ODHADY HUSTOTY

Jaromír ANTOCH, MFF UK Praha

0. Úvod:

Máme-li k disposici pozorování náhodné veličiny X definované na $(R^1, \mathcal{B}_1, \mathcal{L})$ s distribuční funkcí $F(x)$ a hustotou $f(x)$, jednou ze základních statistických úloh se stává nalezení přirozených odhadů $F(x)$ a $f(x)$ na základě nezávislých pozorování X. Cílem tohoto shrnutí je přiblížit některé metody odhadu $f(x)$ a ukázat jejich základní vlastnosti. Zvláštní pozornost je věnována tzv. K-odhadům.

1. Některé druhy odhadů hustoty

Nechť $X_1, \ldots, X_n$, $n \in N$ je posloupnost iidrv's s hustotou $f(x)$. Označme

$$(1.1) \qquad K_n(a,b) = \left\{ \text{počet } X_i \,/ X_i \in (a,b), \quad i=1,\ldots,n \right\} .$$

Chceme-li odhadnout $P(a < X < b) = \int_a^b f(t)\,dt$, můžeme to učinit například pomocí $n^{-1}.K_n(a,b)$, (samozřejmě pro n dosti velké - analogicky jako u empirické distribuční funkce).

Naopak hodnota spojité funkce $f(x)$, $a < x < b$, může být odhadnuta pomocí $(b-a)^{-1}$. $\int_a^b f(t)\,dt$ (je-li naopak (a,b) dostatečně malý interval). Spojíme-li předchozí dva body, dostaneme

$$(1.2) \qquad f(x) \sim (b-a)^{-1} . \int_a^b f(t)\,dt \sim \frac{K_n(a,b)}{n(b-a)} ,$$

kde $\sim$ znamená pouze intuitivní blízkost. Výše uvedený postup a vztah (1.2) pak tvoří ideu většiny definic empirické hustoty. Použijeme-li však tento přístup, dopouštíme se dvou "základních chyb".

A) Odhad $\int_a^b f(t)\,dt$ vztahem $n^{-1}K_n(a,b)$ je přesný, je-li $K_n(a,b)$ dostatečně velké, tj. v důsledku, není-li (a,b) příliš krátké.

B) Odhad $f(x)$ vztahem $(b-a)^{-1} . \int_a^b f(t)\,dt$ pro $x \in (a,b)$ je přesný, pokud je interval (a,b) hodně krátký.

Oba tyto požadavky si navzájem protiřečí a je vždy třeba hledat vhodný kompromis. Dále si uveďme některé možné definice odhadu hustoty $f(x)$.

Nejjednodušším způsobem odhadu $f(x)$ je patrně histogram.

Definice 1.1: Nechť $\ldots x_{-1}(n) < x_0(n) < x_1(n) \ldots$ je některý rozklad R^1 a nechť

$$k_n = x_{i+1}(n) - x_i(n) \qquad i=0, \pm 1, \pm 2, \ldots .$$

Pak definujme odhad $f(x)$ předpisem

$$(1.3) \qquad f_n^{(1)}(x) = \frac{K_n(x_i(n),\, x_{i+1}(n))}{n\, k_n} , \qquad x_i(n) < x < x_{i+1}(n).$$

Základní nevýhodou tohoto přístupu je rozklad R^1 na ekvidistantní intervaly bez ohledu na charakter dat. Tento nedostatek částečně odstraňuje následující definice, jejímž cílem je zdůraznit vliv těch pozorování z výběru, jež jsou blízko danému pevnému x, v němž odhad provádíme.

Definice 1.2 : Definujme odhad $f(x)$ předpisem

$$(1.4) \qquad f_n^{(2)}(x) = \frac{K_n(x - k_n, x + k_n)}{2n\, k_n} \qquad x \in R_1.$$

Není těžké ukázat, že odhad dle předchozí definice je vlastně specielním případem následující obecné třídy tzv. K-odhadů, kterou historicky předcházel.

Definice 1.3 : Nechť $w(x)$ je libovolná hustota na R_1 a $\{k(n)\}$, $n=1,2,\ldots$ posloupnost kladných konstant (závislých na n) taková, že $k(n) \to 0$ pro $n \to \infty$.
Potom definujme odhad $f(x)$ předpisem

$$(1.5) \qquad f_n^{(3)}(x) = \frac{1}{n\, k_n} \sum_{i=1}^{n} w\left[\frac{x - X_i}{k_n}\right] = \frac{1}{k_n} \int_{-\infty}^{\infty} w\left(\frac{x-y}{k_n}\right) dF_n(y),$$

kde $F_n(x)$ je empirická distribuční funkce založená na výběru $X_1, \ldots, X_n$.

