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 „hhovořit 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 šířce vyhlazovacího okna, která řídí hladkost odhadu.

Budeme zabývat následujícími otázkami:

- Jaké jsou statistické vlastnosti jádrových odhadů.
- Jaký vliv má tvar jádra na tvar odhadu.
- 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.

Příslušný toolbox pro program Matlab je dostupný na adrese:

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

V příloze (kapitola 6) 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. Koláčka (jsou dostupná na Elportálu Informačního systému).

Na tomto místě bych ráda poděkovala Kamile Vopatové za pomoc při sazbě tohoto textu a za přispění ke kapitole 5 a podkapitolám o reálných datech.

Obsah

Seznam použitého značení

h	vyhlazovací parametr
$\mathbf{H}$	vyhlazovací matice
$m(\cdot)$	regresní funkce
$f(\cdot)$	hustota pravděpodobnosti
$F(\cdot)$	distribuční funkce
$\hat{h}$	odhad vyhlazovacího parametru h
$\widehat{\sigma}$	odhad směrodatné odchylky σ
$\widehat{m}$	odhad regresní funkce m
$\hat{f}$	odhad hustoty f
$\widehat{F}$	odhad distribuční funkce F
$\int$	značí integrál $\int_{-\infty}^{\infty}$, pokud není uvedeno jinak
$K(\cdot)$	jádrová funkce (jádro)
$V(K)$	$V(K) = \int K^2(x)\,\mathrm{d}x$
$\beta_k(K)$	$\beta_k(K) = \int x^k K(x)\,\mathrm{d}x$
$f * g$	konvoluce funkcí f a g, $(f * g)(x) = \int f(t)g(x-t)\,\mathrm{d}t$
$W(\cdot)$	intergál z jádra, $W(x) = \int^x K(t)\,\mathrm{d}t$
NW	Nadarayův-Watsonův typ odhadu
PC	Priestleyův-Chaův typ odhadu
LL	lokálně lineární typ odhadu
GM	Gasserův-Müllerův typ odhadu

MSE	střední kvadratická chyba (mean square error)
MISE	střední intergální kvadratická chyba (mean integrated square error)
AMISE	asymptotická střední integrální kvadratická chyba (asymptotic mean integrated square error)
AIV	asymptotický tvar integrálu z rozptylu (asymptotic intergated variance)
AISB	asymptotický tvar integrálu z 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

Kapitola 1

Jádrové funkce a jejich vlastnosti

Výstupy z výukové jednotky

Student

- bude znát základní třídy jádrových funkcí, jejich vlastnosti a metody jejich konstrukce.

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:

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

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

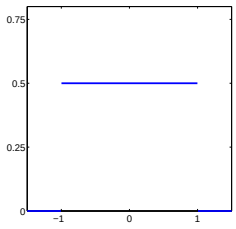
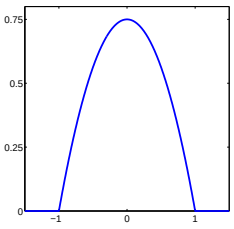
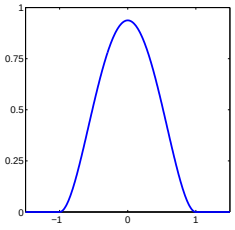
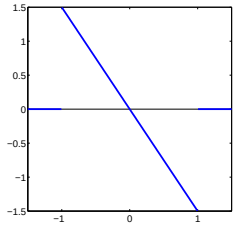
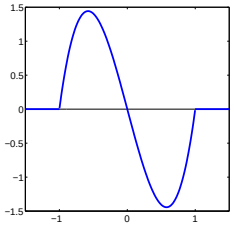
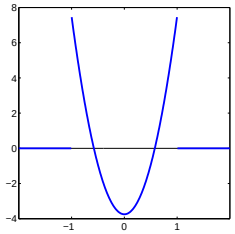
Taková funkce K se nazývá jádro řádu k a třída všech těchto funkcí se označuje $S_{\nu k}$.

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

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

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

Tabulka 1.1: Jádra

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

1.1 Jádra s minimálním rozptylem

Předpokládejme, že $K \in S_{\nu k}$, $0 \le \nu \le k-2$, ν a k jsou obě sudá nebo lichá [1]. Uvažujme funkcionál $V(K) = \int_{-1}^{1} K^2(x)\,\mathrm{d}x$ 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. Mají $k-2$ různých kořenů v intervalu $(-1,1)$. Obecný vztah pro jádra s minimálním rozptylem lze nalézt ve [4].

Příklad 1.2. Jádra s minimálním rozptylem:

$$\begin{aligned} S_{02}\text{:} &\quad K(x) = \tfrac{1}{2} I_{[-1,1]}(x) \\ S_{13}\text{:} &\quad K(x) = -\tfrac{3}{2} x I_{[-1,1]}(x) \\ S_{24}\text{:} &\quad K(x) = -\tfrac{15}{4}(1-3x^2) I_{[-1,1]}(x) \end{aligned}$$

Poznámka 1.1. Jádra s minimálním rozptylem mají skoky v koncových bodech intervalu $[-1,1]$, což negativně ovlivňuje hladkost výsledného odhadu.

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

Tabulka 1.2: Optimální jádra pro $\nu = 0, 1, 2$

$\nu = 0$		
k	δ_{0k}	$K_{opt,0,k}$
2	1,7188	$-\frac{3}{4}(x^2-1)$
4	2,0165	$\frac{15}{32}(x^2-1)(7x^2-3)$
6	2,0834	$-\frac{105}{256}(x^2-1)(33x^4-30x^2+5)$
8	2,1021	$\frac{315}{4096}(x^2-1)(715x^6-1001x^4+385x^2-35)$
$\nu = 1$		
k	δ_{1k}	$K_{opt,1,k}$
3	1,4204	$\frac{15}{4}x(x^2-1)$
5	1,7656	$-\frac{105}{32}x(x^2-1)(9x^2-5)$
7	1,8931	$\frac{315}{32}x(x^2-1)(143x^4-154x^2+35)$
$\nu = 2$		
k	δ_{2k}	$K_{opt,2,k}$
4	1,3925	$-\frac{105}{16}(x^2-1)(5x^2-1)$
6	1,6964	$\frac{315}{64}(x^2-1)(77x^4-58x^2+5)$
8	1,8269	$-\frac{3465}{2048}(x^2-1)(1755x^6-2249x^4+721x^2-35)$

1.2 Optimální jádra

Při vyšetřování statistických vlastností se setkáváme s následujícím funkcionálem

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

který lze zkráceně psát $T(K) = \left(|\beta_k(K)|^{2\nu+1} V(K)^{(k-\nu)}\right)^{2/(2k+1)}$. Jádra, pro která tento funkcionál 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.

Příklad 1.3. Optimální jádra:

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

Přehled vybraných optimálních jader je uveden v tabulce 1.2. Obecný vzorec pro tvar optimálních jader lze nalézt např. v [4].

Užitečný při odvozování statistických vlastností odhadů bude i následující pojem.

Definice 1.2. Pro jádro $K \in S_{\nu k}$ definujeme kanonický faktor

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

Shrnutí
Funkcionály závisející na jádře K $$\beta_k(K) = \int_{-1}^{1} x^k K(x)\,\mathrm{d}x \qquad V(K) = \int_{-1}^{1} K^2(x)\,\mathrm{d}x$$ $$T(K) = \left(|\beta_k(K)|^{2\nu+1} V(K)^{(k-\nu)}\right)^{\frac{2}{2k+1}} \qquad \delta_{\nu k} = \left(\frac{V(K)}{\beta_k^2(K)}\right)^{\frac{1}{2k+1}}$$
Jádra s minimálním rozptylem minimalizují funkcionál $V(K)$. Optimální jádra minimalizují funkcionál $T(K)$.
Podrobnější popis jader a vzorců s nimi souvisejících lze nalézt v toolboxu Matlabu.

Dodatky a cvičení

1. Tvrzení: 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
$$K'_{opt,\nu,k}(x) = K(x), \quad x \in [-1,1]$$
Důkaz viz [4]. Jako příklad lze uvést jádro $K_{opt,0,2}(x) = \frac{3}{4}(1-x^2) \in S_{02}$ a $K'_{opt,0,2}(x) = K_{1,3}(x) = -\frac{3}{2}x \in S_{13}$.

2. Nechť $K \in S_{\nu k}$, $\nu \in \mathbb{N}$, pro $\delta > 0$ položíme
$$K_\delta(\cdot) = \frac{1}{\delta^{\nu+1}} K\left(\frac{\cdot}{\delta}\right).$$
Jádra K, K_δ nazýváme ekvivalentní. Dokažte: Funkcionál $T(K)$ je invariantní vzhledem k této transformaci, tj. $T(K) = T(K_\delta)$.

3. Nechť jsou dány funkce f a g, pro které platí $\int f^2(x)\,\mathrm{d}x < \infty$ a $\int g^2(x)\,\mathrm{d}x < \infty$. Konvoluci $f * g$ definujeme vztahem
$$(f * g)(x) = \int f(t) g(x-t)\,\mathrm{d}t.$$
Vlastnosti konvoluce
   - $f * g = g * f$,
   - $f * (g * h) = (f * g) * h$,
   - $f * (g + h) = f * g + f * h$.

   Ukažte, že pro jádrovou funkci $K \in S_{02}$ platí vztah
$$\int (K * K)(x) x^j\,\mathrm{d}x = \begin{cases} 1 & j = 0, \\ 0 & j = 1, \\ 2\beta_2(K) & j = 2. \end{cases}$$

Kapitola 2

Jádrové odhady regresní funkce

Výstupy z výukové jednotky

Student

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

1 Motivace

Předpokládejme, že pro pevné nebo náhodné hodnoty nezávisle proměnné X máme k dispozici naměřené hodnoty závisle proměnné Y. Chceme-li tato data analyzovat, musíme nalézt vhodný funkční vztah mezi těmito proměnnými.

Jestliže dvojice bodů $(x_i, Y_i), i = 1, \ldots, n$, znázorníme graficky, pak pouhý pohled na takový 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

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

kde m je neznámá regresní funkce, x_i, $i = 1, \ldots, n$, jsou body plánu a ε_i, $i = 1, \ldots, n$, jsou chyby měření, o nichž se předpokládá, že jsou nezávislými, identicky rozdělenými náhodnými veličinami splňujícími podmínky

$$E\varepsilon_i = 0, \quad \operatorname{var} \varepsilon_i = \sigma^2, \quad i = 1, \ldots, n. \tag{2.2}$$

Poznámka 1.1. Jsou-li body plánu uspořádaná nenáhodná čísla, mluvíme o regresním modelu s pevným plánem. V případě, že body plánu $X_1, \ldots, X_n$ jsou náhodné veličiny se stejnou hustotou f, jedná se o regresní model s náhodným plánem. U něj jsou všechny předpoklady podobné jako u modelu s pevným plánem (podrobněji např. [13]).

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

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

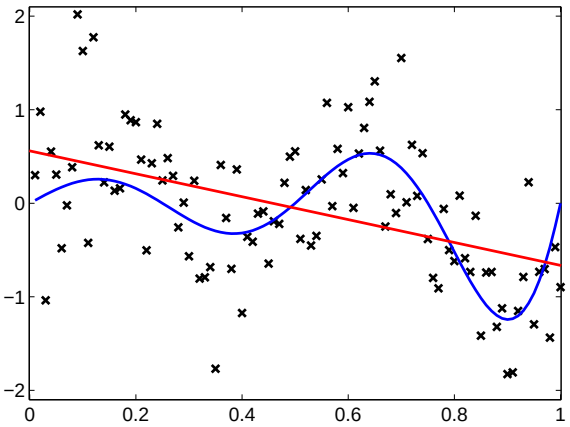
Cílem regresní analýzy je nalézt vhodnou aproximaci $\widehat{m}$ neznámé funkce m. Tento proces odhadu regresní funkce se obvykle nazývá vyhlazování. K tomuto úkolu lze přistoupit dvěma způsoby – parametricky a neparametricky. Příkladem parametrického odhadu regresní funkce je regresní přímka vyjadřující 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 1.1. Obr. 2.1 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 $x_i = 1/100$, $i = 1, \ldots, 100$, a chyby ε_i, $i = 1, \ldots, 100$, mají normální rozdělení $N(0; 0{,}25)$. (Data jsou v tabulce 6.1.)



Obrázek 2.1: Simulovaná data (×) s regresní přímkou (—) a původní funkcí (—)

Předpokládejme, že hledaná křivka je přímka a známou metodou nejmenších čtverců určíme rovnici této přímky. Obr. 2.1 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

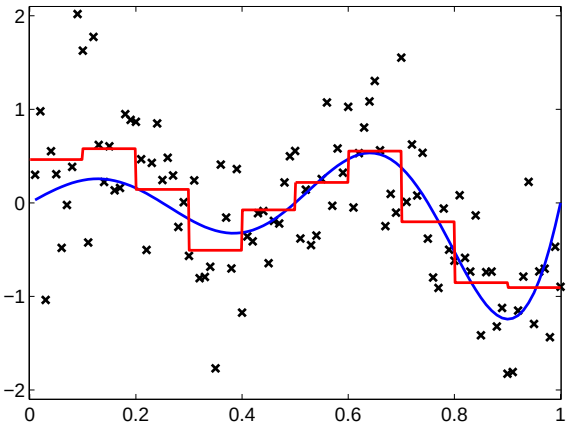
Pokud jde o historii neparametrických metod, připomeňme, že v r. 1857 saský ekonom Engel analyzoval data týkající se nákladů na domácnost a pro vyjádření závislosti použil schodovitou (tj. po částech konstantní funkcí), kterou dnes nazýváme regresogram. Regresogram užívá

stejné základní myšlenky jako histogram pro odhad hustoty. Tato myšlenka spočívá v rozdělení množiny hodnot proměnné X na intervaly B_j, $j = 1, \ldots, J$, a za odhad v bodě $x \in B_j$ se vezme průměr hodnot Y na tomto subintervalu, tj.

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

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

Výsledek aplikace regresogramu na simulovaná data příkladu 1.1 je znázorněn na obr. 2.2. Vidíme, že tento odhad „vhodně“ vystihuje tvar funkce, ale výsledný odhad je příliš hrubý.

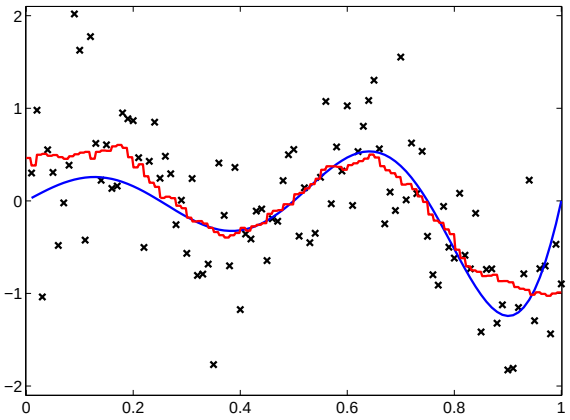


Obrázek 2.2: Regresogram (—) pro simulovaná data (×) z příkladu 1.1 s původní funkcí (—)

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. Přesněji

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

Obr. 2.3 ilustruje aplikaci této metody na simulovaných datech příkladu 1.1. Uvedené metody



Obrázek 2.3: Klouzavý průměr (—) pro simulovaná data z příkladu 1.1

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

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

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

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

1. Nadarayovy-Watsonovy odhady (1964)

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

2. Priestleyovy-Chaovy odhady (1972)

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

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

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

   kde

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

4. Gasserovy-Müllerovy odhady (1979)

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

   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

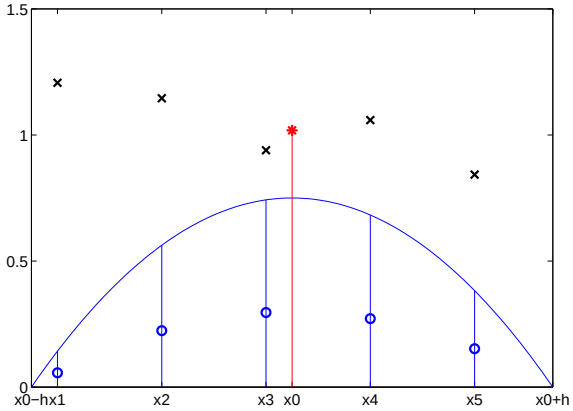
$$\widehat{m}_j(x,h) = \sum_{i=1}^{n} W_i^{(A)}(x,h) Y_i,$$

kde váhy $W_i^{(A)}$ a A značí příslušný typ odhadu NW, PC, LL, GM. V dalším textu budeme zkráceně psát W_i.

