \left(\frac{x-X}{h} \right) = \left(EW \left(\frac{x-X}{h} \right) \right)^2 = (F(x) + o(h))^2$. Počítáme tedy jen integrál (označme jej I_3):

$$\begin{aligned} I_3 &= \frac{1}{n} \int_{-\infty}^{\infty} W^2 \left(\frac{x-y}{h} \right) f(y) dy \\ &= |\text{substituce: } x-y=th| \\ &= \frac{1}{n} \left(\int_{-\infty}^1 W^2(t)f(x-ht) dt + \underbrace{h \int_1^{\infty} f(x-ht) dt}_{=F(x-h)} \right). \end{aligned}$$

První integrál počítáme metodou per partes a máme

$$\begin{aligned} I_3 &= \frac{1}{n} \left[-F(x-ht)W^2(t) \right]_{-1}^1 + \frac{2}{n} \int F(x-ht)W(t)W'(t) dt + \frac{1}{n}F(x-h) \\ &= -\frac{1}{n}F(x-h) + \frac{2}{n} \int F(x-ht)W(t)K(t) dt + \frac{1}{n}F(x-h), \end{aligned}$$

použijeme nyní Taylorův rozvoj funkce F

$$\begin{aligned} &= \frac{2}{n} \int W(t)K(t) (F(x) - htF'(x) + o(h)) dt \\ &= \frac{2}{n}F(x) \underbrace{\int_{-1}^1 W(t)K(t) dt}_{=\frac{1}{2} \text{ (vlastnost 3)}} - \frac{2}{n}hF'(x) \underbrace{\int_{-1}^1 tW(t)K(t) dt}_{=\frac{1}{2}(1-\int_{-1}^1 W^2(x) dx) \text{ (vlastnost 4)}} + o\left(\frac{h}{n}\right), \end{aligned}$$

užitím vlastností funkce W a $F'(x) = f(x)$ dostaneme

$$= \frac{1}{n} \left[F(x) - hf(x) \left(1 - \int_{-1}^1 W^2(t) dt \right) \right] + o\left(\frac{h}{n}\right).$$

Rozptyl je tedy tvaru

$$\begin{aligned} \text{var } \widehat{F}(x, h) &= \frac{1}{n} \left[F(x) - hf(x) \left(1 - \int_{-1}^1 W^2(t) dt \right) \right] + o\left(\frac{h}{n}\right) - \frac{1}{n} (F(x) + o(h))^2 \\ &= \frac{1}{n} F(x)(1-F(x)) - \frac{h}{n} f(x) \left(1 - \int_{-1}^1 W^2(t) dt \right) + o\left(\frac{h}{n}\right). \end{aligned}$$

Výše uvedené výsledky můžeme nyní zformulovat v následující větě:

Věta 4.1. Necht $F \in C^2$, $h \rightarrow 0$, $nh \rightarrow \infty$ pro $n \rightarrow \infty$. Pak

$$\begin{aligned} \text{MSE } \widehat{F}(x, h) &= \frac{1}{n} F(x)(1 - F(x)) - \frac{1}{n} h f(x) \left(1 - \int_{-1}^1 W^2(t) dt \right) \\ &+ \frac{1}{4} (F''(x))^2 h^4 \beta_2^2(K) + o\left(\frac{h}{n} + h^4\right). \end{aligned} \quad (4.4)$$

Globální pohled na kvalitu odhadu lze získat prostřednictvím střední integrální kvadratické chyby (MISE).

Věta 4.2. Necht $F \in C^2$, $V(F'') = \int (F''(x))^2 dx < \infty$, $K \in S_{02}$, $\lim_{n \rightarrow \infty} h = 0$ a $\lim_{n \rightarrow \infty} nh = \infty$. Pak

$$\text{MISE } \widehat{F}(\cdot, h) = \frac{1}{n} \int F(x)(1 - F(x)) dx - c_1 \frac{h}{n} + c_2 h^4 + o\left(\frac{h}{n} + h^4\right), \quad (4.5)$$

kde

$$c_1 = 1 - \int_{-1}^1 W^2(t) dt, \quad c_2 = \frac{1}{4} \beta_2^2(K) V(F'').$$

Naším cílem je nalézt takovou hodnotu vyhlazovacího parametru, pro kterou bude MISE nabývat minimální hodnoty. Ale uvedený tvar MISE není pro takovou analýzu vhodný, a proto (stejně jako při odhadu hustoty a regresní funkce) budeme uvažovat asymptotickou střední integrální kvadratickou chybu AMISE, která v tomto případě je tvaru:

$$\text{AMISE } \widehat{F}(\cdot, h) = \underbrace{\frac{1}{n} \int F(x)(1 - F(x)) dx}_{\text{AIV}(h)} - c_1 \frac{h}{n} + \underbrace{c_2 h^4}_{\text{AISB}(h)}. \quad (4.6)$$

Nyní už lze standardními metodami matematické analýzy nalézt takovou hodnotu h , pro kterou AMISE nabývá minimální hodnoty. Je snadné ukázat, že

$$h_{opt,0,2} = n^{-1/3} \left(\frac{c_1}{4c_2} \right)^{1/3} = O(n^{-1/3}) \quad (4.7)$$

a pak

$$\text{AMISE } \widehat{F}(\cdot, h_{opt,0,2}) = \frac{1}{n} \int F(x)(1 - F(x)) dx - \frac{3}{c_2^{1/3}} \left(\frac{c_1}{4} \right)^{4/3} n^{-4/3}. \quad (4.8)$$

Poznámka 4.2. Optimální hodnota vyhlazovacího parametru pro odhad distribuční funkce je řádu $O(n^{-1/3})$, zatímco pro odhad hustoty s jádrem $K \in S_{02}$ je vyhlazovací parametr řádu $O(n^{-1/5})$.

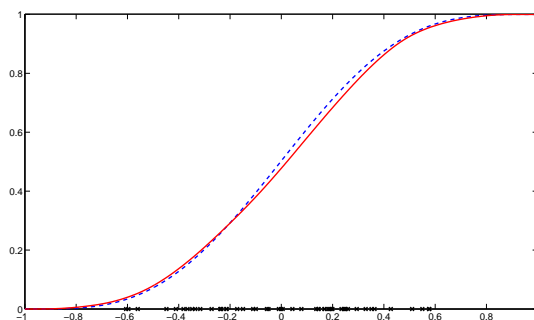
Ukázkový příklad 4.3. Předpokládejme, že známe tvar distribuční funkce $F(x) = \frac{1}{32}(-5x^7 + 21x^5 - 35x^3 + 35x + 16)$ pro $x \in [-1, 1]$. Vypočítejme hodnotu optimálního vyhlazovacího parametru pro odhad s jádrem řádu 2.

Podle vztahu (4.7) potřebujeme spočítat hodnoty c_1 a c_2 . S Epanečnickovým jádrem je

$$\begin{aligned} c_1 &= 1 - \int_{-1}^1 W^2(x) dx = 1 - \int_{-1}^1 \frac{1}{16} (-x^3 + 3x + 2)^2 dx = 0,2571, \\ c_2 &= \frac{1}{4} \beta_2^2(K) V(F'') = \frac{1}{4} \cdot \frac{1}{25} \cdot 3,1818. \end{aligned}$$

Pak $h_{opt,0,2} = 1,2642 \cdot n^{-1/3}$.

Na obrázku 4.5 je odhad s optimálním vyhlazovacím parametrem pro náhodný výběr o 50 pozorování, která pochází z rozdělení s uvedenou distribuční funkcí (data jsou v tabulce 7.4).



Obrázek 4.5: Odhad distribuční funkce z ukázkového příkladu 4.3, odhad (červená, plná) a původní funkce (modrá, čárkovaná) za použití Epanečnikova jádra a $h_{opt,0,2} = 0,3432$

4 Volba jádra

I v tomto případě je volba jádra méně důležitá než volba vyhlazovacího parametru. Lze doporučit jádra třídy S_{02} , např.

- Epanečnikovo $K(x) = \frac{3}{4}(1 - x^2)I_{[-1,1]}(x)$,
- kvartické $K(x) = \frac{15}{16}(1 - x^2)^2I_{[-1,1]}(x)$,
- triweight $K(x) = \frac{35}{32}(1 - x^2)^3I_{[-1,1]}(x)$.

5 Volba vyhlazovacího parametru

5.1 Metody křížového ověřování

Metody křížového ověřování patří k nejužívanějším metodám pro volbu vyhlazovacího parametru. Zde uvedeme pouze metodu navrženou A. Bowmanem (1998). Funkce křížového ověřování je v tomto případě tvaru

$$CV(h) = \frac{1}{n} \sum_{i=1}^n \int \left(I_{(-\infty, x]}(X_i) - \widehat{F}_{-i}(x, h) \right)^2 dx,$$

kde $\widehat{F}_{-i}(x, h)$ je jádrový odhad distribuční funkce s vynecháním bodu X_i . Pak

$$h_{CV} = \arg \min_{h \in H_n} CV(h),$$

přičemž $H_n = [an^{-1/3}, bn^{-1/3}]$ pro vhodná $0 < a < b < \infty$.

5.2 Princip maximálního vyhlazení

Myšlenka této metody je stejná jako pro odhad hustoty. Užijeme-li faktu, že

$$\int (F''(x))^2 dx = \int (f'(x))^2 dx,$$

můžeme aplikovat Terrelovu větu 3.3 pro $k = 1$. V tomto případě je

$$g_1(x) = \frac{15}{16}(1 - x^2)^2I_{[-1,1]}(x),$$

a tedy

$$h_{opt,0,2} = n^{-1/3} \left(\frac{c_1}{\beta_2^2(K)V(f')} \right)^{1/3} \leq n^{-1/3} \left(\frac{c_1}{\beta_2^2(K)} \right)^{1/3} \frac{\sigma}{\sigma_1} V(g'_1)^{-1/3},$$

kde $\sigma_1 = \int x^2 g_1(x) dx = \frac{1}{7}$, $V(g'_1) = \frac{15}{7}$. Odtud plyne, že

$$h_{MS} = n^{-1/3} \left(\frac{7c_1}{15\beta_2^2(K)} \right)^{1/3} \sqrt{7}\hat{\sigma}, \quad (4.9)$$

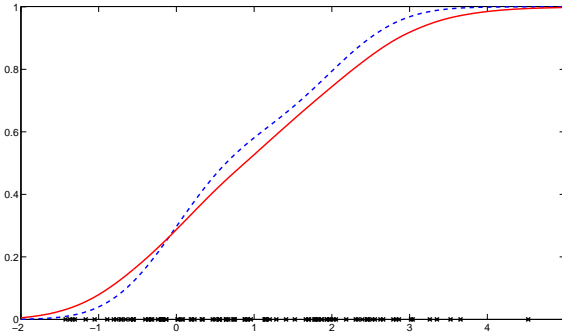
$\hat{\sigma}$ je odhadem σ (viz rovnice (3.13) a (3.14)).

Poznámka 4.3. Hodnota h_{MS} může sloužit jako horní hranice pro množinu vyhlazovacích parametrů volených podle metody křížového ověřování. Tedy $H_n = [h_\ell, h_{MS}]$, kde h_ℓ je nejmenší vzdálenost mezi po sobě jdoucími body $X_i, i = 1, \dots, n$.

Příklad 4.4. Pro data z příkladu 4.1 zvolme Epanečnikovo jádro. Pak hodnoty potřebné pro odhad vyhlazovacího parametru metodou maximálního vyhlazení jsou následující:

$$n = 100, \quad \hat{\sigma} = 1,3426, \quad \beta_2(K) = \frac{1}{5}, \quad c_1 = 1 - \int_{-1}^1 W^2(x) dx = 0,2571.$$

Pak platí $h_{MS} = 1,1037$ a na obrázku 4.6 je zobrazen odhad distribuční funkce.



Obrázek 4.6: Odhad distribuční funkce s $h_{MS} = 1,1037$, odhad (červená, plná), původní funkce (modrá, čárkovaná)

5.3 Plug-in metoda

Společným cílem metod typu plug-in (PI) je odhadnout $V(F'')$. Za předpokladu dostatečné hladkosti funkce f užitím metody per partes dostaneme vztah

$$V(F'') = \int (F''(x))^2 dx = - \int f''(x)f(x) dx.$$

Tudíž se budeme dále zabývat odhadem funkcionálu

$$\psi_1 = \int f''(x)f(x) dx.$$

Je zřejmé, že $\psi_1 = Ef''(X)$, což vede k metodě založené na odhadu druhé derivace hustoty f . Vztah (3.9) použijeme k odhadu druhé derivace s jádrem $K^{(2)} = K_{opt,2,4} \in S_{24}$. Pak

$$\hat{\psi}_1 = n^{-1} \sum_{i=1}^n \hat{f}''(X_i, h) = n^{-2} h^{-3} \sum_{i=1}^n \sum_{j=1}^n K^{(2)} \left(\frac{X_i - X_j}{h} \right),$$

kde podle vztahu (3.11) je

$$h_{opt,2,4} = 10^{1/9} \frac{\delta_{24}}{\delta_{04}} h_{opt,0,4}.$$

Pak

$$\hat{c}_2 = -\frac{1}{4} \beta_2^2(K) \hat{\psi}_1.$$

Shrnutím předchozích úvah dostaneme proceduru pro odhad distribuční funkce F :

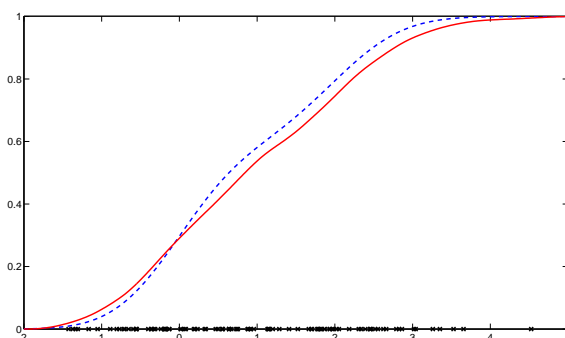
Procedura

1. Najděte optimální vyhlazovací parametr $\hat{h}_{opt,0,4}$ pro odhad hustoty s optimálním jádrem $K_{opt,0,4} \in S_{04}$.
2. Najděte optimální vyhlazovací parametr $\hat{h}_{opt,2,4}$ pro odhad druhé derivace hustoty podle vztahu (3.11) s $k = 4$ a optimálním jádrem $K^{(2)} \in S_{24}$.
3. Vypočítejte odhad funkcionálu $\hat{\psi}_1$ s využitím hodnoty $\hat{h}_{opt,2,4}$ získané v kroku 2.
4. Vyčíslete optimální hodnotu vyhlazovacího parametru

$$h_{PI} = n^{-1/3} \left(\frac{c_1}{-\hat{\psi}_1 \beta_2^2(K)} \right)^{1/3}$$

5. Použijte parametry z předchozích kroků ke konstrukci optimálního jádrového odhadu distribuční funkce $\hat{F}(x, h)$ s daným jádrem $K \in S_{02}$.

Příklad 4.5. S použitím funkce toolboxu zjistíme, že pro data z příkladu 4.1 je vyhlazovací parametr určený plug-in metodou roven $h_{PI} = 0,5717$. Na obrázku 4.7 je odhad distribuční funkce společně se skutečnou distribuční funkcí.

