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, 2].

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* [12, 11, 14]. Histogram zobrazuje relativní četnosti třídicích intervalů jako plochy obdélníků sestrojených nad těmito intervaly. Pak definujeme odhad hustoty četnosti

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

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

Nevýhody histogramu:

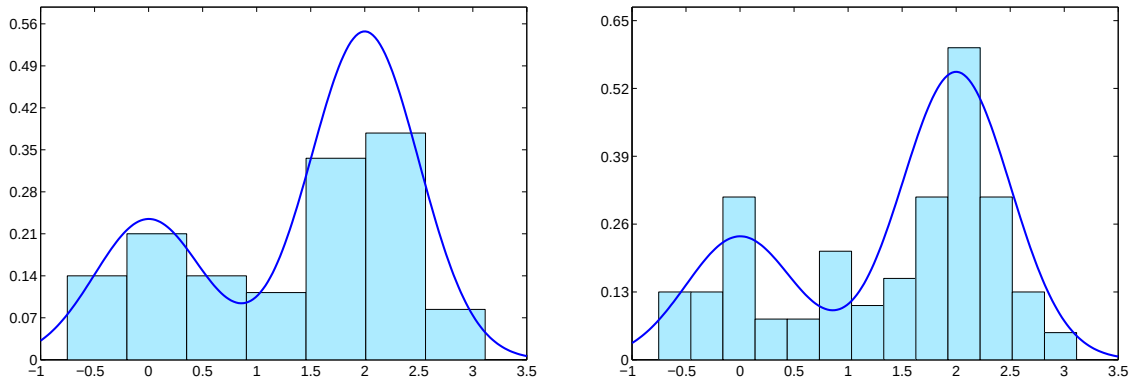
- Histogram je citlivý na počet tříd a jejich šířku.

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

Příklad 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).

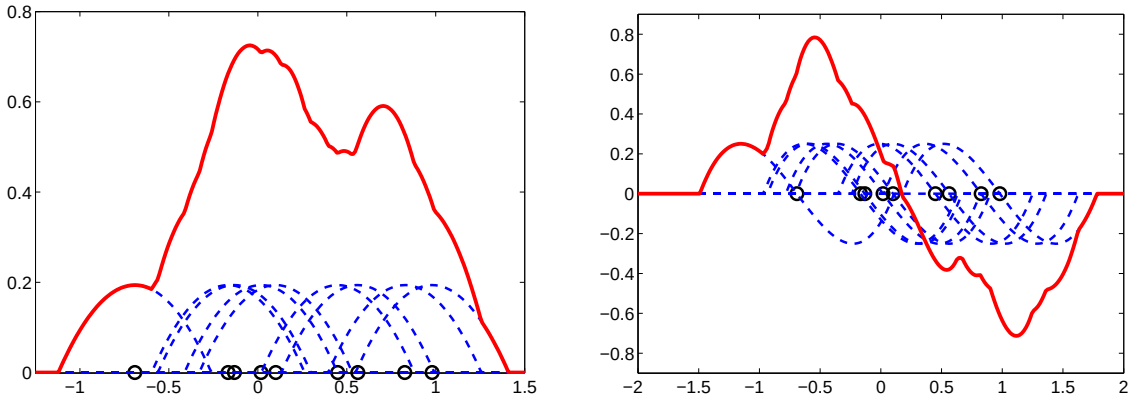
Nyní uvedeme ještě vztah pro jádrový odhad ν-té derivace hustoty. Budeme předpokládat, že $0 \le \nu \le k - 2$ a k, ν jsou stejné parity. Pak

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

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

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

Tedy vychýlení odhadu je tvaru

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

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

Nyní dokážeme vztah pro rozptyl. Víme, že

$$\operatorname{var}\hat{f}(x,h) = \frac{1}{n}\big((K_h^2 * f)(x) - (K_h * f)^2(x)\big)$$

a dále počítáme

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

Tedy

$$\operatorname{MSE}\hat{f}(x,h) = \frac{f(x)}{nh}\int K^2(z)\,\mathrm{d}z + o((nh)^{-1}) + \left(\frac{(-1)^k f^{(k)}(x)}{k!}h^k\beta_k(K) + o(h^k)\right)^2$$

a pak

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

□

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* $\operatorname{var}\hat{f} \to 0$.

