NEPARAMETRICKÉ ODHADY HUSTOT A REGRESNÍCH KŘIVEK

Jaromír Antoch, MFF UK, Praha

Příspěvek je věnován dvěma důležitým úlohám matematické statistiky, totiž, jak na základě dat odhadnout tvar neznámé hustoty, resp. neznámé regresní křivky. Oba tyto problémy jsou v literatuře řazeny mezi neparametrické metody, neboť nejsou obecně vázány na žádný parametrický model. Vlastním parametrem místo toho v nich je neznámá hustota či neznámá regresní křivka. Pozornost je soustředěna především na:

- *histogram*, resp. *regresogram*;
- *jádrové odhady*;
- *odhady pomocí* k_n *nejbližších sousedů*;
- *odhady pomocí ortonormálních funkcí*.

Připomeňme, že uvedené základní postupy nejsou omezeny pouze na popisované dvě úlohy. Naopak, analogicky lze postupovat při odhadu spektrální hustoty, kvantilové funkce, funkce spolehlivosti apod. Touto problematikou se zde nebudeme blíže zabývat, podobně jako nebudeme podrobně diskutovat použití jednotlivých metod pro vyhlazování dat. Zájemce o tyto aplikace najde více informací např. v [1], [11], [16] nebo [21]. Podobně se nebudeme zabývat odhady vícerozměrných hustot. Jejich formální definice jsou totiž v převážné většině případů naprosto analogické definicím uvedeným v odstavci I. a věříme, že potřebnou adaptaci si čtenář snadno v případě potřeby provede sám.

I. NEPARAMETRICKÉ ODHADY HUSTOT

U převážné většiny neparametrických odhadů hustoty se v zásadě vždy vychází z následující úvahy. Nechť $X_1,\dots,X_n$ tvoří náhodný výběr, tj. jedná se o posloupnost nezávislých, stejně rozdělených náhodných veličin s hustotou $f(x)$ a distribuční funkcí $F(x)$, $x\in R^1$. Nechť $-\infty<a<b<+\infty$ jsou pevné konstanty a označme $K_n(a,b]$ = { počet $X_i \mid a < X_i \le b,\ i = 1,\dots,n$ }. Chceme-li odhadnout $P(a < X \le b) = {}_a\int^b f(t)\,dt$, můžeme tak učinit, analogicky jako u empirické distribuční funkce, pomocí hodnoty $n^{-1}.K_n(a,b]$. Naopak, hodnotu spojité funkce $f(x)$ na intervalu $(a,b]$ můžeme odhadnout pomocí $(b-a)^{-1}.{}_a\int^b f(t)\,dt$. Spojením těchto dvou kroků dostáváme

$$(1)\quad f(x) \approx \frac{1}{b-a}\int_a^b f(t)\,dt \approx \frac{K_n(a,b]}{n(b-a)}, \qquad x \in (a,b],$$

kde $\approx$ znamená *přibližnou rovnost.*

Při použití postupu vedoucího k aproximaci typu (1) musíme mít vždy na paměti, že odhad $P(a < X \le b)$ je tím přesnější, čím je $K_n(a,b]$ větší, zatímco odhad $f(x)$ je tím přesnější, čím kratší je interval $(a,b]$. Tyto dva požadavky si bohužel protiřečí, takže při odhadu hustoty je třeba vždy najít optimální kompromis. V našem dalším výkladu se pro jednoduchost omezíme na odhad jednorozměrných hustot. Jak již bylo řečeno, při odhadu vícerozměrných hustot lze postupovat naprosto analogicky. Zájemce o hlubší seznámení se s problematikou odkazujeme na práce [4], [11], [12] a citace v nich uvedené.

I.1. HISTOGRAM

Nejjednodušším a v praxi patrně nejpoužívanějším neparametrickým odhadem hustoty je *histogram.* Nechť $X_1,\ldots,X_n$ je náhodný výběr řídící se hustotou $f(x)$. Nechť $D = \{ t_i \mid i = 0,1,\ldots,m, -\infty = t_0<t_1<\ldots<t_{m-1}<t_m = \infty \}$ je některé dělení R^1 takové, že $d_{n,m}=t_{i+1}-t_i$, $i = 1,\ldots,m-2$, kde $d_{n,m}$ je pevná konstanta. To znamená, že jednotlivé podintervaly dělení D, s výjimkou krajních, jsou ekvidistantní. *Histogram* je definován vztahem

$$(2)\quad \hat{f}_n^{(1)}(x) = \frac{1}{nd_{n,m}} \sum_{j=1}^{m} \sum_{i=1}^{n} I[X_i \in (t_{j-1},t_j]].I[x \in (t_{j-1},t_j]], \qquad x \in R^1,$$

kde $I[.]$ označuje funkci indikátor. To znamená, že při konstrukci histogramu rozložíme R^1 na ekvidistantní intervaly, s výjimkou krajních, a na každém z nich neznámou hustotu odhadneme konstantou rovnou podílu počtu pozorování jež do daného intervalu padnou k hodnotě $nd_{n,m}$.