Jinou metodu, založenou na odhadu Fourierových koeficientů, navrhl Čencov (1962).

Definice 1.4 : Nechť hustota $f(x)$ je taková, že $f \in L^2$ a

$$f(x) > 0 \qquad -\infty \leq A < x < B \leq +\infty$$
$$= 0 \qquad \text{vně } (A,B).$$

Nechť $\varphi = \{\varphi_k(x)\}_{k=1}^{\infty}$ je úplná ortonormální posloupnost definovaná na (A,B). Potom definujme odhad $f(x)$ předpisem

$$(1.6) \qquad f_n^{(4)}(x) = \sum_{k=1}^{n} \hat{c}_k\, \varphi_k(x), \qquad A < x < B,$$

$$\text{kde } \hat{c}_k = \sum_{j=1}^{n} \varphi_k(x_j).$$

Při bližším porovnání předchozích definic brzo zjistíme, že je všechny lze zahrnout do jedné, velice obecné, následovně:

Definice 1.5 : Nechť hustota $f(x)$ splňuje

$$f(x) > 0 \qquad -\infty \leq C < x < D \leq +\infty \quad ;$$
$$= 0 \qquad \text{vně } (C,D) \quad ;$$

a nechť $\Psi = \{\Psi_k(x,y)\}_{k=1}^{\infty}$ je posloupnost Borelovsky měřitelných funkcí na $(A,B)^2$, kde $(C,D) \subseteq (A,B)$. Potom definujme odhad $f(x)$ následovně:

$$(1.7) \qquad f_n^{(5)}(x) = n^{-1} \sum_{k=1}^{n} \Psi(x, x_k) = \int_A^B \Psi_n(x,y)\, dF_n(y), \quad x \in R_1.$$

Výše uvedené definice nejsou samozřejmě jediné, ba právě naopak. Nicméně ukazují nejsilnější a nejužívanější větev, jež se v daném oboru rozvinula. Podrobnou bibliografii lze nalézt např. ve Wertz (1979). Vážní zájemci se pak jistě nejvíce dozvědí z připravované Wertzovi monografie, jež se má objevit v nejbližší době. Z mnoha dalších prací je pak velmi zajímavá např. 6.kapitola knihy Czorgő-Révész, kde čtenář nalezne mnoho velmi zajímavých výsledků a aproximací těchto odhadů na základě metod silných aproximací.

2. K - odhady

Nejpropracovanější oblastí v teorii odhadu neznámé hustoty na základě nezávislých pozorování jsou metody tzv. K-odhadů (Kernel estimators). První odhady tohoto typu navrhl Rosenblatt (1956) a od té doby byly mnohokrát studovány četnými dalšími autory. V literatuře se obvykle definují následovně :

Definice 2.1 : Nechť náhodná veličina X má hustotu $f(x)$, distribuční funkci $F(x)$ a $X_1,\ldots,X_n,\ldots$ jsou její nezávislé kopie. Nechť $w(u)$, váhová funkce, je integrovatelná ohraničená funkce z L_2, $\int_{-\infty}^{\infty} w(u)\,du = 1$. Nechť $\{h(n)\}_{n=1}^{\infty}$ je posloupnost kladných konstant (závislých na n) taková, že $h(n) \to 0$ pro $n \to \infty$. Pak jednorozměrný K-odhad hustoty $f(x)$ určený dvojicí $\left[w(u), \{h(n)\}_{n=1}^{\infty}\right]$, je $\forall x$ definován vztahem

$$f_n(x) = \frac{1}{n\,h(n)} \sum_{j=1}^{n} w\left(\frac{x-X_j}{h(n)}\right) . \tag{2.1}$$

Pozn.: (1) Hovoříme-li o K-odhadu hustoty $f(x)$, máme obvykle na mysli $\forall$ pevné x_o posloupnost odhadů $\widehat{f(x_o)} = \{f_n(x_o)\}_{n=1}^{\infty}$.

(2) Ihned je vidět úzká spojitost s některými odhady z odstavce 1, kde např. tzv. přirozený odhad (1.4) je speciálním případem (2.1) pro váhovou funkci.

$$w(u) = \tfrac{1}{2} \qquad |y| \leq 1$$
$$\quad\;\; = 0 \qquad |y| > 1. \tag{2.2}$$

(3) Díky požadavkům předchozí definice je $w(u)$ vlastně hustota. Tento fakt však není žádným omezením, neb se zde jedná pouze o vhodnou normalizaci váhové funkce $w(u)$.