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

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

kde AMISE je tvaru

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

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

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

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

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

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

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

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

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 odhadhu 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

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

Počítejme derivace AMISE (3.3)

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

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

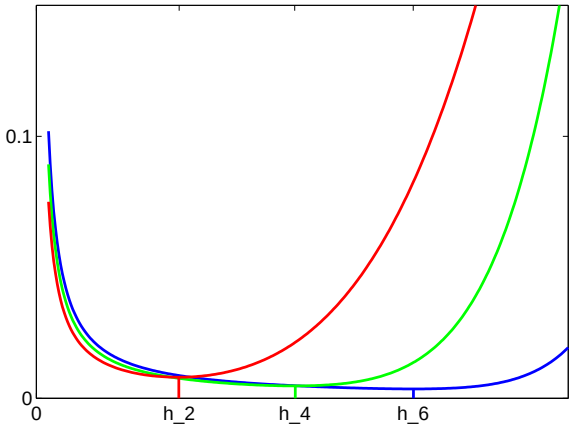
$$h^5 = \frac{V(K)}{n\beta_2^2(K) V(f'')} = \frac{\delta_{02}^5}{nV(f'')} = h_{opt,0,2}^5,$$

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).



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

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

3.1 Odhad derivace hustoty

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

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

Předpokládejme nyní, že platí $0 \le \nu \le k-2$, $\lim_{n\to\infty} h = 0$, $\lim_{n\to\infty} nh^{2\nu+1} = \infty$, $f \in C^{k_0}$ ($k \le k_0$) 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

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

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

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

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.

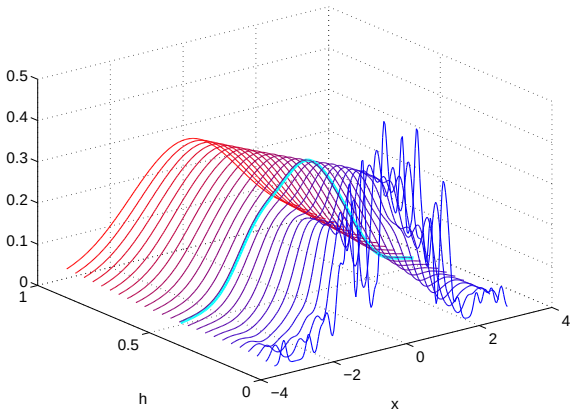
I když tomuto problému byla a je věnována značná pozornost, doposud neexistuje univerzální přístup k řešení tohoto problému. Nejjednodušší metoda je „okometrická". Je účelné „nakreslit" několik křivek s různými vyhlazovacími parametry dříve než užijeme nějakou automatickou proceduru.

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.

V dalších úvahách bude užitečná následující definice:

Definice 5.1. Nechť $\hat{h}$ je odhad $h_{opt,0,k}$. Řekneme, že $\hat{h}$ konverguje k $h_{opt,0,k}$ s relativní rychlostí $n^{-\alpha}$, jestliže

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



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 ($k = 2$)

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

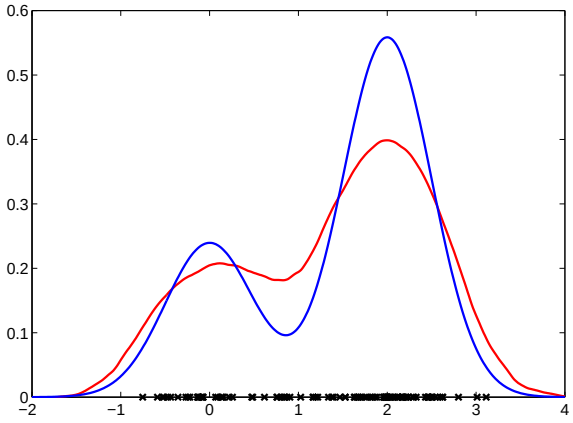
pak dostaneme jednoduchý vztah

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

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 f 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 $\int_{-1}^{1}\left(f^{(k)}(x)\right)^2 \mathrm{d}x$.