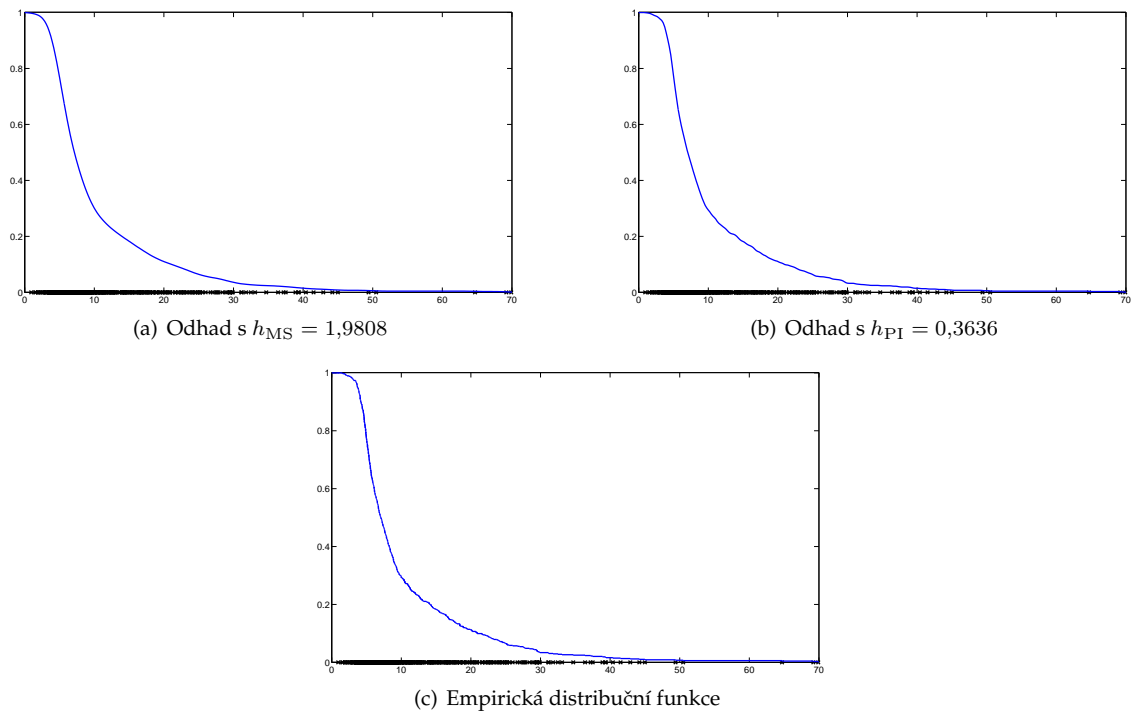


Obrázek 4.7: Odhad distribuční funkce s $h_{PI} = 0,5717$, odhad (červená, plná), původní funkce (modrá, čárkovaná)

6 Aplikace na reálná data

Znovu se podívejme na soubor dat z povodí řeky Dyje (viz tabulky 7.13 a 7.14). Podle metod pro odhad vyhlazovacích parametrů jsme pro Epanečnikovo jádro vypočítali tyto odhady

$$h_{MS} = 1,9808, \quad h_{PI} = 0,3636.$$



Obrázek 4.8: Odhadnuté křivky zabezpečení $\bar{F}(x)$ řeky Dyje v období 1931–1990, na ose x je velikost průtoku v m^3/s

Na obrázku 4.8 jsou zobrazeny odhady křivky zabezpečení ($\bar{F}(x) = 1 - F(x)$) společně s empirickou křivkou zabezpečení. Průběh křivky zabezpečení ukazuje například, že průtok $8 m^3/s$ a vyšší se vyskytuje s pravděpodobností asi 0,3, průtok $20 m^3/s$ a vyšší se vyskytuje s pravděpodobností 0,1. Provoz přehrady tedy stabilizuje měsíční průtok v rozmezí $5 - 8 m^3/s$ s pravděpodobností 0,9 – 0,3.

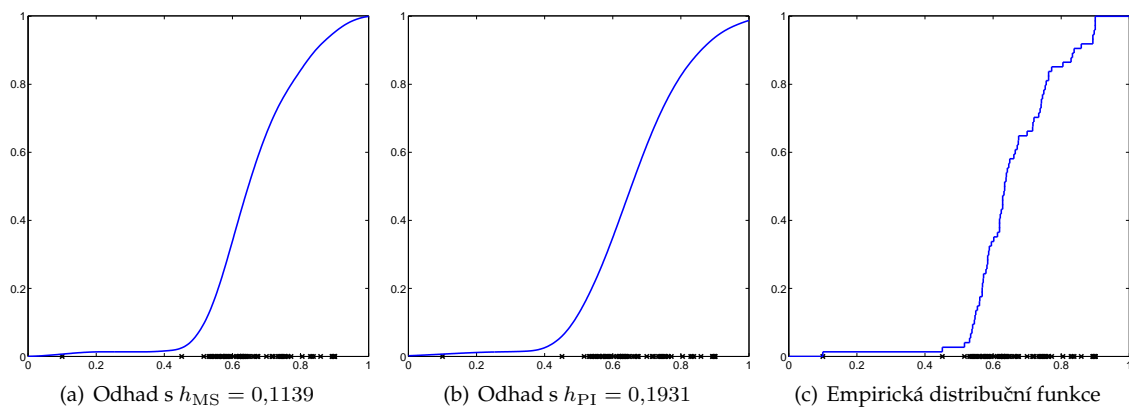
Následující datový soubor pochází z rozsáhlé studie, v níž autoři studovali vliv substituentů v 2,4-diamino-5-(substituovaný benzylo)pyrimidinech. Biologická aktivita při inhibici dihydrofolát reduktázy byla měřena pomocí asociační konstanty. Data jsou v tabulce 7.12 a jsou dostupná na osobních stránkách Dennise D. Boose², kde je také odkaz na původní článek Jonathana D. Hirsta z roku 1994.

Užitím výše uvedených metod jsme (při použití Epanečnikova jádra) dostali následující hodnoty vyhlazovacích parametrů:

$$h_{MS} = 0,1139, \quad h_{PI} = 0,1931.$$

Na obrázku 4.9 jsou uvedeny odhady distribuční funkce s těmito parametry a také je zde pro srovnání uvedena empirická distribuční funkce.

² <http://www4.stat.ncsu.edu/~boos/var.select/pyrimidine.html>



Obrázek 4.9: Odhadnuté distribuční funkce pro vliv substituentů v pyrimidinech

Shrnutí
<p>Odhad distribuční funkce $F(x)$ v bodě x je tvaru</p> $\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^n W\left(\frac{x - X_i}{h}\right), \quad W(x) = \int_{-\infty}^x K(t) dt.$
<p>Asymptotická střední kvadratická chyba jádrového odhadu distribuční funkce je součtem asymptotického tvaru rozptylu (AIV) a druhé mocniny vychýlení (AISB)</p> $\text{AMISE } \widehat{F}(\cdot, h) = \underbrace{\frac{1}{n} \int F(x)(1 - F(x)) dx}_{\text{AIV}(h)} - c_1 \frac{h}{n} + \underbrace{c_2 h^4}_{\text{AISB}(h)},$ <p>kde</p> $c_1 = 1 - \int_{-1}^1 W^2(t) dt, \quad c_2 = \frac{1}{4} \beta_2^2(K) V(F'').$
<p>Optimální vyhlazovací parametr vzhledem k AMISE pro odhad distribuční funkce je tvaru</p> $h_{opt,0,2} = n^{-1/3} \left(\frac{c_1}{4c_2} \right)^{1/3},$ <p>t.j. $h_{opt,0,2} = O(n^{-1/3})$.</p>
<p>Metody pro odhad optimální hodnoty vyhlazovacího parametru h</p> <ul style="list-style-type: none"> metoda křížového ověřování $h_{CV} = \arg \min_{h \in H_n} CV(h)$ $CV(h) = \frac{1}{n} \sum_{i=1}^n \int \left(I_{(-\infty, x]}(X_i) - \widehat{F}_{-i}(x, h) \right)^2 dx,$ <ul style="list-style-type: none"> metoda maximálního vyhlazení $h_{MS} = n^{-1/3} \left(\frac{7c_1}{15\beta_2^2(K)} \right)^{1/3} \sqrt{7\widehat{\sigma}},$ <ul style="list-style-type: none"> plug-in metoda $h_{PI} = n^{-1/3} \left(\frac{c_1}{-\widehat{\psi}_1 \beta_2^2(K)} \right)^{1/3}.$

Výstupy z výukové jednotky

Student

- zná základní typy jádrových odhadů distribuční funkce a jejich statistické vlastnosti
- získal přehled o metodách pro volbu vyhlazovacího parametru
- je schopen navrhnout a implementovat proceduru pro zpracování reálných dat
- se naučil používat příslušný toolbox v Matlabu a dokáže zkonstruovat jádrový odhad distribuční funkce pro daná reálná data

Cvičení

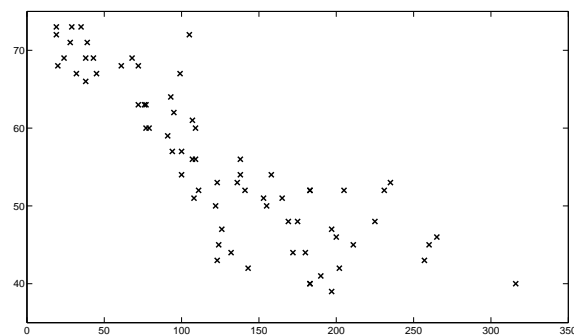
1. Odvoďte tvar funkce $W(x)$ pro kvartické jádro $K(x) = \frac{15}{16}(1 - x^2)^2 I_{[-1,1]}(x)$.
2. Dokažte vlastnosti 2, 3 a 4 funkce W .
3. Dokažte vztahy (4.7) a (4.8).
4. Odvoďte tvar vyhlazovacího parametru podle metody maximálního vyhlazení pro kvartické jádro.
5. Aplikujte metodu maximálního vyhlazení a plug-in metodu na simulovaná data z ukázkového příkladu 4.3.

Kapitola 5

Jádrové odhady dvourozměrných hustot

1 Motivace

Uvažujme soubor dat ze studie The State of the World's Children (UNICEF), který obsahuje míru úmrtnosti dětí od narození do pěti let věku počítanou na 1000 živě narozených dětí a očekávanou délku života narozeného dítěte (s ohledem na úmrtnost v dané populaci v době jeho narození) v 72 zemích, které měly v roce 2001 hrubý národní produkt menší než 1000 amerických dolarů na osobu a rok. Chceme vědět, jaká je pravděpodobnost, že dítě narozené v zemi s vysokou



Obrázek 5.1: Studie UNICEF – míra úmrtnosti (osa x) a očekávaná délka života u dětí (osa y)

mírou úmrtnosti, se dožije vyššího věku. Případně, zda se v datech nevyskytují shluky, tj. zda jde o vícemodální hustotu. V tomto případě se jedná o odhad dvourozměrné hustoty.

Odhady vícerozměrných hustot se budeme zabývat v této kapitole. Ovšem ve vícerozměrném případě nevystačíme s jedním vyhlazovacím parametrem, ale je třeba specifikovat matici vyhlazovacích parametrů. Tato matice řídí jak hladkost, tak i orientaci vícerozměrného vyhlazení. Budeme se zabývat jádrovým odhadem, který je přímým rozšířením jednorozměrného odhadu (3.1) z kapitoly 3, a zaměříme se zejména na odhad dvourozměrné hustoty. Tedy naším cílem je rekonstruovat hustotu pravděpodobnosti z náhodného výběru.

Poznámka 5.1. Jádrové odhady dvourozměrných hustot se obvykle znázorňují pomocí vrstevnic, které umožňují snazší náhled na odhadnutou funkci.

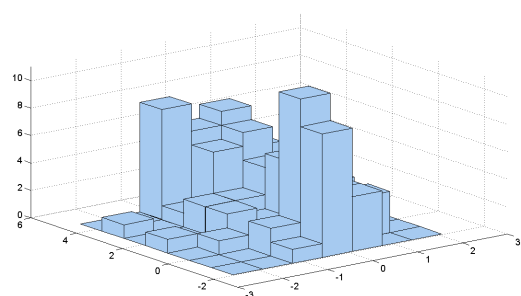
2 Základní typy odhadů

Podobně jako u odhadů hustoty můžeme použít *histogram*, ale ten má zmíněné nevýhody – jde o schodovitou funkci a je citlivý na volbu počtu a šířky třídících obdélníků – viz obrázek 5.2.

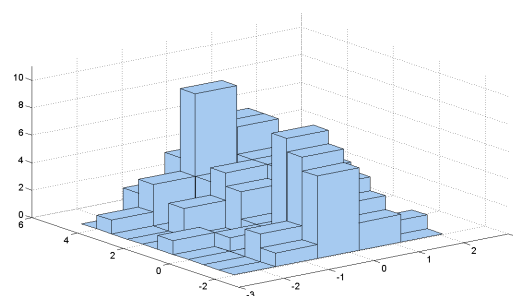
Příklad 5.1. Mějme dán datový soubor o velikosti $n = 100$ generovaný ze směsi tří normálních hustot¹. $N(0, -1; 1/3, 1/3, 0)$, $N(0, 2; 1, 1, 0)$ a $N(0, 4; 1/3, 1/3, 0)$ (Data jsou v tabulce 7.6.)

$$f(x, y) = \frac{1}{3} \frac{3}{2\pi} e^{-\frac{3}{2}(x^2+(y+1)^2)} + \frac{1}{3} \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+(y-2)^2)} + \frac{1}{3} \frac{3}{2\pi} e^{-\frac{3}{2}(x^2+(y-4)^2)}$$

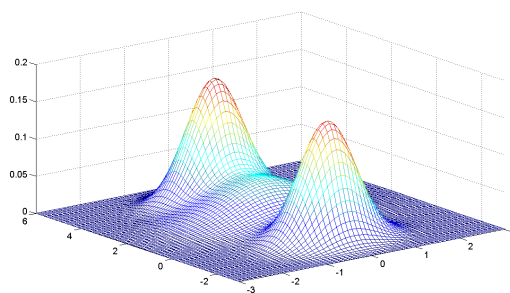
Z obrázků 5.2(a) a 5.2(b) je patrné, že histogram nepostihuje charakteristické rysy hustoty pravděpodobnosti dat, která je zobrazena na obrázku 5.2(c).



(a) 7×7 obdélníků



(b) 5×10 obdélníků



(c) Hustota simulovaných dat

Obrázek 5.2: Histogramy s různými počty třídících obdélníků a hustota pro data z příkladu 5.1

Přejdeme nyní k jádrovým odhadům dvourozměrné hustoty. Předpokládejme, že máme k dispozici náhodný výběr $([X_1, Y_1], \dots, [X_n, Y_n])$ z dvourozměrného spojitého rozdělení s hustotou $f(x, y)$. Jádrový odhad hustoty f v bodě $[x, y] \in \mathbb{R}^2$ je definovaný vztahem

$$\hat{f}(x, y, \mathbf{H}) = \frac{1}{n} \sum_{i=1}^n K_{\mathbf{H}}(x - X_i, y - Y_i) = \frac{1}{n|\mathbf{H}|} \sum_{i=1}^n K(\mathbf{H}^{-1}(x - X_i, y - Y_i)^T), \quad (5.1)$$

K je dvourozměrné jádro a \mathbf{H} je pozitivně definitní matice typu 2×2 a $|\mathbf{H}|$ značí její determinant. Prvky matice \mathbf{H} se nazývají vyhlazovací parametry a matice se pak zkráceně označuje jako *vyhlazovací matice*.

Jádro K je dvourozměrná funkce, kterou můžeme získat pomocí jednorozměrného symetrického jádra K_1 ($K_1 \in S_{02}$). Existují dva typy těchto jader:

- *součinnové jádro* $K^P(x, y) = K_1(x) \cdot K_1(y)$,

¹Používáme zde zkrácený zápis pro dvourozměrnou hustotu normálního rozdělení, a to $N(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2, \rho)$

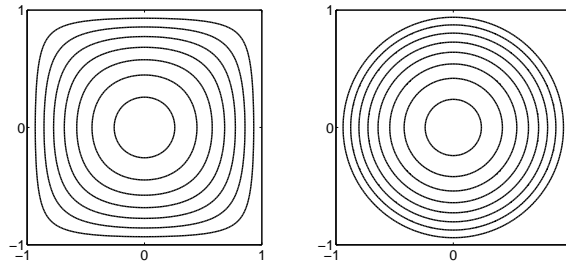
- *sféricky symetrické jádro* $K^S(x, y) = c_k^{-1} K_1(\sqrt{x^2 + y^2})$, $c_k = \iint K_1(\sqrt{x^2 + y^2}) dx dy$.

Příklad 5.2. Epanečnikovo jádro, které je v jednorozměrném případě tvaru $K(x) = \frac{3}{4}(1 - x^2)$, má následující dvourozměrné varianty

$$K^P(x, y) = \frac{9}{16}(1 - x^2)(1 - y^2) \quad \text{pro} \quad -1 \leq x, y \leq 1,$$

$$K^S(x, y) = \frac{2}{\pi}(1 - x^2 - y^2) \quad \text{pro} \quad x^2 + y^2 \leq 1.$$

Na obrázku 5.3 jsou zobrazeny vrstevnice těchto jader.



Obrázek 5.3: Součinnové (vlevo) a sféricky symetrické (vpravo) dvourozměrné Epanečnikovo jádro

Poznámka 5.2. V praxi se často používá Gaussovo jádro

$$K(x, y) = \frac{1}{2\pi} e^{-(x^2 + y^2)/2}, \quad [x, y] \in \mathbb{R}^2.$$

Kromě jeho výhod při výpočtech statistických vlastností odhadu, má navíc zajímavou vlastnost, a to, že je součinnovým i sféricky symetrickým jádrem.

Podívejme se blíže na matici \mathbf{H} . Jde o matici vyhlazovacích parametrů, které řídí hladkost výsledného odhadu. Navíc také udávají orientaci odhadnuté hustoty. Rozlišujeme tři základní třídy vyhlazovacích matic:

- třída \mathcal{S} , která obsahuje matice s jediným vyhlazovacím parametrem,
- třída \mathcal{D} , která zahrnuje diagonální matice,
- třída \mathcal{F} , která obsahuje tzv. plné matice.