V mnoha aplikacích je užitečný zejména Nadarayův-Watsonův odhad $\widehat{m}_{NW}$. 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

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

Obrázek 2.4 ilustruje konstrukci odhadu v bodě x_0, který je založen na pěti pozorováních (x_1, Y_1), $\ldots$, (x_5, Y_5) (černé křížky). Modrá parabola reprezentuje Epanečnikovo jádro K_h a modré kroužky znázorňují hodnoty vah $W_i = K_h(x_i - x_0) / \sum_{i=1}^{5} K_h(x_i - x_0)$ pro $i = 1, \ldots, 5$. Výsledný odhad regresní funkce $\widehat{m}$ v bodě x_0 je označen červenou hvězdičkou.



Obrázek 2.4: Ilustrace Nadarayova-Watsonova odhadu v bodě x_0

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

$$\widehat{m}_{NW}(x_i, h) \to \frac{K(0)Y_i}{K(0)} = Y_i$$

To znamená, že při malé šířce vyhlazovacího okna odhad reprodukuje data (viz obr. 2.5(a)). Dále, pro $h \to \infty$ platí

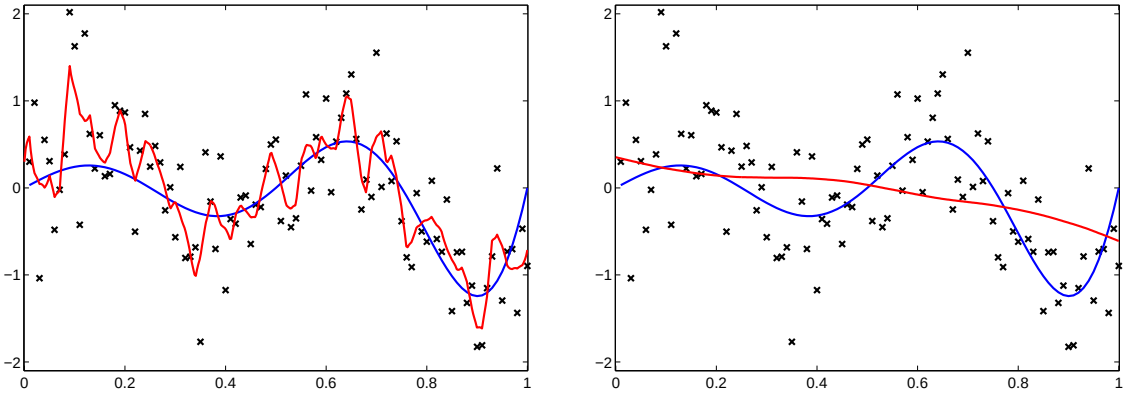
$$\widehat{m}_{NW}(x_i, h) \to \frac{\sum_{j=1}^{n} K(0)Y_j}{\sum_{j=1}^{n} K(0)} = \frac{K(0)\sum_{j=1}^{n} Y_j}{nK(0)} = \frac{1}{n}\sum_{j=1}^{n} Y_j$$

Tedy velká šířka okna vede k přehlazení, a to k průměru dat (viz obr. 2.5(b)).

Na obrázku 2.6 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.

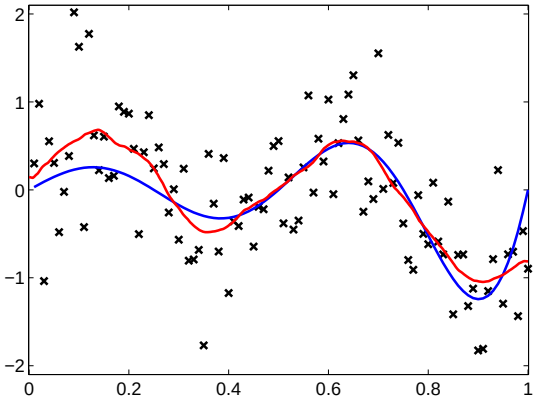
Poznámka 2.1. Intervaly spolehlivosti pro hodnotu regresní funkce m v bodě x jsou užitečné v mnoha aplikacích. Bodový interval spolehlivosti udává interval, v němž s pravděpodobností $1 - \alpha$ leží hodnota funkce m v bodě x. Jsou definovány takto

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



(a) Podhlazený odhad, $h = 0{,}02$ (b) Přehlazený odhad, $h = 0{,}4$

Obrázek 2.5: Podhlazený a přehlazený odhad regresní funkce z příkladu 1.1



Obrázek 2.6: Optimální odhad, $h = 0{,}0816$

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

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

Ukázka intervalu spolehlivosti pro $\alpha = 0{,}05$ je na obrázku 2.7.

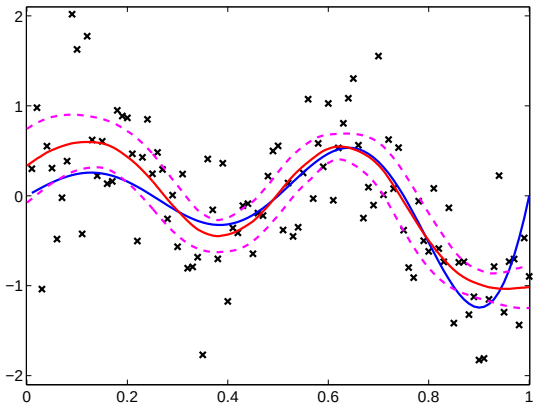
3 Statistické vlastnosti jádrových odhadů

Kvalitu jádrového odhadu lze lokálně popsat pomocí střední kvadratické chyby. Budeme se zabývat asymptotickým tvarem této chyby, neboť pro neparametrické odhady, na rozdíl od odhadů parametrických, neexistuje nevychýlený odhad, tj. takový odhad že, $E\widehat{m} = m$ pro skoro všechna $x \in \mathbb{R}$ [2].

Věta 3.1. *Nechť $K \in S_{0k}, EY^2 < \infty$ a nechť posloupnost vyhlazovacích parametrů $h = h_n$ $(n = 1, 2, \ldots)$ splňuje podmínky: $h_n \to 0$, $nh_n \to \infty$ pro $n \to \infty$. Pak v každém bodě spojitosti funkce m platí*

$$\widehat{m}(x,h) = \sum_{i=1}^{n} W_i(x,h) Y_i \xrightarrow{p} m(x), \tag{2.5}$$

kde $\xrightarrow{p}$ značí konvergenci podle pravděpodobnosti (viz např. [1]). Uvedené odhady $\widehat{m}$ jsou tedy konzistentními odhady m.



Obrázek 2.7: Interval spolehlivosti pro data z příkladu 1.1 při $\alpha = 0{,}05$

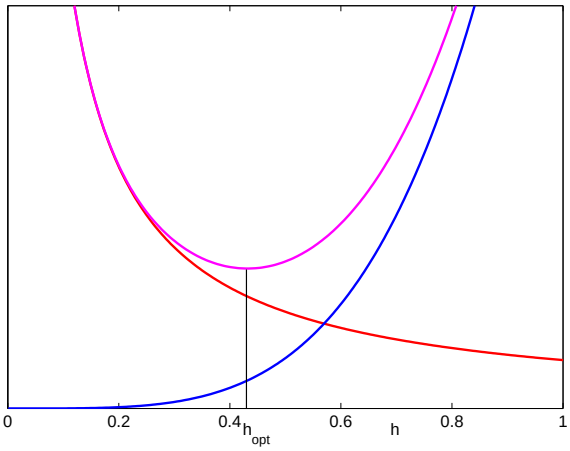
Střední kvadratická chyba MSE odhadu $\widehat{m}$ v bodě x je obecně dána vztahem

$$\mathrm{MSE}\,\widehat{m}(x,h) = E\big(\widehat{m}(x,h) - m(x)\big)^2.$$

Upravíme tento vztah

$$\begin{aligned} &= E\widehat{m}^2(x,h) - 2m(x)E\widehat{m}(x,h) + m^2(x) \\ &= \underbrace{\big(E\widehat{m}(x,h) - m(x)\big)^2}_{\mathrm{bias}^2} + \underbrace{E\widehat{m}^2(x,h) - \big(E\widehat{m}(x,h)\big)^2}_{\mathrm{var}}, \end{aligned} \tag{2.6}$$

což znamená, že střední kvadratická chyba může být vyjádřena jako součet rozptylu odhadu $\mathrm{var}\,\widehat{m}(x,h)$ a čtverce vychýlení $\mathrm{bias}^2\,\widehat{m}(x,h)$ (viz obr. 2.8). Tento rozklad rozptyl-vychýlení usnadňuje analýzu vlastností odhadu, kterou se budeme zabývat.



Obrázek 2.8: MSE (—) jako součet rozptylu (—) a vychýlení (—)

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

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

$$W_i(x,h) = K_h(x - x_i) = \frac{1}{h}K\Big(\frac{x - x_i}{h}\Big).$$

Pro další výpočty budeme předpokládat:

(i) Jádrová funkce K je sudou funkcí na intervalu $[-1,1]$, tj. $K \in S_{02}$.

(ii) Vyhlazovací parametr $h = h(n)$ je posloupností splňující $h \to 0$ a $nh \to \infty$ pro $n \to \infty$.

(iii) Bod x, v němž počítáme odhad, splňuje nerovnost $h < x < 1-h$ pro všechna $n \geq n_0$, kde n_0 je pevné.

Je zřejmé, že pro $n \to \infty$ platí[1]

$$E\widehat{m}(x,h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x-x_i)m(x_i) = \int_0^1 \frac{1}{h}K\Big(\frac{x-t}{h}\Big)m(t)\,\mathrm{d}t + O(n^{-1}). \tag{2.7}$$

Nechť $u = \frac{x-t}{h}$, odtud $\mathrm{d}t = -h\,\mathrm{d}u$ a tedy

$$\begin{aligned} E\widehat{m}(x,h) &= \int_{-(1-x)/h}^{x/h} K(u)m(x-hu)\,\mathrm{d}u + O(n^{-1}) \\ &= \int_{-(1-x)/h}^{x/h} K(u)\Big[m(x) - uhm'(x) + \frac{u^2h^2}{2}m''(x)\Big]\,\mathrm{d}u + o(h^2) + O(n^{-1}). \end{aligned} \tag{2.8}$$

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

$$\begin{aligned} E\widehat{m}(x,h) &= m(x)\int K(u)\,\mathrm{d}u - \int uK(u)\,\mathrm{d}u + \frac{h^2}{2}m''(x)\int u^2K(u)\,\mathrm{d}u + o(h^2) + O(n^{-1}) \\ &= m(x)\int_{-1}^{1} K(u)\,\mathrm{d}u - \int_{-1}^{1} uK(u)\,\mathrm{d}u + \frac{h^2}{2}m''(x)\int_{-1}^{1} u^2K(u)\,\mathrm{d}u + o(h^2) + O(n^{-1}). \end{aligned}$$

Celkem dostaneme

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

Podobně pro rozptyl platí

$$\begin{aligned} \operatorname{var}\widehat{m}(x,h) &= E\big(\widehat{m}(x,h) - E\widehat{m}(x,h)\big)^2 \\ &= E\bigg(\frac{1}{n}\sum_{i=1}^{n} K_h(x-x_i)Y_i - \frac{1}{n}\sum_{=1}^{n} K_h(x-x_i)m(x_i)\bigg)^2 \\ &= \frac{1}{n^2}E\bigg(\sum_{i=1}^{n} K_h(x-x_i)(\underbrace{Y_i - m(x_i)}_{\varepsilon_i})\bigg)^2 \end{aligned}$$

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

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

[1] Jedná se o přibližný výpočet integrálu.

Opět s využitím substituce $u = \frac{x-t}{h}$ můžeme pro $n \to \infty$ psát

$$\operatorname{var} \widehat{m}(x,h) = \frac{\sigma^2}{nh} \int K^2(u)\,\mathrm{d}u + o((nh)^{-1}) = \frac{\sigma^2}{nh} \int_{-1}^{1} K^2(u)\,\mathrm{d}u + o((nh)^{-1}) \approx \frac{\sigma^2}{nh} V(K).$$

Tímto jsme dokázali následující větu o tvaru střední kvadratické chyby.

Věta 3.2. *Nechť jsou splněny předpoklady $(i) - (iii)$, pak střední kvadratická chyba nabývá tvaru*

$$\operatorname{MSE} \widehat{m}(x,h) = \underbrace{\frac{\sigma^2}{nh} V(K)}_{\text{var}} + \underbrace{h^4 \beta_2^2(K) \frac{1}{4} \big(m''(x)\big)^2}_{\text{bias}^2} + o(h^4 + (nh)^{-1}). \tag{2.9}$$

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

$$\operatorname{MISE}(h) = \int_0^1 \operatorname{MSE}(x,h)\,\mathrm{d}x = \operatorname{AMISE}(h) + o(h^4 + (nh)^{-1}).$$

AMISE je tvaru

$$\operatorname{AMISE}(h) = \operatorname{AMISE} \widehat{m}(\cdot,h) = \underbrace{\frac{\sigma^2}{nh} V(K)}_{\text{AIV}} + \underbrace{\frac{h^4}{4} \beta_2^2(K) V(m'')}_{\text{AISB}}, \tag{2.10}$$

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

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

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

položíme ji rovnu nule a vyjádříme h

$$h_{opt,0,2}^5 = \frac{\sigma^2 V(K)}{n \beta_2^2(K) V(m'')}. \tag{2.11}$$

Poznámka 3.1. Tento výpočet vede k nalezení minima AMISE, protože platí

$$\frac{\mathrm{d}^2 \operatorname{AMISE}}{\mathrm{d}h^2} > 0.$$

Poznámka 3.2. Jestliže jádro K náleží do třídy $\S_{0k}$, pak AMISE je tvaru

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

a pro optimální vyhlazovací parametr h platí

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

kde $V(m^{(k)}) = \int_0^1 \big(m^{(k)}(x)\big)^2\,\mathrm{d}x$, podrobněji např. [8].

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

Lemma 3.1. *Pro $h_{opt,0,k}$ platí*

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

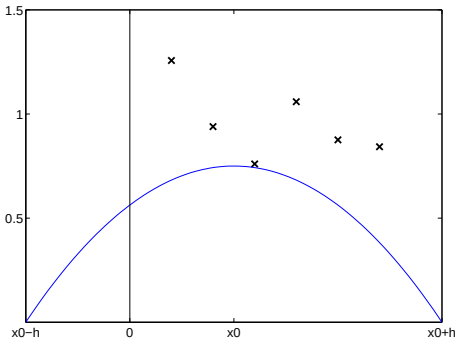
Důkaz. Viz cvičení. □

Vztah (2.13) pro optimální šířku vyhlazovacího okna ukazuje, že řád konvergence optimální šířky vyhlazovacího okna $h_{opt,0,2}$ závisí na řádu jádra k, tedy pro $k = 2$ je $O\big(n^{-1/5}\big)$. Rovnice (2.13) má pouze teoretický charakter, protože hodnota $h_{opt,0,k}$ závisí na neznámých veličinách σ^2 a $m''(x)$, a tedy není užitečná pro praktické účely. Na druhou stranu nám umožní vyjádřit asymptotickou rychlost konvergence $\mathrm{AMISE}(h)$. Dosadíme-li (2.13) do vztahu (2.10) pro AMISE, dostaneme

$$\mathrm{AMISE}(h_{opt,0,2}) = n^{-4/5}\big(\sigma^2 V(K)\big)^{4/5}\big(\beta_2^2(K)(k!)^{-2}V(m'')\big)^{1/5}. \tag{2.14}$$

Odtud plyne, že řád konvergence pro jádro řádu k je $n^{-\frac{2k}{2k+1}}$, tedy 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. O těchto problémech pojednáme podrobněji v dalším odstavci. Nevýhodou jader vyšších řádů je fakt, že pro tato jádra je optimální šířka okna větší, což může mít negativní dopad na hraniční efekty. 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 je plošší.

Poznámka 3.3. 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 nesplňuje 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. 2.9. Hraniční efekty jsou také patrné na obrázcích 2.6 a 2.12,



Obrázek 2.9: Hraniční efekt

zejména u pravého okraje intervalu. Problém okrajových efektů lze překonat např. použitím hraničních jader [9] nebo reflexní metodou [4].

4 Volba jádra

Volba jádra není z asymptotického hlediska podstatná, jak je zřejmé z faktu (2.14). Je vhodné zvolit optimální jádro, které minimalizuje fukcionál $T(K)$, neboť tato jádra jsou spojitá na $\mathbb{R}$ a odhadovaná regresní funkce tak „zdědí“ hladkost jádra. Vhodná jsou jádra třídy S_{02} a S_{04}, lze je vybrat z tabulek 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:

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

Funkce křížového ověřování je definována takto

$$\mathrm{CV}(h) = \frac{1}{n} \sum_{i=1}^{n} \big(\widehat{m}_{-i}(x_i, h) - Y_i\big)^2 \tag{2.15}$$

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

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

Při praktických úlohách hledáme minimum na intervalu $H_n = [h_\ell, h_u]$, přičemž pro ekvidistantní body plánu byl na základě zkušeností odvozen interval $[\frac{1}{n}, 2]$.

Poznámka 5.1. Někdy se místo chyby MISE používá průměrná střední kvadratická chyba AMSE (*average mean square error*)

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

Využívá se zejména v případech, kdy je nevyhovující použít numerické integrování související s chybou MISE.