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

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

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

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

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

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

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

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

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

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

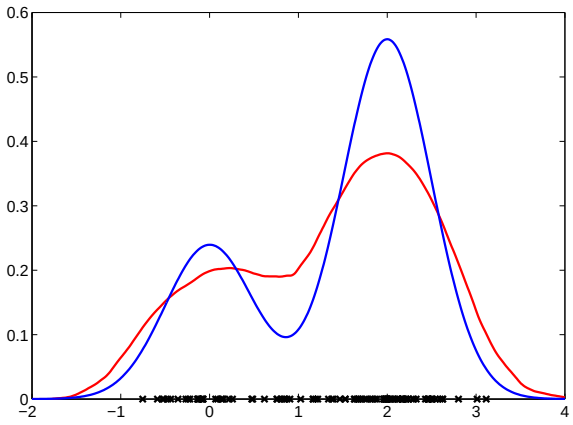
Pak pro $K_{opt,0,k}$

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

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

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

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 (—)

Poznámka 5.2. Hodnota h_{MS} může sloužit jako horní hranice pro množinu vyhlazovacích parametrů volených podle jiné metody, např. metody křížového ověřování. Tedy $H_n = [h_\ell, h_{\mathrm{MS}}]$, kde h_ℓ je nejmenší vzdálenost mezi po sobě jdoucími body $X_i, i = 1, \ldots, n$.

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:

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

Definujme funkci křížového ověřování

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

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

Věta 5.2. *Platí*

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

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

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

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

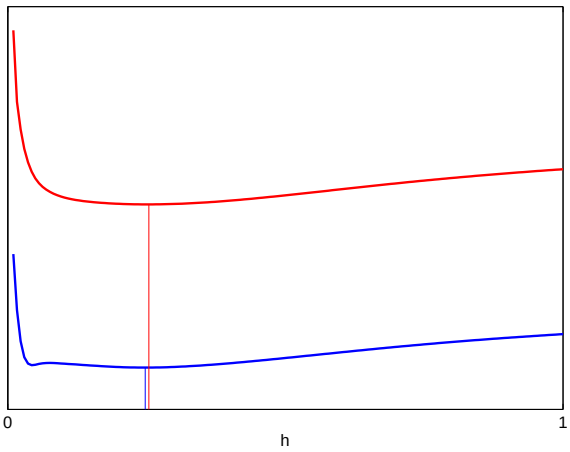
$$E\operatorname{CV}(h) = E\int\hat{f}^2(x,h)\,\mathrm{d}x - 2E\int\hat{f}(x,h)f(x)\,\mathrm{d}x.$$

□

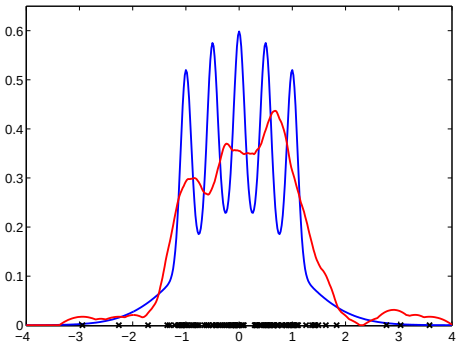
Odhad $h_{opt,0,k}$ je dán vztahem $h_{\mathrm{CV}} = \arg\min_{h\in H_n}\operatorname{CV}(h)$. Odtud plyne, že $\operatorname{CV}(h)+\int f^2(x)\,\mathrm{d}x$ je pro každé h nevychýleným odhadem $\operatorname{MISE}(h)$. Protože $\int f^2(x)\,\mathrm{d}x$ nezávisí na h, minimalizace $E\operatorname{CV}(h)$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

Poznámka 5.3. 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.

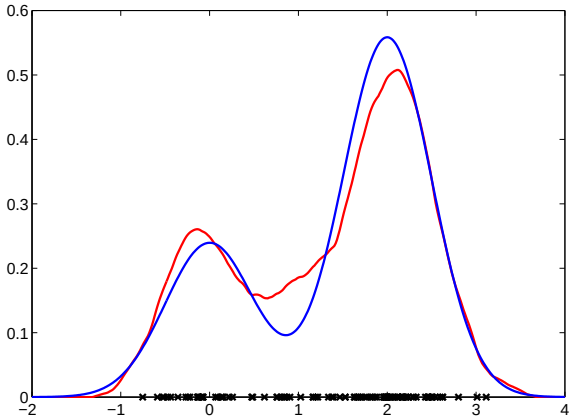
Příklad 5.4. Jádrový odhad hustoty dat z příkladu 2.1 je zobrazen na obr. 3.9. Pro rekonstrukci bylo použito Epanečnikovo jádro a vyhlazovací parametr určený metodou křížového ověřování $h_{\mathrm{CV}} = 0{,}5628$.