Rozdíly mezi jednotlivými maticemi jsou patrné z tabulky 5.1, kde jsou zobrazeny vrstevnice sféricky symetrického Epanečnikova jádra v závislosti na třídě matic.

Budeme se zabývat jádrovými odhady s diagonální vyhlazovací maticí. Jádrový odhad s maticí třídy \mathcal{S} dává ve všech směrech stejnou míru vyhlazení, což neponechává příliš mnoho prostoru pro zachycení variability dat. Na druhou stranu při použití matice třídy \mathcal{F} je potřeba odhadnout větší počet parametrů, což znamená vyšší výpočetní náročnost.

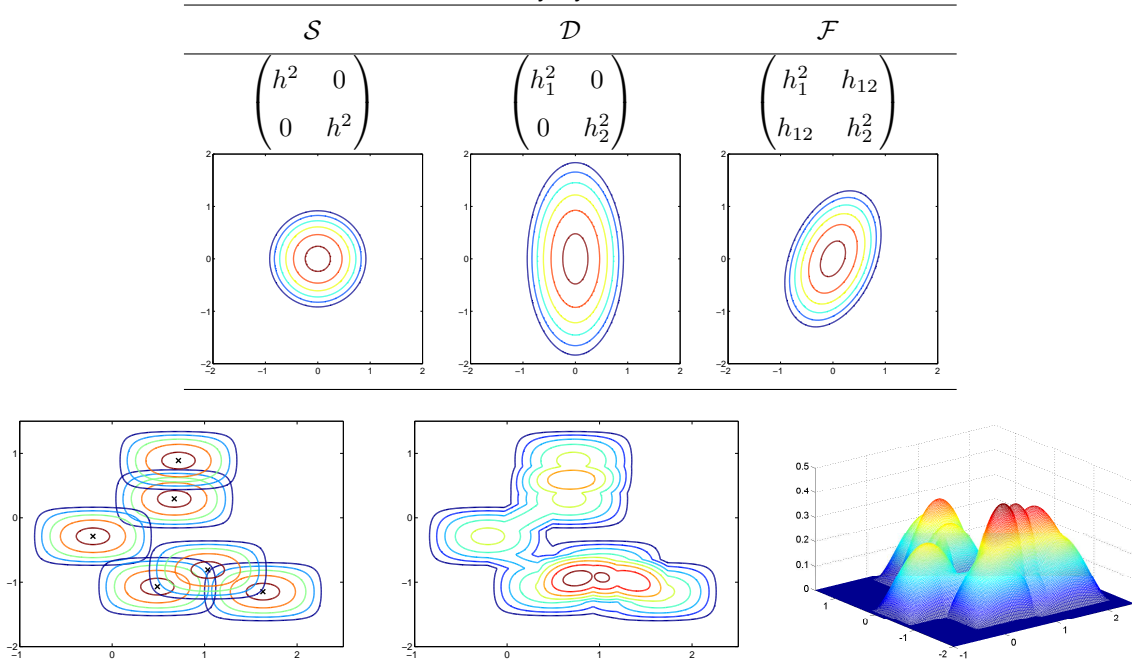
Konstrukce jádrového odhadu je analogická konstrukci jednorozměrného odhadu. Tedy v každém bodě $[X_i, Y_i]$ sestrojíme jádro $K_{\mathbf{H}}$ a odhad v bodě $[x, y]$ je průměr n hodnot jader v tomto bodě – viz obrázek 5.4.

3 Statistické vlastnosti jádrových odhadů hustoty

Stejně jako u jádrových odhadů jednorozměrných hustot můžeme kvalitu jádrového odhadu hustoty popsat lokálně pomocí střední kvadratické chyby:

$$\text{MSE } \hat{f}(x, y, \mathbf{H}) = \underbrace{\frac{1}{n} \left((K_{\mathbf{H}}^2 * f)(x, y) - (K_{\mathbf{H}} * f)^2(x, y) \right)}_{\text{var}} + \underbrace{\left((K_{\mathbf{H}} * f)(x, y) - f(x, y) \right)^2}_{\text{bias}},$$

Tabulka 5.1: Třídy vyhlazovacích matic



Obrázek 5.4: Konstrukce jádrového odhadu hustoty

nebo globálně pomocí střední integrální kvadratické chyby

$$\text{MISE } \hat{f}(\cdot, \mathbf{H}) = \iint \text{MSE } \hat{f}(x, y, \mathbf{H}) \, dx \, dy.$$

Optimální vyhlazovací matice minimalizuje MISE. Je zřejmé, že tyto optimální hodnoty vyhlazovacích parametrů není možné z MISE přímo vyjádřit. Stejně jako u odhadu jednorozměrných hustot se budeme zabývat asymptotickou střední integrální kvadratickou chybou AMISE.

Výpočty pro matici třídy \mathcal{F} jsou velmi náročné, a proto se v dalších úvahách omezíme na odhady s diagonální maticí.

Věta 5.1. Předpokládejme, že funkce f , jádro K a diagonální matice vyhlazovacích parametrů $\mathbf{H} = \text{diag}(h_1^2, h_2^2) = \begin{pmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{pmatrix}$ splňují následující předpoklady:

- (i) Necht $\mathbf{H} = \mathbf{H}_n$ je posloupnost matic takových, že $(nh_1^2 h_2^2)^{-1}$ a prvky h_1^2 a h_2^2 konvergují k nule pro $n \rightarrow \infty$.
- (ii) Dále necht všechny druhé parciální derivace funkce f jsou ohraničené, spojitě a integrovatelné se čtvercem.
- (iii) Jádro K splňuje

$$\begin{aligned} \iint xK(x, y) \, dx \, dy &= \iint yK(x, y) \, dx \, dy = 0, \\ \iint x^2 K(x, y) \, dx \, dy &= \iint y^2 K(x, y) \, dx \, dy = \beta_2(K). \end{aligned}$$

Pak platí

$$\text{MISE } \hat{f}(\cdot, \mathbf{H}) = \text{AMISE } \hat{f}(\cdot, \mathbf{H}) + o(h_1^4 + h_2^4) + o((h_1 h_2 n)^{-1}),$$

kde

$$\text{AMISE } \hat{f}(\cdot, \mathbf{H}) = \underbrace{\frac{V(K)}{nh_1h_2}}_{\text{AIV}(\mathbf{H})} + \underbrace{\frac{1}{4}\beta_2^2(K)(h_1^4V(f_{xx}) + 2h_1^2h_2^2V(f_{xy}) + h_2^4V(f_{yy}))}_{\text{AISB}(\mathbf{H})}, \quad (5.2)$$

přičemž označení je ve shodě s předchozími kapitolami, tj. $V(g) = \iint g^2(x, y) dx dy$.

Důkaz věty o tvaru AMISE je založen na Taylorově rozvoji funkce $f(x, y)$ a lze jej nalézt např. v knize [14].

Hodnoty parametrů h_1, h_2 , pro které AMISE nabývá minimální hodnoty, jsou dány vztahy:

$$h_{1,opt} = \left(\frac{V^{3/4}(f_{yy})V(K)}{n\beta_2^2(K)V^{3/4}(f_{xx})[V(f_{xy}) + V^{1/2}(f_{xx})V^{1/2}(f_{yy})]} \right)^{1/6}, \quad (5.3)$$

$$h_{2,opt} = \left(\frac{V^{3/4}(f_{xx})V(K)}{n\beta_2^2(K)V^{3/4}(f_{yy})[V(f_{xy}) + V^{1/2}(f_{xx})V^{1/2}(f_{yy})]} \right)^{1/6}. \quad (5.4)$$

Z těchto vztahů plyne, že množina přípustných vyhlazovacích parametrů je tvaru $a_1n^{-1/6} \leq h_1 \leq b_1n^{-1/6}$, $a_2n^{-1/6} \leq h_2 \leq b_2n^{-1/6}$ pro vhodné konstanty $0 < a_1 < b_1 < \infty$, $0 < a_2 < b_2 < \infty$.

Ukázkový příklad 5.3. Předpokládejme, že známe tvar hustoty $f(x, y) = \frac{3}{2\pi} e^{-3(x^2+y^2)/2}$ pro $[x, y] \in \mathbb{R}^2$. Vypočítejme hodnoty optimálních vyhlazovacích parametrů pro odhad se součinným Epanečnickovým jádrem.

Podle vztahů (5.3) a (5.4) potřebujeme spočítat výrazy²

$$f(x, y) = \frac{3}{2\pi} e^{-3(x^2+y^2)/2},$$

$$V(f_{xx}) = \iint f_{xx}^2(x, y) dx dy = \iint f_{xxxx}(x, y)f(x, y) dx dy = 3^4 \cdot (2\pi)^{-1},$$

$$V(f_{xy}) = \iint f_{xy}^2(x, y) dx dy = \iint f_{xxyy}(x, y)f(x, y) dx dy = 3^3 \cdot (2\pi)^{-1},$$

$$V(f_{yy}) = \iint f_{yy}^2(x, y) dx dy = \iint f_{yyyy}(x, y)f(x, y) dx dy = 3^4 \cdot (2\pi)^{-1}.$$

Pro součinné Epanečnickovo jádro $K(x, y) = \frac{9}{16}(1-x^2)(1-y^2)$ platí

$$V(K) = \iint K^2(x, y) dx dy = 0,36,$$

$$\beta_2(K) = \iint x^2K(x, y) dx dy = \iint y^2K(x, y) dx dy = 0,2.$$

Pak pro optimální vyhlazovací parametry máme

$$h_{1,opt}^6 = \frac{(3^4 \cdot (2\pi)^{-1})^{3/4} \cdot 0,36}{n \cdot 0,2^2(3^4 \cdot (2\pi)^{-1})^{3/4} [3^3 \cdot (2\pi)^{-1} + (3^4 \cdot (2\pi)^{-1})^{1/2}(3^4 \cdot (2\pi)^{-1})^{1/2}]},$$

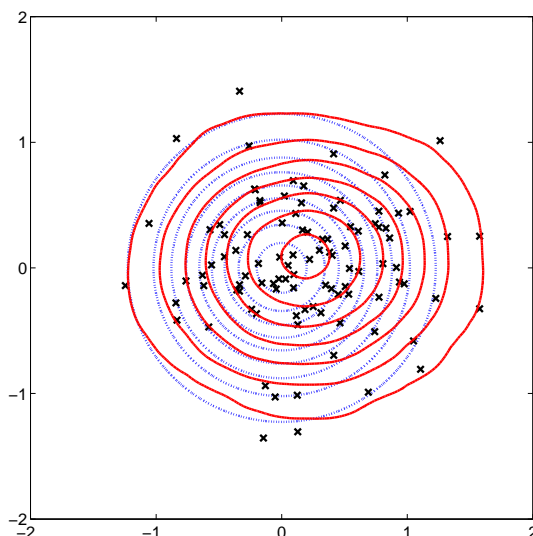
$$h_{2,opt}^6 = \frac{(3^4 \cdot (2\pi)^{-1})^{3/4} \cdot 0,36}{n \cdot 0,2^2(3^4 \cdot (2\pi)^{-1})^{3/4} [3^3 \cdot (2\pi)^{-1} + (3^4 \cdot (2\pi)^{-1})^{1/2}(3^4 \cdot (2\pi)^{-1})^{1/2}]}.$$

Tedy $h_1 = 0,8978 \cdot n^{-1/6}$, $h_2 = 0,8978 \cdot n^{-1/6}$.

Oba parametry jsou totožné, což se dalo očekávat, protože hustota $f(x, y)$ je symetrická jak podle osy x , tak podle osy y .

Tedy pro soubor o 100 prvcích bude vyhlazovací matice rovna $\mathbf{H}_{opt,0,2} = \text{diag}(h_1^2, h_2^2) = \text{diag}(0,776, 0,776)$. Odhad s optimální vyhlazovací maticí pro tento datový soubor (viz tabulku 7.7) jsou na obrázku 5.5.

²Použili jsme metodu „per partes“ pro vícerozměrné integrály.



Obrázek 5.5: Odhad hustoty z ukázkového příkladu 5.3, odhad (červená, plná) a původní funkce (modrá, čárkovaná) při použití součinného Epanečnikova jádra

4 Volba jádra

Podobně jako u odhadu jednorozměrné hustoty není volba jádra podstatná. Je vhodné zvolit součinný tvar optimálního jádra. Tím zajistíme jistou hladkost výsledného odhadu a navíc výpočty s využitím součinných jader jsou jednodušší.

Poznámka 5.3. V literatuře se také využívá Gaussovo jádro $K(x, y) = (2\pi)^{-1} e^{-(x^2+y^2)/2}$, které se zdá být výhodnější při studiu asymptotických vlastností jádrového odhadu. Na druhou stranu má nevýhodu, že jeho nosičem je celá reálná osa, což způsobuje „nedokonalost“ při odhadech hustot s omezeným definičním oborem.

5 Volba vyhlazovacího parametru

5.1 Metoda referenční hustoty

Předpokládejme, že náhodný výběr $([X_1, Y_1], \dots, [X_n, Y_n])$ pochází z normálního rozdělení s hustotou

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{x^2}{2\sigma_1^2} - \frac{y^2}{2\sigma_2^2}}$$

a jádro K je dvourozměrnou standardizovanou normální hustotou, tj.

$$K(x, y) = \frac{1}{2\pi} e^{-\frac{x^2}{2} - \frac{y^2}{2}}.$$

Pak podle metody referenční hustoty lze získat tento odhad vyhlazovací matice

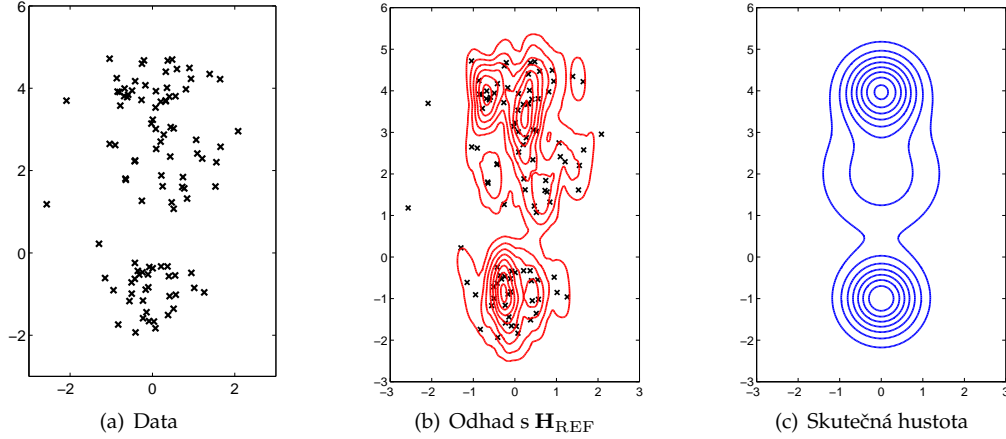
$$\mathbf{H}_{\text{REF}} = \begin{pmatrix} \hat{\sigma}_1^2 n^{-1/3} & 0 \\ 0 & \hat{\sigma}_2^2 n^{-1/3} \end{pmatrix},$$

kde $\hat{\sigma}_i^2$ jsou odhady rozptylů pro jednotlivé proměnné. Obecnější tvar tohoto vztahu, který je znám také jako Scottovo pravidlo, lze nalézt v [13].

Příklad 5.4. Pro simulovaná data z příkladu 5.1 je matice vyhlazovacích parametrů podle metody referenční hustoty rovna s využitím Gaussova jádra takto:

$$\mathbf{H}_{\text{REF}} = \begin{pmatrix} 0,1292 & 0 \\ 0 & 0,9945 \end{pmatrix}.$$

Na obrázku 5.6 je vykreslen odhad hustoty s pomocí Gaussova jádra a porovnání odhadu se skutečnou hustotou.



Obrázek 5.6: Jádrový odhad dvourozměrné hustoty z příkladu 5.1 s Gaussovým jádrem – metoda referenční hustoty

5.2 Metoda křížového ověřování

Metoda křížového ověřování je založena na odhadu hustoty v bodě $[X_i, Y_i]$ s vynecháním tohoto pozorování. Funkci metody křížového ověřování CV můžeme zapsat ve tvaru

$$CV(\mathbf{H}) = \iint (\hat{f}(x, y, \mathbf{H}))^2 dx dy - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i, Y_i, \mathbf{H}).$$

kde

$$\hat{f}_{-i}(X_i, Y_i, \mathbf{H}) = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n K_{\mathbf{H}}(X_i - X_j, Y_i - Y_j)$$

Někdy se metoda CV nazývá nevyhýlená metoda křížového ověřování (*unbiased cross-validation*), důvodem je jednoduchý vztah mezi CV a MISE, který uvádí následující věta.