Základní nevýhodou histogramu je, že poskytuje značně hrubý odhad neznámé hustoty. Na druhé straně je výpočetně velmi jednoduchý, zvláště máme-li hodnoty X_i uspořádané, a poskytuje alespoň základní představu o typu rozdělení dat, jejich šikmosti, špičatosti apod. Za dosti obecných podmínek lze dokázat konzistenci odhadu $\hat{f}_n^{(1)}(x)$ a jeho asymptotickou normalitu. Více informací zájemce nalezne např. v [11].

I.2. JÁDROVÉ ODHADY HUSTOTY

Nejlépe prostudovaným typem odhadů neznámé hustoty na základě náhodného výběru jsou tak zvané *jádrové odhady*. První odhady tohoto druhu byly původně navrženy již počátkem 50.let pro odhad spektrální hustoty a od té doby byly postupně zdokonalovány. Základní myšlenky si později našly řadu analogických použití i v jiných oblastech matematické statistiky. Výchozím stavebním kamenem pro ně je pojem *jádra*, jímž rozumíme libovolnou funkci $K:(R^1,B^1) \to [0,+\infty)$, jež je symetrická, ohraničená a pro níž

$$(3) \qquad \int_{-\infty}^{\infty} K(x)\,dx = 1 \qquad \text{a} \qquad \lim_{x\to\infty} x.K(x) = 0.$$

Tabulka 1. Přehled nejdůležitějších jader.

Název jádra	$K(x)$	
1. Epanechnikovo	$\frac{3}{4\sqrt{5}} - \frac{3x^2}{20\sqrt{5}}$	$\lvert x\rvert \le \sqrt{5}$
	0	$\lvert x\rvert > \sqrt{5}$
2. Kosinové	$(1/2).\cos x$	$\lvert x\rvert \le \pi/2$
	0	$\lvert x\rvert > \pi/2$
3. Trojúhelníkové	$1-\lvert x\rvert$	$\lvert x\rvert \le 1$
	0	$\lvert x\rvert > 1$
4. Klouzavé okénko	$1/2$	$\lvert x\rvert \le 1$
	0	$\lvert x\rvert > 1$
5. Normální	$[1/\sqrt{(2\pi)}].\exp(-x^2/2)$	$x \in R^1$
6. Laplaceovo	$(1/2).\exp(-\lvert x\rvert)$	$x \in R^1$
7. Cauchyho	$[\pi.(1+x^2)]^{-1}$	$x \in R^1$

Nechť $X_1,\ldots,X_n$ je náhodný výběr řídící se hustotou $f(x)$. Nechť $(h_n,\ n = 1,\ldots)$ je posloupnost kladných čísel taková, že $h_n \to 0$ pro $n \to \infty$ a $K(x)$ je některé jádro. *Jádrový odhad hustoty* $f(x)$ je definován vztahem

$$(4) \qquad \hat{f}_n^{(2)}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left[\frac{x-X_i}{h_n}\right], \qquad x \in R^1.$$

To znamená, že jádrový odhad není nic jiného než vážený průměr z těch pozorování, jež padnou do některého symetrického okolí bodu v němž odhadujeme, přičemž váhy jsou určeny jádrem $K(x)$. Jak vidíme z tabulky 1. shrnující nejpoužívanější typy jader, většina z nich dle očekávání preferuje pozorování ležící blízko bodu v němž odhadujeme. Konstanty h_n v (4) hrají roli "parametru měřítka" umožňujícího pružně měnit tvar jádra. Je přitom zřejmé, že čím menší bude hodnota h_n, tím více bude odhad koncentrován na pozorování ležící blízko bodu v němž neznámou hustotu odhadujeme.

Nejčastěji se v praxi používá jádro typu 4. vedoucí na tak zvaný *odhad typu klouzavého okénka*. Tento odhad v každém bodě x v němž odhadujeme je v podstatě váženým průměrem (s týmiž vahami) z těch pozorování, jež padnou do $[x-h_n, x+h_n]$. Je přímým zobecněním histogramu a navíc eliminuje jeho základní nevýhodu spočívající v rozkladu R^1 na ekvidistantní intervaly. Často se též v literatuře doporučuje používat Epanechnikovo jádro typu 1. Máme-li však za úkol odhadnout např. hustotu rovnoměrného rozdělení či jiného rozdělení s ohraničeným nosičem J, $J = \{ x \mid f(x) > 0\}$, či majícího výrazné skoky, ukazuje se, že pro tento případ je mnohem výhodnější volit jádra typu 5.-7. Dosáhne se tím přesnějšího odhadu především na krajích intervalu J či na okolí skoků.

Otázce volby nejvhodnější konstanty h_n v (4) byla v literatuře věnována velká pozornost, neboť podstatným způsobem ovlivňuje kvalitu odhadů. Převážná většina doporučení vychází z volby vhodné míry kvality (přesnosti) odhadů, jako je například (integrální) střední čtvercová chyba, maximum vychýlení apod., a úvah o její minimalizaci. Jejich společnou nevýhodou je potřeba znalosti mnoha apriorních informací o odhadované hustotě. Tak například odhad s Epanechnikovovým okénkem minimalizuje pro $n \to \infty$ integrální střední čtvercovou chybu mezi všemi jádrovými odhady typu (4) s jádrem vyhovujícím vztahu (3). V praxi se k této úloze přistupuje často metodou pokusu a omylu, tj., odhad se vypočte pro různé hodnoty h_n a za optimální se zvolí ta hodnota, pro niž je výsledná křivka "opticky nejhladší". Jinou možností je užití *metody křížového ověřování*. Popišmě si stručně přístup vycházející z metody maximální věrohodnosti, neboť jej lze analogicky užít i pro jiné typy neparametrických odhadů hustoty. V podstatě se jedná o to nalézt tu konstantu $\hat{h}$, jež maximalizuje