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



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 (—)

5.4 Iterační metoda

Připomeňme nyní vztah (3.6):

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

Přepíšeme tuto rovnici

$$\frac{V(K)}{nh} - 2kh^{2k}\frac{\beta_2^2(K)}{(k!)^2}V(f^{(k)}) = 0. \tag{3.16}$$

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

$$\begin{aligned}\widehat{\mathrm{var}}\hat{f}(x,h) &= \frac{1}{nh}\int K^2(y)\hat{f}(x-hy)\,\mathrm{d}y\\ \widehat{\mathrm{bias}}\hat{f}(x,h) &= \overline{\mathrm{bias}}\hat{f}(x,h) = (K*\hat{f})(x,h) - \hat{f}(x,h)\\ &= \hat{f}(x-hy,h)K(y)\,\mathrm{d}y - \hat{f}(x,h).\end{aligned}$$

Odtud plyne

$$\widehat{\mathrm{AIV}}\hat{f}(\cdot,h) = \frac{V(K)}{nh}\hat{f}(\cdot,h)$$

a

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

výraz lze upravit pomocí konvolucí a dostaneme

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

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

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

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

$$\widehat{\mathrm{AISB}}\hat{f}(\cdot,h) = \frac{1}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\Lambda_h(X_i-X_j).$$

Místo rovnice (3.16) řešíme rovnici

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

Tuto rovnice lze také přepsat ve tvaru:

$$h = \frac{nV(K)}{2k \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \Lambda_h(X_i - X_j)}. \tag{3.17}$$

Uvedenou rovnici lze řešit Newtonovou metodou, neboť derivaci funkce lze snadno spočítat užitím konvolucí. Řešení této rovnice označíme $\hat{h}_{\mathrm{IT}}$.

Řešení rovnice (3.17) lze považovat za vhodnou aproximaci $h_{opt,0,k}$. Tato skutečnost je dokázána v následující větě [3].

Věta 5.3. *Nechť*

$$P(h) = \frac{V(K)}{nh} - 2kh^{2k} \frac{\beta_2^2(K)}{(k!)^2} V(f^{(k)})$$

a

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

Pak platí

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

Poznámka 5.4. Rychlost konvergence pro metodu křížového ověřování:

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

Rychlost konvergence pro iterační metodu

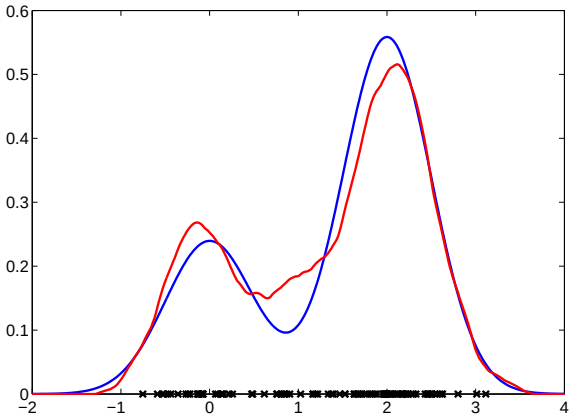
$$\frac{\hat{h}_{\mathrm{IT}} - h_{opt,0,2}}{h_{opt,0,2}} = O(n^{-1/10})$$

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

Příklad 5.5. Jádrový odhad hustoty dat z příkladu 2.1 s Epanečnikovým jádrem a vyhlazovacím parametrem určeným iterační metodou je uveden na obr. 3.10.