Věta 5.2. Funkce CV je nevyhýleným odhadem MISE, tj. platí

$$E CV(\mathbf{H}) = \text{MISE } \hat{f}(\cdot, \mathbf{H}) - \iint f^2(x, y) dx dy.$$

Důkaz. Vypočtěme střední hodnotu CV:

$$\begin{aligned} E CV(\mathbf{H}) &= E \iint \hat{f}^2(x, y, \mathbf{H}) dx dy - \frac{2}{n} \sum_{i=1}^n E \hat{f}_{-i}(X_i, Y_i, \mathbf{H}) \\ &= E \iint \hat{f}^2(x, y, \mathbf{H}) dx dy - 2EK_{\mathbf{H}}(X_1 - X_2, Y_1 - Y_2) \\ &= E \iint \hat{f}^2(x, y, \mathbf{H}) dx dy - 2 \iiint K_{\mathbf{H}}(x - u, y - v) f(x, y) f(u, v) dx dy du dv \end{aligned}$$

a úpravou MISE

$$\begin{aligned}
 \text{MISE } \hat{f}(\cdot, \mathbf{H}) &= E \iint (\hat{f}(x, y, \mathbf{H}) - f(x, y))^2 dx dy \\
 &= E \iint \hat{f}^2(x, y, \mathbf{H}) dx dy - 2E \int \hat{f}(x, y, \mathbf{H}) f(x, y) dx dy + \iint f^2(x, y) dx dy \\
 &= E \iint \hat{f}^2(x, y, \mathbf{H}) dx dy - 2 \iiint K_{\mathbf{H}}(x - u, y - v) f(x, y) f(u, v) dx dy du dv \\
 &\quad + \iint f^2(x, y) dx dy.
 \end{aligned}$$

Porovnáním upravených výrazů dostaneme tvrzení. □

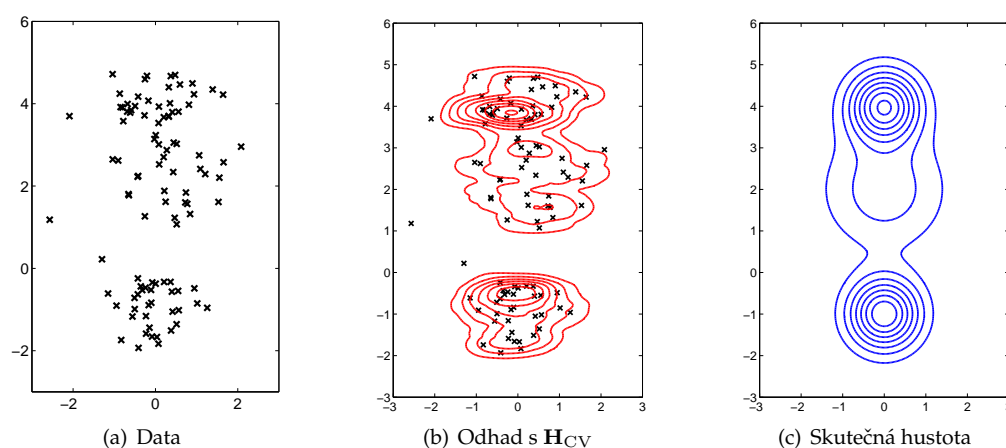
Optimální matici vyhlazovacích parametrů vzhledem k metodě CV označíme \mathbf{H}_{CV} , tj.

$$\mathbf{H}_{CV} = \arg \min_{\mathbf{H} \in \mathcal{D}} CV(\mathbf{H}).$$

Příklad 5.5. Použijeme-li součinnové Epanečnikovo jádro, pak pro simulovaná data z příkladu 5.1 dostanem matici vyhlazovacích parametrů určenou podle metody křížového ověřování v následujícím tvaru:

$$\mathbf{H}_{CV} = \begin{pmatrix} 1,4532 & 0 \\ 0 & 0,1431 \end{pmatrix}.$$

Na obrázku 5.7 je vykreslen odhad hustoty s touto maticí.

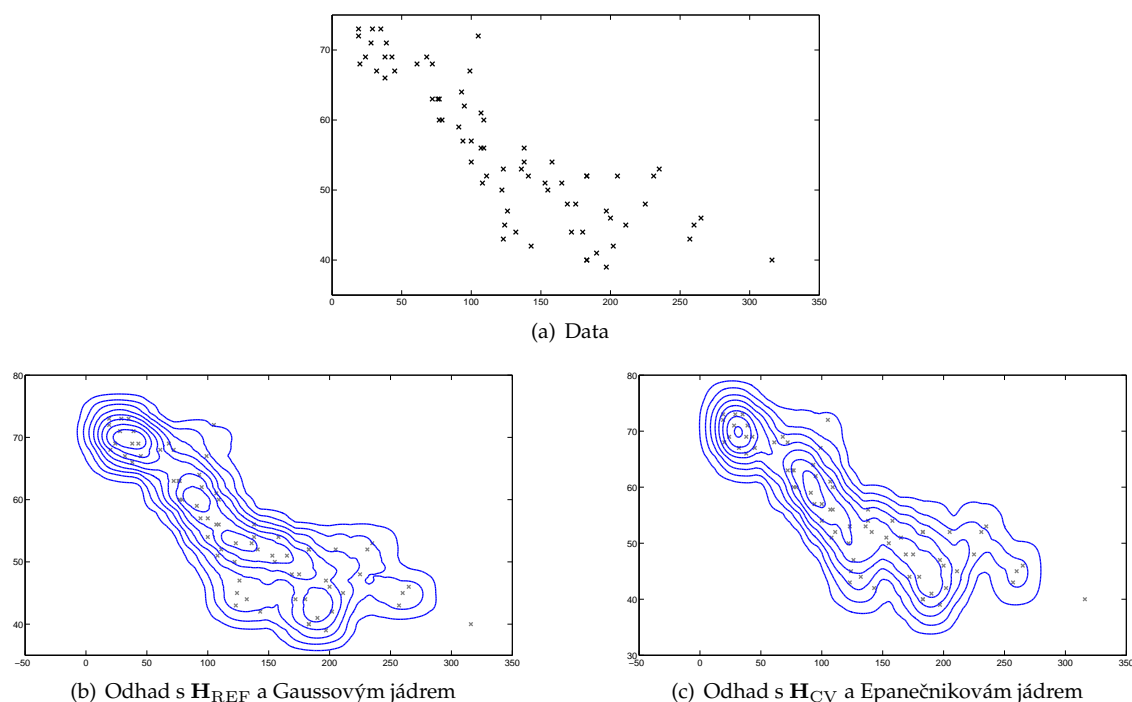


Obrázek 5.7: Jádrový odhad dvourozměrné hustoty z příkladu 5.1 se součinným Epanečnikovým jádrem – metoda křížového ověřování

6 Aplikace na reálná data

Vraťme se k datovému souboru z motivačního odstavce. Jak už jsme zmínili, data pochází ze studie The State of the World's Children (UNICEF) a obsahuje míru úmrtnosti dětí od narození do pěti let věku počítanou na 1000 živě narozených dětí a očekávanou délku života narozeného dítěte (s ohledem na úmrtnost v dané populaci v době jeho narození) v 72 zemích, které měly v roce 2001 hrubý národní produkt menší než 1000 amerických dolarů na osobu a rok.³ Data jsou shrnuta v tabulce 7.15.

³<http://www.unicef.org/sowc03/tables/table1.html>



Obrázek 5.8: Vrstevníkové grafy odhadnutých hustot pro data ze studie UNICEF – na ose x je míra úmrtnosti dětí do pěti let (na 1000 živě narozených dětí) a na ose y očekávaná délka života narozeného dítěte

Odhady vyhlazovacích matic podle uvedených metod jsou tyto

$$\mathbf{H}_{\text{REF}} = \begin{pmatrix} 1145,2 & 0 \\ 0 & 24,9 \end{pmatrix} \quad \mathbf{H}_{\text{CV}} = \begin{pmatrix} 674,14 & 0 \\ 0 & 59,35 \end{pmatrix}$$

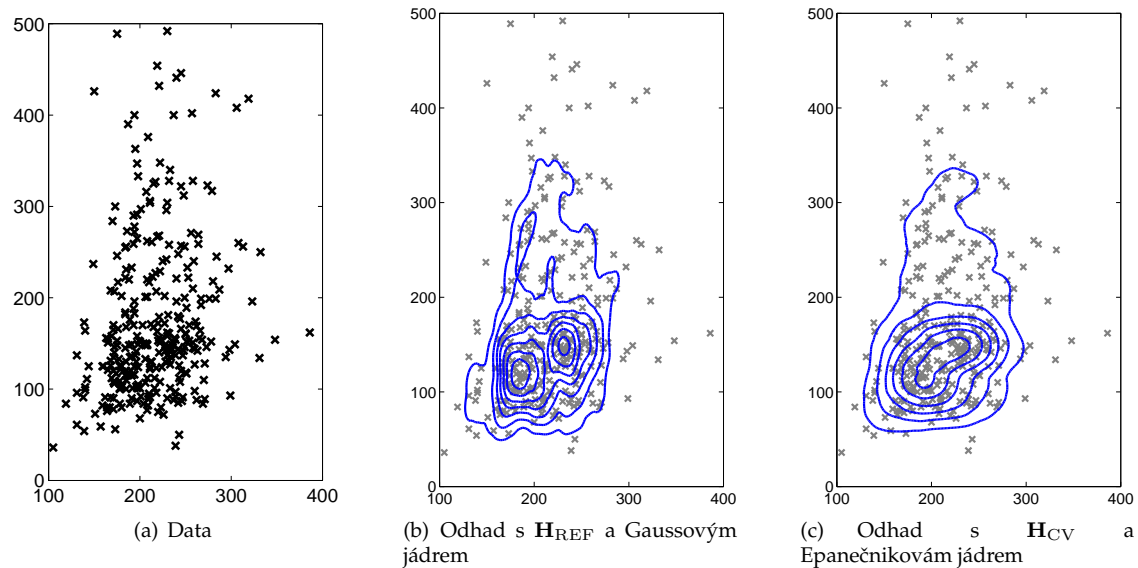
Na obrázku 5.8 jsou znázorněna data, vrstevníkové jádrového odhadu s Gaussovým jádrem a maticí určenou metodou referenční hustoty a vrstevníkové odhadu se součinným Epanečnikovým jádrem a maticí určenou podle metody křížového oěřování. jádrem.

Následující datový soubor pochází ze studie koncentrace lipidů v krevní plazmě, která vyšla v časopise *Circulation* v roce 1980. Výběrový soubor, který jsme převzali z [11] a s nímž zde pracujeme, obsahuje měření množství cholesterolu a triglyceridu v krevní plazmě u 320 pacientů, kteří si stěžovali na bolest v hrudníku a u nichž bylo potvrzeno zúžení tepen. Data jsou shrnuta v tabulkách 7.16 a 7.17.

Maticе vyhlazovacích parametrů určené podle metody referenční hustoty a metody křížového ověřování (s Epanečnikovým jádrem) jsou následující:

$$\mathbf{H}_{\text{REF}} = \begin{pmatrix} 270,5 & 0 \\ 0 & 1516,5 \end{pmatrix} \quad \mathbf{H}_{\text{CV}} = \begin{pmatrix} 1797,2 & 0 \\ 0 & 892,7 \end{pmatrix}$$

Na obrázku 5.9(a) jsou znázorněna data, na obrázku 5.9(b) jsou vrstevníkové jádrového odhadu s Gaussovým jádrem a maticí určenou metodou referenční hustoty a na obrázku 5.9(c) jsou zobrazeny vrstevníkové odhadu se součinným Epanečnikovým jádrem a maticí určenou podle metody křížového oěřování. jádrem.



Obrázek 5.9: Vrstevnicové grafy odhadnutých hustot pro koncentraci lipidů – na ose x je vynešeno množství cholesterolu (v miligramech na 100 ml plazmy) a na ose y množství triglyceridu v krevní plazmě (mg/100 ml)

Shrnutí
<p>Odhad dvourozměrné hustoty pravděpodobnosti $f(x, y)$ v bodě $[x, y]$ je tvaru</p> $\hat{f}(x, y, \mathbf{H}) = \frac{1}{n \mathbf{H} } \sum_{i=1}^n K\left(\mathbf{H}^{-1}(x - X_i, y - Y_i)^T\right)$
<p>Dva typy jader:</p> <ul style="list-style-type: none"> • součinnové jádro: $K^P(x, y) = K_1(x) \cdot K_1(y)$, • sféricky symetrické jádro: $K^S(x, y) = c_k^{-1} K_1(\sqrt{x^2 + y^2})$, $c_k = \iint K_1(\sqrt{x^2 + y^2}) dx dy$.
<p>Asymptotická střední integrální kvadratická chyba dvourozměrného jádrového odhadu</p> $\text{AMISE } \hat{f}(\cdot, \mathbf{H}) = \frac{V(K)}{nh_1 h_2} + \frac{1}{4} \beta_2^2(K) (h_1^4 V(f_{xx}) + 2h_1^2 h_2^2 V(f_{xy}) + h_2^4 V(f_{yy})).$
<p>Optimální vyhlazovací parametry vzhledem k AMISE</p> $h_{1,opt} = \left(\frac{V^{3/4}(f_{yy})V(K)}{n\beta_2^2(K)V^{3/4}(f_{xx})[V(f_{xy}) + \sqrt{V(f_{xx})V(f_{yy})}]} \right)^{1/6},$ $h_{2,opt} = \left(\frac{V^{3/4}(f_{xx})V(K)}{n\beta_2^2(K)V^{3/4}(f_{yy})[V(f_{xy}) + \sqrt{V(f_{xx})V(f_{yy})}]} \right)^{1/6}.$
<p>Metody pro odhad optimálních hodnot matice vyhlazovacích parametrů $\mathbf{H} = \text{diag}(h_1, h_2)$</p> <ul style="list-style-type: none"> • metoda referenční hustoty $h_{i,\text{REF}} = \hat{\sigma}_i n^{-1/6}, \quad i = 1, 2,$ <ul style="list-style-type: none"> • metoda křížového ověřování $\mathbf{H}_{\text{CV}} = \arg \min_{\mathbf{H} \in \mathcal{D}} \text{CV}(\mathbf{H}), \quad \text{CV}(\mathbf{H}) = \iint (\hat{f}(x, y, \mathbf{H}))^2 dx dy - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i, Y_i, \mathbf{H}).$

Výstupy z výukové jednotky

Student

- zná součinnová a sférická dvourozměrná jádra pro odhady dvourozměrných hustot
- porozuměl principu vyhlazování dvourozměrných hustot
- pochopil nejjednodušší metody pro volbu prvků diagonální vyhlazovací matice
- zvládne použití příslušného toolboxu v Matlabu pro simulační studii i pro zpracování reálných dat

Dodatek

Symetrická matice A je pozitivně definitní právě tehdy, když všechny její hlavní minory (subdeterminanty) jsou kladné, tj. platí

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \quad M_{11} = |a_{11}| > 0, M_{22} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \dots, M_{nn} = |A| > 0.$$

Cvičení

1. Určete součinnové a sféricky symetrické dvourozměrné jádro odvozené z kvartického jádra $K(x) = \frac{15}{16}(1 - x^2)^2$.
2. Odvoďte vztahy (5.3) a (5.4) pro optimální vyhlazovací parametry.
3. Ukažte, že pro optimální vyhlazovací matici vzhledem k AMISE $\hat{f}(\cdot, \mathbf{H})$ platí

$$\text{AIV}(\mathbf{H}_{\text{AMISE}}) = 2 \cdot \text{AISB}(\mathbf{H}_{\text{AMISE}}),$$

kde $\mathbf{H}_{\text{AMISE}} = \arg \min_{\mathbf{H} \in \mathcal{D}} \text{AMISE}$.