$$L(h) = \sum_{i=1}^{n} \hat{f}_{ni}^{(2)}(X_i),$$

kde

$$\hat{f}_{ni}^{(2)}(x) = \frac{1}{(n-1)h} \sum_{\substack{j=1 \\ j=/=i}}^{n} K\left[\frac{x - X_j}{h}\right], \qquad x \in R^1,$$

a užít ji v (4) místo h_n. Vedle maximální věrohodnosti byla pro křížové ověřování navržena i jiná kritéria, například (integrální) střední čtvercová chyba apod., srovnej též odstavec II. 2. Analogicky lze použít křížové ověřování i při konstrukci histogramu a dalších neparametrických odhadů hustoty.

Řada autorů se zabývala otázkou nalézt podmínky, za nichž jsou jádrové odhady konzistentní, asymptoticky nestranné, normální apod. Více informací čtenář nalezne např. v [4], [11] nebo [12].

I.3. ODHADY HUSTOTY POMOCÍ k_n NEJBLIŽŠÍCH SOUSEDŮ

Nechť $X_1,\ldots,X_n$ je náhodný výběr z rozdělení s hustotou $f(x)$. Nechť $\{ k_n,\ n = 1,2,\ldots\}$ je posloupnost přirozených čísel taková, že $k_n \to \infty$ & $k_n/n \to 0$ pro $n \to \infty$. Nechť pro každé $x \in R^1$ je $J_n(x) = \{ i \mid X_i$ je některý z k_n nejbližších sousedů k x mezi $X_1,\ldots,X_n\}$. Nechť $K(x)$ je některé jádro definované vztahem (3) a $D_n = max\{ d_i \mid d_i = |x - X_i|,\ i \in J_n(x) \}$. *Odhad hustoty* $f(x)$ *pomocí* k_n *nejbližších sousedů* je definován vztahem

$$(5) \qquad \hat{f}_n^{(3)}(x) = \frac{1}{nD_n} \sum_{i=1}^{n} K\left[\frac{x - X_i}{D_n}\right] . I[\, X_i \in J_n(x)\,], \qquad x \in R^1.$$

Přes svou podobnost s jádrovými odhady hustoty se odhad $\hat{f}_n^{(3)}(x)$ od nich podstatně liší v tom, že je vždy počítán na základě pevného počtu pozorování padnoucích do proměnlivě velkého okolí bodu x v němž odhadujeme. Jádrové odhady byly naopak počítány z různého počtu pozorování jež padly do některého pevného symetrického okolí bodu x. V praxi se v převážné většině případů volí opět jádro typu 4. Po výpočetní stránce jsou odhady pomocí k_n nejbližších sousedů jednoduché, zvláště máme-li předem uspořádaná data. Co se týče spotřeby času počítače při jejich výpočtu, jsou proti jádrovým odhadům méně náročné. Simulační studie ukázaly, že z hlediska přesnosti jsou zpravidla jen o málo horší než jádrové odhady.

Asymptotické chování je podobné jako u jádrových odhadů. Více informací nalezne zájemce v [4] nebo [11], kde je též podrobně diskutována volba optimální konstanty k_n v závislosti na zvoleném typu jádra a apriorních předpokladech o modelu. V praxi k volbě k_n přistupujeme podobně jako u jádrových odhadů, tj., buď metodou pokusu a omylu nebo pomocí metody křížového ověřování.

I.4. ODHADY HUSTOTY POMOCÍ ORTONORMÁLNÍCH FUNKCÍ

Nechť $X_1,\ldots,X_n$ je náhodný výběr z rozdělení s hustotou $f(x)$. Nechť $P = \{ P_i(x),\ i = 1,2,\ldots \}$ je úplná ortonormální posloupnost funkcí v $L_2(R^1)$. Položme

$$c_i = \sum_{j=1}^{n} P_i(X_j), \qquad i = 1,2,\ldots \tag{6}$$

Odhad hustoty $f(x)$ *pomocí posloupnosti ortonormálních funkcí* je definován vztahem

$$\hat{f}_n^{(4)}(x) = \sum_{i=1}^{\infty} c_i.P_i(x), \qquad x \in R^1. \tag{7}$$

V praxi se samozřejmě užívá pouze prvních L_n členů v (7), tj. za odhad se bere

$$\hat{f}_n^{(4')}(x) = \sum_{i=1}^{L_n} c_i.P_i(x), \qquad x \in R^1. \tag{8}$$

Úloha určení optimálního počtu sčítanců je značně složitá a daleko překračuje meze našeho textu, blíže viz práce citované v [12]. Za systém P se doporučuje volit například posloupnost Legendrových či Hermitových polynomů. Celkově lze říci, že odhady tohoto druhu jsou výpočetně dosti složité, velmi náročné na spotřebu času počítače a v praxi se používají pouze zřídka.

