JÁDROVÉ ODHADY ZÁKLADNÍ VLASTNOSTI A OPTIMÁLNÍ VÝBĚR PARAMETRŮ ODHADU

JAROSLAV MICHÁLEK

Katedra aplikované matematiky PřF MU Brno

1. Úvod

Cílem příspěvku je uvést základní vlastnosti jádrových odhadů neznámé regresní funkce. Jádrové odhady patří mezi neparametrické metody a ve srovnání s parametrickými odhady jsou zvlášť užitečné v situacích, kdy není není k dispozici dostatečná informace o tvaru parametrické regresní funkce, v situacích kdy neznámá parametrická funkce závisí ve srovnání s rozsahem dat na velkém počtu parametrů, které pak není možné dostatečně kvalitně odhadnout a v situacích kdy iterativní numerické metody odhadu parametrů nekonvergují nebo když jejich konvergence závisí na počátečních odhadech parametrů (viz [3]).

V příspěvku budou popsány základní vlastnosti jádrových odhadů a budou uvedeny některé přístupy k optimálnímu výběru parametrů odhadu (volba jádra, šířky vyhlazovacího okna). Počítačová implementace dále uvedených metod bude popsána v příspěvku V. Veselého v tomto sborníku (viz [17]) a tam budou také uvedeny konkrétní příklady jádrových odhadů na simulovaných a reálných datech.

2. Jádrové odhady regresní funkce pro model s náhodným plánem

Jádrový odhad regresní funkce je přirozeným zobecněním jádrového odhadu hustoty. Motivace zavedení jádrových odhadů hustoty včetně jejich vlastností je přehledně popsána v práci J. Antocha [1], proto zde připomeneme jenom základní pojmy a označení.

Nechť $X_1, \ldots, X_n$ je náhodný výběr rozsahu n z absolutně spojitého rozdělení o hustotě $f_1(x)$, $K(x)$, $x \in \mathbf{R}$ sudá ohraničená funkce,

$$\int_{-\infty}^{\infty} K(x)dx = 1$$

a $\{h_n\}_{n=1}^{\infty}$ posloupnost kladných čísel taková, že $\lim_{n\to\infty} h_n = 0$. Položme pro $x \in \mathbf{R}$

$$K_{h_n}(x) = \frac{1}{h_n} K(\frac{x}{h_n}) \tag{1}$$

a

$$\hat{f}_{h_n}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{h_n}(x - X_i). \tag{2}$$

Pak funkci $\hat{f}_{h_n}(x)$ nazýváme jádrovým odhadem hustoty $f_1(x)$, funkci $K(x)$ jádrem a h_n šířkou vyhlazovacího okna.

Typickými příklady jader jsou následující funkce (v uvedených vzorcích je $I_{\langle -1,1\rangle}(x)$ indikátor intervalu $\langle -1,1\rangle$).

$K(x) = \frac{1}{2} I_{\langle -1,1\rangle}(x)$	obdélníkové jádro
$K(x) = (2\pi)^{-1/2} e^{-x^2/2}$	gausovské jádro
$K(x) = \frac{3}{4}(1-x^2).I_{\langle -1,1\rangle}(x)$	Epanenčnikovo jádro
$K(x) = (1-\lvert x\rvert) I_{\langle -1,1\rangle}(x)$	trojúhelníkové jádro
$K(x) = \frac{15}{16}(1-x^2)^2 I_{\langle -1,1\rangle}(x)$	kvartické jádro.

Uvažujme dále dvourozměrný náhodný vektor (X, Y) z absolutně spojitého rozdělení o hustotě $f(x, y)$ a nechť střední hodnost $E|Y| < \infty$, označme $m(x) = E(Y|X = x)$ regresní funkci, $f_1(x) = \int_{-\infty}^{\infty} f(x, y)dy$ marginální hustotu X a předpokládejme, že je dán náhodný výběr $(X_1, Y_1), \ldots, (X_n, Y_n)$ z rozdělení o hustotě $f(x, y)$. Za těchto předpokladů zvolíme pevné $x \in \mathbf{R}$ a odhad $m(x)$ budeme v intervalu $\langle x - h_n, x + h_n\rangle$ hledat váženou metodou nejmenších čtverců. Budeme-li v prvním přiblížení předpokládat, že funkci m lze na intervalu $\langle x - h_n, x + h_n\rangle$, jehož délka pro $n \to \infty$ konverguje k nule, aproximovat konstantou, kterou označíme Θ, lze odhad Θ nalézt minimalizací výrazu

$$\sum_{i=1}^{n} K_{h_n}(x - X_i)(Y_i - \Theta)^2, \tag{3}$$

kde $K_h(x - X_i)$ jsou příslušné váhové koeficienty. Snadno nahlédneme, že výraz (3) nabývá minimální hodnoty v bodě

$$\hat{\Theta} = \sum_{i=1}^{n} Y_i K_{h_n}(x - X_i) / \sum_{i=1}^{n} K_{h_n}(x - X_i),$$

když $\sum_{i=1}^{n} K_{h_n}(x - X_i) \neq 0$.