4. Aplikujte metodu referenční hustoty a metodu křížového ověřování na simulovaná data z ukázkového příkladu 5.3.
5. Ověřte, že matice A je pozitivně definitní maticí

$$A = \begin{pmatrix} 6 & -1 & 2 \\ -1 & 4 & -3 \\ 2 & -3 & 9 \end{pmatrix}.$$

Kapitola 6

Návody ke cvičením a odpovědi na otázky

Kapitola 1

Cv. 1 Je zřejmé, že jádro K_γ má nosič $[-\gamma, \gamma]$. Do funkcionálu T dosadíme jádro $K_\gamma(t)$:

$$\begin{aligned} T(K_\delta) &= \left[\left(\int_{-\delta}^{\delta} K_\delta^2(t) dt \right)^{k-\nu} \cdot \left| \int_{-\delta}^{\delta} t^k K_\delta(t) dt \right|^{2\nu+1} \right]^{\frac{2}{2k+1}} \\ &= \left[\left(\int_{-\delta}^{\delta} \frac{1}{\delta^{2\nu+2}} K^2\left(\frac{t}{\delta}\right) dt \right)^{k-\nu} \cdot \left| \int_{-\delta}^{\delta} t^k \frac{1}{\delta^{\nu+1}} K\left(\frac{t}{\delta}\right) dt \right|^{2\nu+1} \right]^{\frac{2}{2k+1}} \\ &= \left| \frac{t}{\delta} = x \right| \\ &= \left[\left(\int_{-1}^1 \frac{1}{\delta^{2\nu+2}} \delta K^2(x) dx \right)^{k-\nu} \cdot \left| \int_{-1}^1 x^k \frac{\delta^k}{\delta^{\nu+1}} \delta K(x) dx \right|^{2\nu+1} \right]^{\frac{2}{2k+1}} \\ &= \left[\left(\int_{-1}^1 K^2(x) dx \right)^{k-\nu} \cdot \left| \int_{-1}^1 x^k K(x) dx \right|^{2\nu+1} \right]^{\frac{2}{2k+1}} \cdot \left(\frac{1}{\delta^{(2\nu+1)(k-\nu)}} \delta^{(k-\nu)(2\nu+1)} \right)^{\frac{2}{2k+1}} \\ &= T(K) \end{aligned}$$

Cv. 4 Výraz

$$\int (K * K)(x) x^j dx = \iint K(z) K(x-z) x^j dz dx$$

spočítáme postupně pro jednotlivé hodnoty j .

$$\begin{aligned} j = 0 : \quad & \iint K(z) K(x-z) dz dx = \int K(z) \int K(x-z) dx dz \\ &= | \text{substituce vnitřního integrálu: } x-z=w | = \int K(z) \int K(w) dw dz \\ &= \int K(z) dz \cdot \int K(w) dw = 1 \end{aligned}$$

Podobně pro $j = 1$ a $j = 2$, přičemž využíváme vlastností jádra (definice 1.1).

Cv. 5 Ověříme podmínky jádra třídy S_{02} , tj.

- $\int_{-1}^1 K(x) dx = 1 \rightarrow$ odtud spočítáme hodnotu konstanty $A = \frac{\pi}{4}$,
- $\int_{-1}^1 xK(x) dx = 0 \rightarrow$ metodou per partes spočítáme nebo úvahou o integrování liché funkce přes symetrický interval odvodíme, že vztah platí,
- $\int_{-1}^1 x^2 K(x) dx = 0 \rightarrow$ dvojnásobným použitím metody per partes vypočítáme hodnotu $\beta_2(K) = 1 - \frac{8}{\pi^2} \doteq 0,1894$, která je nenulová, tudíž jde o jádro třídy S_{02} .

Nakonec, s využitím vztahu pro druhou mocninu funkce $\cos x$, tj. $\cos^2 x = (1 + \cos 2x)/2$, spočítáme hodnotu $V(K) = \int_{-1}^1 K^2(x) dx = \frac{\pi^2}{16} \doteq 0,6169$.

Kapitola 2

Otázka na str. 15 Obdélníkové jádro: $K_h(z) = \frac{1}{2h}$ pro $z \in [-h, h]$. Váhy Nadarayova-Watsonova odhadu:

$$\begin{aligned} W_i(x_0, h) &= \frac{K_h(x_0 - x_i)}{\sum_{j=1}^n K_h(x_0 - x_j)} = \frac{\frac{1}{2h} I_{[-h, h]}(x_0 - x_i)}{\sum_{j=1}^n \frac{1}{2h} I_{[-h, h]}(x_0 - x_j)} \\ &= \frac{\frac{1}{2h} I_{[-h, h]}(x_0 - x_i)}{\frac{n}{2h} I_{[-h, h]}(x_0 - x_j)} = \frac{I_{[-h, h]}(x_0 - x_i)}{n I_{[-h, h]}(x_0 - x_j)} \end{aligned}$$

Tedy NW odhad s obdélníkovým jádrem je totožný s klouzavým průměrem (viz obr. 2.7(b)).

Cv. 1 Obdélníkové jádro: $K_h(z) = \frac{1}{2h}$ pro $z \in [-h, h]$. Priestleyův-Chaův odhad:

$$\hat{m}_{PC}(x, h) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) Y_i = \frac{1}{n} \sum_{i=1}^n \frac{1}{2h} I_{[-h, h]}(x - x_i) Y_i = \frac{1}{2nh} \sum_{i=1}^n Y_i I_{[-h, h]}(x - x_i).$$

Princip je podobný jako u Nadaraya-Watsonových odhadů (str. 15).

Cv. 2

$$\begin{aligned} \hat{m}_{NW}(x, h, K) &= \frac{\sum_{i=1}^n \frac{1}{h} K\left(\frac{x-x_i}{h}\right) Y_i}{\sum_{i=1}^n \frac{1}{h} K\left(\frac{x-x_i}{h}\right)} = |h = h^* \delta| = \frac{\sum_{i=1}^n \frac{1}{h^* \delta} K\left(\frac{x-x_i}{h^* \delta}\right) Y_i}{\sum_{i=1}^n \frac{1}{h^* \delta} K\left(\frac{x-x_i}{h^* \delta}\right)} \\ &= \frac{\sum_{i=1}^n \frac{1}{h^*} \frac{1}{\delta} K\left(\frac{x-x_i}{\delta}\right) Y_i}{\sum_{i=1}^n \frac{1}{h^*} \frac{1}{\delta} K\left(\frac{x-x_i}{\delta}\right)} = \frac{\sum_{i=1}^n \frac{1}{h^*} K_\delta\left(\frac{x-x_i}{h^*}\right) Y_i}{\sum_{i=1}^n \frac{1}{h^*} K_\delta\left(\frac{x-x_i}{h^*}\right)} \end{aligned}$$

Cv. 3 Do vztahu (2.13) dosadíme hodnoty $V(K) = 1,250$ a $\beta_4(K) = -0,0476$ (z tabulky 1.2) a z regresního modelu

$$\begin{aligned} h_{opt,0,k}^{2k+1} &= \frac{\sigma^2 V(K) (k!)^2}{2kn \beta_k^2(K) V(m^{(k)})}, \\ h_{opt,0,4}^9 &= \frac{0,25 \cdot 1,25 \cdot (4!)^2}{2 \cdot 4 \cdot 100 \cdot (-0,0476)^2 \cdot 32\pi^8} = 3,2705 \cdot 10^{-4}, \\ h_{opt,0,4} &= 0,41. \end{aligned}$$

Cv. 5 Ve vztahu (2.7) použijeme Taylorův rozvoj

$$m(x - hu) = m(x) - hum'(x) + \dots + \frac{(-1)^k}{k!} h^k u^k m^{(k)}(x) + o(h^k).$$

Využijeme vlastnosti jádra $K \in S_{0k}$ a dostaneme

$$E\hat{m}(x, h) = m(x) + (-1)^k \frac{h^k}{k!} \beta_k(K) m^{(k)}(x) + o(h^k) + O(n^{-1}).$$

Dále postupujeme jako při odvození vztahu (2.11).

Cv. 6 Spočítejme nejdříve derivaci AMISE $\hat{m}(\cdot, h)$ (viz vztah (2.12))

$$\frac{d\text{AMISE } \hat{m}(\cdot, h)}{dh} = -\frac{\sigma^2 V(K)}{nh^2} + \frac{2k}{(k!)^2} h^{2k-1} \beta_k^2(K) V(m^{(k)}).$$

Jelikož hledáme minimum funkce AMISE $\hat{m}(\cdot, h)$, položíme tuto derivaci rovnu nule

$$-\frac{\sigma^2 V(K)}{nh^2} + \frac{2k}{(k!)^2} h^{2k-1} \beta_k^2(K) V(m^{(k)}) = 0.$$

Nechť $h_{opt,0,k}$ je řešením této rovnice. Dále vynásobíme tuto rovnici parametrem $h_{opt,0,k}$

$$-\frac{\sigma^2 V(K)}{nh_{opt,0,k}} + \frac{2k}{(k!)^2} h_{opt,0,k}^{2k} \beta_k^2(K) V(m^{(k)}) = 0,$$

neboli

$$\frac{\sigma^2 V(K)}{nh_{opt,0,k}} = 2k \frac{1}{(k!)^2} h_{opt,0,k}^{2k} \beta_k^2(K) V(m^{(k)})$$

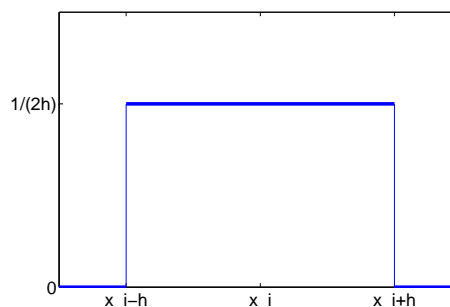
$$\text{AIV}(h_{opt,0,k}) = 2k \text{AISB}(h_{opt,0,k}).$$

Kapitola 3

Otázka na str. 31 Obdélníkové jádro: $K_h(z) = \frac{1}{2h}$ pro $z \in [-h, h]$. Odhad hustoty

$$\hat{f}(x, h) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i) = \frac{1}{n} \sum_{i=1}^n \frac{1}{2h} I_{[-h, h]}(x - X_i) = \frac{1}{2nh} \sum_{i=1}^n I_{[-h, h]}(x - X_i).$$

Tedy v každém bodě x_i sestrojíme obdélník se šířkou $2h$ a výškou $(2nh)^{-1}$, tyto odhady se pak sečtou. Tomuto typu odhadu se také říká *naivní odhad*. Naivní odhad v jistém smyslu osvobozuje histogram od volby polohy třídicích intervalů. Tento odhad „klouže“ přes data – jako histogram pro odhad hustoty odpovídá regresogramu pro odhad regresní funkce, tak výsledek odhadu hustoty s obdélníkovým jádem odpovídá klouzavému průměru u regrese.



Cv. 1

$$\begin{aligned} \text{AMISE } \hat{f}(\cdot, h) &= \frac{V(K)}{nh} + \frac{1}{(k!)^2} h^{2k} \beta_k^2(K) V(f^{(k)}) \\ &= T(K) \left(\frac{V(K)}{nhT(K)} + \frac{1}{(k!)^2} h^{2k} \frac{\beta_k^2(K)}{T(K)} V(f^{(k)}) \right) \end{aligned}$$

Pomocné výpočty: $T(K) = \delta_{0k}^{2k} \beta_k^2(K)$, $\delta_{0k}^{2k+1} = \frac{V(K)}{\beta_k^2(K)}$

$$\begin{aligned} \frac{V(K)}{T(K)} &= \frac{V(K)}{\delta_{0k}^{2k} \beta_k^2(K)} = \frac{\delta_{0k}^{2k+1}}{\delta_{0k}^{2k}} = \delta_{0k}, \\ \frac{\beta_k^2(K)}{T(K)} &= \frac{\beta_k^2(K)}{\delta_{0k}^{2k} \beta_k^2(K)} = \frac{1}{\delta_{0k}^{2k}}. \end{aligned}$$

Celkem

$$\text{AMISE } \hat{f}(\cdot, h) = T(K) \left(\frac{\delta_{0k}}{nhT(K)} + \frac{1}{(k!)^2} h^{2k} \frac{V(f^{(k)})}{\delta_{0k}^{2k}} \right).$$

Cv. 2 Pro danou hustotu $f(x) = 20x(1-x)^3$ spočítáme její čtvrtou derivaci: $f^{(4)}(x) = -480$, pak $V(f^{(4)}) = 480^2$. Dosazením do vztahu (3.5) s využitím vlastností jádra $K \in S_{04}$ dostaneme

$$\begin{aligned} h_{opt,0,k}^{2k+1} &= \frac{\delta_{0k}^{2k+1} (k!)^2}{2knV(f^{(k)})}, \\ h_{opt,0,4}^9 &= \frac{\delta_{04}^9 (4!)^2}{2 \cdot 4 \cdot 50 \cdot 480^2} = \frac{1,25}{(-0,0476)^2} \frac{(4!)^2}{2 \cdot 4 \cdot 50 \cdot 480^2} = 0,0034 \\ h_{opt,0,4} &= 0,5326 \end{aligned}$$

Tato hodnota je příliš velká na to, aby byla optimální hodnotou vyhlazovacího parametru pro hustotu, která je definována na intervalu $[0, 1]$. Ve výpočtu samotném není chyba, avšak byl porušen předpoklad o spojitosti derivací funkce až do řádu 4 včetně – viz větu 3.2. Už první derivace této funkce hustoty není spojitá (nakreslete si její graf).

Cv. 3 Postupujeme podobně jako v ukázkovém příkladě na str. 36, $h_{opt,0,2} = 0,9221 \cdot n^{-1/5}$.

Cv. 4 a) $\sigma_k^2 = \int_{-1}^1 x^2 g_k(x) dx$, $g_k(x) = A_k(1-x^2)^{k+1}$ a $A_k = \frac{(2k+3)!}{(k+1)!2^{2k+3}}$ pak

$$\begin{aligned} \sigma_k^2 &= A_k \int_{-1}^1 x^2 (1-x^2)^{k+1} dx = A_k \int_{-1}^1 x \cdot x (1-x^2)^{k+1} dx \\ &= \left| \begin{array}{ll} u = x & u' = 1 \\ v' = x(1-x^2)^{k+1} & v = \frac{-1}{2(k+2)} (1-x^2)^{k+2} \end{array} \right| \\ &= A_k \left[x \cdot \frac{-1}{2(k+2)} (1-x^2)^{k+2} \right]_{-1}^1 + A_k \int_{-1}^1 \frac{1}{2(k+2)} (1-x^2)^{k+2} dx \\ &= A_k \frac{1}{2(k+2)} \int_{-1}^1 (1-x^2)^{k+2} dx. \end{aligned}$$

Víme, že

$$\int_{-1}^1 \underbrace{A_{k+1} (1-x^2)^{k+2}}_{=g_{k+1}(x)} dx = 1 \quad (g_{k+1} \text{ je hustota}).$$

Odtud plyne, že

$$\int_{-1}^1 (1-x^2)^{k+2} dx = \frac{1}{A_{k+1}},$$

a tedy

$$\begin{aligned} \sigma_k^2 &= A_k \cdot \frac{1}{2(k+2)} \cdot \frac{1}{A_{k+1}} \\ &= \frac{(2k+3)!}{(k+1)!2^{2k+3}} \cdot \frac{1}{2(k+2)} \cdot \frac{((k+1)+1)!2^{2(k+1)+3}}{(2(k+1)+3)!} \\ &= \frac{(2k+3)!(k+2)!2^{2k+5}}{(k+1)!2^{2k+3}2(k+2)(2k+5)!} = \frac{1}{2k+5}. \end{aligned}$$

b)

$$\begin{aligned} \int x^2 \frac{\sigma_g}{\sigma} g\left(\frac{\sigma_g}{\sigma} x\right) dx &= \int \left(\frac{\sigma_g}{\sigma} x\right)^2 \frac{\sigma}{\sigma_g} g\left(\frac{\sigma_g}{\sigma} x\right) dx = \left| \frac{\sigma_g}{\sigma} x = t, \frac{\sigma_g}{\sigma} dx = dt \right| \\ &= \int t^2 \frac{\sigma}{\sigma_g} g(t) \frac{\sigma}{\sigma_g} dt = \frac{\sigma^2}{\sigma_g^2} \underbrace{\int t^2 g(t) dt}_{=\sigma_g^2} = \sigma^2. \end{aligned}$$

(Podrobněji např. [5].)