Shrneme-li na závěr doposud vypočítané hodnoty vyhlazovacích parametrů pro simulovaná data z příkladu 2.1, můžeme vizuálně porovnat jednotlivé odhady – viz obrázek 3.11.

$$h_{opt,0,2} = 0{,}5122 \quad h_{\mathrm{REF}} = 0{,}9639 \quad h_{\mathrm{MS}} = 1{,}0409 \quad h_{\mathrm{CV}} = 0{,}5628 \quad h_{\mathrm{IT}} = 0{,}5314$$



Obrázek 3.10: Odhad hustoty s $h = 0{,}5314$, odhad (—), původní funkce (—)

6 Automatická procedura

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

Vyjdeme ze vztahu

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

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

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

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}.

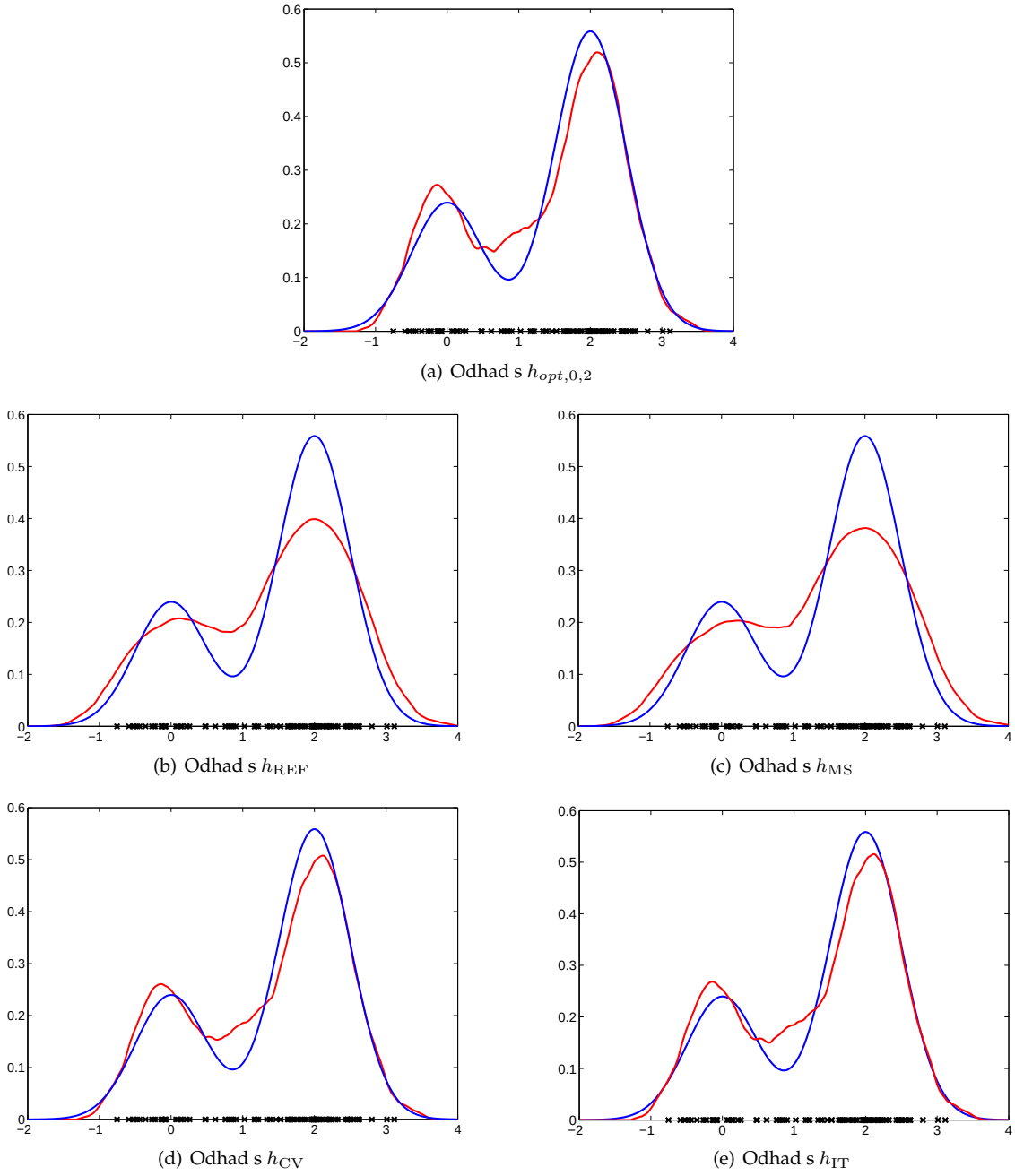
Krok 2 Pro $k \in I(k_0)$ a $K_{opt,0,k} \in S_{0k}$ najděte optimální vyhlazovací parametr $\hat{h}_{opt,0,k}$.