Hodnotu $\hat{\Theta}$ tedy můžeme považovat za odhad $m(x)$ v bodě x, tento odhad označíme $\hat{m}_{h_n}(x)$ a pokud nebude nebezpečí nedorozumění, budeme index h_n vynechávat. Pak, užijeme-li vzorce (2), můžeme odhad $\hat{m}_{h_n}(x)$ psát ve tvaru

$$\hat{m}_{h_n}(x) = \hat{m}(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i K_{h_n}(x - X_i) / \hat{f}_{h_n}(x) = \frac{1}{n} \sum_{i=1}^{n} W_{h_n i}(x) Y_i, \tag{4}$$

kde

$$W_{h_n i}(x) = K_{h_n}(x - X_i) / \hat{f}_{h_n}(x). \tag{5}$$

Váhové koeficienty $W_{h_n i}(x)$ dané vzorcem (5) byly navrženy Nadarayou [13] a Watsonem [18], proto se odhad (4) někdy nazývá Nadarayovým-Watsonovým odhadem. Pravá část

vzorce (4) představuje obecný tvar jádrových odhadů neznámé regresní funkce, v literatuře lze nalézt různé tvary váhových funkcí $W_{h_n i}(x)$ zde se zmíníme o následujících, často používaných váhových funkcích:

1. $W^{(1)}_{h_n i}(x) = K_{h_n}(x - X_i)/f_1(x)$ pro $f_1(x) > 0$ ... viz např. Greblicki a Krzyzak [7] a Johnston [9]

2. $W^{(2)}_{h_n i}(x) = n(X_i - X_{i-1})K_{h_n}(x - X_i)$ ... viz např. Pristley and Chao [14]

3. $W^{(3)}_{h_n i}(x) = n\int_{s_i-1}^{s_i} K_{h_n}(x-u)du$, kde $X_{(i-1)} \le S_{i-1} \le X_{(i)}$ a $X_{(i)}$ je i-tá pořádková statistika ... viz např. Gasser a Müller [5].

Ve vzorcích uvedených v bodě 2 a 3 se X_0 volí definitoricky s ohledem na interval možných hodnot náhodné veličiny X.

Popsaná experimentální situace se nazývá regresní model s náhodným plánem, protože pozorování nezávisle proměnné X v regresní funkci jsou náhodné veličiny. Konzistence odhadu (4) s vahami (5) je dána následující větou.

Věta 1. *Nechť je dán regresní model s náhodným plánem a nechť platí*

1. $EY^2 < \infty$
2. $\int_{-\infty}^{\infty} |K(u)|du < \infty$, $\lim_{|u|\to\infty} uK(u) = 0$
3. $\lim_{n\to\infty} h_n = 0$ a $\lim_{n\to\infty} nh_n = \infty$.
4. *Bod* $x \in \mathbf{R}$ *je bodem spojitosti funkcí* $f_1(x)$, $m(x)$, $\sigma^2(x) = D(Y|X = x)$ a $f_1(x) > 0$.

Pak $\hat{m}_{h_n}(x) \xrightarrow{P} m(x)$ *přičemž* $\xrightarrow{P}$ *značí konvergenci podle pravděpodobnosti pro* $n \to \infty$.

Důkaz viz [8].

3. Jádrové odhady regresní funkce pro model s pevným plánem

V mnoha experimentálních situacích jsou hodnoty nezávisle proměnné X voleny experimentátorem (častá je např. ekvidistantní volba). Potom mluvíme o regresním modelu s pevným plánem. Zapíšeme jej ve tvaru

$$Y_i = m(x_i) + e_i, \qquad i = 1, \ldots, n$$

kde m je neznámá odhadovaná regresní funkce, e_i jsou nezávislé náhodné veličiny, $Ee_i = 0$, $De_i = \sigma^2$, $i = 1, \ldots, n$ a $x_1, \ldots, x_n$ jsou dané nenáhodné hodnoty nezávisle proměnné a tvoří plán experimentu. Y_i je i-té měření $m(x_i)$ zatížené chybou e_i, $i = 1, \ldots, n$.

Hodnoty x_i lze zadat nezápornou integrovatelnou funkcí $f_1(x)$ s vlastností $\int_{-\infty}^{\infty} f_1(x)dx = 1$ předpisem

$$\int_{-\infty}^{\infty} f_1(x)dx = \frac{i-1}{n-1}, \quad i = 1, \ldots, n.$$

Pak funkci $f_1(x)$ nazýváme hustotou generující plán experimentu $x_1, \ldots, x_n$. Skutečnost, že je označena stejně jako marginální hustota v modelu s náhodným plánem nebude působit těžkosti, protože z kontextu bude vždy jasné o jaký model se jedná a role funkce f_1 je v modelu s náhodným plánem a v modelu s pevným plánem podobná.