Věta 5.1. *Střední hodnota funkce* $\mathrm{CV}(h)$ *je*

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

Důkaz. Funkci křížového ověřování lze rozepsat

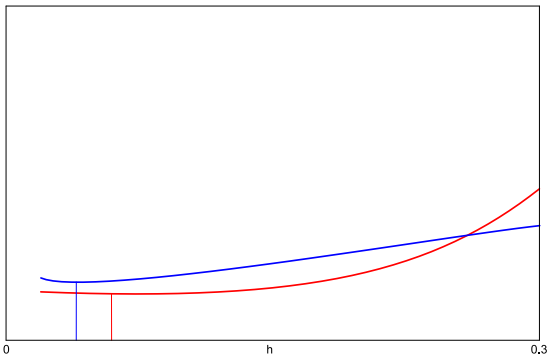
$$\begin{aligned}
\mathrm{CV}(h) &= \frac{1}{n} \sum_{i=1}^{n} \big(\widehat{m}_{-i}(x_i, h) - m(x_i) - (\underbrace{Y_i - m(x_i)}_{\varepsilon_i})\big)^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \big(\widehat{m}_{-i}(x_i, h) - m(x_i)\big)^2 - 2\frac{1}{n} \sum_{i=1}^{n} \big(\widehat{m}_{-i}(x_i, h) - m(x_i)\big)\varepsilon_i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

Pak střední hodnota $E\,\mathrm{CV}(h)$ je rovna součtu tří hodnot

$$\frac{1}{n}E\sum_{i=1}^{n}\big(\widehat{m}_{-i}(x_i,h)-m(x_i)\big)^2=\mathrm{AMSE},$$
$$-\frac{2}{n}E\sum_{i=1}^{n}\varepsilon_i\Big(\sum_{\substack{\ell=1\\ \ell\neq i}}^{n}W_\ell(x_i,h)Y_\ell-m(x_i)\Big)=0,$$
$$\frac{1}{n}E\sum_{i=1}^{n}\varepsilon_i^2=\sigma^2.$$

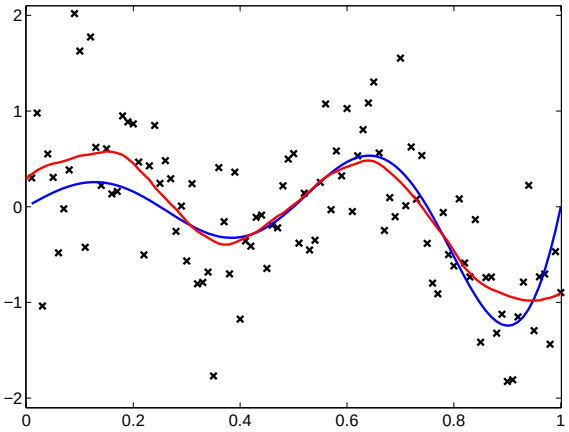
Druhý a třetí vztah plyne z vlastností (2.2). □

Tento výsledek znamená, že minimalizace $E\,\mathrm{CV}(h)$ odpovídá minimalizaci AMSE. Jestliže tedy předpokládáme, že minimum $\mathrm{CV}(h)$ je blízko minima $E\,\mathrm{CV}(h)$, pak tato minimalizace dává dobrou volbu vyhlazovacího parametru – viz ilustrace na obr. 2.10.



Obrázek 2.10: Porovnání minima AMSE (červenou) a minima funkce křížového ověřování CV (modrou) pro simulovaná data

Příklad 5.1. Použijeme metodu křížového ověřování pro nalezení vyhlazovacího parametru pro data z příkladu 1.1. Při použití Epanečnikova jádra získáme vyhlazovací parametr $h_{\mathrm{CV}}=0{,}1158$. Na obrázku 2.11 je zobrazen odhad regresní funkce s tímto parametrem.



Obrázek 2.11: Simulovaná data (×) s jádrovým odhadem regresní funkce ($h_{\mathrm{CV}}=0{,}1158$) (—) a původní funkcí (—)

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. [4, 7]).

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 [4]), která odhadne všechny parametry tak, aby byla minimalizována AMISE. Procedura byla původně odvozena pro odhad hustoty pravděpodobnosti ([5]), ale lze ji aplikovat i pro odhad regresní funkce. Uvedeme zde její zjednodušenou verzi.

Podle poznámky 3.2 je známo, že AMISE a $h_{opt,0,k}$ jsou tvaru

$$\mathrm{AMISE}(h) = \frac{\sigma^2}{nh} V(K) + \frac{1}{(k!)^2} h^{2k} \beta_k^2(K) V(m^{(k)}), \qquad h_{opt,0,k}^{2k+1} = \frac{\sigma^2 V(K)(k!)^2}{2kn\beta_k^2(K)V(m^{(k)})}.$$

Dosadíme-li vyhlazovací parametr $h_{opt,0,k}$ do chyby AMISE a použijeme vztahy

$$\delta_{0k} = \Big(\frac{V(K)}{\beta_k^2(K)}\Big)^{\frac{1}{2k+1}} \qquad T(K) = \big(\beta_k(K)V^k(K)\big)^{\frac{2}{2k+1}}$$

dostaneme vyjádření AMISE

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

v němž je vidět separace parametrů K, h a k, což umožňuje vybrat tyto parametry simultánně, Právě tento vztah je základem automatické procedury. Označme

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

a množinu vhodných řádů k označme

$$I(k_0) = \Big\{2j, j = 0, \ldots, \Big[\frac{k_0}{2}\Big]\Big\},$$

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

Krok 1 Pro $k \in I(k_0)$ najděte optimální jádro $K_{opt,0,k} \in S_{0k}$, které je dáno tabulkou 1.2 a vypočtěte kanonický faktor δ_{0k}.

Krok 2 Pro $k \in I(k_0)$ a $K_{opt,0,k} \in S_{0k}$ najděte optimální vyhlazovací parametr $\hat{h}_{opt,0,k}$.

Krok 3 Pro $k \in I(k_0)$ vypočtěte hodnotu výběrového kriteria $L(k)$ s využitím hodnot získaných v krocích 1 a 2.

Krok 4 Vypočtěte optimální hodnotu řádu $\hat{k}$, které minimalizuje funkcionál $L(k)$.

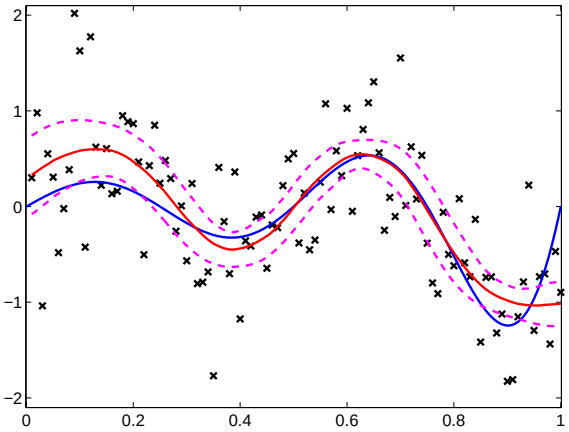
Krok 5 Použijte parametry z předchozích kroků k získání optimálního jádrového odhadu regresní funkce, tj.

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

Příklad 6.1. Aplikace procedury na data z příkladu 1.1. 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}	h	$L(k)$
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 2.12.



Obrázek 2.12: Simulovaná data (×) s jádrovým odhadem regresní funkce při použití procedury (—) a skutečnou funkcí (—) společně s 95% intervalem spolehlivosti

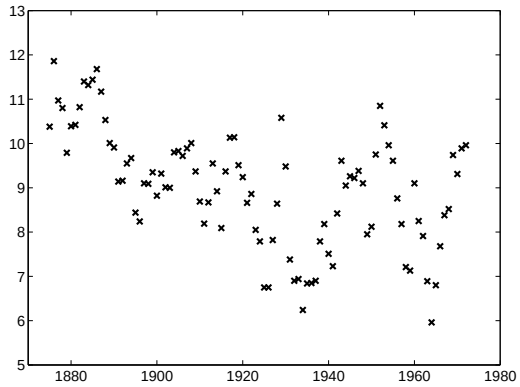
7 Aplikace na reálná data

První 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 6.5 a na obrázku 2.13.

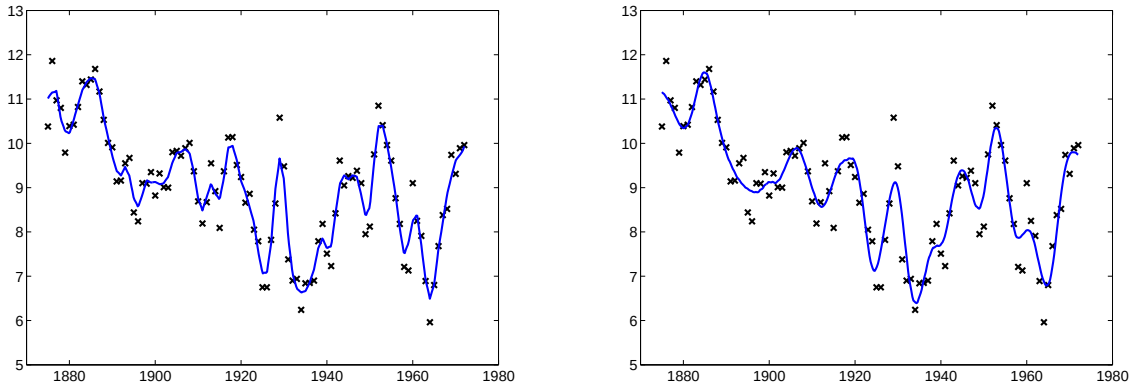
Vyhlazovací parametry jsme odhadli pomocí metody křížového ověřování (za použití Epanečnikova jádra) a také pomocí automatické procedury. Hodnoty vyhlazovacích parametrů jsou následující:

$$h_{\mathrm{CV}} = 0{,}0204 \quad (K_{opt,0,2}), \qquad h_{\mathrm{proc}} = 0{,}1525 \quad (K_{opt,0,12}).$$

Výsledné odhady na obrázku 2.14 ukazují, že metoda křížového ověřování spíše podhlazuje.



Obrázek 2.13: Úroveň hladiny Huronského jezera



(a) CV (b) procedura

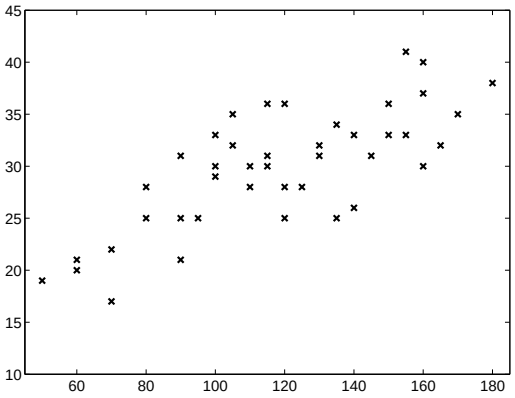
Obrázek 2.14: Odhadnuté regresní funkce – na ose x jsou roky a na ose y je hladina vody ve stopách (snížená o 570 stop – viz poznámka u dat)

Druhým datovým souborem jsou měření axiální délky krystalů ledu. Měření byla prováděna v místnosti s konstantní teplotou $-5\,^{\circ}\mathrm{C}$ v časových intervalech 50 až 180 vteřin po přinesení krystalu do místnosti. V tomto případě nejde o ekvidistantní data, protože hodnoty se liší o pět či deset vteřin. Data jsou v tabulce 6.6 a na obrázku 2.15.

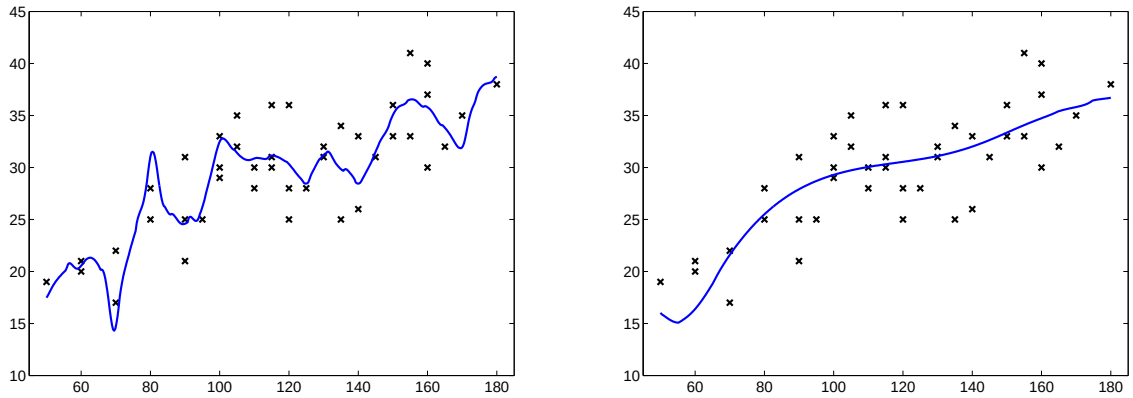
Odhady vyhlazovacích parametrů podle metody křížového ověřování a pomocí automatické procedury:

$$h_{\mathrm{CV}} = 0{,}1865 \quad (K_{opt,0,2}), \qquad h_{\mathrm{proc}} = 0{,}8826 \quad (K_{opt,0,10}).$$

Výsledné odhady regresní funkce jsou na obrázku 2.16. Je vidět, že CV metoda dává spíš podhlazený odhad. Na druhou stranu, odhad pomocí procedury se může zdát již přehlazený.



Obrázek 2.15: Axiální délka krystalů ledu



(a) CV (b) procedura

Obrázek 2.16: Odhadnuté regresní funkce – na ose x je vynesen čas ve vteřinách a na ose y délka krystalu v mikrometrech

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

Dodatky a cvičení

1. Pro odhady $\widehat{m}_{NW}, \widehat{m}_{LL}, \widehat{m}_{GM}$ 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. $$\widehat{m}(x,h,K) = \widehat{m}(x,h^*,K_\delta).$$

2. Dokažte vztah pro střední hodnotu odhadu (2.7).

3. Dokažte vztah (2.14) pro obecné k, tj. že platí $$\operatorname{AMISE}(h_{opt,0,k}) = n^{-2k/(2k+1)}\Big(\sigma^2 V(K)\Big)^{2k/(2k+1)}\Big(\beta_k^2(K)(k!)^{-2}V(m^{(k)})\Big)^{1/(2k+1)}.$$

   Návod: Předpokládejme, že funkce m je spojitá a má spojité derivace až do řádu k včetně, tj. $m \in C^k[0,1]$. Užitím Taylorova (až do řádu k) rozvoje upravíme vztah pro střední hodnotu odhadu podobně jako ve vztahu (2.8). Užijte poznámky 3.2.

4. Dokažte vztah z lemmatu 3.1.

5. Dokažte, že příspěvek jádra $K_{\delta_{0k}}$, $\delta_{0k} = \left(\frac{V(K)}{\beta_k^2(K)}\right)^{1/(2k+1)}$, $K \in S_{0k}$, k oběma částem chyby AMISE je stejný, tj. platí rovnost

$$\int_{-\delta_{0k}}^{\delta_{0k}} K_{\delta_{0k}}^2(t)\,\mathrm{d}t = \left(\int_{-\delta_{0k}}^{\delta_{0k}} t^k K_{\delta_{0k}}(t)\,\mathrm{d}t\right)^2.$$

6. Aplikujte metodu křížového ověřování a automatickou proceduru na simulovaná i reálná data.

Kapitola 3

Jádrové odhady hustoty

Výstupy z výukové jednotky

Student

- bude znát tvar jádrových odhadů hustoty pravděpodobnosti.
- bude schopen analyzovat statistické vlastnosti těchto odhadů.
- se seznámí s metodami pro volbu vyhlazovacího parametru.
- porozumí 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.

1 Motivace

Hustota pravděpodobnosti je základním pojmem ve statistice [1, 3].

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 $X_1, \ldots, X_n$, 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 [11, 10, 13]. 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:

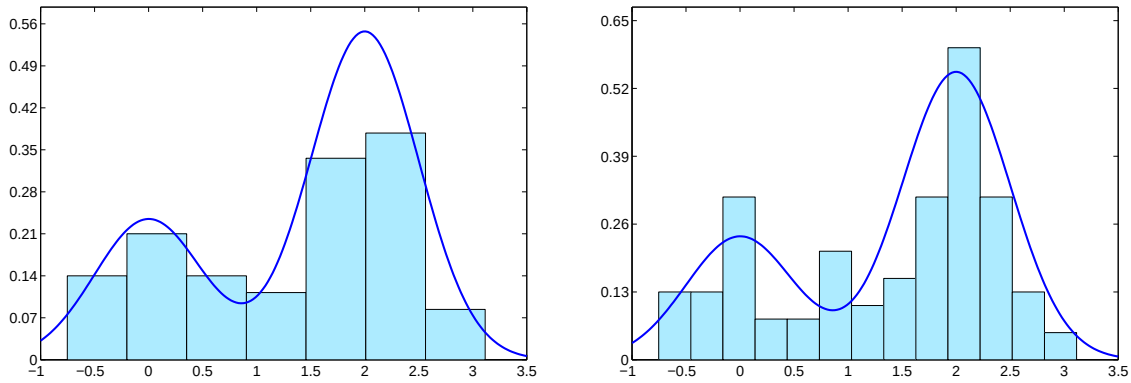
- Histogram je citlivý na počet tříd a jejich šířku.