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.

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



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

(b) Odhad s h_{REF}

(c) Odhad s h_{MS}

(d) Odhad s h_{CV}

(e) Odhad s h_{IT}

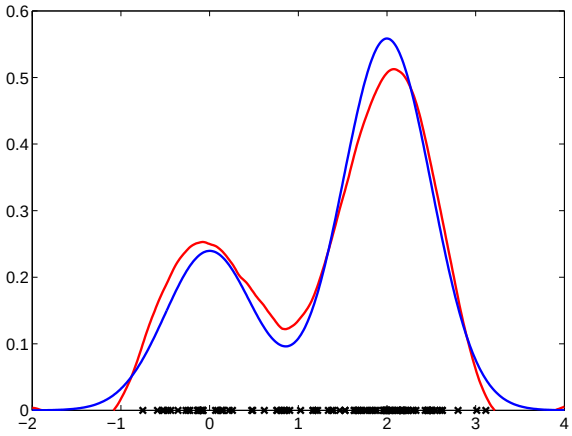
Obrázek 3.11: Srovnání odhadů pro data z příkladu 2.1

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. [3]). 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 3.12.



Obrázek 3.12: 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 3.13. 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

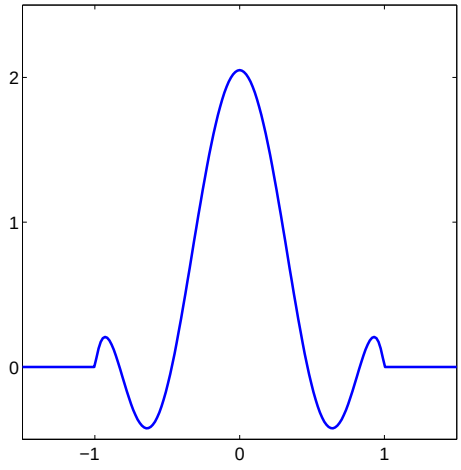
Datový soubor obsahuje morfologická měření padesáti exemplářů od obojího pohlaví a obou barevných forem (oranžová a modrá) krabů rodu *Leptograpsus*.[1] Pro odhad hustoty nám postačuje jeden druh měření, vybrali jsme délku podél středové osy krunýře, která byla měřena v milimetrech. Data jsou uvedena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

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

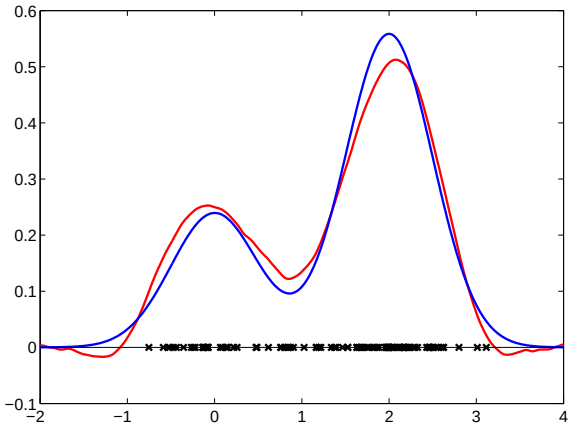
$$h_{\mathrm{REF}}=5{,}7856, \quad h_{\mathrm{MS}}=6{,}2480, \quad h_{\mathrm{CV}}=8{,}1317, \quad h_{\mathrm{IT}}=6{,}8263.$$

U automatické procedury je v toolboxu implicitně nastavena iterační metoda pro odhad vyhlazovacího parametru, proto jsme tuto volbu ponechali i zde, ať má čtenář možnost porovnání při vlastních výpočtech. Při použití procedury vyjde vyhlazovací parametr roven $h_{\mathrm{proc}}=31{,}7329$ s optimálním jádrem $K_{opt,0,8}$. Výsledné odhady hustoty na jsou zachyceny na obrázku 3.14.