Dále budeme pro jednoduchost předpokládat, že $0 \le x_1 \le x_2 \le \cdots \le x_n \le 1$ a jádrovým odhadem funkce $m(x)$ budeme analogicky jako v modelech s náhodným plánem

rozumět statistiku

$$\hat{m}(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i W_{Ki}(x), \tag{6}$$

kde funkce $W_{Ki}(x)$ jsou vlastně váhové koeficienty závisející na jádře K. V této práci za ně budeme volit některou z váhových funkcí $W^{(j)}_{h_n i}(x)$, $j = 1, 2, 3$ uvedenou v předchozím modelu s náhodným plánem, přičemž ve vzorcích pro jejich vyjádření místo X_i píšeme x_i.

Podobně jako v případě modelu s náhodným plánem jsou jádrové odhady (6) za určitých předpokladů konzistentním odhadem regresní funkce m. Dříve než uvedeme příslušnou větu, zavedeme toto označení:

$\mathrm{Lip}(\langle a, b\rangle)$ značí třídu lipschitzovsky spojitých funkcí na intervalu $\langle a, b\rangle$, které jsou rovny 0 vně intervalu $\langle a, b\rangle$

$C^k(\langle a, b\rangle)$ značí třídu k-krát spojitě diferencovatelných funkcí na intervalu $\langle a, b\rangle$

$m^{(k)}$ značí k-tou derivaci funkce m

$$\mathcal{M}_k = \{K \in \mathrm{Lip}(\langle -1, 1\rangle : \int_{-1}^{1} x^j K(x)dx = \begin{cases} 0 & j = 1, 2, \ldots, k-1 \\ 1 & j = 0 \\ \beta_k \neq 0, & j = k \end{cases}\}$$

Je-li K jádro, $K \in \mathcal{M}_k$, budeme říkat, že jádro je řádu k, přičemž k je celé číslo, $k \geq 0$.

Věta 2. *Nechť je dán regresní model s pevným plánem a předpokládejme, že platí:*

1. $f_1 \in Lip(\langle a, b\rangle)$
2. $m \in C^k(\langle a, b\rangle)$, $k > 0$
3. $K \in \mathcal{M}_k$
4. $\lim_{n\to\infty} h_n = 0$ a $\lim_{n\to\infty} nh_n = \infty$.

Položme $s_0 = 0$, $s_i = (x_{i+1} + x_i)/2$ *pro* $i = 1, \ldots, n-1$, $s_n = 1$. *Pak pro každé* $x \in (0, 1)$ *je*

$$\hat{m}(x) = \frac{1}{n}\sum_{i=1}^{n} W^{(3)}_{h_n i}(x) Y_i = \frac{1}{h_n}\sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{h_n}\right) du \tag{7}$$

konsistentním odhadem $m(x)$.

Dále pro rozptyl tohoto odhadu platí

$$D\hat{m}_{h_n}(x) = \frac{\sigma^2}{nh_n f_1(x)}(C_K + o(1)), \tag{8}$$

$$kde \quad C_K = \int_{-1}^{1} K^2(x)dx \tag{9}$$

a pro jeho vychýlení platí

$$B(x) = E\hat{m}_{h_n}(x) - m(x) = h_n^k m^{(k)}(x) B_K + o(1), \tag{10}$$

kde

$$B_K = (-1)^k \beta_k / k!. \tag{11}$$

Navíc, jestliže $\lim_{n\to\infty} nh^{2k+1} = d^2 > 0$ *pro nějaké* $d > 0$, *pak statistika* $\sqrt{nh_n}(\hat{m}_{h_n}(x) - m(x))$ *má asymptoticky normální rozdělení*

$$N(dm^{(k)}(x)B_K, \sigma^2 C_K).$$

Důkaz viz [12].

Lze ukázat, že tvrzení analogické (8) a (10) platí za určitých specifických předpokladů i pro model s náhodným plánem, když f_1 nahradíme marginální hodnotou a parametr $\sigma^2/2$ podmíněným rozptylem $D(Y|X = x)$ (viz [12]).

4. Jádrové odhady a vážené lokální regresní odhady

V odstavci 2 byl jádrový odhad zaveden jako vážný lokální odhad regresní funkce, která byla ve vyhlazovacím okně $(x - h_n, x + h_n)$ aproximována konstantou, přičemž váhové koeficienty byly zadány pomocí dané jádrové funkce. V tomto odstavci ukážeme že lokální regresní odhady jsou asymptoticky ekvivalentní s jádrovými odhady.