- Histogram je schodovitá funkce, ale přitom předpokládáme, že neznámá hustota je spojitá.

Příklad 2.1. Mějme dán datový soubor generovaných ze směsi dvou normálních hustot $N(0; 0{,}25)$ a $N(2; 0{,}25)$, který má rozsah $n = 100$.

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

(Data jsou v tabulce 6.2.) Na obr. 3.1 je patrné, že histogram nevystihuje hustotu pravděpodobnosti dat.



(a) 7 intervalů o šířce $h = 0{,}55$ (b) 13 intervalů o šířce $h = 0{,}29$

Obrázek 3.1: 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 (Rosenblatt (1956), Parzen (1962))

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

$K \in S_{0k}$ 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 – viz obrázek 3.2(a).

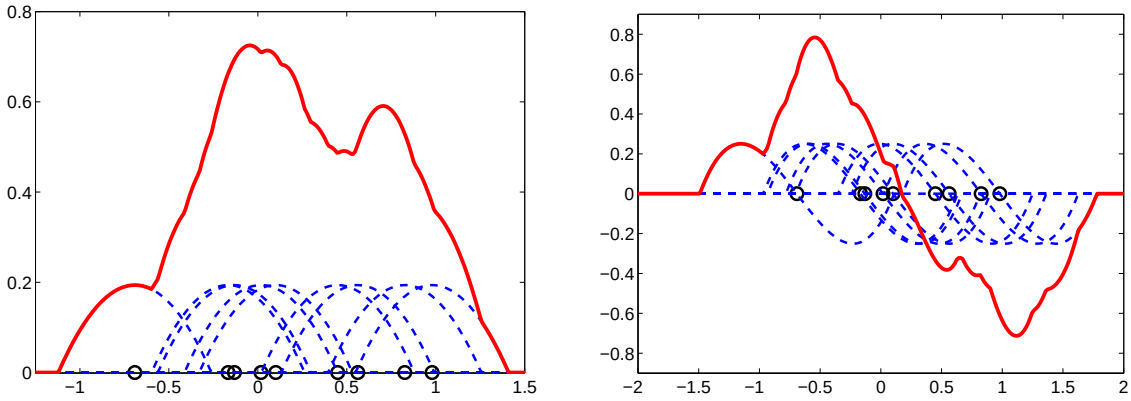
Nyní uvedeme ještě vztah pro jádrový odhad ν-té derivace hustoty. Budeme předpokládat, že $0 \le \nu \le k - 2$ a k, ν jsou stejné parity. Pak

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

Konstrukce jádrového odhadu derivace je stejná jako konstrukce odhadu hustoty – obr. 3.2(b) – pouze místo jádra $K \in S_{0k}$ používáme jádro třídy $S_{\nu k}$.

3 Statistické vlastnosti jádrových odhadů hustoty

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



(a) Hustota (b) Derivace hustoty

Obrázek 3.2: Konstrukce jádrového odhadu hustoty a její derivace

Věta 3.1.

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

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

$$E\hat{f}(x,h) = E\frac{1}{n}\sum_{i=0}^{n} K_h(x - X_i) = EK_h(x - X) = \int K_h(x-y)f(y)\,\mathrm{d}y = (K_h * f)(x).$$

Vychýlení (bias) pak bude mít tvar $\operatorname{bias} \hat{f}(x,h) = E\hat{f}(x,h) - f(x) = (K_h * f)(x) - f(x)$. Dále upravíme vztah pro rozptyl

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

□

Důsledek.

$$\operatorname{MISE} = \frac{1}{n}\Big(\int (K_h^2 * f)(x)\,\mathrm{d}x - \int (K_h * f)^2(x)\,\mathrm{d}x\Big) + \int \big((K_h * f)(x) - f(x)\big)^2\,\mathrm{d}x.$$

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 spojité derivace až do řádu k_0 (tj. $f \in C^{k_0}$) pro $0 < k \le k_0$, $K \in S_{0k}$ a $\int \big(f^{(k)}(x)\big)^2\,\mathrm{d}x < \infty$, dále předpokládejme $h \to 0$ a $nh \to \infty$ pro $n \to \infty$. Pak platí*

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

kde $V(f^{(k)}) = \int \big(f^{(k)}(x)\big)^2\,\mathrm{d}x$.

Důkaz. Nejprve vypočteme střední hodnotu

$$\begin{aligned} E\hat{f}(x,h) = (K_h * f)(x) &= \int K_h(x-y)f(y)\,\mathrm{d}y = \int \frac{1}{h}K\Big(\frac{x-y}{h}\Big)f(y)\,\mathrm{d}y \\ &= \int K(z)f(x-hz)\,\mathrm{d}z \end{aligned}$$

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

$$\begin{aligned} &= \int K(z)\big[f(x) - f'(x)hz + \cdots + \tfrac{(-1)^k}{k!}f^{(k)}(x)h^k z^k + o(h^k)\big]\,\mathrm{d}z \\ &= f(x) + \frac{(-1)^k f^{(k)}(x)}{k!}h^k\beta_k(K) + o(h^k). \end{aligned}$$

Nyní dokážeme vztah pro rozptyl. Víme, že platí vztahy

$$\begin{aligned} \operatorname{var}\hat{f}(x,h) &= \frac{1}{n}\big((K_h^2 * f)(x) - (K_h * f)^2(x)\big) \\ E\hat{f}(x,h) &= f(x) + \frac{(-1)^k f^{(k)}(x)}{k!}h^k\beta_k(K) + o(h^k) = f(x) + o(1), \end{aligned}$$

tedy

$$\begin{aligned} \operatorname{var}\hat{f}(x,h) &= \frac{1}{n}\int \frac{1}{h^2}K^2\Big(\frac{x-y}{h}\Big)f(y)\,\mathrm{d}y - \frac{1}{n}(f(x)+o(1))^2 \\ &= \frac{1}{nh}\int K^2(z)\underbrace{f(x-hz)}_{=f(x)+o(1)}\,\mathrm{d}z - \frac{1}{n}(f(x)+o(1))^2 \\ &= \frac{1}{nh}\int K^2(z)(f(x)+o(1))\,\mathrm{d}z + \frac{1}{n}(f(x)+o(1))^2 \\ &= \frac{f(x)}{nh}\int K^2(z)\,\mathrm{d}z + o((nh)^{-1}). \end{aligned}$$

□

Důsledek. *Nechť $h \to 0$, $nh \to \infty$ pro $n \to \infty$, pak $\hat{f}$ je konzistentním odhadem f, tj. $E\hat{f} \to f$ a* $\operatorname{var}\hat{f} \to 0$.

Vztah mezi chybami MSE, MISE a AMISE je podobný jako u odhadu regresní funkce, tedy platí

$$\operatorname{MISE}(h) = \int \operatorname{MSE}(x,h)\,\mathrm{d}x = \operatorname{AMISE}(h) + o(h^{2k} + (nh)^{-1}),$$

kde AMISE je tvaru

$$\operatorname{AMISE}\hat{f}(\cdot,h) = \operatorname{AMISE}(h) = \frac{V(K)}{nh} + \frac{1}{(k!)^2}h^{2k}\beta_2^2(K)V(f^{(k)}). \tag{3.3}$$

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

$$\operatorname{AIV}(h) = \frac{V(K)}{nh} \qquad \operatorname{AISB}(h) = \frac{1}{(k!)^2}h^{2k}\beta_2^2(K)V(f^{(k)}),$$

tedy $\operatorname{AMISE}(h) = \operatorname{AIV}(h) + \operatorname{AISB}(h)$.

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

$$\mathrm{AMISE}(h) = T(K)\left(\frac{\delta_{0k}^{2k+1}}{nh} + \frac{h^{2k} V(f^{(k)})}{\delta_{0k}(k!)^2}\right). \tag{3.4}$$

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

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

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

Vypočtěme hodnotu AMISE při dosazení optimálního parametru $h_{opt,0,k}$:

$$\mathrm{AMISE}(h_{opt,0,k}) = T(K) V(f^{(k)})^{1/(2k+1)} n^{-2k/(2k+1)} \frac{2k+1}{(2k(k!)^2)^{1/(2k+1)}}, \tag{3.5}$$

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

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

$$\mathrm{AIV}(h_{opt,0,k}) = 2k\,\mathrm{AISB}(h_{opt,0,k}). \tag{3.6}$$

Nyní uvedeme zajímavou vlastnost vyhlazovacího parametru.

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

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

Počítejme

$$\frac{\mathrm{d}^2\mathrm{AMISE}}{\mathrm{d}h^2} = \frac{2V(K)}{nh^3} + 3h^2\beta_2^2(K)V(f'')$$
$$\frac{\mathrm{d}^3\mathrm{AMISE}}{\mathrm{d}h^3} = \frac{-6V(K)}{nh^4} + 6h\beta_2^2(K)V(f'').$$

Řešením rovnice $\mathrm{d}^3\mathrm{AMISE} \,/\, \mathrm{d}h^3 = 0$ je

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

tj. $h_{opt,0,2}$ také realizuje minimum $\mathrm{d}^2\mathrm{AMISE} \,/\, \mathrm{d}h^2$.

Obecně lze ukázat, že

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

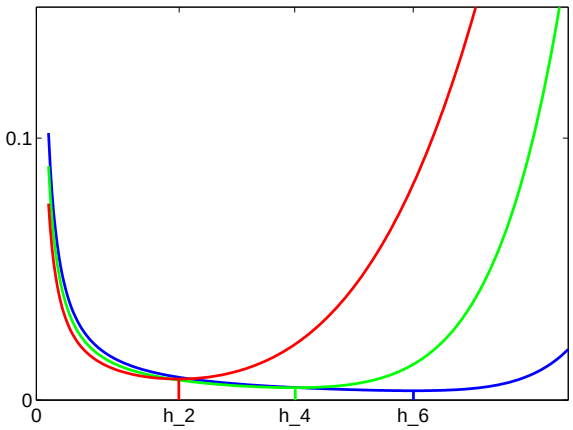
a to znamená, že pro jádra vyšších řádů je minimum AMISE plošší a tedy volba h blízká optimální hodnotě $h_{opt,0,k}$ nevede k velkému růstu AMISE (obr. 3.3).

Vztah pro optimální hodnotu vyhlazovacího parametru poskytuje informaci, že asymptoticky je $h \approx n^{-1/(2k+1)}$. Ale vztah má pouze teoretický charakter, protože optimální parametr závisí na neznámé hustotě f.

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, $0 < a_k < b_k < \infty$.



Obrázek 3.3: AMISE pro jádra vyšších řádů s vyznačenými minimálními hodnotami pro jádra řádu $2, 4, 6$

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

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

Předpokládejme nyní, že platí $0 \leq \nu \leq k-2$, $\lim_{n\to\infty} h = 0$, $\lim_{n\to\infty} nh^{2\nu+1} = \infty$, $f \in C^{k_0}$ ($k \leq k_0$) a $V(f^{(k)}) = \int \big(f^{(k)}(x)\big)^2 \,\mathrm{d}x < \infty$. 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.

Optimální hodnota vyhlazovacího parametru je dána vztahem

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

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

$$\frac{h_{opt,\nu,k}}{h_{opt,0,k}} = \left(\frac{(2\nu+1)k}{k-\nu}\right)^{1/(2k+1)} \frac{\delta_{\nu k}}{\delta_{0k}}. \tag{3.7}$$

Pro ν a k lichá platí

$$\frac{h_{opt,\nu,k}}{h_{opt,1,k}} = \left(\frac{(2\nu+1)(k-1)}{3(k-\nu)}\right)^{1/(2k+1)} \frac{\delta_{\nu k}}{\delta_{1k}}. \tag{3.8}$$

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

$$h_{opt,2,4} = 10^{1/9} \frac{\delta_{24}}{\delta_{04}} h_{opt,0,4}, \tag{3.9}$$

přičemž

$$K_{opt,0,4} = \frac{15}{32}(x^2-1)(7x^2-3), \qquad \delta_{04} = 2{,}0165,$$
$$K_{opt,2,4} = \frac{105}{16}(1-x^2)(5x^2-1), \qquad \delta_{24} = 1{,}3925.$$

4 Volba jádra

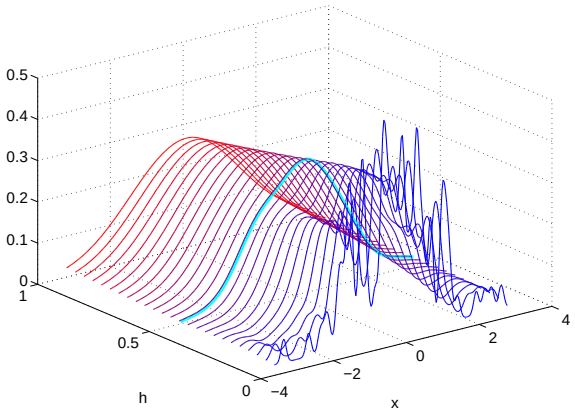
Volba jádra není z asymptotického hlediska podstatná, jak je zřejmé z faktu (3.5). Je vhodné zvolit optimální jádro, které minimalizuje fukcionál $T(K)$, neboť tato jádra jsou spojitá na $\mathbb{R}$ a hladkost jádra také „zdědí“ odhadovaná hustota.

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 užijeme 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 3.4, na němž jsou zobrazeny odhady pro simulovaná data ($n = 100$) 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 až k přehlazení.



Obrázek 3.4: Volba vyhlazovacího parametru

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

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

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

$$h_{\mathrm{REF}} = \left(\frac{2^{2k}(k!)^3\sqrt{\pi}}{(2k)!k} \right)^{\frac{1}{2k+1}} \delta_{0k}\sigma n^{-\frac{1}{2k+1}}, \tag{3.10}$$

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

$$\widehat{\sigma}_{SD} = \Big(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\Big)^{1/2}, \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{3.11}$$

$$\widehat{\sigma}_{IQR} = \frac{X_{[3n/4]} - X_{[n/4]}}{\Phi^{-1}(\frac{3}{4}) - \Phi^{-1}(\frac{1}{4})}, \tag{3.12}$$

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 $\min\{\widehat{\sigma}_{SD}, \widehat{\sigma}_{IQR}\}$.

Poznámka 5.1. Pokud za jádro K zvolíme Gaussovo jádro ($k = 2$)

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

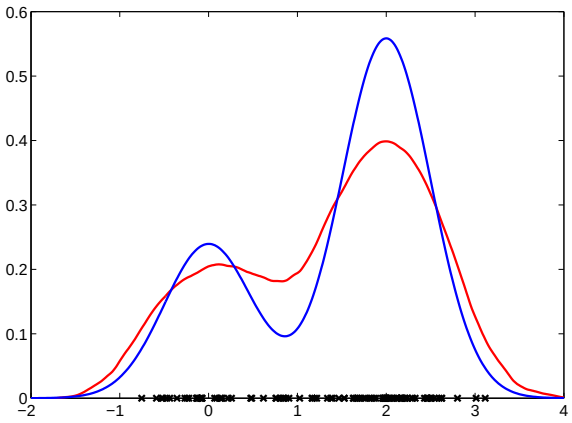
pak dostaneme jednoduchý vztah

$$h_{\mathrm{REF}} = \Big(\frac{4}{3n}\Big)^{1/5}\sigma. \tag{3.13}$$

Příklad 5.1. Použijme odhad vyhlazovacího parametru pro data z příkladu 2.1 metodou referenční hustoty. Pro Epanečnikovo jádro, které je řádu $k = 2$, se vztah (3.10) zjednoduší na tvar

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

Dále odhadneme směrodatnou odchylku: $\widehat{\sigma}_{SD} = 1{,}0325$, $\widehat{\sigma}_{IQR} = 1{,}3344$, tedy $\widehat{\sigma} = 1{,}0325$. Po dosazení počtu prvků $n = 100$ a parametru $\delta_{02} = 1{,}7188$ získáme hodnotu vyhlazovacího parametru pro odhad hustoty $h_{\mathrm{REF}} = 0{,}9639$. Na obrázku 3.5 je vykreslen odhad hustoty s tímto parametrem.



Obrázek 3.5: Odhad hustoty s $h_{\mathrm{REF}} = 0{,}9639$, odhad (—), původní funkce (—)

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

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

nejmenší hodnotu integrálu $\int_{-1}^{1}\big(f^{(k)}(x)\big)^2\,\mathrm{d}x$.