II. NEPARAMETRICKÉ ODHADY REGRESNÍCH KŘIVEK

Mezi nejrozšířenější postupy při statistické analýze dat patří metody regresní analýzy. Především se jedná o odhady neznámých parametrů v lineárním a nelineárním modelu, predikci, stanovení intervalů spolehlivosti apod. V praxi jsme však často postaveni před problém zpracovat data, s nimiž jsme se doposud nesetkali a pro něž model, byť jen přibližný, k dispozici *a priori* nemáme. V takovém případě je výhodnější než zkoušet modely náhodně použít nejprve některý neparametrický odhad a získat pomocí něj alespoň základní představu o průběhu neznámé regresní křivky. Teprve na

základě této prvotní analýzy volíme v druhém kroku vhodnou třídu parametrických modelů a v ní se snažíme nalézt model optimální. Cílem tohoto odstavce je seznámit čtenáře s hlavními representanty neparametrických odhadů regresních křivek, jejich vlastnostmi, chováním a možnostmi uplatnění.

Nechť (X,Y) je náhodný vektor definovaný na (R^{p+1}, B^{p+1}), $p \geq 1$, kde $X=(X_1,\ldots,X_p)$. Předpokládejme , že rozdělení $P_{X,Y}$ vektoru (X,Y) je absolutně spojité vůči Lebesquově míře a označme $p(x,y)$ hustotu vektoru (X,Y), resp. $f(x)$ marginální hustotu vektoru X. Předpokládejme dále, že :

(A) existuje $U \in R^p$ taková, že pro všechna $x \in U$ je $f(x) =/= 0$;
(B) $E\ Y^2 < \infty$;
(C) $V(X) = E\ [\{Y - E\ (Y|X = x)\}^2 | X = x] < \infty$ pro všechna $x \in R^p$.

Splnění podmínek (A)-(C) je základním předpokladem potřebným pro důkazy nejdůležitějších vlastností navržených odhadů.

Naším cílem je konstruovat odhady regresní funkce

(9) $r(x) = E\ (\ Y \mid X = x\ ),\qquad x \in R^p,$

na základě nezávislých pozorování $((X_i,Y_i),\ i = 1,\ldots,n\ )$ vektoru (X,Y) v případě, kdy její skutečný tvar neznáme a víme jen, že existuje. Dále popsané metody se v literatuře opět nazývají metodami neparametrickými, neboť se nevztahují na odhad parametrů určujících tvar $r(x)$. Tento název však i tentokrát není úplně přesný, protože parametr který odhadujeme je ve skutečnosti $r(x)$. Pro jednoduchost se omezíme na případ $p = 1$, tzv. že jak X tak Y budou reálné náhodné veličiny. Uvidíme však, že zobecnění pro případ $p>1$ je zpravidla evidentní a nevyvolává v žádném z dále uvedených případů principiální problémy.

Vedle svého hlavního uplatnění nabízí neparametrický přístup k odhadu neznámé regresní křivky i některé další možnosti:
- umožňuje provádět předpověď (predikci) aniž bychom byli vázáni na určitý parametrický model;
- poskytuje kontrolu o výskytu odlehlých pozorování v datech;
- umožňuje vyhlazení dat apod.

Neparametrickými odhady regresních křivek se v uplynulých dvaceti pěti letech zabývala řada autorů. Zájemcům doporučujeme přehlednou Collombovu bibliografii [16] pokrývající široké spektrum těchto prací. Přitom je zajímavé, že doposud s vyjímkou sborníku [19] nebyla vydána jediná monografie věnovaná této oblasti.

II.1. REGRESOGRAM

Nejjednodušším neparametrickým odhadem neznámé regresní křivky je *regresogram*, který je přímou analogií histogramu. Nechť $((X_i,Y_i),\ i = 1,\ldots,n)$, jsou nezávislé kopie náhodného vektoru (X,Y). Nechť $D = \{t_i \mid i = 0,1,\ldots,m,\ -\infty = t_0<t_1<\ldots<t_{m-1}<t_m = \infty\}$ je některé dělení R^1 takové, že $d_{n,m}=t_{i+1}-t_i$, $i=1,\ldots,m-2$, kde $d_{n,m}$ je konstanta. To znamená, že jednotlivé podintervaly dělení D, s výjimkou krajních, jsou ekvidistantní. *Regresogram* je definován vztahem

$$(10)\quad \hat{r}_n^{(1)}(x) = \sum_{j=1}^{m} \left[\frac{\sum_{i=1}^{n} Y_i . I[X_i \epsilon (t_{j-1},t_j]] . I[x \epsilon (t_{j-1},t_j]]}{\sum_{i=1}^{n} I[X_i \epsilon (t_{j-1},t_j]]} \right],$$

pokud $\sum_{i=1}^{n} I[X_i \epsilon (t_{j-1},t_j]] \neq 0$,

$= 0$, jinak.