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

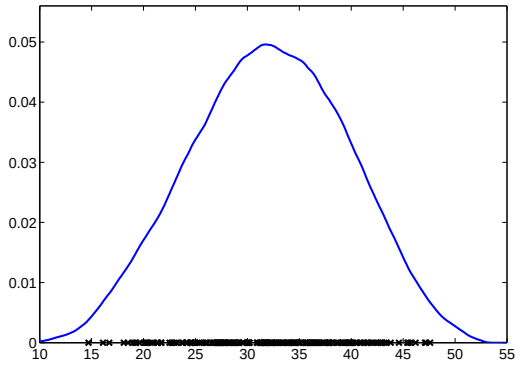


(a) Jádro $K_{opt,0,6}$

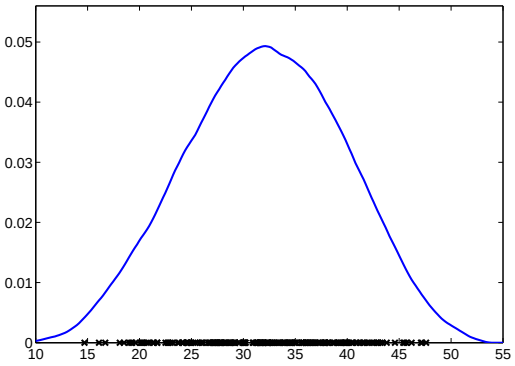


(b) Odhad hustoty

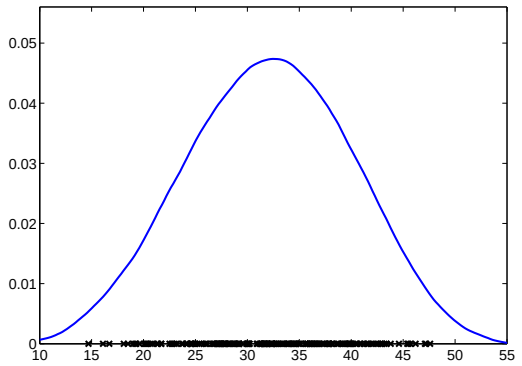
Obrázek 3.13: Jádro třídy S_{06} a k němu příslušný odhad hustoty při použití procedury (—) a původní funkcí (—)



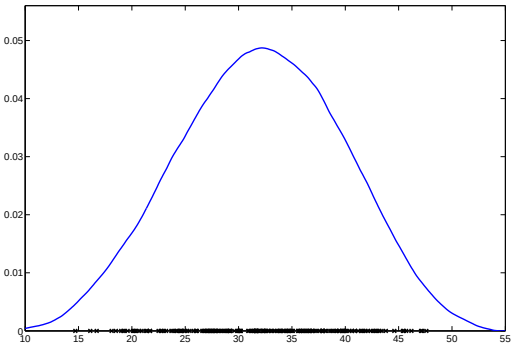
(a) REF



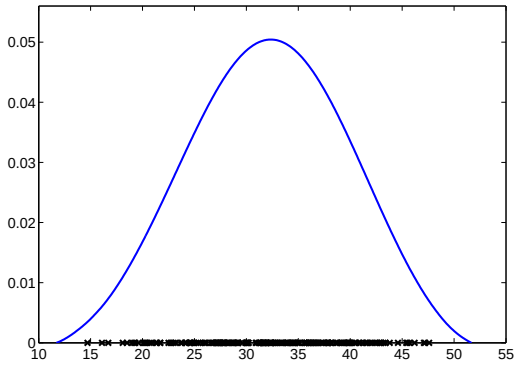
(b) MS



(c) CV



(d) IT



(e) procedura

Obrázek 3.14: Grafy odhadnutých hustot pro délku krunýře

Shrnutí

Odhad hustoty pravděpodobnosti f v bodě x je tvaru

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

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

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

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

- iterační metoda

$$h_{\mathrm{IT}} = \text{ pevný bod funkce: } h = \frac{nV(K)}{2k\sum_{\substack{i,j=1\\ i\neq j}}^{n}\Lambda_h(X_i-X_j)}.$$

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.