Lze ukázat, že platí

1. $\sigma_k^2 = \int_{-1}^{1} x^2 g_k(x)\,\mathrm{d}x = \frac{1}{2k+5}$.
2. Pro $r>0$, $\int\big(rg^{(k)}(rx)\big)^2\,\mathrm{d}x = r^{2k+1}\int\big(g^{(k)}(x)\big)^2\,\mathrm{d}x$ pro každou hustotu, pro kterou integrál existuje.
3. Jestliže hustota g má rozptyl σ_g^2, pak hustota $\frac{\sigma_g}{\sigma}g\big(\frac{\sigma_g}{\sigma}x\big)$ má rozptyl σ^2. (Podrobněji např. [6].)

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

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

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 [6]:

$$V(g_k^{(k)}) = \int_{-1}^{1}\big(g_k^{(k)}(x)\big)^2\,\mathrm{d}x = \frac{(2k+3)!(2k+2)!}{2^{2k+2}(2k+1)(2k+5)(k+1)!^2}. \tag{3.14}$$

Použijeme-li poslední vyjádření (3.14), dostaneme horní hranici pro vyhlazovací parametry

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

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 5.2. Určeme hodnotu h_{MS} pro odhad hustoty s jádrem řádu $k=2$, tj. $K \in S_{0k}$.
Podle věty 5.1 je hustota g_2 tvaru

$$g_2(x) = \frac{35}{32}(1-x^2)^3$$

a dále z vlastností funkce g_2 a ze vztahu (3.14) plyne

$$\sigma_2^2 = \int_{-1}^{1} x^2 g_2(x)\,\mathrm{d}x = \frac{1}{9} \quad \text{a} \quad \int_{-1}^{1}\big(g''(x)\big)^2\,\mathrm{d}x = 35.$$

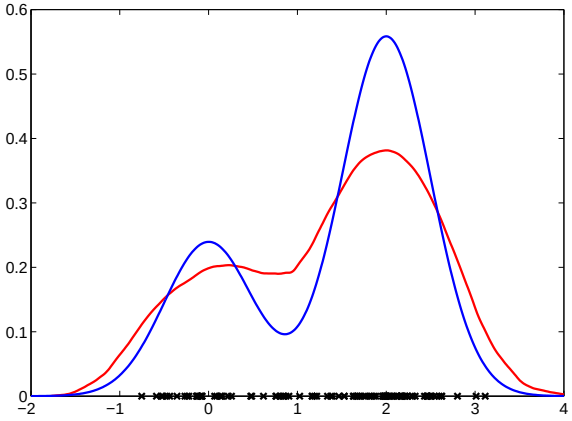
Pak pro $K_{opt,0,k}$

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

Příklad 5.3. Pro data z příkladu 2.1 bude vyhlazovací parametr určený metodou maximálního vyhlazení s Epanečnikovým jádrem ($k = 2$, $V(K) = 3/5$, $\beta_2(K) = 1/5$) roven

$$h_{\mathrm{MS}} = \widehat{\sigma} n^{-1/5} \Big(\frac{3/5}{1/25} \cdot \frac{243}{35} \Big)^{1/5} = 1{,}0409,$$

protože $\widehat{\sigma} = 1{,}0325$ a $n = 100$. Výsledný odhad je vidět na obr. 3.6.



Obrázek 3.6: Odhad hustoty s $h_{\mathrm{MS}} = 1{,}0409$, odhad (—), původní funkce (—)

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:

$$\begin{aligned} \mathrm{MISE}\, \hat{f}(\cdot, h) &= E \int \big(\hat{f}(x,h) - f(x) \big)^2 \, \mathrm{d}x \\ &= E \int \hat{f}^2(x,h) \, \mathrm{d}x - 2E \int \hat{f}(x,h) f(x) \, \mathrm{d}x + \int f^2(x) \, \mathrm{d}x. \end{aligned}$$

$$\mathrm{MISE}\, \hat{f}(x,h) - \int f^2(x) \, \mathrm{d}x = E \Big(\int \hat{f}^2(x,h) \, \mathrm{d}x - 2 \int \hat{f}(x,h) f(x) \, \mathrm{d}x \Big).$$

Definujme funkci křížového ověřování

$$\mathrm{CV}(h) = \int \hat{f}^2(x,h) \, \mathrm{d}x - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i, h), \tag{3.15}$$

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

Věta 5.2. *Platí*

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

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

$$E \Big(\int \hat{f}^2(x,h) \, \mathrm{d}x - 2 \int \hat{f}(x,h) f(x) \, \mathrm{d}x \Big).$$

Důkaz. Střední hodnota prvního členu rovnice (3.15) je zřejmá, potřebujeme spočítat střední hodnotu druhého členu tohoto vyjádření.

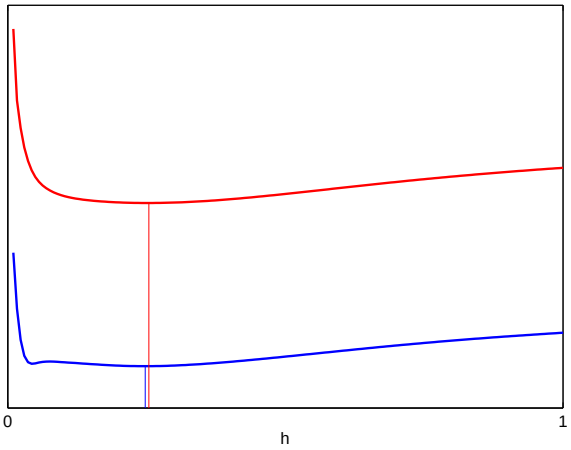
$$\begin{aligned} E\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i,h) &= E\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n-1}\sum_{\substack{j=1\\j\neq i}}^{n}K_h(X_i-X_j) \\ &= E\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}K_h(X_i-X_j) = EK_h(X_1-X_2) \\ &= \iint \underbrace{K_h(x-y)f(y)}_{E\hat{f}(x,h)}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

Odtud

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

□

Odhad $h_{opt,0,k}$ je dán vztahem $h_{\mathrm{CV}} = \arg\min_{h\in H_n}\mathrm{CV}(h)$. Odtud plyne, že $\mathrm{CV}(h)+\int f^2(x)\,\mathrm{d}x$ je pro každé h nevychýleným odhadem $\mathrm{MISE}(h)$. Protože $\int f^2(x)\,\mathrm{d}x$ nezávisí na h, minimalizace $E\,\mathrm{CV}(h)$ odpovídá minimalizaci MISE. Jestliže předpokládáme, že $\min\mathrm{CV}(h) \sim \min E\,\mathrm{CV}(h)$, dostaneme dobrou aproximaci optimální hodnoty h.



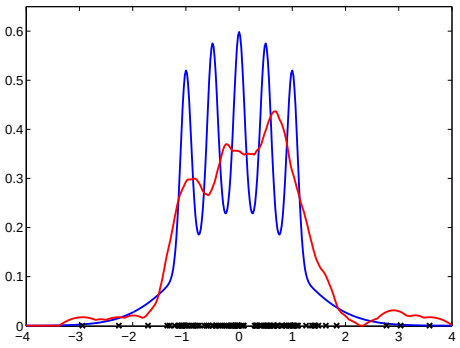
Obrázek 3.7: Porovnání minima MISE (červenou) a minima funkce křížového ověřování CV (modrou) pro simulovaná data

Poznámka 5.2. Předpokládejme, že $k = 2$. Pak vychýlení odhadu může být velké, jestliže f'' 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 3.8.

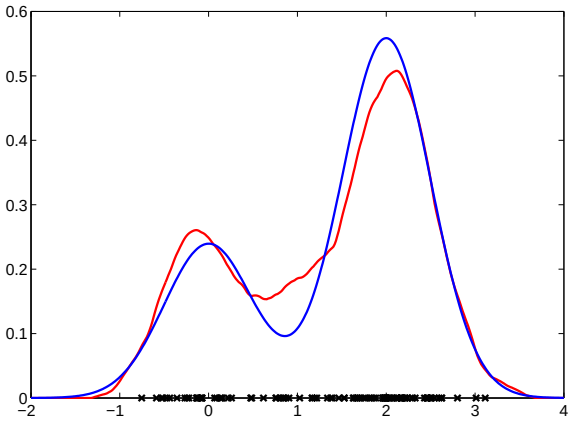
Příklad 5.4. Jádrový odhad hustoty dat z příkladu 2.1 je zobrazen na obr. 3.9. Pro rekonstrukci bylo použito Epanečnikovo jádro a vyhlazovací parametr určený metodou křížového ověřování $h_{\mathrm{CV}} = 0{,}5628$.

Shrneme-li na závěr doposud vypočítané hodnoty vyhlazovacích parametrů pro simulovaná data z příkladu 2.1, můžeme vizuálně porovnat jednotlivé odhady – viz obrázek 3.10.

$$h_{opt,0,2} = 0{,}5122 \quad h_{\mathrm{REF}} = 0{,}9639 \quad h_{\mathrm{MS}} = 1{,}0409 \quad h_{\mathrm{CV}} = 0{,}5628$$



Obrázek 3.8: Zahlazení vrcholů a údolí při odhadu hustoty směsi normálních rozdělení



Obrázek 3.9: Odhad hustoty s $h_{\mathrm{CV}} = 0{,}5628$, odhad (—), původní funkce (—)

6 Automatická procedura

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

Vyjdeme ze vztahu

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

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

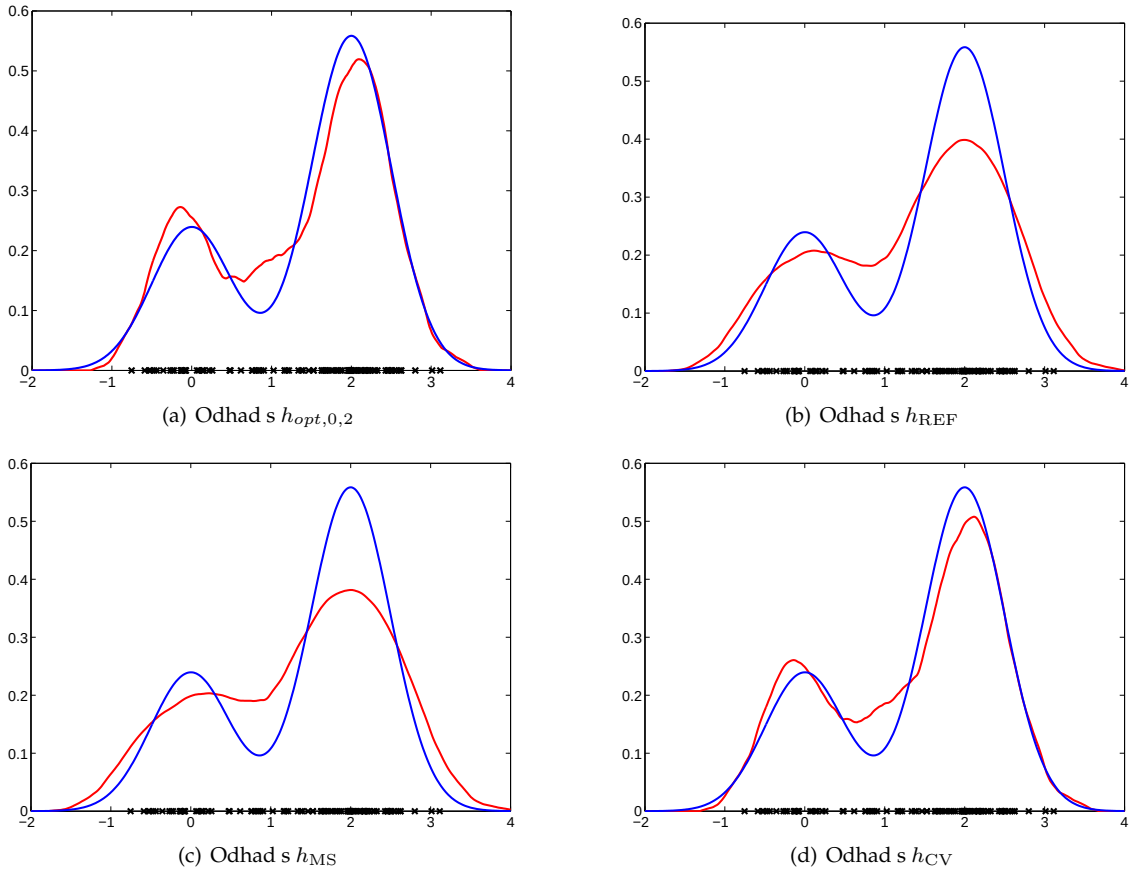
$$V(f^{(k)}) = \frac{\delta_{0k}^{2k+1}(k!)^2}{2knh_{opt,0,k}^{2k+1}}$$

a tuto hodnotu dosadíme do předchozího vztahu:

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

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

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



(a) Odhad s $h_{opt,0,2}$

(b) Odhad s h_{REF}

(c) Odhad s h_{MS}

(d) Odhad s h_{CV}

Obrázek 3.10: Srovnání odhadů pro data z příkladu 2.1

Krok 1 Pro $k \in I(k_0)$ najděte optimální jádro $K_{opt,0,k} \in S_{0k}$, které je dáno tabulkou 1.2, k němu příslušný kanonický faktor δ_{0k}.

Krok 2 Pro $k \in I(k_0)$ a $K_{opt,0,k} \in S_{0k}$ najděte optimální vyhlazovací parametr $\hat{h}_{opt,0,k}$.

Krok 3 Pro $k \in I(k_0)$ vypočtěte hodnotu výběrového kriteria $L(k)$ s využitím hodnot získaných v krocích 1 a 2.

Krok 4 Vypočtěte optimální hodnotu řádu $\hat{k}$, které minimalizuje funkcionál $L(k)$.

Krok 5 Použijte parametry z předchozích kroků ke konstrukci optimálního jádrového odhadu hustoty, tj.

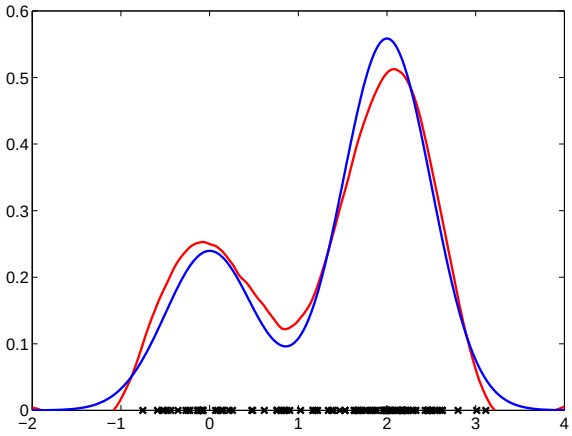
$$\hat{f}(x, \hat{h}_{opt,0,\hat{k}}) = \frac{1}{n\hat{h}_{opt,0,\hat{k}}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{\hat{h}_{opt,0,\hat{k}}}\right).$$

Příklad 6.1. 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 toolboxu Matlabu, který je doprovodným materiálem těchto skript, je jako implicitní metoda pro odhad vyhlazovacího parametru automatickou procedurou použita iterační metoda (podrobněji např. [4]). Proto při výpočtu optimálních parametrů použijeme mezivýpočty z tohoto toolboxu.

k	$K_{opt,0,k}$	δ_{0k}	h	$L(K)$
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 3.11.



Obrázek 3.11: Simulovaná data (×) s jádrovým odhadem hustoty při použití procedury (—) a původní funkcí (—)

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 3.12. 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: $K(x) = \frac{35}{32}(1-x^2)^3 I_{[-1,1]}(x)$.

7 Aplikace na reálná data

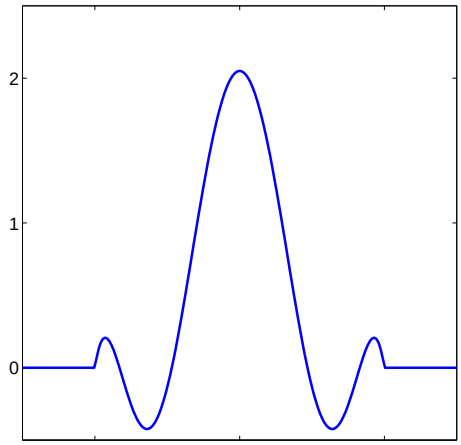
Datový soubor obsahuje morfologická měření padesáti exemplářů od obojího pohlaví a obou barevných forem (oranžová a modrá) krabů rodu *Leptograpsus*.[1] Pro odhad hustoty nám postačuje jeden druh měření, vybrali jsme délku podél středové osy krunýře, která byla měřena v milimetrech. Data jsou uvedena v tabulce 6.7.