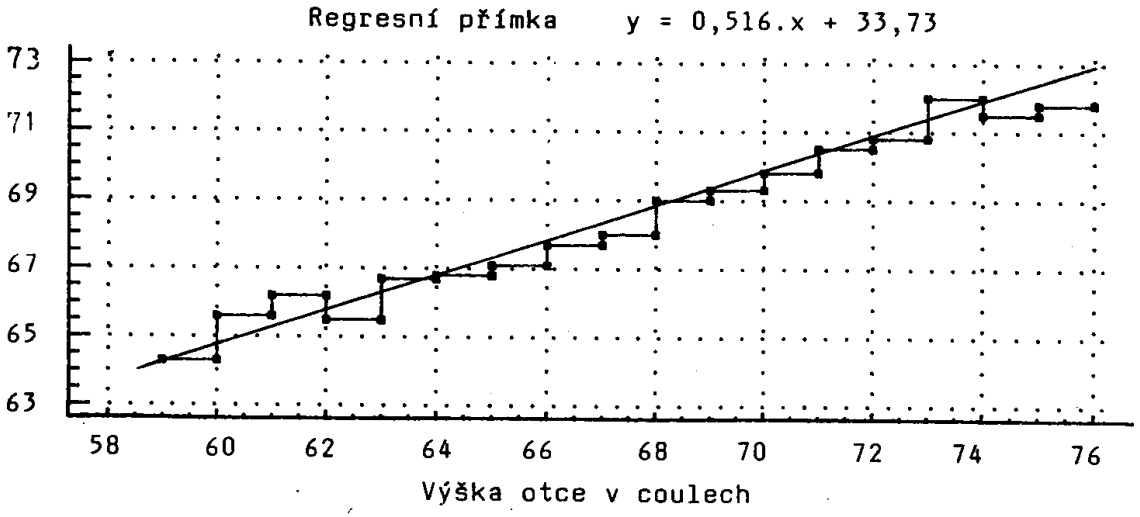
Jinými slovy, leží-li $x \epsilon (t_{j-1},t_j]$, potom za odhad $r(.)$ v bodě x užijeme aritmetický průměr z těch hodnot Y_i, pro něž odpovídající X_i leží v intervalu $(t_{j-1},t_j]$. Je zajímavé si všimnout, že odhad tohoto druhu použil již Pearson v roce 1903. Teoretické vlastnosti odhadu $\hat{r}_n^{(1)}(x)$ podrobně zkoumali například Collomb a Tukey, viz [16].

Regresogram má řadu nevýhod. Vyplývají především z toho, že je konstantní na předem daných intervalech a poskytuje pouze hrubou představu o průběhu $r(x)$. Dále lze snadno ukázat, že se jedná o odhad vychýlený, se značným rozptylem a vysoce nerobustní. Jeho nerobustnost je úzce spojena s nerobustností výběrového průměru. Přes všechny tyto nevýhody jej lze doporučit především pro prvotní ohledání dat, neboť se velice snadno a rychle počítá. Ve srovnání s jinými neparametrickými odhady je výpočetně mnohem jednodušší a je nenáročný na potřebu paměti počítače. Přitom často dává dostatečnou představu o datech. Jeho vysoká nerobustnost může být dokonce na prospěch, neboť příliš velké oscilace $\hat{r}_n^{(1)}(x)$ nás informují o možném výskytu odlehlých pozorování. Vzhledem ke všem těmto důvodům se doporučuje použití regresogramu především během prvních fází průzkumové analýzy dat.

Příklad 1.

Pearson a Lee v [23] studovali závislost výšky syna na výšce otce. Na základě více než 1000 pozorování rozdělených do tříd po jednom palci se snažil dokázat platnost obecného zákona regrese vysloveného Galtonem v následující formě " ... *Each pecularity in a man is shared by his kinsman, but on the average in a less degree.*" *Z historického hlediska je zajímavé poznamenat, že právě v této Pearsonově práci je slovo regrese poprvé použito k označení podmíněného očekávání, stejně jako je* poprvé *užit výraz regresní přímka.* Výsledky shrnuje obrázek 1. Je zajímavé si přitom všimnout, že v tomto případě regresogram přináší pro extrémní (ať již malé či velké) výšky otců více informace než regresní přímka proložená daty.

Obr.1. Závislost výšky syna na výšce otce.



Jak je zřejmé z (10), ve většině praktických případů padne do jednotlivých podintervalů rozkladu D různý počet hodnot X_i, což samozřejmě zvyšuje nepřesnost odhadu, zvláště při malých rozsazích výběru. Proto byl navržen a studován tzv. regresogram s náhodným krokem. Tato modifikace spočívá v tom, že interval do nějž hodnoty X_i padnou rozdělíme na m disjunktních podintervalů tak, aby do každého z nich padl přibližně týž počet bodů X_i, tj. cca $[m/n]$ bodů. Odhad je pak definován opět vztahem (10).