Cv. 5 Vztah (3.12) pro $k = 2$:

$$h_{\text{REF}} = \left(\frac{2^4 (2!)^3 \sqrt{\pi}}{(4)! 2} \right)^{\frac{1}{5}} \delta_{02} \sigma n^{-\frac{1}{5}}.$$

Dále pro Gaussovo jádro $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ platí $V(K) = \frac{1}{2\sqrt{\pi}}$, $\beta_2(K) = 1$ a $\delta_{02} = \left(\frac{V(K)}{\beta_2(K)}\right)^{1/5} = \left(\frac{1}{2\sqrt{\pi}}\right)^{1/5}$. Pak

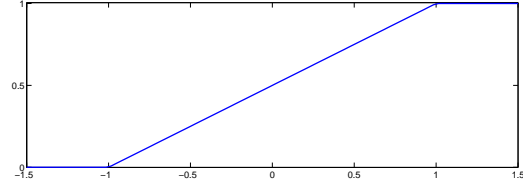
$$h_{\text{REF}} = \left(\frac{8\sqrt{\pi}}{3}\right)^{\frac{1}{5}} \left(\frac{1}{2\sqrt{\pi}}\right)^{1/5} \sigma n^{-\frac{1}{5}} = \left(\frac{4}{3n}\right)^{1/5} \sigma.$$

Kapitola 4

Otázka na str. 55 Obdélíkové jádro $K(x) = \frac{1}{2}$, tedy funkce

$$W(x) = \int_{-\infty}^x \frac{1}{2} dt = \int_{-1}^x \frac{1}{2} dt = \left[\frac{t}{2} \right]_{-1}^x = \frac{x+1}{2}.$$

$$W(x) = \begin{cases} 0 & x < -1, \\ \frac{x+1}{2} & x \in [-1, 1] \\ 1 & x > 1. \end{cases}$$



$\hat{F}(x, h) = \frac{1}{n} \sum_{i=1}^n \frac{x+1-x_i}{2} = \frac{1}{2} + \frac{1}{n} \sum_{i=1}^n \frac{x-x_i}{2}$. Vlastnosti $W(x)$:

$$\int_{-1}^1 W^2(x) dx = \int_{-1}^1 \frac{(x+1)^2}{4} dx = \frac{1}{4} \left[\frac{(x+1)^3}{3} \right]_{-1}^1 = \frac{2}{3},$$

$$\int_{-1}^1 W(x) dx = \frac{x+1}{2} dx = \frac{1}{2} \left[\frac{(x+1)^2}{2} \right]_{-1}^1 = 1.$$

Cv. 1 Podle definice stačí integrovat polynom, který odpovídá kvartickému jádru, tj.

$$\begin{aligned} W(x) &= \int_{-1}^x \frac{15}{16} (1-t^2)^2 dt = \frac{15}{16} \int_{-1}^x (1-2t^2+t^4) dt \\ &= \frac{15}{16} \left[t - \frac{2}{3}t^3 + \frac{1}{5}t^5 \right]_{-1}^x = \frac{15}{16} \left(x - \frac{2}{3}x^3 + \frac{1}{5}x^5 + \frac{8}{15} \right) = \frac{1}{16} (3x^5 - 10x^3 + 15x + 8). \end{aligned}$$

Cv. 2 • Vlastnost 2: platí $W'(x) = K(x)$,

$$\begin{aligned} \int_{-1}^1 W(x) dx &= \left| \begin{array}{ll} u' = 1 & u = x \\ v = W(x) & v' = W'(x) = K(x) \end{array} \right| \\ &= [xW(x)]_{-1}^1 - \underbrace{\int_{-1}^1 xK(x) dx}_{=0} = 1 \cdot W(1) - (-1) \cdot W(-1) = 1. \end{aligned}$$

Protože $0 \leq W(x) \leq 1$, pak platí $\int_{-1}^1 W^2(x) dx \leq \int_{-1}^1 W(x) dx = 1$.

• Vlastnost 3:

$$\underbrace{\int_{-1}^1 W(x)K(x) dx}_{=T} = \left| \begin{array}{ll} u = W(x) & u' = K(x) \\ v' = K(x) & v = W(x) \end{array} \right| = [W^2(x)]_{-1}^1 - \underbrace{\int_{-1}^1 K(x)W(x) dx}_{=T},$$

$$W(1) = 1 \text{ a } W(-1) = 0$$

$$T = 1 - T \Rightarrow T = \int_{-1}^1 W(x)K(x) dx = \frac{1}{2}.$$

•

$$\begin{aligned} \int_{-1}^1 xW(x)K(x) dx &= \left| \begin{array}{ll} u = x & u' = 1 \\ v' = K(x)W(x) & v = \frac{1}{2}W^2(x) \end{array} \right| \\ &= [x\frac{1}{2}W^2(x)]_{-1}^1 - \int_{-1}^1 \frac{1}{2}W^2(x) dx = \frac{1}{2} - \frac{1}{2} \int_{-1}^1 W^2(x) dx \\ &= \frac{1}{2} \left(1 - \int_{-1}^1 W^2(x) dx \right). \end{aligned}$$

Cv. 3 Vztah (4.6) pro AMISE zderivujeme vzhledem k h , položíme roven nule a vypočítáme h :

$$\frac{d\text{AMISE } \hat{F}(\cdot, h)}{dh} = -c_1 \frac{1}{n} + 4c_2 h^3 = 0.$$

Toto vypočítané h dosadíme do rovnice (4.6) a upravíme.

Cv. 4 Kvartické jádro: $K(x) = \frac{15}{16}(1 - x^2)^2$, $\beta_2(K) = \frac{1}{7}$ a $\int_{-1}^1 W^2(x) dx = 0,7835$. Dosadíme do vztahu (4.9) a dostaneme

$$h_{\text{MS}} = n^{-1/3} \left(\frac{7 \cdot (1 - 0,7835)}{15 \cdot \frac{1}{49}} \right)^{1/3} \sqrt{7} \cdot \hat{\sigma} = 4,5089 \cdot \hat{\sigma} \cdot n^{-1/3}.$$

Kapitola 5

Cv. 1 Hodnota $c_k = \iint K_1(\sqrt{x^2 + y^2}) dx dy$ pro kvartické jádro je $\pi/3$.

Cv. 2 Zderivujeme vztah (5.2) pro AMISE a dostaneme

$$0 = \frac{d\text{AMISE } \hat{f}(\cdot, \mathbf{H})}{dh_1} = -\frac{V(K)}{nh_1^2 h_2} + \frac{1}{4} \beta_2^2(K) (4h_1^3 V(f_{xx}) + 4h_1 h_2^2 V(f_{xy})) \quad (6.1)$$

$$0 = \frac{d\text{AMISE } \hat{f}(\cdot, \mathbf{H})}{dh_2} = -\frac{V(K)}{nh_1 h_2^2} + \frac{1}{4} \beta_2^2(K) (4h_1^2 h_2 V(f_{xy}) + 4h_2^3 V(f_{yy})) \quad (6.2)$$

První rovnici vynásobíme h_1 , druhou rovnici vynásobíme h_2 a odečteme je

$$0 = \frac{1}{4} \beta_2^2(K) (4h_1^4 V(f_{xx}) + 4h_1^2 h_2^2 V(f_{xy})) - \frac{1}{4} \beta_2^2(K) (4h_1^2 h_2^2 V(f_{xy}) + 4h_2^4 V(f_{yy})),$$

odtud plyne

$$4h_1^4 V(f_{xx}) + 4h_1^2 h_2^2 V(f_{xy}) = 4h_1^2 h_2^2 V(f_{xy}) + 4h_2^4 V(f_{yy}),$$

tj.

$$\frac{h_1^4}{h_2^4} = \frac{V(f_{yy})}{V(f_{xx})}.$$

Nyní dosadíme do rovnice (6.1) $h_1^4 = h_2^4 \cdot \frac{V(f_{yy})}{V(f_{xx})}$, vypočítáme h_2 a pak h_1 .

Cv. 3 Rovnici (6.1) vynásobíme h_1 , rovnici (6.2) vynásobíme h_2 a rovnice sečteme

$$\underbrace{\frac{2V(K)}{nh_1h_2}}_{2 \cdot \text{AIV}} = 4 \cdot \underbrace{\frac{1}{4} \beta_2^2(K) (h_1^4 V(f_{xx}) + 2h_1^2 h_2^2 V(f_{xy}) + h_2^4 V(f_{yy}))}_{\text{AISB}}.$$

Odtud už plyne tvrzení.

Cv. 5 Ověříme, že hlavní minory jsou kladné:

$$M_{11} = |6| = 6 > 0, \quad M_{22} = \begin{vmatrix} 6 & -1 \\ -1 & 4 \end{vmatrix} = 23 > 0, \quad M_{33} = \begin{vmatrix} 6 & -1 & 2 \\ -1 & 4 & -3 \\ 2 & -3 & 9 \end{vmatrix} = 149 > 0.$$

Kapitola 7

Datové soubory

Tabulka 7.1: Hodnoty simulovaných dat z příkladu 2.1 – regrese

x	y	x	y	x	y	x	y	x	y
0,01	0,3002	0,02	0,9792	0,03	-1,0372	0,04	0,5519	0,05	0,3070
0,06	-0,4816	0,07	-0,0225	0,08	0,3848	0,09	2,0187	0,10	1,6268
0,11	-0,4240	0,12	1,7734	0,13	0,6198	0,14	0,2228	0,15	0,6049
0,16	0,1343	0,17	0,1602	0,18	0,9489	0,19	0,8871	0,20	0,8664
0,21	0,4661	0,22	-0,5034	0,23	0,4270	0,24	0,8499	0,25	0,2444
0,26	0,4820	0,27	0,2926	0,28	-0,2577	0,29	0,0068	0,30	-0,5664
0,31	0,2407	0,32	-0,8052	0,33	-0,7914	0,34	-0,6835	0,35	-1,7689
0,36	0,4086	0,37	-0,1572	0,38	-0,7018	0,39	0,3614	0,40	-1,1739
0,41	-0,3584	0,42	-0,4120	0,43	-0,1105	0,44	-0,0875	0,45	-0,6454
0,46	-0,1926	0,47	-0,2206	0,48	0,2188	0,49	0,4979	0,50	0,5546
0,51	-0,3811	0,52	0,1413	0,53	-0,4521	0,54	-0,3496	0,55	0,2547
0,56	1,0735	0,57	-0,0313	0,58	0,5820	0,59	0,3219	0,60	1,0265
0,61	-0,0495	0,62	0,5318	0,63	0,8049	0,64	1,0842	0,65	1,3028
0,66	0,5615	0,67	-0,2485	0,68	0,0955	0,69	-0,1043	0,70	1,5522
0,71	0,0104	0,72	0,6246	0,73	0,0784	0,74	0,5351	0,75	-0,3824
0,76	-0,7979	0,77	-0,9098	0,78	-0,0599	0,79	-0,5005	0,80	-0,6184
0,81	0,0816	0,82	-0,5874	0,83	-0,7349	0,84	-0,1342	0,85	-1,4160
0,86	-0,7407	0,87	-0,7340	0,88	-1,3214	0,89	-1,1229	0,90	-1,8261
0,91	-1,8089	0,92	-1,1517	0,93	-0,7882	0,94	0,2239	0,95	-1,2950
0,96	-0,7328	0,97	-0,7041	0,98	-1,4370	0,99	-0,4688	1,00	-0,8973

Tabulka 7.2: Hodnoty simulovaných dat z ukázkového příkladu 2.2 – regrese

x	y	x	y	x	y	x	y	x	y
0	0,1848	0,0204	0,4321	0,0408	-0,0322	0,0612	-0,4980	0,0816	-0,1456
0,1020	0,4379	0,1224	-0,1272	0,1429	0,4285	0,1633	0,2718	0,1837	0,6568
0,2041	-0,1325	0,2245	0,3708	0,2449	0,1820	0,2653	1,2750	0,2857	0,8176
0,3061	1,0174	0,3265	0,4667	0,3469	0,6689	0,3673	0,7680	0,3878	1,1553
0,4082	0,8496	0,4286	1,1259	0,4490	0,4616	0,4694	0,9023	0,4898	0,7931
0,5102	0,6047	0,5306	1,1178	0,5510	1,0450	0,5714	0,9589	0,5918	0,5856
0,6122	1,1626	0,6327	0,9237	0,6531	0,7113	0,6735	0,7370	0,6939	0,6072
0,7143	0,1737	0,7347	0,4766	0,7551	0,2761	0,7755	0,1754	0,7959	0,0686
0,8163	0,1642	0,8367	-0,2599	0,8571	0,4293	0,8776	0,2708	0,8980	0,0943
0,9184	0,0556	0,9388	-0,1630	0,9592	0,2710	0,9796	-0,0292	1,0000	-0,1786

Tabulka 7.3: Hodnoty simulovaných dat z příkladu 3.1 – hustota

0,0916	-0,5149	0,4746	0,1535	0,0676	0,2576	0,1307	-0,4707
-0,0812	-0,0730	-0,2660	0,8411	-0,4379	-0,2419	-0,3560	-0,5871
-0,0961	-0,1370	0,7650	-0,1245	-0,5321	0,8017	0,6173	-0,1148
-0,7531	-0,2223	-0,0780	0,1380	-0,1306	0,2217	2,1959	1,3747
1,5260	1,6294	1,7461	1,8397	2,0062	0,4854	1,7715	2,6212
1,4666	2,4669	2,1752	1,9855	2,0912	1,2175	1,9577	2,8020
2,0492	2,0207	1,6329	1,9846	2,1162	2,2132	1,8136	1,8818
3,0118	0,8708	3,1147	2,1688	2,5000	1,1679	1,7050	1,8610
2,2114	1,1649	2,2358	1,3936	2,0331	2,3262	2,1635	2,5413
2,5030	1,6745	2,1285	1,5278	1,3391	2,4624	2,0000	1,9725
2,4556	2,2973	2,1751	2,6251	2,4649	2,1199	1,6548	1,6742
2,5961	1,1941	1,9878	1,0256	2,5102	2,4309	2,0006	1,9646
0,7569	2,2906	0,9038	0,8404				

Tabulka 7.4: Hodnoty simulovaných dat z ukázkového příkladu 3.2 a 4.3 – hustota/distribuční funkce

0,3636	0,5101	0,1681	0,2509	0,2348	-0,0995	0,1452	-0,3361
0,1957	-0,6054	-0,2134	-0,5599	-0,4479	0,3291	0,1843	-0,1722
0,5500	-0,5962	-0,0562	-0,1096	0,2596	0,2940	-0,3160	-0,0094
-0,0499	0,1363	0,1991	0,2475	-0,2144	0,1691	0,1448	-0,3475
-0,4110	-0,0015	0,5774	-0,1492	0,0784	-0,2717	0,2437	-0,2367
0,0054	0,1886	0,4272	0,5760	0,0432	-0,3811	-0,3658	-0,2341
0,3520	-0,2376						

Tabulka 7.5: Hodnoty simulovaných dat z příkladu 4.1 – distribuční funkce

0,5603	-0,5920	0,0667	-0,3630	0,2023	-0,4002	0,3266	0,4929
1,1413	-0,1294	-1,4256	-0,5597	0,9031	-0,7148	0,6406	0,0827
0,9578	-1,3073	-0,1318	-0,8052	1,9387	0,5501	0,9193	-0,7055
-0,3124	-0,1816	0,7323	-0,1852	0,4677	-1,3679	-0,2359	-0,5491
-1,0514	0,3386	0,1880	0,0223	-0,8891	0,7517	0,2335	-0,1994
0,0153	-0,1747	-1,1668	-0,1904	-0,5542	-0,6528	-0,7709	-0,3557
-1,3351	0,6428	2,5201	1,9800	1,9652	1,2018	3,0187	1,8668
1,2855	3,3514	1,7752	1,4110	1,7062	1,1521	0,8799	4,5260
3,6555	2,3075	0,7429	1,1345	1,8235	2,7914	0,6680	-0,3299
0,5509	2,3335	2,3914	2,4517	1,8697	2,1837	1,5238	2,8620
0,6383	2,4550	1,1513	1,6651	2,5528	3,0391	0,8824	3,2607
2,6601	1,9321	1,8048	1,7824	1,6969	2,0230	2,0513	2,8261
3,5270	2,4669	1,7903	2,6252				