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

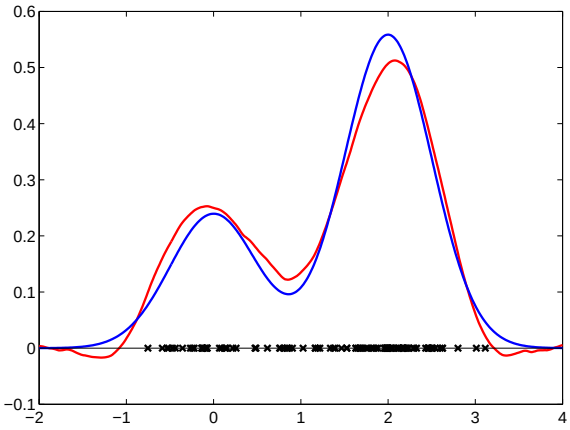
$$h_{\mathrm{REF}} = 5{,}7856, \quad h_{\mathrm{MS}} = 6{,}2480, \quad h_{\mathrm{CV}} = 8{,}1317.$$

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 $h_{\mathrm{proc}} = 31{,}7329$ s optimálním jádrem $K_{opt,0,8}$. Výsledné odhady hustoty na jsou zachyceny na obrázku 3.13.

[1] Celý datový soubor je dostupný v programu R.

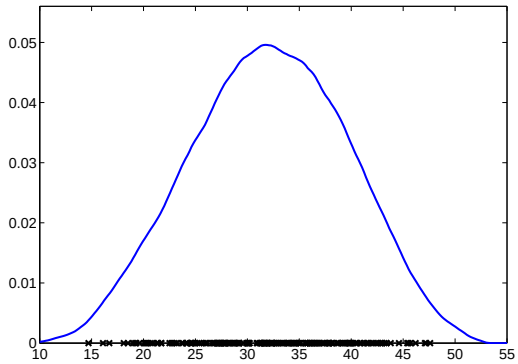


(a) Jádro $K_{opt,0,6}$

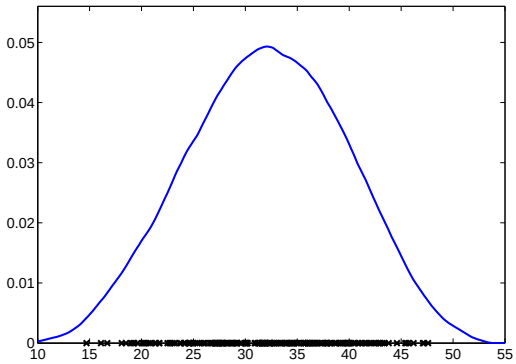


(b) Odhad hustoty

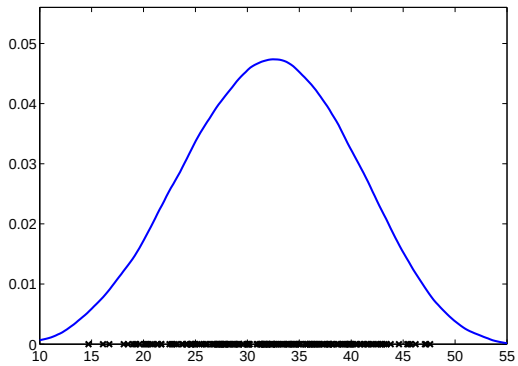
Obrázek 3.12: Jádro třídy S_{06} a k němu příslušný odhad hustoty při použití procedury (—) a původní funkcí (—)



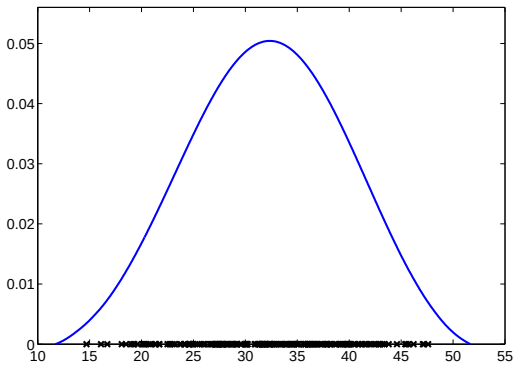
(a) REF



(b) MS



(c) CV



(d) procedura

Obrázek 3.13: Grafy odhadnutých hustot pro délku krunýře

Shrnutí

Odhad hustoty pravděpodobnosti f v bodě x je tvaru

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

Asymptotická střední kvadratická chyba jádrového odhadu hustoty pravděpodobnosti s jádrem řádu k je součtem asymptotického tvaru rozptylu (AIV) a druhé mocniny vychýlení (AISB)

$$\mathrm{AMISE}(h) = \underbrace{\frac{V(K)}{nh}}_{\mathrm{AIV}} + \underbrace{\frac{1}{(k!)^2} h^{2k} \beta_2^2(K) V(f^{(k)})}_{\mathrm{AISB}}.$$

Optimální vyhlazovací parametr vzhledem k AMISE pro odhad hustoty pravděpodobnosti s jádrem řádu k je tvaru

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

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

Metody pro odhad optimální hodnoty vyhlazovacího parametru h

- metoda referenční hustoty

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

- metoda maximálního vyhlazení

$$h_{\mathrm{MS}} = \widehat{\sigma} n^{-1/(2k+1)} b_k,$$

- metoda křížového ověřování

$$h_{\mathrm{CV}} = \arg \min_{h \in H_n} \quad \mathrm{CV}(h) = \int \hat{f}^2(x,h)\, \mathrm{d}x - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i, h).$$

Automatická procedura pro simultánní volbu optimálního jádra, vyhlazovacího parametru a řádu jádra je dostupná v toolboxu Matlabu.

Dodatky a cvičení

1. Dokažte vztah (3.4) pro tvar chyby AMISE.
2. Dokažte vztah (3.13).
3. Spočítejte (3.10).
4. Spočítejte h_{MS} pro
   - Epanečnikovo jádro $K(x) = \frac{3}{4}(1 - x^2)$, je-li $\frac{V(K)}{\beta_2^2(K)} = \frac{3/5}{(1/5)^2} = 15$,
   - kvartické jádro $K(x) = \frac{15}{16}(1 - x^2)^2$, je-li $\frac{V(K)}{\beta_2^2(K)} = \frac{5/7}{(1/7)^2} = 35$.

5. Aplikujte metody pro odhad vyhlazovacího parametru a automatickou proceduru na simulovaná i reálná data.

Kapitola 4

Jádrové odhady distribuční funkce

Výstupy z výukové jednotky

Student

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

1 Motivace

Distribuční funkce popisuje rozložení pravděpodobnosti náhodné veličiny (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.

Nejužívanějším neparametrickým odhadem distribuční funkce F 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 $\widehat{F}$, který se získá integrací známého jádrového odhadu hustoty (3.1)

2 Základní typy neparametrických odhadů

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

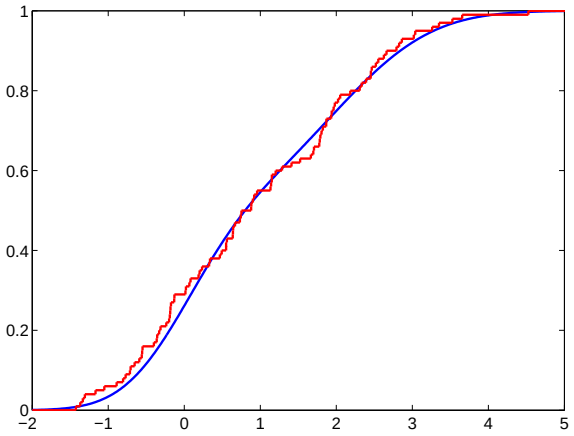
$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i).$$

Tento odhad má sice dobré statistické vlastnosti, ale je to schodovitá funkce (viz obr. 4.1, a proto se budeme zabývat postupy, které umožní zkonstruovat „hladký“ odhad distribuční funkce F.

Příklad 2.1. Mějme dán náhodný výběr o velikosti $n = 100$ ze směsi dvou normálních hustot $N(0; 4/9)$ a $N(2; 1)$ s hustotou

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

(Data jsou v tabulce 6.3.) Z obrázku 4.1 je patrné, že schodovitá funkce nevystihuje plně charakter distribuční funkce.



Obrázek 4.1: Empirická distribuční funkce (červeně) a skutečná distribuční funkce (modře) pro data z příkladu 2.1

Nejznámější postup spočívá v integraci jádrového odhadu hustoty, t.j.

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

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

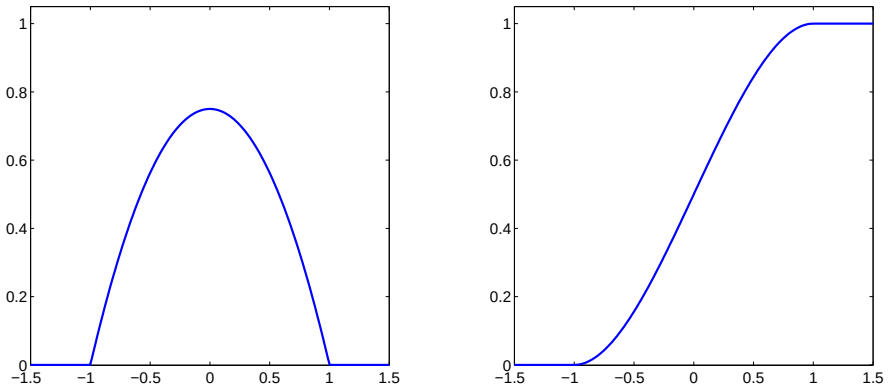
$$\widehat{F}(x,h) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\frac{x-X_i}{h}} K(y)\,\mathrm{d}y = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x-X_i}{h}\right).$$

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

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

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

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

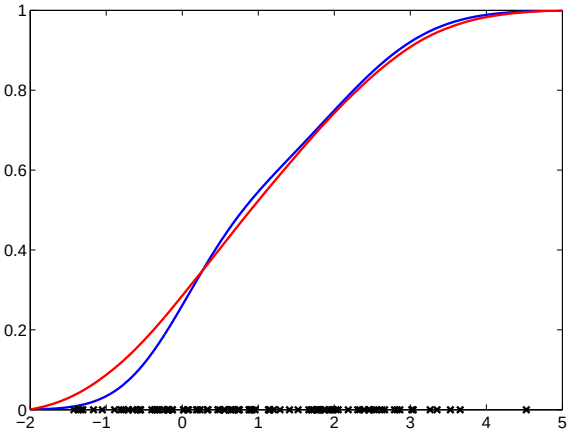


Obrázek 4.2: Epanečnikovo jádro K (vlevo) a k němu příslušná funkce W (vpravo)

Příklad 2.2. Použijeme-li Epanečnikovo jádro $K(x) = \frac{3}{4}(1-x^2)I_{[-1,1]}(x)$, pak funkce W je tvaru

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

Pro data z příkladu 2.1 je jádrový odhad distribuční funkce zachycen na obrázku 4.3.



Obrázek 4.3: Jádrový odhad distribuční funkce s parametrem $h = 1{,}5$

3 Statistické vlastnosti odhadu

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

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

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

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

Označme první integrál I_1 a druhý I_2. Spočítejme první z nich. S využitím vlastností funkce W dostaneme

$$I_1 = -F(x-h) + \int_{-1}^{1} K(t)F(x-ht)\,\mathrm{d}t. \tag{4.2}$$

Dále použijeme Taylorův rozvoj

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

tedy

$$I_1 = -F(x-h) + F(x) + \frac{1}{2}F''(x)h^2\beta_2(K) + o(h^2).$$

Je zřejmé, že pro integrál I_2 platí

$$I_2 = F(x-h). \tag{4.3}$$

Vychýlení odhadu je tedy tvaru

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

Poznámka 3.1. Vztahy (4.2) a (4.3) dávají zajímavý vztah pro vychýlení

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

Výpočet tvaru pro rozptyl je komplikovanější, a proto ho nebudeme uvádět. Platí

$$\operatorname{var}\widehat{F}(x,h) = \frac{1}{n}F(x)(1-F(x)) - \frac{1}{n}hf(x)\left(1 - \int_{-1}^{1} W^2(t)\,\mathrm{d}t\right) + o\left(\frac{h}{n}\right).$$

Důkaz tvaru rozptylu lze najít např. v [4].

Výše uvedené výsledky můžeme nyní zformulovat v následující větě:

Věta 3.1. *Nechť $F \in C^2$, $h \to 0, nh \to \infty$ pro $n \to \infty$. Pak*

$$\operatorname{MSE}\widehat{F}(x,h) = \frac{1}{n}F(x)(1-F(x)) - \frac{1}{n}hf(x)\left(1 - \int_{-1}^{1} W^2(t)\,\mathrm{d}t\right) + \frac{1}{4}(F''(x))^2h^4\beta_2^2(K) + o\left(\frac{h}{n} + h^4\right). \tag{4.4}$$

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

Věta 3.2. *Nechť $F \in C^2$, $V(F'') = \int (F''(x))^2\,\mathrm{d}x < \infty$, $K \in S_{02}$, $\lim_{n\to\infty} h = 0$ a $\lim_{n\to\infty} nh = \infty$. Pak*

$$\operatorname{MISE}\widehat{F}(\cdot,h) = \frac{1}{n}\int F(x)(1-F(x))\,\mathrm{d}x - c_1\frac{h}{n} + c_2h^4 + o\left(\frac{h}{n} + h^4\right), \tag{4.5}$$

kde

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

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:

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

Nyní už lze standardními metodami matematické analýzy nalézt takovou hodnotu h, pro kterou $\mathrm{AMISE}(h)$ nabývá minimální hodnoty. Je snadné ukázat, že

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

a pak

$$\mathrm{AMISE}(h_{opt,0,2}) = \frac{1}{n}\int F(x)(1-F(x))\,\mathrm{d}x - \frac{3}{c_2^{1/3}}\left(\frac{c_1}{4}\right)^{4/3} n^{-4/3}. \tag{4.8}$$

Poznámka 3.2. Optimální hodnota vyhlazovacího parametru pro odhad distribuční funkce je řádu $n^{-1/3}$, zatímco pro odhad hustoty s jádrem $K \in S_{02}$ je vyhlazovací parametr řádu $n^{-1/5}$.

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

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

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

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

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

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

$$\int (F''(x))^2 \,\mathrm{d}x = \int (f'(x))^2 \,\mathrm{d}x,$$

můžeme aplikovat Terrelovu větu 5.1 pro $k = 1$. V tomto případě je

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

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

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

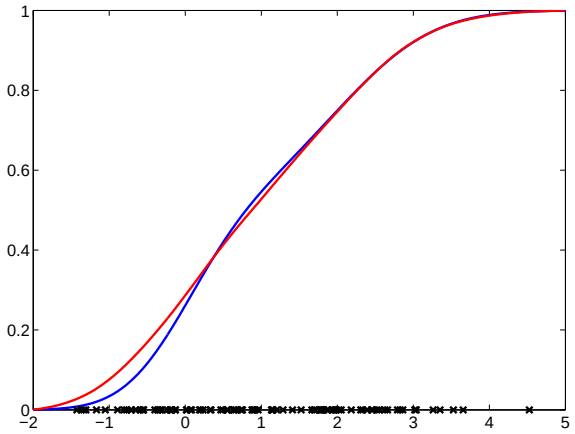
$\widehat{\sigma}$ je odhadem σ (viz rovnice (3.11) a (3.12)).

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

Příklad 5.1. Pro data z příkladu 2.1 zvolme Epanečnikovo jádro. Pak hodnoty potřebné pro odhad vyhlazovacího parametru metodou maximálního vyhlazení jsou následující:

$$n = 100, \qquad \widehat{\sigma} = 1{,}3426, \qquad \beta_2(K) = \frac{1}{5}, \qquad c_1 = 1 - \int_{-1}^{1} W^2(x) \,\mathrm{d}x = 0{,}2571.$$

Pak platí $h_{\mathrm{MS}} = 1{,}1037$ a na obrázku 4.4 je zobrazen odhad distribuční funkce.



Obrázek 4.4: Odhad distribuční funkce s $h_{\mathrm{MS}} = 1{,}1037$, odhad (—), původní funkce (—)

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

$$V(F'') = \int (F''(x))^2 \,\mathrm{d}x = -\int f''(x) f(x) \,\mathrm{d}x.$$

Tudíž se budeme dále zabývat odhadem funkcionálu

$$\psi_1 = \int f''(x) f(x)\,\mathrm{d}x.$$

Je zřejmé, že $\psi_1 = Ef''(X)$, což vede k metodě založené na odhadu druhé derivace hustoty f. Vztah (3.7) použijeme k odhadu druhé derivace s jádrem $K^{(2)} \in S_{24}$. Pak

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

kde podle vztahu (3.9) je

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

a tedy

$$\widehat{c_2} = -\frac{1}{4}\beta_2^2(K)\widehat{\psi}_1.$$

Shrnutím předchozích úvah dostaneme proceduru pro odhad distribuční funkce F:

Krok 1 Najděte optimální vyhlazovací parametr $\hat{h}_{opt,0,4}$ pro odhad hustoty s optimálním jádrem $K_{opt,0,4} \in S_{04}$.

Krok 2 Najděte optimální vyhlazovací parametr $\hat{h}_{opt,2,4}$ pro odhad druhé derivace hustoty podle vztahu (3.9) s $k = 4$ a optimálním jádrem $K^{(2)}_{opt,2,4} \in S_{24}$.