Předpokládejme, že je dán regresní model s pevným plánem $Y_i = m(x_i) + e_i$, $i = 1, \ldots, n$, uvažujme pro dané $x \in \mathrm{R}$ vyhlazovací okno $(x - h_n, x + h_n) \subset \langle 0, 1)$ a za předpokladu, že regresní funkce m je v okolí bodu x dostatečně hladká, můžeme ji ve zvoleném vyhlazovacím okně aproximovat polynomem $\varrho(u) = \sum_{j=1}^{k} \Theta_j (u - x)^{j-1}$ stupně $k - 1$. Označíme-li $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ vektor pozorování závislé proměnné Y, $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_k)'$ vektor neznámých parametrů, $\mathbf{Q} = (q_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,k}}$ matici plánu experimentu odpovídající zvolenému vyhlazovacímu oknu, tedy $q_{ij} = (x_i - x)^{j-1}$, pak vážený odhad vektoru $\boldsymbol{\Theta}$ lze ve zvoleném vyhlazovacím okně získat váženou metodou nejmenších čtverců minimalizaci výrazu

$$(\mathbf{Y} - \mathbf{Q}\boldsymbol{\Theta})' \mathbf{C}^{-1} (\mathbf{Y} - \mathbf{Q}\boldsymbol{\Theta}), \tag{12}$$

kde $\mathbf{C}$ je vhodná symetrická a pozitivně definitní váhová matice. Označíme-li $\tilde{\boldsymbol{\Theta}}$ odhad $\boldsymbol{\Theta}$ získaný minimalizací (12), pak vážený lokální odhad $\tilde{m}(x)$ regresní funkce $m(x)$ lze psát ve tvaru

$$\tilde{m}(x) = \tilde{\varrho}(x) = \tilde{\Theta}_1, \text{ kde } \tilde{\Theta}_1 \text{ je první souřadnice vektoru } \tilde{\boldsymbol{\Theta}}.$$

Cleveland v [2] doporučuje volit matici $\mathbf{C}$ diagonální s diagonálními prvky $c_{ii} = [G(\frac{x_i - x}{h_n})]^{-1}$, kde $G \in \mathrm{Lip}((-1, 1))$ je vhodná váhová funkce, kladná na intervalu $(-1, 1)$. Uvedený odhad $\tilde{m}(x)$ je potom poměrně robustní, jeho vlastnosti jsou demonstrovány na simulovaných i reálných datech v [10], kde je popsána i jeho další modifikace pro případ modelu s nehomogenním rozptylem.

Zapíšeme-li odhad $\tilde{\boldsymbol{\Theta}}$ v jeho analytickém tvaru, dostaneme

$$\tilde{\boldsymbol{\Theta}} = (\mathbf{Q}'\mathbf{C}^{-1}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{C}^{-1}\mathbf{Y} \tag{13}$$

a odtud

$$\tilde{m}(x) = \tilde{\Theta}_1 = \frac{1}{n} \sum_{i=1}^{n} W_{G,i}(x) Y_i, \tag{14}$$

kde $W_{G_i}(x)$, $i = 1,\ldots,n$ jsou vhodné váhové koeficienty (prvky prvního řádku matice $(\mathbf{Q}'\mathbf{C}^{-1}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{C}^{-1}$, jak plyne z vyjádření (13)) a prostřednictvím $\mathbf{C}$ závisí na váhové funkci G.

Výraz (14) formálně odpovídá zápisu jádrového odhadu

$$\hat{m}(x) = \frac{1}{n}\sum_{i=1}^{n} W_{Ki}(x)Y_i, \tag{15}$$

kde $W_{Ki}(x)$ jsou váhové koeficienty typu $W^{(3)}_{h_n i}(x)$ odpovídající jádru K. Asymptotické ekvivalence odhadů (14) a (15) je popsána v následující větě.

Věta 3. *Nechť (14) je lokální vážený odhad regresní funkce m v modelu s pevným plánem získaný lokálním odhadem polynomu stupně $k-1$ a nechť $G \in Lip((-1,1))$ je váhová funkce, kladná na intervalu $\langle -1,1\rangle$.*

Pak existuje odpovídající jádrový odhad (15) s jádrem K_G tak, že

$$\lim_{n\to\infty} \sup_{1\le i\le n} \left|\frac{W_{Gi}}{W_{Ki}} - 1\right| = 0$$

(kde klademe $\frac{0}{0} = 1$), přičemž K_G je jediné jádro řádu k, které minimalizuje funkcionál

$$\int_{-1}^{1} \frac{K^2(x)}{G(x)}dx \tag{16}$$

pro $K_G \in \mathcal{M}_k$.

Důkaz viz [11].

Z uvedené věty plyne, že vážené lokální regresní odhady mohou být považovány za speciální jádrové odhady, kde řád jádra je o jedničku zvětšený stupeň polynomu, který byl lokálně daty proložen. Tvar jádra řádu k odpovídá tvaru váhové funkce G v tom smyslu, že minimalizuje (16).