Tabulka 7.6: Hodnoty simulovaných dat z příkladu 5.1 – dvourozměrná hustota

$[x; y]$	$[x; y]$	$[x; y]$	$[x; y]$
$[-0,9466; -0,9065]$	$[0,3206; 3,7036]$	$[-0,1005; -0,8364]$	$[0,3740; 4,6725]$
$[-1,2990; 0,2211]$	$[1,5530; 2,2048]$	$[-0,0255; 3,1416]$	$[-0,6542; 1,8057]$
$[-0,1060; -0,5221]$	$[-0,4199; 2,2447]$	$[1,0924; 2,4151]$	$[0,0067; 3,2366]$
$[-0,2661; 3,7164]$	$[0,5607; 3,8076]$	$[-2,5687; 1,1808]$	$[0,3246; 3,6851]$
$[-0,2423; -0,4698]$	$[-0,3203; -0,5281]$	$[0,3535; 4,0112]$	$[0,5172; 1,0718]$
$[0,4428; 3,0552]$	$[1,2592; -0,9592]$	$[-0,2283; -1,5876]$	$[0,3677; -0,3296]$
$[-0,4947; 3,9449]$	$[0,3992; -0,5683]$	$[0,4789; 4,7000]$	$[0,8149; 3,9766]$
$[0,7421; 1,8418]$	$[0,5522; -0,5483]$	$[-0,7812; 3,5773]$	$[0,9459; -0,4850]$
$[-0,2062; 4,6773]$	$[-0,6864; 3,8091]$	$[-0,4114; -1,9323]$	$[1,3919; 4,3463]$
$[-0,5084; -0,7163]$	$[-0,0837; -1,6576]$	$[0,3252; 4,4022]$	$[0,5677; -1,0168]$
$[1,0639; 2,7468]$	$[-0,4263; -0,2424]$	$[0,0356; -1,6677]$	$[-0,3552; -0,4372]$
$[0,8428; 1,3214]$	$[-0,6772; 3,9948]$	$[0,0065; -0,3730]$	$[-1,1498; -0,6113]$
$[0,9380; 4,2303]$	$[-1,0416; 4,7185]$	$[0,0741; -1,8323]$	$[0,4332; 2,3411]$
$[1,6562; 2,5772]$	$[0,2473; 1,6168]$	$[1,0169; -0,8521]$	$[-0,1442; -1,4423]$
$[-0,9036; 2,6207]$	$[0,2175; 1,8817]$	$[1,2094; 2,2929]$	$[-0,1639; -0,8925]$
$[-0,2548; 4,5990]$	$[-0,0761; -0,3419]$	$[1,6432; 4,2209]$	$[-0,2305; -1,1597]$
$[-0,5887; 3,8457]$	$[-0,8657; 4,2465]$	$[-0,4201; 4,1738]$	$[0,7746; 1,5665]$
$[0,0917; 2,5265]$	$[-0,1711; 4,0729]$	$[0,0750; 3,5324]$	$[-0,4325; 2,2264]$
$[-0,8325; -1,7409]$	$[0,4678; 1,2271]$	$[-0,4280; -0,6307]$	$[0,5144; -1,3583]$
$[-1,0391; 2,6486]$	$[1,5309; 1,6144]$	$[0,0852; 3,9299]$	$[0,2001; 2,7036]$
$[2,0835; 2,9561]$	$[-0,6129; 3,7847]$	$[-0,6465; 1,7759]$	$[-0,2576; 1,2650]$
$[-0,5027; -0,9906]$	$[0,4180; -1,0515]$	$[0,9001; 4,4942]$	$[0,7250; 1,6023]$
$[0,0828; 3,0127]$	$[0,2756; 2,8724]$	$[0,4105; 3,7973]$	$[0,3790; -1,5109]$
$[-0,5589; -1,1707]$	$[-0,8010; 3,9085]$	$[0,5950; 4,4687]$	$[0,5173; 3,0249]$
$[-2,0882; 3,6987]$	$[-0,8476; 3,9186]$	$[0,2027; 3,6723]$	$[0,2096; -0,3292]$

Tabulka 7.7: Hodnoty simulovaných dat z ukázkového příkladu 5.3 – dvourozměrná hustota

$[x; y]$	$[x; y]$	$[x; y]$	$[x; y]$
[0,7747; -0,2338]	[-0,5707; 0,3048]	[1,0496; -0,5814]	[-0,2162; 0,6287]
[-0,8382; 1,0305]	[-0,3572; -0,1754]	[0,5395; -0,0050]	[0,6096; 0,2925]
[0,0925; 0,6947]	[0,1659; 0,3014]	[0,3654; 0,2292]	[-0,8424; -0,2788]
[-0,3359; -0,1337]	[-1,0566; 0,3541]	[-0,2593; 0,9716]	[0,5481; 0,3282]
[0,4142; -0,6963]	[1,3209; 0,2500]	[0,0963; -0,0532]	[-1,2451; -0,1409]
[0,9754; -0,1265]	[0,7403; -0,5079]	[-0,3364; -0,1852]	[0,1285; -0,4529]
[0,4500; -0,2107]	[0,2221; 0,0677]	[0,4020; 0,1007]	[-0,0651; -0,1245]
[-0,0224; -0,0881]	[0,0509; 0,0194]	[-0,4559; 0,2646]	[0,8215; 0,7400]
[0,0037; 0,3580]	[0,3963; -0,1655]	[-0,4936; 0,3453]	[-0,6208; -0,1418]
[-0,0525; -1,0281]	[-0,1459; -1,3552]	[0,6898; -0,9893]	[0,3499; -0,1369]
[0,3121; -0,3577]	[-0,8342; -0,4158]	[-0,5587; 0,0235]	[0,1167; -0,3805]
[-0,2008; -0,3640]	[0,7448; 0,3520]	[0,7743; 0,4517]	[-0,3353; 1,4068]
[0,5053; 0,1746]	[0,8057; 0,0337]	[0,1853; -0,3315]	[0,9373; -0,1127]
[0,6134; -0,0292]	[0,1236; -1,0137]	[0,5062; -0,1486]	[0,1122; 0,4327]
[-0,2395; -0,3295]	[0,2070; 0,2853]	[0,0209; 0,5724]	[-0,2105; 0,6219]
[1,0225; 0,4485]	[0,1278; -1,3047]	[1,5764; -0,3258]	[-0,1710; 0,5205]
[0,3258; 0,2279]	[0,9137; 0,0028]	[1,5757; 0,2523]	[0,1753; 0,6524]
[-0,4563; 0,0888]	[0,4638; -0,4380]	[-0,7620; -0,1040]	[-0,1581; -0,1200]
[0,1570; 0,5177]	[0,8600; 0,2380]	[0,8297; 0,3161]	[-0,0159; 0,0854]
[0,5334; -0,2092]	[-0,1855; 0,0353]	[0,3817; 0,1251]	[1,1058; -0,8072]
[0,0905; 0,1033]	[-0,1735; 0,5355]	[-0,2887; -0,0636]	[0,4137; 0,9078]
[0,7721; 0,3236]	[1,2273; -0,2427]	[0,0312; -0,0889]	[0,0941; -0,1589]
[-0,3653; 0,1392]	[0,9307; 0,4357]	[-0,0436; -0,1685]	[-0,2732; 0,2647]
[1,2611; 1,0134]	[0,4676; 0,5378]	[0,4136; 0,4765]	[-0,5806; -0,4704]
[0,2506; -0,3084]	[0,3003; 0,1400]	[-0,6306; -0,0581]	[-0,1304; -0,9382]

Tabulka 7.8: Hodnoty reálných dat z kapitoly 2, odstavce 7 – Huronské jezero

x	y	x	y	x	y	x	y	x	y
1875	10,38	1876	11,86	1877	10,97	1878	10,80	1879	9,79
1880	10,39	1881	10,42	1882	10,82	1883	11,40	1884	11,32
1885	11,44	1886	11,68	1887	11,17	1888	10,53	1889	10,01
1890	9,91	1891	9,14	1892	9,16	1893	9,55	1894	9,67
1895	8,44	1896	8,24	1897	9,10	1898	9,09	1899	9,35
1900	8,82	1901	9,32	1902	9,01	1903	9,00	1904	9,80
1905	9,83	1906	9,72	1907	9,89	1908	10,01	1909	9,37
1910	8,69	1911	8,19	1912	8,67	1913	9,55	1914	8,92
1915	8,09	1916	9,37	1917	10,13	1918	10,14	1919	9,51
1920	9,24	1921	8,66	1922	8,86	1923	8,05	1924	7,79
1925	6,75	1926	6,75	1927	7,82	1928	8,64	1929	10,58
1930	9,48	1931	7,38	1932	6,90	1933	6,94	1934	6,24
1935	6,84	1936	6,85	1937	6,90	1938	7,79	1939	8,18
1940	7,51	1941	7,23	1942	8,42	1943	9,61	1944	9,05
1945	9,26	1946	9,22	1947	9,38	1948	9,10	1949	7,95
1950	8,12	1951	9,75	1952	10,85	1953	10,41	1954	9,96
1955	9,61	1956	8,76	1957	8,18	1958	7,21	1959	7,13
1960	9,10	1961	8,25	1962	7,91	1963	6,89	1964	5,96
1965	6,80	1966	7,68	1967	8,38	1968	8,52	1969	9,74
1970	9,31	1971	9,89	1972	9,96				

Tato data jsou přístupná i v programu R, kde jsou původní hodnoty úrovně hladiny. V této tabulce je hodnota úrovně hladiny snížena o 570 stop.

Tabulka 7.9: Hodnoty reálných dat z kapitoly 2, odstavce 7 – krystaly ledu

x	y	x	y	x	y	x	y	x	y
50	19	60	20	60	21	70	17	70	22
80	25	80	28	90	21	90	25	90	31
95	25	100	29	100	30	100	33	105	32
105	35	110	28	110	30	110	30	115	30
115	31	115	36	120	25	120	28	120	36
125	28	130	31	130	32	135	25	135	34
140	26	140	33	145	31	150	33	150	36
155	33	155	41	160	30	160	37	160	40
165	32	170	35	180	38				

Tabulka 7.10: Hodnoty reálných dat z kapitoly 3, odstavce 7 – krabi

16,1	18,1	19,0	20,1	20,3	23,0	23,8	24,5	24,2	25,2
27,3	26,8	27,7	27,2	27,4	26,8	28,2	28,3	27,8	29,2
31,3	31,9	31,4	32,4	32,5	32,3	33,0	35,8	34,0	33,8
34,9	36,0	35,6	35,7	38,1	36,2	37,3	36,4	36,7	37,6
38,7	39,7	39,2	42,1	41,6	40,9	41,9	43,2	42,4	47,1
14,7	19,3	18,5	19,2	19,6	20,4	20,9	21,3	21,7	22,5
22,5	22,8	24,7	24,6	23,7	24,9	26,0	24,6	25,4	26,1
27,1	26,7	27,9	27,3	27,6	27,9	28,4	28,6	30,0	30,1
30,1	31,7	32,8	31,8	31,9	31,7	33,9	32,6	32,4	33,4
32,8	33,9	33,6	34,5	34,5	34,2	36,6	38,2	38,6	40,9
16,7	20,2	20,7	22,7	23,2	24,2	26,0	27,1	26,6	27,5
29,2	28,9	29,1	28,7	28,7	27,8	29,2	29,9	29,0	30,2
30,9	30,2	31,7	32,3	31,6	35,0	36,1	34,4	34,6	36,0
36,9	36,7	38,8	37,9	37,8	36,9	37,2	39,2	39,1	39,8
40,6	42,8	42,9	45,5	45,7	43,4	45,4	44,6	47,2	47,6
21,4	21,7	24,1	25,0	25,8	27,0	28,8	28,1	29,6	30,0
30,1	31,2	31,6	31,0	31,0	31,6	31,4	31,6	32,3	33,1
34,5	34,5	33,3	34,0	34,7	37,9	35,1	35,6	36,5	37,0
34,7	35,8	36,3	37,8	37,9	39,9	39,4	40,1	40,4	39,8
39,4	40,0	41,5	39,9	43,8	41,2	41,7	42,6	43,0	46,2

Tabulka 7.11: Hodnoty reálných dat z kapitoly 3, odstavce 7 – cholesterol

184	215	221	210	208	197	250	180	212	297	168	208	180	268	219
319	250	285	221	227	224	172	181	215	179	245	193	242	172	262
243	211	219	173	308	249	294	266	169	260	267	270	213	131	218
225	263	233	131	251	284	216	243	208	193	232	197	220	254	248
159	171	196	184	204	197	209	174	191	228	218	191	332	175	190
211	261	249	233	260	227	258	167	217	204	199	228	188	178	233
194	280	185	212	211	175	231	230	175	386	230	150	417	191	191
245	200	194	298	228	276	196	223	192	185	245	279	207	194	138
144	178	185	209	220	258	168	194	208	249	184	207	187	160	172
269	252	185	271	221	232	185	171	265	200	236	169	239	172	119
176	171	233	244	306	171	165	193	278	221	206	186	234	248	195
244	194	331	171	177	348	131	178	140	208	218	206	206	304	218
198	170	184	163	173	239	313	184	258	197	240	230	181	178	240
171	283	239	232	236	175	229	211	211	251	283	210	242	264	139
243	206	105	235	222	165	194	168	164	187	185	245	198	210	140
257	222	149	203	216	230	168	240	198	164	230	185	188	189	242
191	179	253	196	189	260	251	195	264	185	140	178	226	201	237
246	271	191	201	267	231	299	230	208	151	171	159	242	242	229
209	259	238	194	239	222	231	176	198	230	233	213	200	180	323
217	208	220	247	237	254	256	214	245	157	197	185	219	239	162
247	197	223	193	227	258	274	250	287	165	221	222	262	189	232
198	139	273	142	232	195	170	234	158	219	155	243	168	237	150
222	167	266	207	209	207	205	200	116	217	190	238	221	201	228
157	234	156	168	178	190	169	194	190	187	210	289	180	178	130
178	149	208	201	193	251	206	265	147	160	168				

Tabulka 7.12: Hodnoty reálných dat z kapitoly 4, odstavce 6 – pyrimidin

0,571	0,900	0,833	0,582	0,587	0,549	0,742	0,634
0,639	0,100	0,547	0,568	0,516	0,900	0,538	0,531
0,763	0,619	0,613	0,619	0,859	0,540	0,893	0,838
0,897	0,745	0,560	0,584	0,900	0,893	0,674	0,569
0,579	0,642	0,720	0,619	0,632	0,451	0,572	0,738
0,561	0,763	0,624	0,534	0,554	0,628	0,638	0,829
0,584	0,602	0,628	0,595	0,646	0,545	0,675	0,568
0,589	0,621	0,628	0,634	0,649	0,661	0,665	0,671
0,700	0,716	0,717	0,734	0,741	0,749	0,753	0,756
0,772	0,805						

Tabulka 7.13: Hodnoty reálných dat z kapitoly 4, odstavce 6 – Dyje – část 1.

16,9	13,2	11,4	15,4	40,4	18,9	5,44	2,15	1,96	2,20
5,74	6,15	8,69	10,4	23,2	3,53	5,66	5,68	3,55	5,59
3,14	3,02	0,937	1,49	4,33	4,90	5,41	5,00	5,03	5,13
14,4	15,1	11,1	11,7	6,34	4,82	7,02	7,86	5,61	8,87
33,1	14,5	8,09	4,79	5,06	5,28	9,12	12,0	8,14	10,2
12,6	14,1	7,34	6,14	5,00	4,75	4,82	8,31	25,5	6,74
10,9	11,0	16,6	18,0	19,4	12,4	20,7	29,4	14,7	17,8
15,3	17,4	24,1	49,4	16,4	18,4	36,4	31,8	29,5	12,1
15,6	12,0	8,29	7,87	7,48	7,24	9,30	14,5	64,7	69,2
26,0	29,9	16,5	13,9	13,7	8,62	22,9	13,0	20,1	22,0
15,0	45,0	16,2	14,4	10,9	9,43	11,5	6,06	5,17	5,14
6,09	6,96	6,13	5,42	5,61	5,93	5,97	6,56	5,28	4,07
4,24	4,34	4,09	4,98	8,58	31,1	14,3	16,8	12,1	8,91
7,87	6,48	9,65	17,6	15,2	21,7	29,6	16,4	8,52	6,23
4,75	3,82	4,09	4,87	7,26	7,56	7,31	20,2	29,1	5,64
4,03	4,25	4,10	4,38	7,11	5,43	6,71	7,09	8,84	16,1
39,4	27,7	8,03	7,22	5,35	4,95	5,37	3,98	3,89	5,38
20,9	50,5	20,9	9,01	4,90	4,90	4,53	5,57	5,39	9,48
9,31	5,34	5,10	3,97	3,10	2,49	1,96	2,69	2,67	8,43
10,4	12,8	8,67	4,95	7,11	6,07	5,85	7,34	6,66	8,69
6,57	3,46	1,59	1,92	5,56	5,27	10,2	15,8	12,2	6,41
28,4	6,89	8,04	5,36	9,47	11,2	14,0	3,94	3,75	6,55
19,4	14,2	11,3	6,15	4,28	3,65	3,87	4,10	4,84	5,34
15,4	9,31	6,93	4,87	3,69	5,21	7,20	10,3	9,35	5,14
4,87	4,94	4,87	5,06	4,39	4,06	4,78	4,94	10,6	4,67
5,14	5,24	7,29	9,40	23,90	10,50	18,40	25,20	9,01	5,57
6,01	15,9	9,10	5,90	5,88	6,26	9,41	20,1	16,8	13,0
17,9	8,87	6,81	6,38	7,58	7,36	9,97	8,72	9,18	14,8
12,6	13,0	7,66	6,30	5,72	6,38	7,82	7,43	8,51	8,91
7,94	6,69	9,70	18,4	8,60	6,84	5,87	5,48	6,97	5,99
8,55	10,9	5,92	7,49	6,71	5,87	5,83	6,55	9,04	11,8
8,11	8,11	6,58	6,15	7,59	6,35	20,9	12,8	4,97	9,47
6,22	25,0	9,99	13,1	11,3	8,80	10,5	18,6	11,3	8,41
8,02	24,0	9,17	7,40	7,15	8,41	6,92	9,77	7,67	25,2

Tabulka 7.14: Hodnoty reálných dat z kapitoly 4, odstavce 6 – Dyje – část 2.