Krok 3 Vypočtěte odhad funkcionálu $\widehat{\psi}_1$ s využitím hodnot $\hat{h}_{opt,0,4}$ a $\hat{h}_{opt,2,4}$ získaných v krocích 1 a 2.

Krok 4 Vyčíslete optimální hodnotu vyhlazovacího parametru

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

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

Příklad 5.2. S použitím funkce toolboxu zjistíme, že pro data z příkladu 2.1 je vyhlazovací parametr určený plug-in metodou roven $h_{\mathrm{PI}} = 0{,}5717$. Na obrázku 4.5 je odhad distribuční funkce společně se skutečnou distribuční funkcí.

6 Aplikace na reálná data

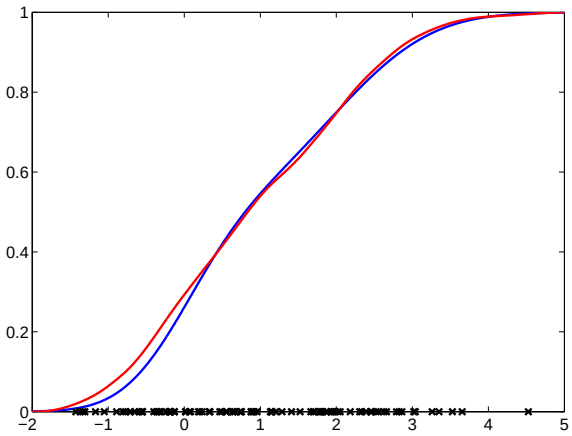
Datový soubor pochází z rozsáhlé studie, v níž autoři studovali vliv substituentů v 2,4-diamino-5-(substituovaný benzyl)pyrimidinech. Biologická aktivita při inhibici dihydrofolát reduktázy byla měřena pomocí asociační konstanty. Data jsou v tabulce 6.8 a jsou dostupná na osobních stránkách Dennise D. Boose[1], 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:

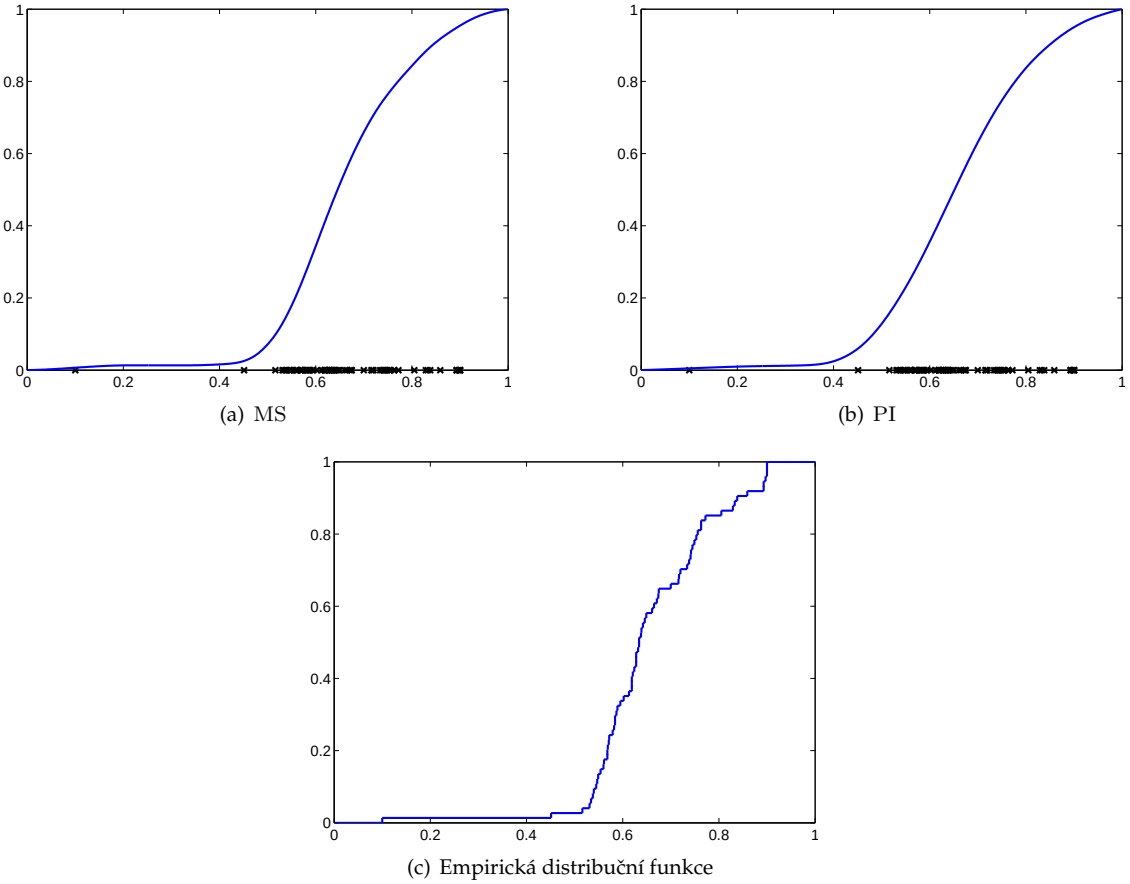
$$h_{\mathrm{MS}} = 0{,}1139, \quad h_{\mathrm{PI}} = 0{,}1931.$$

Na obrázku 4.6 jsou uvedeny odhady distribuční funkce s těmito parametry a také je zde pro srovnání uvedena empirická distribuční funkce.

[1] `http://www4.stat.ncsu.edu/~boos/var.select/pyrimidine.html`



Obrázek 4.5: Odhad distribuční funkce s $h_{\mathrm{PI}} = 0{,}5717$, odhad (—), původní funkce (—)



(a) MS

(b) PI

(c) Empirická distribuční funkce

Obrázek 4.6: Odhadnuté distribuční funkce

Shrnutí

Odhad distribuční funkce $F(x)$ v bodě x je tvaru

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

Asymptotická střední kvadratická chyba jádrového odhadu distribuční funkce je součtem asymptotického tvaru rozptylu (AIV) a druhé mocniny vychýlení (AISB)

$$\mathrm{AMISE}(h) = \underbrace{\frac{1}{n}\int F(x)(1 - F(x))\,\mathrm{d}x - c_1\frac{h}{n}}_{\mathrm{AIV}} + \underbrace{c_2 h^4}_{\mathrm{AISB}},$$

kde

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

Optimální vyhlazovací parametr vzhledem k AMISE pro odhad distribuční funkce je tvaru

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

t.j. $h_{opt,0,2} = O(n^{-1/3})$.

Metody pro odhad optimální hodnoty vyhlazovacího parametru h

- metoda křížového ověřování $h_{\mathrm{CV}} = \arg\min_{h \in H_n} \mathrm{CV}(h)$

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

- metoda maximálního vyhlazení

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

- plug-in metoda

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

Dodatky a 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 Epanečnikovo jádro.

5. Odvoďte tvar vyhlazovacího parametru podle plug-in metody pro Epanečnikovo a pro kvartické jádro.

6. Aplikujte metodu maximálního vyhlazení a plug-in metodu na simulovaná i reálná data.

Kapitola 5

Jádrové odhady dvourozměrných hustot

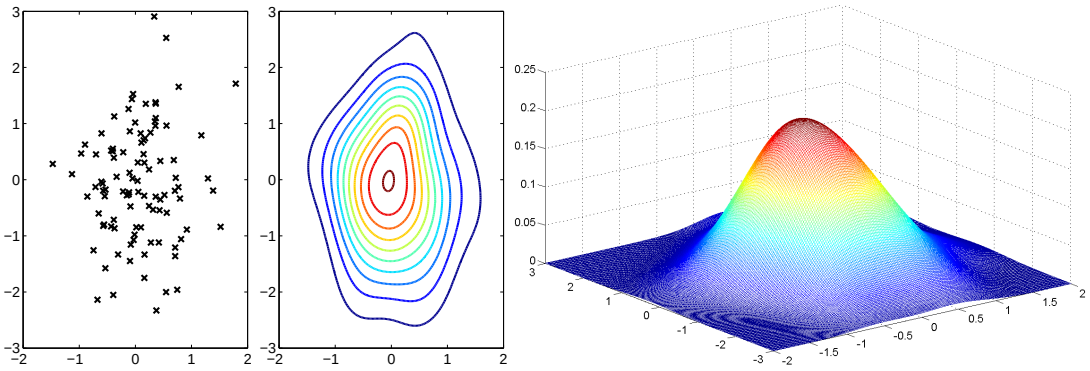
Výstupy z výukové jednotky

Student

- bude znát součinová a sférická dvourozměrná jádra pro odhady dvourozměrných hustot.
- porozumí principu vyhlazování dvourozměrných hustot.
- pochopí nejjednodušší metody pro volbu prvků diagonální vyhlazovací matice.
- zvládne použití příslušného toolboxu v Matlabu pro simulační studii i pro zpracování reálných dat.

1 Motivace

V této kapitole se budeme zabávat rozšířením jádrových odhadů pro jednorozměrné hustoty na odhad vícerozměrných hustot. 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) v kapitole 3, a zaměříme se zejména na odhad dvourozměrné hustoty.



Obrázek 5.1: Náhodný výběr a jeho jádrový odhad

Poznámka 1.1. Jádrové odhady dvourozměrných hustot se obvykle znázorňují pomocí vrstevnic, které umožňují snazší náhled na odhadnutou funkci.

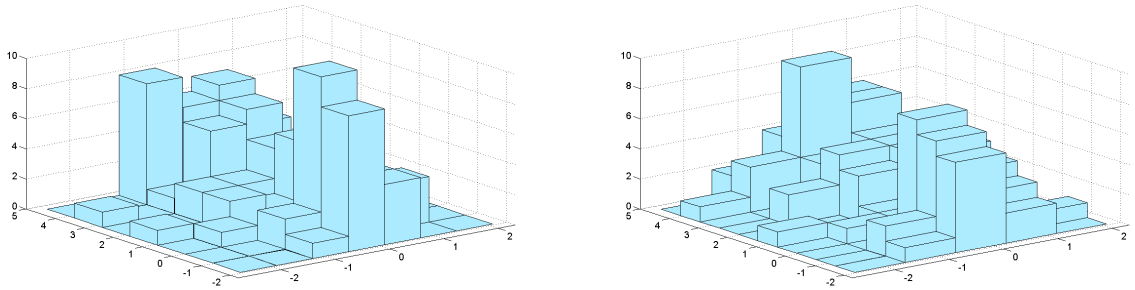
2 Základní typy odhadů

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 třídicích obdélníků – viz obrázek 5.2.

Příklad 2.1. Mějme dán datový soubor o velikosti $n = 100$ generovaný ze směsi tří normálních hustot $N(0, -1; 1/3, 1/3, 0)$, $N(0, 2; 1, 1, 0)$ a $N(0, 4; 1/3, 1/3, 0)$[1].

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

(Data jsou v tabulce 6.4.) Z obrázku 5.2 je patrné, že histogram nepostihuje charakteristické rysy hustoty pravděpodobnosti dat.



(a) 7×7 obdélníků (b) 5×10 obdélníků

Obrázek 5.2: Histogramy s různými počty třídicích obdélníků

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

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

přičemž $\mathbf{H}$ je matice vyhlazovacích parametrů a K je dvourozměrné jádro.

Jádro K je dvourozměrná funkce, kterou můžeme získat pomocí jednorozměrného symetrického jádra K_1. Existují dva typy těchto jader:

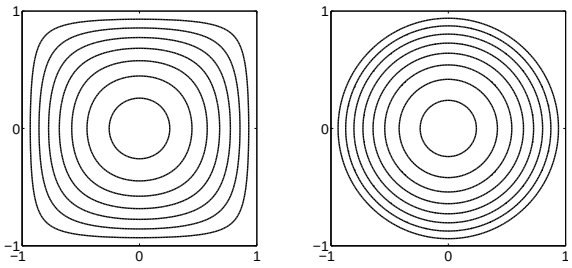
- *součinové jádro* $K^P(x, y) = K_1(x) \cdot K_1(y)$,
- *sféricky symetrické jádro* $K^S(x, y) = c_k^{-1} K_1\left(\sqrt{x^2 + y^2}\right)$, $c_k = \iint K_1\left(\sqrt{x^2 + y^2}\right) \mathrm{d}x\,\mathrm{d}y$.

Příklad 2.2. Epanečnikovo jádro, které je v jednorozměrném případě tvaru $K(x) = \frac{3}{4}(1-x^2)I_{[-1,1]}(x)$, má následující dvourozměrné varianty

$$K^P(x_1, x_2) = \frac{9}{16}(1 - x^2)(1 - y^2) \quad \text{pro} \quad -1 \le x, y \le 1,$$
$$K^S(x_1, x_2) = \frac{2}{\pi}(1 - x^2 - y^2) \quad \text{pro} \quad x^2 + y^2 \le 1.$$

Na obrázku 5.3 jsou zobrazeny vrstevnice těchto jader.

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



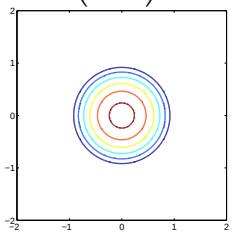
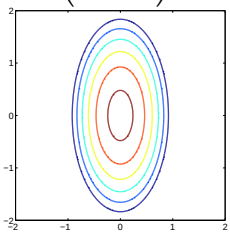
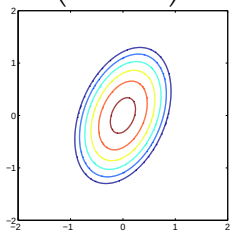
Obrázek 5.3: Součinové (vlevo) a sféricky symetrické (vpravo) dvourozměrné Epanečnikovo jádro

Podívejme se blíže na vyhlazovací 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 součinového Epanečnikova jádra v závislosti na třídě matic.

Tabulka 5.1: Třídy vyhlazovacích matic

$\mathcal{S}$	$\mathcal{D}$	$\mathcal{F}$
$\begin{pmatrix} h & 0 \\ 0 & h \end{pmatrix}$	$\begin{pmatrix} h_1 & 0 \\ 0 & h_2 \end{pmatrix}$	$\begin{pmatrix} h_1 & h_{12} \\ h_{12} & h_2 \end{pmatrix}$
		

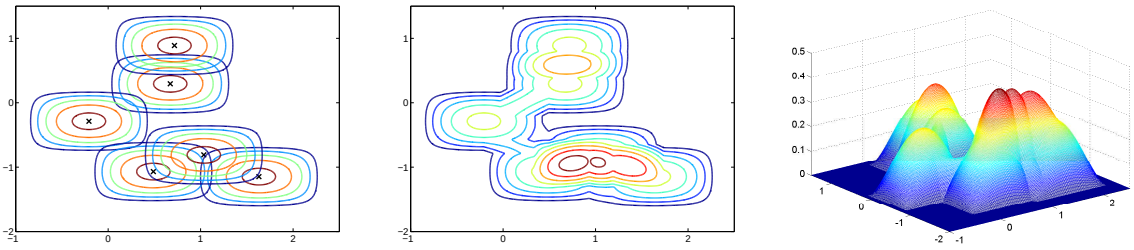
Budeme se zabývat jádrovými odhady s diagonální vyhlazovací maticí. Jádrový odhad s maticí třídy $\mathcal{S}$ dává ve všech směrech stejnou míru vyhlazení, což neponechává příliš mnoho prostoru pro zachycení variabilty dat. Na druhou stranu při použití matice třídy $\mathcal{F}$ je potřeba odhadnout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

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



Obrázek 5.4: Konstrukce jádrového odhadu hustoty

nebo globálně pomocí střední integrální kvadratické chyby

$$\mathrm{MISE}\,\hat{f}(x,y;\mathbf{H}) = \iint \mathrm{MSE}\,\hat{f}(x,y;\mathbf{H})\,\mathrm{d}x\,\mathrm{d}y.$$

Poznámka 3.1. Podobně jako v jednorozměrném případě se definuje konvoluce funkcí dvou proměnných. Nechť jsou dány funkce f a g, pro které platí $\iint f^2(x,y)\,\mathrm{d}x\,\mathrm{d}y < \infty$ a $\iint g^2(x,y)\,\mathrm{d}x\,\mathrm{d}y < \infty$. Konvoluci $f * g$ definujeme vztahem

$$(f*g)(x,y) = \iint_{-\infty}^{\infty} f(t,u)g(x-t,y-u)\,\mathrm{d}t\,\mathrm{d}u.$$

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.

Věta 3.1. *Předpokládejme, že funkce f, jádro K a vyhlazovací matice $\mathbf{H} = \mathrm{diag}(h_1,h_2)$[2] splňují následující předpoklady.*

(i) Nechť $\mathbf{H} = \mathbf{H}_n$ je posloupnost vyhlazovacích matic takových, že $(n|\mathbf{H}|)^{-1}$ a prvky matice $\mathbf{H}$ konvergují k nule pro $n \to \infty$.