5. Optimální volba šířky vyhlazovacího okna h_n

Kvalitu odhadu $\hat{m}(x)$ v bodě x lze lokálně popsat jako střední kvadratickou chybu (MSE = mean squared error), kterou označíme $d_M(x,h_n) = E(\hat{m}(x) - m(x))^2$. Po jednoduché úpravě, lze MSE zapsat pomocí rozptylu odhadu $D\hat{m}(x)$ a vychýlení odhadu $B(x)$ ve tvaru

$$d_M(x,h_n) = D\hat{m}(x) + B^2(x). \tag{17}$$

Dosadíme-li do (17) za rozptyl $D\hat{m}(x)$ a vychýlení $B(x)$ z (18) a (10) dostaneme pro model s pevným plánem za předpokladů Věty 2 vyjádření

$$d_M(x,h_n) = \sigma^2(C_K + o(1))/nh_n f_1(x) + (h_n^k m^{(k)}(x)B_K + o(1))^2. \tag{18}$$

Zároveň je ze vzorců (8) a (10) vidět, že s rostoucím h_n klesá rozptyl odhadu a roste kvadrát jeho vychýlení a naopak. Je proto optimální volba h_n spojena s hledáním vhodného kompromisu mezi rozptylem odhadu a jeho vychýlením. Jednoduchým kritériem pro

nalezení optimální šířky vyhlazovacího okna h_n může být minimalizace MSE (17). Lze se snadno přesvědčit, že tato minimalizace vede na řešení rovnice $\frac{\partial d_M(x,h)}{\partial h} = 0$, kde místo h_n píšeme pouze h. Označíme-li h^* jediné řešení této rovnice, pak snadno nahlédneme, že optimální volba h je

$$h^* = \left[\frac{1}{2k}\frac{\sigma^2 C_K}{f_1(x)(B_K m^{(k)}(x))^2}\cdot\frac{1}{n}\right]^{1/(2k+1)} \tag{19}$$

a této volbě h odpovídá MSE

$$d_M(x,h^*) = n^{-\frac{2k}{2k+1}}\left[c(k)(m^{(k)}(x)B_K)^{\frac{2}{2k+1}}(C_K\sigma^2/f_1(x))^{\frac{2k}{2k+1}} + o(1)\right], \tag{20}$$

kde $c(k) = (2k)^{-\frac{2k}{2k+1}} + (2k)^{\frac{1}{2k+1}}$.

Pro $k = 2$ dostaneme ze (19) a (20)

$$h^* = a_1(k, m(x), \sigma^2, f_1(x), K).n^{-1/5}$$
$$d_M(x,h) = a_2(k, m(x), \sigma^2, f_1(x), K)n^{-4/5},$$

kde a_1, a_2 jsou vhodné funkce uvedených parametrů nezávislé na n. Vzhledem k tomu, že závisí na odhadovaných neznámých parametrech $\sigma^2, m(x)$, nelze vzorce (19) v praktické situaci bezprostředně využít. Odhadem optimální šířky vyhlazovacího okna h^* se budeme zabývat později.

Jak bylo řečeno výše, optimální volba h daná vzorcem (19) získaná minimalizací MSE je lokální a je otázka, zda přechodem ke globálnímu popisu chyby odhadu se celá situace nezjednoduší.

Označme

$$d_I(h) = \int_0^1 d_M(x,h)dx \tag{21}$$

integrální střední kvadratickou chybu (IMSE = integrated MSE) odhadu. Pak s ohledem na vzorce (18) a (21) snadno zjistíme, že optimální globální volba h, která minimalizuje IMSE (21) je tvaru

$$h^{**} = \left[\frac{1}{2k}\frac{\sigma^2 C_K}{B_K^2}\left(\int_0^1 f^{-1}(x)dx \Big/ \int_0^1 m^{(k)}(x)dx\right)\frac{1}{n}\right]^{1/(2k+1)} \tag{22}$$

a tedy z praktického hlediska přechod od lokální volby optimální šířky okna ke globální volbě nepřináší žádná zjednodušení.

6. Optimální volba jádra

Z předchozího odstavce plyne, že rozptyl a vychýlení jádrového odhadu, MSE a IMSE odhadu závisí na jádře prostřednictvím konstant C_K a B_K, které jsou dány vzorci (9) a (11). Proto optimální volba jádrové funkce K bude spojena s vyšetřováním chování funkcionálů C_K a B_K v závislosti na jádře K. V tomto odstavci budeme předpokládat, že jádro $K \in \mathcal{M}_k$ a řád jádra K je $k > 0$, k sudé.

Nejdříve se zabývejme otázkou nalezení jádra K, které minimalizuje rozptyl odhadu $D\hat{m}_{h_n}(x)$ v regresním modelu s pevným plánem. Ze vzorce (8) ihned plyne, že jádro minimalizující rozptyl je jádro, které minimalizuje funkcionál C_K daný vzorcem (8) pro $K \in \mathcal{M}_k$.

Řešení tohoto minimalizačního problému se odvodí pomocí Legendrových polynomů a je dáno následující větou.