II.2. JÁDROVÉ ODHADY

Mějme k dispozici nezávislé kopie $((X_i,Y_i),\ i=1,\ldots,n)$, náhodného vektoru (X,Y). Náš cíl budiž týž jako v předchozím odstavci, tj. odhadnout neznámou regresní křivku $r(x)$ na základě těchto dat. Jak jsme viděli, základním nedostatkem regresogramu je rozklad prostoru hodnot vysvětlující proměnné X na ekvidistantní intervaly, na každém z nichž je $r(x)$ odhadnuta konstantou. Pro překonání tohoto nedostatku Nadaraja a Watson, viz [22] a [26], navrhli postupovat analogicky jako v případě jádrových odhadů hustoty a za odhad $r(x)$ zvolit vážený průměr z jednotlivých pozorování. Přesněji, použít statistiku

$$(11)\quad \hat{r}_n^{(2)}(x) = \sum_{i=1}^{n} Y_i . W_{ni}(x), \qquad x \in R^1,$$

kde

$$(12)\quad W_{ni}(x) = \frac{(nh_n)^{-1}.K[(x-X_i)/h_n]}{(nh_n)^{-1}.\Sigma_{i=1}^{n} K[(x-X_i)/h_n]}, \qquad \Sigma_{i=1}^{n} K(\ldots) \neq 0,$$

$$= 0, \qquad \Sigma_{i=1}^{n} K(\ldots) = 0.$$

Přitom $\{h_n,\ n = 1,2,\ldots\}$ je posloupnost kladných konstant taková, že $h_n \to 0$ pro $n \to \infty$ a $K(x)$ je některé jádro splňující (3). Odtud pochází i název *jádrové odhady regresní křivky*.

Prohlédneme-li si pečlivě tvar jmenovatele v (12), vidíme, že se nejedná o nic jiného než o jádrový odhad $\hat{f}_n^{(2)}(x)$ marginální hustoty $f(x)$ náhodné veličiny X. Z (11)-(12) navíc vidíme, že zatímco tvar vah je určen zvoleným jádrem $K(x)$, jejich velikost je opět *parametrizována* pomocí konstant h_n. Tato normalizace má za účel adaptovat váhy vzhledem k lokálnímu výskytu hodnot X_i na okolí bodu v němž odhadujeme.

Úloha jádra je analogická jako v případě jádrových odhadů hustoty, tj., preferovat pro odhad v bodě x ta pozorování Y_i, pro něž odpovídající X_i leží blízko x. Pro praxi se doporučuje na základě zkušeností i výsledků simulačních experimentů užívat především jádra typu 1.,4. a 5. z tabulky 1. K otázce volby optimální hodnoty h_n se vrátíme později.

Pro lepší pochopení jádrových odhadů je zajímavé si všimnout jejich následující vlastnosti. Pro pevné $x \in R^1$ si totiž vždy můžeme $\hat{r}_n^{(2)}(x)$ s kladnými váhami $W_{ni}(x)$ představit jako řešení úlohy

$$\min_{t\in R^1} \sum_{i=1}^{n} K\left[\frac{x-X_i}{h_n}\right].(Y_i - t)^2 = \sum_{i=1}^{n} K\left[\frac{x-X_i}{h_n}\right].(Y_i - \hat{r}_n^{(2)}(x))^2.$$

Jinými slovy to znamená, že $\hat{r}_n^{(2)}(x)$ minimalizuje na určitém okolí bodu x daném volbou jádra $K(x)$ a konstanty h_n vážený součet čtverců reziduí. Vzhledem k vysoké nerobustnosti metody nejmenších čtverců odtud mimo jiné vyplývá i vysoká citlivost jádrových odhadů na případná odlehlá pozorování Y_i. Všimněme si, že váhy $W_{ni}(x)$ jsou počítány pouze na základě hodnot X_i a nikterak neberou v úvahu hodnoty pozorování Y_i. Proto byly v literatuře navrženy a studovány robustní verze jádrových odhadů. V podstatě se jedná opět o odhady typu (11), kde jsou navíc přidány některé další váhy V_{ni} tak, aby omezily vliv případných odlehlých pozorování mezi těmi Y_i, pro něž odpovídající X_i leží blízko bodu v němž odhadujeme. V literatuře lze nalézt i některé další přístupy k jádrovým odhadům, například pomocí lokálního prokládání polynomy apod. Více se o nich lze dočíst například v [21].

Vraťme se nyní k otázce volby optimální hodnoty konstanty h_n v (12). I zde můžeme samozřejmě postupovat metodou pokusu a omylu, tj. spočíst odhad (11) pro několik hodnot h_n a zvolit tu hodnotu, jež nám poskytuje "opticky nejhladší" křivku. Mnohem přesnější výsledky, podložené navíc dobrými asymptotickými vlastnostmi získaných odhadů, poskytuje opět metoda křížového ověřování. Popišme si pro změnu stručně přístup založený na mimimalizaci chyby predikce. V podstatě se jedná o to nalézt tu konstantu $\hat{h}$, jež minimalizuje funkci

$$CV(h) = \sum_{i=1}^{n} [Y_i - \hat{r}_{ni}^{(2)}(X_i)]^2 W_{ni}(X_i),$$

kde

$$\hat{r}_{ni}^{(2)}(X_i) = \sum_{\substack{j=1 \\ j\neq i}}^{n} Y_j . W_{nj}(X_i) \qquad \text{a} \qquad W_{nj}(X_i) \text{ jsou dány vztahem (12).}$$

Zhruba řečeno, nejedná se vlastně o nic jiného než o nalezení toho $\hat{h}$, pro nějž je minimalizován vážený součet čtverců chyb predikce pro napozorovaná data.