7,60	6,09	24,2	11,8	7,63	6,52	7,88	8,27	6,56	9,86
5,25	3,92	5,32	21,6	11,3	8,90	6,84	4,53	4,71	4,74
4,78	4,53	4,67	4,64	4,57	5,04	5,49	5,02	4,82	4,70
4,58	5,17	5,71	6,70	7,88	7,00	34,7	37,2	37,5	69,9
18,4	7,79	5,37	4,65	6,16	6,20	10,0	31,2	8,34	7,36
5,33	5,45	5,42	11,4	22,2	7,26	5,90	14,3	22,4	23,0
24,8	14,3	5,55	27,8	7,97	4,40	4,62	5,34	5,05	4,58
17,5	16,7	12,7	9,37	5,05	6,10	4,61	4,17	5,26	5,84
5,68	7,01	5,65	5,56	23,7	18,9	7,60	8,16	6,12	4,42
4,33	4,20	4,40	11,5	5,28	4,70	17,5	27,5	7,72	6,47
5,48	5,23	5,14	5,19	4,89	10,5	8,18	8,68	8,00	8,05
6,82	8,88	6,67	5,71	5,74	4,86	5,13	5,07	5,45	4,90
4,87	9,28	24,5	9,70	10,0	7,18	4,17	4,55	4,71	4,66
4,49	4,18	4,67	12,0	8,80	5,61	4,40	4,70	4,70	4,08
3,60	3,92	3,69	3,62	3,60	4,67	4,14	3,48	8,85	5,65
3,88	4,78	4,10	27,6	18,4	9,33	3,64	8,10	8,28	4,62
19,9	5,55	4,38	3,66	4,07	4,10	32,5	11,1	10,4	6,39
5,21	5,58	5,23	4,61	4,69	4,35	5,14	15,5	8,87	41,5
38,9	14,0	8,39	6,73	6,40	5,81	5,35	4,96	4,91	4,16
4,65	6,23	15,5	6,79	6,93	6,09	5,54	6,03	6,51	5,11
3,82	3,85	4,42	4,00	9,98	27,0	11,0	6,18	7,21	6,77
5,30	6,65	5,76	21,6	9,67	19,2	4,55	26,7	18,9	7,75
6,29	9,09	8,43	7,83	6,80	4,55	4,66	11,2	16,4	6,72
4,97	5,27	4,59	5,81	4,11	7,52	12,5	22,5	19,5	22,8
17,2	14,6	6,88	5,96	5,69	6,64	6,05	5,57	5,48	4,24
7,19	16,7	17,7	23,3	9,60	9,37	8,98	7,41	4,96	3,90
3,47	2,85	3,03	3,67	2,95	6,80	10,4	5,67	5,50	5,66
5,39	4,91	5,14	5,98	5,69	9,23	22,5	20,2	20,7	16,9
8,35	28,9	8,81	6,25	5,02	16,7	22,7	12,7	17,7	14,0
8,13	13,6	11,1	8,27	7,68	5,41	4,73	4,46	29,8	25,2
39,3	42,9	24,9	29,5	24,5	10,5	7,57	7,90	8,77	12,5
10,6	9,64	44,2	29,5	8,45	8,42	8,02	7,04	5,83	4,13
3,83	4,99	12,7	5,56	6,72	6,44	10,1	7,63	8,41	7,66
7,28	6,14	4,79	3,64	2,85	3,15	3,93	7,51	6,70	7,76
6,60	4,88	3,44	3,77						

Tabulka 7.15: Hodnoty reálných dat z kapitoly 5, odstavce 6 – studie UNICEF

$[x; y]$	$[x; y]$	$[x; y]$	$[x; y]$	$[x; y]$	$[x; y]$	$[x; y]$
[19; 73]	[235; 53]	[43; 69]	[257; 43]	[109; 56]	[225; 48]	[205; 52]
[28; 71]	[99; 67]	[77; 63]	[38; 66]	[39; 71]	[143; 42]	[72; 63]
[19; 72]	[20; 68]	[108; 51]	[153; 51]	[45; 67]	[105; 72]	[95; 62]
[175; 48]	[29; 73]	[24; 69]	[94; 57]	[155; 50]	[35; 73]	[68; 69]
[132; 44]	[260; 45]	[123; 53]	[123; 43]	[138; 54]	[93; 64]	[107; 61]
[109; 60]	[38; 69]	[76; 63]	[169; 48]	[32; 67]	[79; 60]	[77; 60]
[158; 54]	[183; 52]	[122; 50]	[107; 56]	[126; 47]	[202; 42]	[100; 54]
[100; 57]	[183; 52]	[61; 68]	[124; 45]	[138; 56]	[141; 52]	[165; 51]
[180; 44]	[136; 53]	[91; 59]	[183; 40]	[197; 47]	[197; 39]	[231; 52]
[200; 46]	[111; 52]	[72; 68]	[183; 40]	[265; 46]	[211; 45]	[316; 40]
[172; 44]	[190; 41]					

Tabulka 7.16: Hodnoty reálných dat z kapitoly 5, odstavce 6 – koncentrace lipidů – část 1.

$[x; y]$	$[x; y]$	$[x; y]$	$[x; y]$	$[x; y]$
[184; 145]	[215; 168]	[221; 432]	[210; 92]	[208; 112]
[197; 87]	[250; 118]	[180; 80]	[212; 130]	[297; 232]
[168; 208]	[208; 262]	[180; 102]	[268; 154]	[219; 454]
[319; 418]	[250; 161]	[285; 930]	[221; 268]	[227; 146]
[224; 124]	[172; 106]	[181; 119]	[215; 325]	[179; 126]
[245; 166]	[193; 290]	[242; 179]	[172; 207]	[262; 88]
[243; 126]	[211; 306]	[219; 163]	[173; 300]	[308; 260]
[249; 146]	[294; 135]	[266; 164]	[169; 158]	[260; 98]
[267; 192]	[270; 110]	[213; 261]	[131; 96]	[218; 567]
[225; 240]	[263; 142]	[233; 340]	[131; 137]	[251; 189]
[284; 245]	[216; 112]	[243; 50]	[208; 220]	[193; 188]
[232; 328]	[197; 291]	[220; 75]	[254; 153]	[248; 312]
[159; 125]	[171; 78]	[196; 130]	[184; 255]	[204; 150]
[197; 265]	[209; 82]	[174; 117]	[191; 233]	[228; 130]
[218; 123]	[191; 90]	[332; 250]	[175; 246]	[190; 120]
[211; 304]	[261; 174]	[249; 256]	[233; 101]	[260; 127]
[227; 172]	[258; 145]	[167; 80]	[217; 227]	[204; 84]
[199; 153]	[228; 149]	[188; 148]	[178; 125]	[233; 141]
[194; 278]	[280; 218]	[185; 115]	[212; 171]	[211; 124]
[175; 148]	[231; 181]	[230; 90]	[175; 489]	[386; 162]
[230; 158]	[150; 426]	[417; 198]	[191; 115]	[191; 136]
[245; 120]	[200; 152]	[194; 183]	[298; 143]	[228; 142]
[276; 199]	[196; 103]	[223; 80]	[192; 101]	[185; 130]
[245; 257]	[279; 317]	[207; 316]	[194; 116]	[138; 91]
[144; 125]	[178; 84]	[185; 100]	[209; 89]	[220; 153]
[258; 151]	[168; 126]	[194; 196]	[208; 201]	[249; 200]
[184; 182]	[207; 150]	[187; 115]	[160; 125]	[172; 146]
[269; 84]	[252; 233]	[185; 110]	[271; 128]	[221; 140]
[232; 258]	[185; 256]	[171; 165]	[265; 156]	[200; 68]
[236; 152]	[169; 112]	[239; 154]	[172; 140]	[119; 84]
[176; 217]	[171; 108]	[233; 127]	[244; 108]	[306; 408]
[171; 120]	[165; 121]	[193; 170]	[278; 152]	[221; 179]

Tabulka 7.17: Hodnoty reálných dat z kapitoly 5, odstavce 6 – koncentrace lipidů – část 2.

$[x; y]$	$[x; y]$	$[x; y]$	$[x; y]$	$[x; y]$
[206; 133]	[186; 273]	[234; 135]	[248; 142]	[195; 363]
[244; 177]	[194; 125]	[331; 134]	[171; 90]	[177; 133]
[348; 154]	[131; 61]	[178; 101]	[140; 99]	[208; 148]
[218; 207]	[206; 148]	[206; 107]	[304; 149]	[218; 96]
[198; 103]	[170; 284]	[184; 184]	[163; 156]	[173; 56]
[239; 97]	[313; 256]	[184; 222]	[258; 210]	[197; 158]
[240; 196]	[230; 162]	[181; 104]	[178; 100]	[240; 441]
[171; 170]	[283; 424]	[239; 92]	[232; 131]	[236; 148]
[175; 153]	[229; 242]	[211; 91]	[211; 122]	[251; 152]
[283; 199]	[210; 217]	[242; 85]	[264; 269]	[139; 173]
[243; 112]	[206; 201]	[105; 36]	[235; 144]	[222; 229]
[165; 151]	[194; 400]	[168; 91]	[164; 80]	[187; 390]
[185; 231]	[245; 322]	[198; 124]	[210; 95]	[140; 102]
[257; 402]	[222; 348]	[149; 237]	[203; 170]	[216; 101]
[230; 304]	[168; 131]	[240; 221]	[198; 149]	[164; 76]
[230; 146]	[185; 116]	[188; 220]	[189; 84]	[242; 144]
[191; 115]	[179; 126]	[253; 222]	[196; 141]	[189; 135]
[260; 144]	[251; 117]	[195; 137]	[264; 259]	[185; 120]
[140; 164]	[178; 109]	[226; 72]	[201; 297]	[237; 88]
[246; 87]	[271; 89]	[191; 149]	[201; 92]	[267; 199]
[231; 161]	[299; 93]	[230; 137]	[208; 77]	[151; 73]
[171; 135]	[159; 82]	[242; 180]	[242; 248]	[229; 296]
[209; 376]	[259; 240]	[238; 156]	[194; 272]	[239; 38]
[222; 151]	[231; 145]	[176; 166]	[198; 333]	[230; 492]
[233; 142]	[213; 130]	[200; 101]	[180; 202]	[323; 196]
[217; 327]	[208; 149]	[220; 172]	[247; 137]	[237; 400]
[254; 170]	[256; 271]	[214; 223]	[245; 446]	[157; 59]
[197; 101]	[185; 168]	[219; 267]	[239; 137]	[162; 91]
[247; 91]	[197; 347]	[223; 154]	[193; 259]	[227; 202]
[258; 328]	[274; 323]	[250; 160]	[287; 209]	[165; 155]
[221; 156]	[222; 108]	[262; 169]	[189; 176]	[232; 583]
[198; 105]	[139; 54]	[273; 146]	[142; 111]	[232; 161]

Kapitola 8

Přílohy

Tabulka 8.1: Optimální jádra pro $\nu = 0$ a $\nu = 1$

$\nu = 0$	
k	$K_{opt,0,k}$
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	$\frac{315}{4096}(x^2 - 1)(715x^6 - 1001x^4 + 385x^2 - 35)$
10	$-\frac{3465}{65536}(x^2 - 1)(4199x^8 - 7956x^6 + 4914x^4 - 1092x^2 + 63)$
12	$\frac{9009}{524288}(x^2 - 1)(52003x^{10} - 124355x^8 + 106590x^6 - 39270x^4 + 5775x^2 - 231)$

$\nu = 1$	
k	$K_{opt,1,k} = K^{(1)}$
3	$\frac{15}{4}x(x^2 - 1)$
5	$-\frac{105}{32}x(x^2 - 1)(9x^2 - 5)$
7	$\frac{315}{32}x(x^2 - 1)(143x^4 - 154x^2 + 35)$
9	$-\frac{3465}{4096}x(x^2 - 1)(1105x^6 - 1755x^4 + 819x^2 - 105)$
11	$\frac{45045}{65536}x(x^2 - 1)(6783x^8 - 14212x^6 + 10098x^4 - 2772x^2 + 231)$
13	$-\frac{45045}{524288}x(x^2 - 1)(260015x^{10} - 676039x^8 + 646646x^6 - 277134x^4 + 51051x^2 - 3003)$

Tabulka 8.2: Optimální jádra pro $\nu = 2$ a $\nu = 3$

$\nu = 2$	
k	$K_{opt,2,k} = K^{(2)}$
4	$-\frac{105}{16}(x^2 - 1)(5x^2 - 1)$
6	$\frac{315}{64}(x^2 - 1)(77x^4 - 58x^2 + 5)$
8	$-\frac{3465}{2048}(x^2 - 1)(1755x^6 - 2249x^4 + 721x^2 - 35)$
10	$\frac{45045}{8192}(x^2 - 1)(3553x^8 - 6392x^6 + 3618x^4 - 672x^2 + 21)$
12	$-\frac{45045}{262144}(x^2 - 1)(676039x^{10} - 1562351x^8 + 1271974x^6 - 429726x^4 + 52899x^2 - 1155)$
14	$\frac{765765}{1048576}(x^2 - 1)(884925x^{12} - 2495270x^{10} + 2653027x^8 - 1315028x^6 + 301587x^4 - 26598x^2 + 429)$

$\nu = 3$	
k	$K_{opt,3,k}$
5	$\frac{945}{16}x(x^2 - 1)(7x^2 - 3)$
7	$-\frac{10395}{64}x(x^2 - 1)(39x^4 - 38x^2 + 7)$
9	$\frac{135135}{2048}x(x^2 - 1)(935x^6 - 1405x^4 + 597x^2 - 63)$
11	$-\frac{135135}{8192}x(x^2 - 1)(29393x^8 - 59432x^6 + 40018x^4 - 10032x^2 + 693)$
13	$\frac{2297295}{262144}x(x^2 - 1)(382375x^{10} - 969703x^8 + 895622x^6 - 364078x^4 + 61347x^2 - 3003)$
15	$-\frac{43648605}{1048576}x(x^2 - 1)(510255x^{12} - 1554570x^{10} + 1825625x^8 - 1034540x^6 + 288145x^4 - 35178x^2 + 1287)$

Literatura

- [1] Anděl, J.: Základy matematické statistiky. Matfyzpress, Praha (2005) ISBN 80-86732-40-1
- [2] Forbelská, M., Koláček, J.: Pravděpodobnost a statistika I. Elektronický učební text, Masarykova univerzita, Brno (2012) <http://is.muni.cz/elportal/?id=1130308>
- [3] Horová, I., Koláček, J., Zelinka, J.: Kernel Smoothing in Matlab. Theory and Practise of Kernel Smoothing. World Scientific, Singapur (2012) ISBN 978-981-4405-48-5
- [4] Horová, I., Vieu, P., Zelinka, J.: Optimal Choice of Nonparametric Estimates of a Density and of its Derivatives. *Statistics & Decisions* 20, 355–378 (2002)
- [5] Horová, I., Zelinka, J.: Contribution to the bandwidth choice for kernel density estimates, *Comput. Stat.* 22, 31–47 (2007)
- [6] Köhler, M., Schindler, A., Sperlich, S.: A Review and Comparison of Bandwidth Selection Methods for Kernel Regression. Discussion Paper No. 95, Georg-August-Universität Göttingen (2011)
- [7] Koláček, J.: Jádrové odhady regresní funkce. Disertační práce, Masarykova univerzita, Brno (2005) http://is.muni.cz/th/19999/prif_d/
- [8] Müller, H.-G.: Nonparametric regression analysis of longitudinal data. Springer, New York (1988) ISBN 978-0-387-96844-5
- [9] Müller, H.-G.: Smooth optimum kernel estimators near endpoints. *Biometrika* 78, 521–530 (1991)
- [10] Scott, D.W., Gorry, G.A., Hoffman, R.G., Barboriak, J.J., Gotto, A.M.: A new approach for evaluating risk factors in coronary artery disease: a study of lipid concentrations and severity of disease in 1847 males, *Circulation* 62, 477–484 (1980)
- [11] Scott, D.W.: Multivariate density estimation: Theory, practise, and visualization. Wiley, New York (1992) ISBN 0-471-54770-0
- [12] Silverman, B.W.: Density estimation for statistics and data analysis. Chapman and Hall, London (1986) ISBN 0-412-24620-1
- [13] Vopatová, K.: Volba vyhlazovacích parametrů pro jádrové odhady hustot. Disertační práce, Masarykova univerzita, Brno (2011) http://is.muni.cz/th/63985/prif_d/
- [14] Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman and Hall, London (1995) ISBN 0-412-55270-1