Věta 4. *Uvažujme regresní model s pevným plánem. Potom jádro $K \in \mathcal{M}_k$, které minimalizuje rozptyl odhadu $D\hat{m}(x)$ je polynom stupně $k-2$ s $k-2$ různými kořeny v $(-1,1)$ daný vzorcem*

$$K(x) = \sum_{i=0}^{k-2} \lambda_i x^i I_{(-1,1)}(x),$$

kde $\lambda_i = 0$ pro i liché a

$$\lambda_i = (-1)^{i/2} k!(k+i)!k(k-i)/i!(i+1).2^{2k+1}(\frac{k}{2}!)^2 \frac{k-i}{2}!\frac{k+i}{2}! \text{ pro } i \text{ sudé}.$$

Důkaz viz [6].

Pro $k = 2$ je optimální jádro minimalizující rozptyl obdélníkové jádro z odstavce 2. Toto jádro má v bodech -1 a 1 nespojitosti, které obecně mají vliv na špatnou kvalitu odhadu pro konečné výběry.

Další přístup k optimální volbě jádra je nalezení takového jádra $K \in \mathcal{M}_k$, které by minimalizovalo MSE odhadu. Vzhledem k tomu, že MSE je funkcí h, budeme hledat jádro K, které minimalizuje $d_M(x, h^*)$, tedy minimální MSE při optimální volbě h. Ze vzorce (20) je zřejmé, že takové jádro získáme minimalizací funkcionálu

$$T(K) = C_K^k|\beta_k| = \left(\int_{-1}^{1} K^2(x)dx\right)^k \left|\int_{-1}^{1} x^k K(x)dx\right| \tag{23}$$

pro $K \in \mathcal{M}_k$. Výsledek této optimalizace je popsán následující větou.

Věta 5. *Optimální jádro K, které minimalizuje funkcionál $T(K)$, $K \in \mathcal{M}_k$ daný vzorcem (23) a mající nejvýše $k-2$ změn znamení, které vyžadují momentové podmínky kladené na jádro, je polynom*

$$K(x) = \sum_{i=0}^{k} \lambda_i x^i I_{(-1,1)}(x), \tag{24}$$

kde

$\lambda_i = 0$ *pro i liché*

$\lambda_i = (-1)^{i/2} k(k+2)!(k+i)!(k+2-i)/i!(i+1)2^{2k+3}\frac{k}{2}!\frac{k+2-i}{2}!\frac{k+i}{2}!\frac{k+2}{2}!$ *pro i sudé.*

Důkaz viz [11].

Pro $k = 2$ dostaneme z věty 5, že optimální jádro je Epanečnikovo (viz odstavec 2) a pro $k = 4$ dostaneme z (24) optimální jádro ve tvaru

$$K(x) = \frac{15}{32}(7x^4 - 10x^2 + 3).$$

Vliv jádra na jádrový odhad $\hat{m}(x)$ v regresním modelu s pevným plánem demonstruje následující tabulka pro $k = 2$. V této tabulce $\varrho(K, K_{\text{opt}}) = T(K)/T(K_{\text{opt}})$ a K_{opt} je Epanečnikovo jádro.

Jádro	Epanečnikovo	kvartické	trojúhelníkové	gausovské	obdélníkové
$\varrho(K, K_{\text{opt}})$	1	1,005	1,011	1,041	1,060

Vzhledem k tomu, že vliv jádra na MSE odhadu závisí pouze na velikosti funkcionálu $T(K)$, plyne z uvedené tabulky, že tento vliv není nikterak podstatný.

Závěr tohoto odstavce věnujeme optimální volbě jádra, která minimalizuje IMSE odhadu. Ze vzorců (20) a (21) je vidět, že MSE závisí při optimální volbě šířky okna na funkcionálu $T(K)$ stejným způsobem jak IMSE nebo vyjádřeno jinak, podíl MSE/IMSE na funkcionálu $T(K)$ nezávisí. Proto optimální volba jádra (24), která minimalizuje MSE odhadu zároveň minimalizuje také IMSE odhadu a naopak.

7. Hraniční efekty

Kvalita jádrových odhadů regresní funkce $m(x)$ při konečném rozsahu výběru n a při dané šířce vyhlazovacího okna h_n a pro pevný plán $x_1, \ldots, x_n$ takový, že $0 \le x_1 \le \cdots \le x_n \le 1$, je v bodech $x \in \langle 0, h_n) \cup (1 - h_n, 1\rangle$ ovlivněna skutečností, že vyhlazovací okno $(x - h_n, x + h_n\rangle$ není podmnožinou intervalu $\langle 0, 1\rangle$ možných hodnot nezávislé proměnné. Podobná situace nastává i pro jádrové odhady regresní funkce v modelu s náhodným plánem. Mluvíme potom o vlivu hraničních efektů na sledovaný odhad. Řešení tohoto problému bylo věnováno mnoho úsilí (viz např. [12], [8], [16] apod.). Müller v [12] např. ukazuje, že za určitých obecných předpokladů platí pro vychýlení odhadu $B(x)$, který je založen na jádře řádu k, asymptotické vzorce

$$\int_0^1 B^2(x)dx = O(h_n^3) \text{ a } \int_h^{1-h} B^2(x)dx = O(h_n^{2k}), \text{ kde } O \text{ značí funkci "velké } O\text{"}.$$

Tedy už pro řád jádra $k = 2$ je řád vychýlení odhadu určen hraničním efektem.