Jádrové odhady tvoří patrně nejlépe prostudovanou třídu neparametrických odhadů regresních křivek a lze je doporučit pro většinu běžných aplikací. Čtenář zajímající se o podmínky regularity za nichž jsou příslušné odhady konzistentní, asymptoticky nestranné, normální apod., nalezne informace např. v [16], [19], [21], [22], [26] a pracech v nich citovaných.

II.3. ODHADY POMOCÍ k_n NEJBLIŽŠÍCH SOUSEDŮ

Oba předchozí postupy mají společnou nevýhodu v tom, že výpočet odhadu je pro různé body x v nichž odhadujeme zpravidla založen na různém počtu pozorování. Po zkušenostech z odstavce I. však je zřejmé, že i zde bude možné definovat odhady, jejichž výpočet bude založen na určitém předem pevně daném počtu těch pozorování Y_i, pro něž odpovídající hodnoty X_i leží blízko bodu v němž odhadujeme.

Nechť $\{ (X_i,Y_i),\ i = 1,...,n \}$, jsou nezávislé kopie náhodného vektoru (X,Y). Nechť pro každé $x \in R^1$ $J_n(x) = \{ i \mid X_i$ je některý z k_n nejbližších sousedů k x mezi $X_1,...,X_n \}$, kde $\{ k_n,\ n = 1,2,..\}$ je posloupnost přirozených čísel taková, že $k_n \to \infty$ a $k_n/n \to 0$ pro $n \to \infty$. *Odhad funkce $r(x)$ pomocí k_n nejbližších sousedů* je definován vztahem

$$(13)\quad \hat{r}_n^{(3)}(x) = \frac{1}{k_n} \sum_{i=1}^{n} Y_i . I[\ i \in J_n(x)\ ], \qquad x \in R^1.$$

Jak vyplývá z (13), $\hat{r}_n^{(3)}(x)$ není vlastně nic jiného než výběrový průměr z těch pozorování Y_i, pro něž X_i tvoří k_n nejbližších sousedů k bodu x v němž odhadujeme.

Zobecnění vztahu (14) jdou v podstatě dvěma směry. První z nich vychází z jádrových odhadů a snaží se preferovat v (13) ta pozorování Y_i, pro něž odpovídající X_i leží blíže bodu v němž odhadujeme. Nechť $\{ v_{nj} \mid j = 1,...,n \ \&\ n = 1,2,...\}$ je trojúhelníkové schéma nezáporných konstant splňující pro všechna n podmínku

$$v_{n1} \geq \ ... \geq v_{nn} \quad \& \quad \max_j v_{nj} \to 0 \quad \text{pro} \quad n \to \infty \quad \& \quad \sum_{j=k_n+1}^{n} v_{nj} \to 0 \quad \text{pro} \quad n \to \infty.$$

Pro daný bod x zkonstruujme nové trojúhelníkové schéma $\{ w_{ni}(x) \mid i = 1,...,n \ \&\ n = 1,2,...\}$ tak, že pro každé n je $\Sigma_i\, w_{ni} = 1$ a $w_{ni}(x) = v_{nj}$, kde j je pořadí $|x-X_i|$ uvnitř množiny $\{|x-X_k|\ |\ k = 1,...,n\}$. *Odhad funkce $r(x)$ pomocí vážených k_n nejbližších sousedů* je definován vztahem

$$(14)\quad \hat{r}_n^{(3')}(x) = \sum_{i=1}^{n} Y_i . w_{ni}(x), \qquad x \in R^1.$$

Položíme-li $v_{nj} = k_n^{-1}$ pro $j = 1,...,k_n$ a $v_{nj} = 0$ pro $j = k_n+1,...,n$, potom odhad (14) splývá s (13).

Druhý přístup si všímá vysoké nerobustnosti jak odhadu (13), tak (14), a podobně jako u jádrových odhadů se snaží nahradit výběrový průměr některým odhadem robustním vzhledem k možnému výskytu odlehlých hodnot Y_i, například *M*- či *L*- odhadem.

Z výpočetní stránky jsou odhady pomocí k_n nejbližších sousedů jednoduché pouze pro variantu (13). V tomto případě je lze počítat velice rychle a úsporně. Zvolíme-li však některou z modifikací zahrnujících váhy, ať již (14) či některou z robustních verzí, výpočet se komplikuje a může být velice náročný na spotřebu času počítače. S volbou optimálního k_n jsou podobné problémy jako u jádrových odhadů. V praxi se postupuje opět buď pomocí křížového ověřování, nebo se odhad spočte pro několik různých hodnot k_n a za řešení se zvolí hodnota poskytující "opticky nejhladší" křivku. Existují též některá doporučení vycházející z asymptotických úvah, viz např. [22].