(ii) Dále nechť všechny druhé parciální derivace funkce f jsou ohraničené, spojité a integrovatelné se čtvercem.

(iii) Jádro K splňuje

$$\iint xK(x,y)\,\mathrm{d}x\,\mathrm{d}y = \iint yK(x,y)\,\mathrm{d}x\,\mathrm{d}y = 0$$
$$\iint x^2K(x,y)\,\mathrm{d}x\,\mathrm{d}y = \iint y^2K(x,y)\,\mathrm{d}x\,\mathrm{d}y = \beta_2(K).$$

Pak platí

$$\mathrm{MISE}(\mathbf{H}) = \mathrm{AMISE}(\mathbf{H}) + o(h_1^2+h_2^2) + o\big((h_1h_2n)^{-1}\big),$$

kde

$$\mathrm{AMISE}(\mathbf{H}) \equiv \mathrm{AMISE}\,\hat{f}(\cdot,\mathbf{H}) = \frac{V(K)}{nh_1h_2} + \frac{1}{4}\beta_2^2(K)\big(h_1^4V(f_{xx}) + 2h_1^2h_2^2V(f_xf_y) + h_2^2V(f_{yy})\big), \quad (5.2)$$

přičemž označení je ve shodě s předchozími kapitolami, tj. $V(g) = \iint g^2(x,y)\,\mathrm{d}x\,\mathrm{d}y$.

[2] Užíváme zde zkráceného zápisu: $\mathrm{diag}(h_1,h_2) = \begin{pmatrix} h_1 & 0 \\ 0 & h_2 \end{pmatrix}$

Důkaz věty o tvaru AMISE je založen na Taylorově rozvoji funkce $f(x, y)$ a lze jej nalézt např. v knize [13].

Hodnoty parametrů h_1, h_2, pro které $\mathrm{AMISE}(h_1, h_2)$ nabývá minimální hodnoty, jsou dány vztahy:

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

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

4 Volba jádra

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

Poznámka 4.1. V literatuře se také využívá Gaussovo jádro $K(x, y) = (2\pi)^{-1} \mathrm{e}^{-(x^2+y^2)/2}$, které se zdá být výhodnějším při studiu asymptotických vlastností jádrové odhadu. Na druhou stranu má nevýhodu, že jeho nosičem je celá reálná osa, což způsobuje „nedokonalosť" při odhadech hustot s omezeným definičním oborem.

5 Volba vyhlazovacího parametru

5.1 Metoda referenční hustoty

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

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

a jádro K je dvourozměrnou standardizovanou normální hustotou, tj.

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

Pak podle metody referenční hustoty lze získat tyto odhady vyhlazovacích parametrů

$$h_{i,\mathrm{REF}} = \widehat{\sigma}_i n^{-1/6}, \quad i = 1, 2.$$

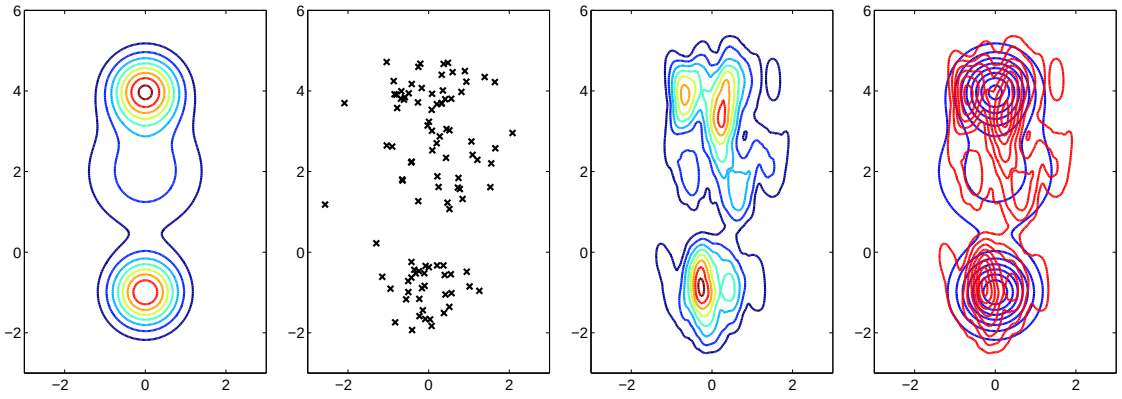
Tento vztah je také znám jako Scottovo pravidlo ([10]).

Příklad 5.1. Pro simulovaná data z příkladu 2.1 vychází vyhlazovací matice podle metody referenční hustoty s využitím součinového Epanečnikova jádra takto:

$$\mathbf{H}_{\mathrm{REF}} = \begin{pmatrix} 0{,}3595 & 0 \\ 0 & 0{,}9972 \end{pmatrix}.$$

Na obrázku 5.5 je vykreslen odhad hustoty s touto vyhlazovací maticí a porovnání odhadu se skutečnou hustotou.

Poznámka 5.1. V toolboxu, který doplňuje tato skripta, se uvádí druhá mocnina vyhlazovací matice, tedy $\mathbf{H}^2$.



(a) Skutečná hustota (b) Data (c) Odhad s $\mathbf{H}_{\mathrm{REF}}$ (d) Porovnání odhadu (—) se skutečnou hustotou (—)

Obrázek 5.5: Jádrový odhad dvourozměrné hustoty – referenční hustota

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

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

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

Věta 5.1. *Funkce* CV *je nevychýleným odhadem* MISE, *tj. platí*

$$E\,\mathrm{CV}(\mathbf{H}) = \mathrm{MISE}(\mathbf{H}) - \iint f^2(x, y)\, \mathrm{d}x\, \mathrm{d}y.$$

Důkaz. Vypočtěme střední hodnotu CV:

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

a úpravou MISE

$$\begin{aligned}\mathrm{MISE}(\mathbf{H}) &= E\iint \left(\hat{f}(x,y,\mathbf{H}) - f(x,y)\right)^2 \mathrm{d}x\,\mathrm{d}y \\ &= E\iint \hat{f}^2(x,y,\mathbf{H})\,\mathrm{d}x\,\mathrm{d}y - 2E\int \hat{f}(x,y,\mathbf{H}) f(x,y)\,\mathrm{d}x\,\mathrm{d}y + \iint f^2(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= E\iint \hat{f}^2(x,y,\mathbf{H})\,\mathrm{d}x\,\mathrm{d}y - 2\iiiint K_{\mathbf{H}}(x-u,y-v) f(x,y) f(u,v)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}u\,\mathrm{d}v \\ &\quad + \iint f^2(x,y)\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

Porovnáním upravených výrazů dostaneme tvrzení. □

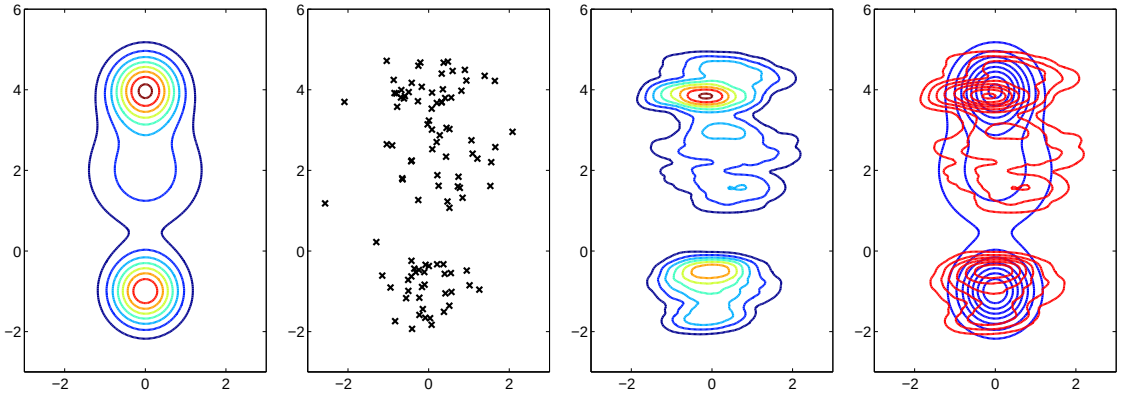
Optimální vyhlazovací parametry vzhledem k metodě CV označíme $\mathbf{H}_{\mathrm{CV}}$, tj.

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

Příklad 5.2. Použijeme-li součinové Epanečnikov jádro, pak pro simulovaná data z příkladu 2.1 dostanem vyhlazovací matici určenou podle metody křížového ověřování v následujícím tvaru:

$$\mathbf{H}_{\mathrm{CV}} = \begin{pmatrix} 1{,}2055 & 0 \\ 0 & 0{,}3783 \end{pmatrix}.$$

Na obrázku 5.6 je vykreslen odhad hustoty s touto vyhlazovací maticí.



(a) Skutečná hustota (b) Data (c) Odhad s $\mathbf{H}_{\mathrm{CV}}$ (d) Porovnání odhadu (—) se skutečnou hustotou (—)

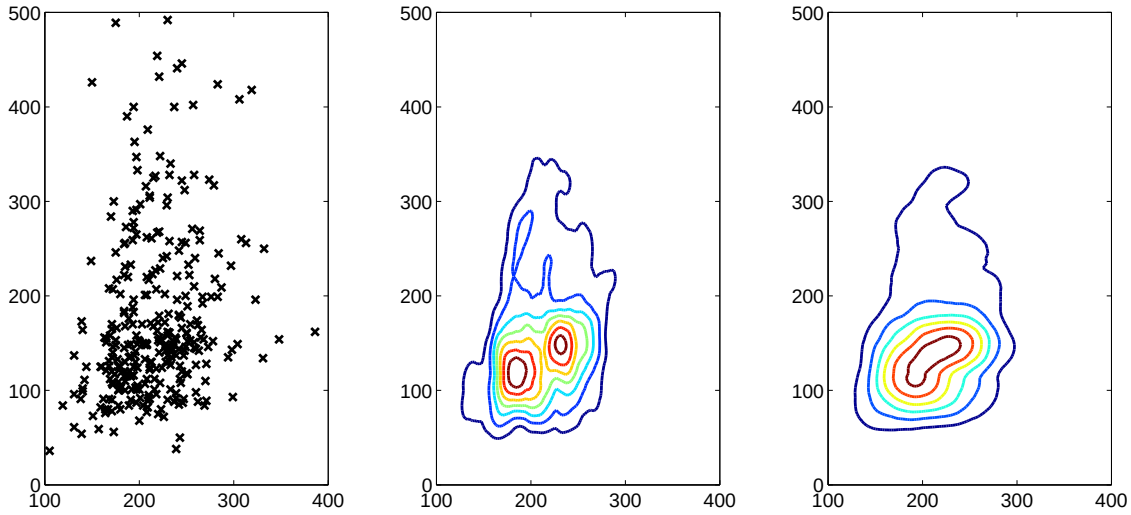
Obrázek 5.6: Jádrový odhad dvourozměrné hustoty – metoda křížového ověřování

6 Aplikace na reálná data

Tento datový soubor pochází ze studie koncentrace lipidů v krevní plazmě, která vyšla v časopise Circulation v roce 1980. Výběrový soubor, který jsme převzali z [10] a s nímž zde pracujeme, obsahuje měření množství cholesterolu a triglyceridu v krevní plazmě u 320 pacientů, kteří si stěžovali na bolest v hrudníku. Data jsou shrnuta v tabulkách 6.9 a 6.10.

Vyhlazovací matice určené podle metody referenční hustoty a metody křížového ověřování jsou následující:

$$\mathbf{H}_{\mathrm{REF}} = \begin{pmatrix} 16{,}45 & 0 \\ 0 & 38{,}94 \end{pmatrix} \qquad \mathbf{H}_{\mathrm{CV}} = \begin{pmatrix} 42{,}39 & 0 \\ 0 & 29{,}88 \end{pmatrix}$$



(a) Data (b) REF (c) CV

Obrázek 5.7: Vrstevnicové grafy odhadnutých hustot pro koncentraci lipidů – na ose x je vyneseno množství cholesterolu (v miligramech na 100 ml plazmy) a na ose y množství triglyceridu v krevní plazmě (mg/100 ml)

Na obrázku 5.7 jsou znázorněna data a vrstevnice jádrového odhadu s Epanečnikovým součinovým jádrem.

Shrnutí

Odhad dvourozměrné hustoty pravděpodobnosti $f(x, y)$ v bodě $[x, y]$ je tvaru

$$\hat{f}(x, y, h_1, h_2) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h_1}, \frac{y - Y_i}{h_2}\Big)$$

Dva typy jader:

- součinové jádro: $K^P(x, y) = K_1(x) \cdot K_1(y)$,
- sféricky symetrické jádro: $K^S(x, y) = c_k^{-1} K_1\big(\sqrt{x^2 + y^2}\big)$, $c_k = \iint K_1\big(\sqrt{x^2 + y^2}\big)\,\mathrm{d}x\,\mathrm{d}y$.

Asymptotická střední integrální kvadratická chyba dvourozměrného jádrového odhadu

$$\mathrm{AMISE}(\mathbf{H}) = \frac{V(K)}{nh_1h_2} + \frac{1}{4}\beta_2^2(K)\big(h_1^4 V(f_{xx}) + 2h_1^2h_2^2 V(f_x f_y) + h_2^2 V(f_{yy})\big).$$

Optimální vyhlazovací parametry vzhledem k AMISE

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

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

Metody pro odhad optimálních hodnot vyhlazovacích parametrů $\mathbf{H} = \mathrm{diag}(h_1, h_2)$

- metoda referenční hustoty

$$h_{i,\mathrm{REF}} = \widehat{\sigma}_i n^{-1/6}, \quad i = 1, 2,$$

- metoda křížového ověřování

$$\mathbf{H}_{\mathrm{CV}} = \arg\min_{\mathbf{H}\in\mathcal{D}} \mathrm{CV}(\mathbf{H}), \quad \mathrm{CV}(\mathbf{H}) = \iint \big(\hat{f}(x, y, \mathbf{H})\big)^2\,\mathrm{d}x\,\mathrm{d}y - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i, Y_i, \mathbf{H}).$$

Dodatky a cvičení

1. Určete součinové 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. Aplikujte metodu referenční hustoty a metodu křížového ověřování na simulovaná i reálná data.

Kapitola 6

Přílohy

1 Symbolika O, o

Symboly O a o se často používají pro vyjádření chyb matematických výrazů

Nechť f je funkce definovaná v okolí bodu a. Symbol

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

značí, že

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

Podobně symbol

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

značí, že

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

Dodatek „pro $x \to a$“ 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 $h \to 0$.

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.

2 Datové soubory

Tabulka 6.1: Hodnoty simulovaných dat z příkladu 1.1 – 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 6.2: Hodnoty simulovaných dat z příkladu 2.1 – 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 6.3: Hodnoty simulovaných dat z příkladu 2.1 – 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 6.4: Hodnoty simulovaných dat z příkladu 2.1 – 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 6.5: Hodnoty reálných dat z kapitolky 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 6.6: Hodnoty reálných dat z kapitolky 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 6.7: Hodnoty reálných dat z kapitolky 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 6.8: Hodnoty reálných dat z kapitolky 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 6.9: Hodnoty reálných dat z kapitolky 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 6.10: Hodnoty reálných dat z kapitolky 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]

Literatura

[1] Anděl, J.: Základy matematické statistiky. Matfyzpress, Praha (2005) ISBN 80-86732-40-1

[2] Collomb, G.: Estimation non paramétrique de la régression par la méthode du noyau. Disertační práce, Univerzita Paula Sabatiera, Toulouse (1976)

[3] Forbelská, M., Koláček, J.: Pravděpodobnost a statistika I. Elektronický učební text, Masarykova unverzita, Brno (2012) http://is.muni.cz/el/1431/podzim2012/M3121/um/

[4] 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

[5] Horová, I., Vieu, P., Zelinka, J.: Optimal Choice of Nonparametric Estimates of a Density and of its Derivatives. Statistics & Decisions 20, 355–378 (2002)

[6] Horová, I., Zelinka, J.: Contribution to the bandwidth choice for kernel density estimates, Comput. Stat. 22, 31–47 (2007)

[7] 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)

[8] Koláček, J.: Jádrové odhady regresní funkce. Disertační práce, Masarykova univerzita, Brno (2005) http://is.muni.cz/th/19999/prif_d/

[9] Müller, H.-G.: Smooth optimum kernel estimators near endpoints. Biometrika 78, 521–530 (1991)

[10] Scott, D.W.: Multivariate density estimation: Theory, practise, and visualization. Wiley, New York (1992) ISBN 0-471-54770-0

[11] Silverman, B.W.: Density estimation for statistics and data analysis. Chapman and Hall, London (1986) ISBN 0-412-24620-1

[12] Terrell, G.R.: The maximal smoothing principle in density estimation. J. Am. Stat. Assoc. 85, 470–477 (1990)

[13] Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman and Hall, London (1995) ISBN 0-412-55270-1