Eliminace hraničních efektů je poměrně komplikovaný problém, zde se zmíníme o způsobu doporučeném v [16]. Budeme uvažovat regresní model s pevným ekvidistantním plánem, $x_i = \frac{i}{n}$, $i = 0, 1, \ldots, n$ a jádrové odhady $\hat{m}(x)$ s váhovou funkcí $W_{h_n i}^{(2)}(x)$. Tedy ze (4) dostaneme odhad

$$\hat{m}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{h_n}(x - x_i) Y_i, \tag{25}$$

který má střední hodnotu

$$E\hat{m}(x) = \int_{(x-1)/h}^{x/h} K(n) m(x - uh_n)du + O((nh_n)^{-1}), \tag{26}$$

za předpokladu, že $K \in \text{Lip}((-1, 1))$ a $m \in \text{Lip}(\langle 0, 1\rangle)$.

Položme $\varrho = \varrho(x) = x/h_n$ a je-li $\varrho < 1$ dochází (k levému) a je-li $\varrho - \frac{1}{h} \ge -1$ (k pravému) hraničnímu efektu. Po dosazení ϱ za x/h_n do (26) a po rozvinutí $E\hat{m}(x) = E\hat{m}(\varrho h_n)$ do Taylorovy řady podle h_n, lze modifikovat jádro K tak, aby v tomto rozvoji vymizely členy,

které mají výrazný vliv na vychýlení odhadu (které způsobují hraniční efekty). Výsledkem tohoto postupu (viz [16]) je jádro

$$K_\varrho^J(x) = (1-R)K(x)/\omega_K(0,\varrho) + (R/\alpha)K(x/\alpha)/\omega_K(0,\varrho/\alpha),$$

kde

$$\omega_K(j,\varrho) = \int_{-1}^{\varrho} t^j K(t)dt, \quad j = 0,1,2$$

$$R = [\omega_K(1,\varrho)/\omega_K(0,\varrho)]/[\omega_K(1,\varrho)/\omega_K(0,\varrho) - \alpha\omega_K(1,\varrho/\alpha)/\omega_K(0,\varrho/\alpha)]$$

a doporučená volba $\alpha = 2 - \varrho$.

8. Odhad šířky vyhlazovacího okna h_n

V odstavci 5 byla nalezena optimální šířka vyhlazovacího okna, která minimalizovala MSE respektive IMSE. V tomto odstavci budeme pracovat s diskrétním protějškem IMSE.

Označme

$$d_A(h) = \frac{1}{n}\sum_{i=1}^{n}(\hat{m}(x_i) - m(x_i))^2$$

průměrnou střední chybu (ASE = averaged squared error) jádrového odhadu (25) s váhovou funkcí $W_{hni}^{(2)}(x)$ v modelu s pevným ekvidistantním plánem. Protože ASE $d_A(h)$ závisí na neznámé regresní funkci m, je potřeba hodnoty $m(x_i)$ odhadnout. Zvolíme-li za odhad hodnot $m(x_i)$ přímo veličiny Y_i dostaneme odhad $d_A(h)$, který označíme $p(h)$. Tedy

$$p(h) = \frac{1}{n}\sum_{i=1}^{n}(\hat{m}(x_i) - Y_i)^2 \tag{27}$$

a je to průměrný reziduální součet čtverců (RSS) používaný pro odhad $\hat{m}(x_i)$. Po dosazení $Y_i = m(x_i) + e_i$ do výrazu pro RSS, dostaneme po jednoduché úpravě

$$p(h) = \frac{1}{n}\sum_{j=1}^{n} e_j^2 + d_A(h) + C_{1n}(h_n). \tag{28}$$

kde $C_{1n} = -\frac{2}{n}\sum_{j=1}^{n} e_j(\hat{m}_{h_n}(x_j) - m(x_j))$.