II.4. ODHADY POMOCÍ ORTONORMÁLNÍCH FUNKCÍ

Nechť $\{ (X_i, Y_i),\ i = 1,\dots,n \}$, jsou nezávislé kopie náhodného vektoru (X,Y). Nechť $\mathbf{P} = \{ P_i(x),\ i = 1,2,\dots \}$ je úplná ortonormální posloupnost funkcí v $L_2(R^1)$. Položme

$$(15)\quad d_i = \sum_{j=1}^{n} Y_j \cdot \int_{A_j} P_i(x)\, dx,$$

kde $\{ A_j,\ j = 1,\dots,n \}$ tvoří takový rozklad R^1, že $\bigcup_j A_j = R^1$ a $X_j \in A_j$, $j = 1,\dots,n$. *Odhad funkce r(x) pomocí posloupnosti ortonormálních funkcí* $\mathbf{P}$ je definován vztahem

$$(16)\quad \hat{r}_n^{(4)}(x) = \sum_{i=1}^{\infty} d_i . P_i(x), \qquad x \in R^1.$$

I zde v praxi užíváme pro odhad pouze prvních L_n členů v (16), tj. volíme

$$(17)\quad \hat{r}_n^{(4')}(x) = \sum_{i=1}^{L_n} d_i . P_i(x), \qquad x \in R^1.$$

Úloha určení L_n je ještě složitější než v případě odhadu hustoty. Výpočetní složitost je též vyšší, neboť výpočet vah d_i je složitější než výpočet konstant c_i v (6).

Za systém $\mathbf{P}$ se doporučuje volit například posloupnost Legendrových či Hermitových polynomů apod., blíže viz [16] nebo [21]. Odhady tohoto typu jsou pro svoji výpočetní složitost a problémy s určením optimální hodnoty L_n v praxi velmi málo používány.

LITERATURA

I. Vybrané práce týkající se neparametrických odhadů hustoty.

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

[1] Antoch J. a Cipra T. (1990), Neparametrické odhady spolehlivostních charakteristik. Výzkumná zpráva pro ŠKODA, k.p.

[2] Bean S.J. a Tsakas C.P. (1980), Developments in nonparametric density estimation. *International Statistical Review*, 48, 267-287.

[3] Beneš V. (1988), Odhady hustoty pravděpodobnosti useknutého rozdělení. Sborník *ROBUST 88*, 14-17, JČSMF, Praha.

[4] Devroye L. a Györfi L. (1985), Nonparametric density estimation : The L_1-view. J. Wiley, New York.

[5] Eryer J. (1977), Revue of some nonparametric methods of density estimation. *Journal of Inst. Math. Applic.*, 20, 335-354.

[6] Nadaraja A.E. (1980), Něparametričeskije ocenki plotnosti verojatnostěj i krivych regressii. Izdatělstvo Tbilisskogo Universitetà, Tbilisi.

[7] Parzen E. (1962), On estimation of probability density function and mode. *Annals of Mathematical Statistics*, 33, 1065-1076.

[8] Rosenblatt M. (1956), Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27, 642-669.

[9] Sheater S.J. (1986), An improved data-based algorithm for choosing the window width when estimating the density at a point. *Computational Statistics and Data Analysis*, 4, 61-65.

[10] Silvermann J. (1982), AS 176. *Applied Statistics.*

[11] Wertz W. (1978), Statistical density estimation : A survey. Vandenboeck & Rupprecht, Götingen.

[12] Wertz W. a Scheiner B. (1979), Statistical density estimation : A bibliography. *International Statistical Review*, 47, 155-175.

II. Vybrané práce týkající se neparametrických odhadů regresních křivek.

[13] Ahmad I.A. a Lin P.E. (1976), Nonparametric sequential estimation of multiple regression function. *Bull. Math. Statistics*, 17, 63-75.

[14] Antoch J. (1986), Neparametrické odhady regresních křivek. Sborník *ROBUST 86*, 1-20, JČSMF, Praha.

[15] Bhattacharya P.K. a Partasarathy K.R. (1961), Some limit theorems in regression theory. *Sankhya*, A 23, 91-102.

[16] Collomb G. (1985), Nonparametric regresion: An up-to-date bibliography. *Statistics*, 16, 309-324.

[17] Devroye L.P. (1979), The uniform convergence of the Nadaraja-Watson regression function estimate. *Canadian Journal of Statistics*, 6, 179-191.

[18] Epanechnikov B.A. (1969), Něparametričeskaja ocenka mnogomernoj plotnosti verojatnostěj. *Těorija verojatnostěj i jejo primeněnija*, 15, 156-161.

[19] Gasser T. and Rosenblatt M. (1979), Smoothing techniques for curve estimation. *Lecture Notes in Mathematics*, 757, Springer, Berlin.

[20] Hall P. (1984), Asymptotic properties of integrated square error and crossvalidation for kernel estimation of a regression function. *Zeitschrift fur Wahrsscheinlichkeitstheorie und verw. Gebiete*, 67, 175-196.

[21] Härdle W. (1990), Applied nonparametric regression. Springer, Berlin.

[22] Nadaraja A.E. (1964), Něparametričeskaja ocenka regressii. *Těoria verojatnostěj i jejo primeněnija*, 9, 141-142.

[23] Pearson K. a Lee A. (1903), On the laws of inheritance in a man. *Biometrika*, 2, 357-362.

[24] Stone C.J. (1977), Consistent nonparametric regression. *Annals of Statistics*, 2, 595-645.

[25] Tukey J.W. (1961), Curves as parameters and touch estimation. *Sbornik 4. Berkleyského Symposia*, 681-694.

[26] Watson G.S. (1964), Smooth regression analysis. *Sankhya*, A 26, 359-372.

[27] Wang W.W. (1983), On the consistency of cross-validation in kernel nonparametric regression. *Annals of Statistics*, 11, 1136-1141.