Dále snadno nahlédneme, že pro odhad (25) platí

$$EC_{1n} = -2\sigma^2 K(0)/nh_n.$$

Pak

$$Ep(h) = \sigma^2 + Ed_A(h) - 2\sigma^2 K(0)/nh_n$$

a $p(h)$ není nestranným odhadem $Ed_A(h)$. Vhodnou modifikací odhadu $Ed_A(h)$ se jeví veličina

$$R(h) = p(h) - \hat{\sigma}^2 + 2\hat{\sigma}^2 K(0)/nh_n, \tag{29}$$

kde $\hat{\sigma}^2$ je vhodný odhad rozptylu σ^2, v [15] je doporučená volba

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2$$

nebo

$$\hat{\sigma}^2 = \frac{1}{3(n-2)} \sum_{i=2}^{n-1} (Y_i - \frac{1}{2}(Y_{i-1} + Y_{i+1}))^2.$$

Odhad šířky vyhlazovacího okna h_n pak zvolíme tak, aby kriteriální funkce $R(h)$ daná vzorcem (29) byla co nejblíže nule (v [17] je na toto kritérium odkaz jako $K1$ kritérium optimality).

Odhadu šířky okna h_n je v literatuře věnována velká pozornost (viz např. [8], [12]). Základní přístupy vycházejí z nalezení vhodného odhadu ASE. Sem patří tzv. "cross-validation" přístup, jehož idea je v lineární regresi a je asymptoticky ekvivalentní uvedenému postupu a pak tzv. metoda penalizačních funkcí, která odhad $p(h)$ daný vzorcem (27) modifikuje násobením jednotlivých sčítanců pravé strany (27) vhodnou váhovou funkcí s cílem, asymptoticky vynulovat člen typu $C_{1n}(h)$ ve výrazu (28).

Uvedené postupy odhadu šířky vyhlazovacího okna h_n jsou založeny na odhadu ASE, jejíž střední hodnota je diskrétním protějškem IMSE a jde tedy o přístup globální. Lokální přístup by vycházel z MSE a vedl by k odhadům s proměnlivou šířkou vyhlazovacího okna závislou na nezávislé proměnné x. Odhady získané lokálním přístupem k volbě vyhlazovacího okna jsou pak velmi adaptivní. Zde se jimi už nebudeme zabývat, viz např. [4].

Literatura

[1] Antoch, J.: (1982), *Odhady hustoty*, Sborník JČSMF ROBUST 82, s. 1-9.

[2] Cleveland, W.S.: (1979), *Robust locally weighted regression and smoothing scatterplots*, JASA 74, p. 829-836.

[3] Deuflhard, P. and Apostolescu, V.: (1980), *A study of the Gauss-Newton method for the solution of nonlinear least squares problems*, In: Special Topics of Applied Mathematics, Ed. Frehse, Pallaschke, Trottenberg, North Holland, Amsterdam, p. 129-150.

[4] Fan, J. and Gijbels, I.: (1992), *Variable bandwidth and local linear regression smoothers*, Ann. Statist. 20, p. 2008-2036.

[5] Gasser, T. and Müller. H.G.: (1979), *Kernel estimation of regression functions*, In: Smoothing Techniques for Curve Estimation, eds. Gasser and Rosenblatt, Heidelberg: Springer-Verlag.

[6] Gasser, T., Müller, H.G. and Mammitzsch, V.: (1985), *Kernels for nonparametric curve estimation*, J. Roy. Statist. Soc. B 47, p. 238-252.

[7] Greblicki, W. and Krzyzak, A.: (1980), *Asymptotic properties of kernel estimates of a regression function*, Journal of Statistical Planning and Inference, 4, p. 81-90.

[8] Härdle, W.: (1990), *Applied nonparametric regression*, Cambridge University Press. Cambridge.

[9] Johnston, G.J.: (1979), *Smooth nonparametric regression analysis*, Institute of Statistics Mimeo series 1253, University of North Carolina, Chapel Hill, NC.

[10] Michálek, J., Budíková, M. and Brázdil, R.: (1993), *Metody odhadu trendu časové řady na příkladu středoevropských teplotních řad*, Národní klimatologický program, Praha.

[11] Müller, H.G.: (1987), *Weighted local regression and kernel methods for nonparametric curve fitting*, JASA 82, p. 231-238.

[12] Müller, H.G.: (1988), *Nonparametric Regression Analysis of Longitudinal Data*, Lecture notes in Statistics 46, Springer-Verlag.
[13] Nadaraya, E.A.: (1964), *On estimating regression*, Theory Prob. Appl. 10, p. 186-190.
[14] Pristley, M.B. and Chao, M.T.: (1972), *Nonparametric function fitting*, Series B, 34, p.385-392.
[15] Rice, J.A.: (1984), *Bandwidth choice for nonparametric kernel regression*, Ann. Statist. 12, p. 1215-1230.
[16] Rice, J.A.: (1984), *Boundary modification for kernel regression*, Communications in statistics, Serie A, 13, p. 893-900.
[17] Veselý, V.: (1994), *Jádrové odhady - implementace, ilustrativní příklady, aplikace*, Sborník ROBUST 94.
[18] Watson, G.S.: (1964), *Smooth regression analysis*, Sankhya, Series A, 26, p. 359-372.

Tato práce byla vypracována za finanční podpory GA ČR reg. číslo grantu 201/93/2408.