Problémy, spojené s definicí 2.1, jež nás především zajímají, jsou následující:

(I) Za jakých podmínek jsou K-odhady asymptoticky nestranné a asymptoticky konzistentní.

(II) Asymptotická normalita K-odhadů a těsnost této aproximace.

(III) Posouzení lokálních a globálních kvalit K-odhadů.

(IV) Volba optimální váhové funkce a optimální posloupnosti konstant $\{h(n)\}_{n=1}^{\infty}$.

Pokusme se na ně nyní, alespoň z části, odpovědět. Odpověď na (I) lze nalézt ve větě 2.1, odpověď na (II) ve větě 2.2.

Věta 2.1 Nechť $X_1, X_2, \ldots$ je posloupnost nezávislých kopií téže náhodné veličiny X s hustotou $f(x)$.

(a) Nechť váhová funkce $w(u)$ a posloupnost konstant $\{h(n)\}_{n=1}^{\infty}$ splňuje podmínky Definice 2.1 a navíc

$$\lim_{y \to \infty} |y\,w(y)| = 0. \tag{2.3}$$

Potom K-odhady typu (2.1) jsou asymptoticky nestranné ve všech bodech spojitosti hustoty $f(x)$, tj.

$$E\,f_n(x) \to f(x)\,, \qquad n \to \infty .$$

(b) Nechť jsou splněny podmínky ad (a) a navíc

$$\lim_{n \to \infty} n\,h(n) = +\infty . \tag{2.4}$$

Potom odhady typu (2.1) jsou asymptoticky konzistentní podle kvadratického středu, tj.

$$E|f_n(x) - f(x)|^2 \longrightarrow 0\,, \qquad n \to \infty .$$

Dk : Rosenblatt (1971).

Všimněme si, že $\forall$ pevné $n \in N$ lze vztah (2.1) přepsat ve tvaru

(2.5) $$f_n(x) = \frac{1}{n} \sum_{k=1}^{n} V_{nk} ,$$

kde

$$V_{nk} = (h(n))^{-1} \cdot w\left(\frac{x-X_k}{h(n)}\right), \qquad k=1,\dots,n,$$

jsou nezávislé kopie náhodné veličiny $V_n = (h(n))^{-1} \cdot w\left(\frac{x-X}{h(n)}\right)$. V obdrženém trojúhelníkovém schematu není obtížné ověřit podmínku Ljapunoveva typu pro platnost CLT a dostaneme:

Věta 2.2 Nechť sudá funkce $w(u)$ a posloupnost konstant $\{h(n)\}_{n=1}^{\infty}$ splňuje podmínky (a), (b) Věty 2.1. Potom odhady typu (2.1) jsou asymptoticky normální, tj. $\forall c \in R^1$

(2.6) $$\lim_{n\to\infty} P\left[\frac{f_n(x) - E(f_n(x))}{\sigma(f_n(x))} \leq c\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c} e^{-\frac{y^2}{2}} dy = \Phi(c).$$

Pozn. Z Berry-Essenovy nerovnosti dostáváme následující těsnost výše uvedené normální aproximace, tj.

(2-7) $$\sup_{-\infty < a < +\infty} \left| P\left[\frac{f_n(x) - E\, f_n(x)}{\sigma(f_n(x))} \leq a\right] - \Phi(a) \right| \leq \frac{C.E|V_n|^3}{\sqrt{n}\, \sigma^3(V_n)} \sim \frac{1}{\sqrt{n\, h(n) f(x)}} \cdot \frac{\int_{-\infty}^{\infty} |w(y)|^3 dy}{\left(\int_{-\infty}^{\infty} w^2(y) dy\right)^{3/2}} .$$

Při odpovědi na otázku (III) se musíme především dohodnout na vhodných mírách pro posouzení kvality odhadu. Kvalitu musíme přitom posuzovat jak z lokálního, tak globálního hlediska. Nejčastěji užívanými mírami jsou např.

(2.8)
$$\begin{array}{lll} \text{rozptyl} & \dots\dots & \sigma^2(f_n(x)) \\ \text{MSE} & \dots\dots & E|f_n(x)-f(x)|^2 \\ \text{IMSE} & \dots\dots & \int_{-\infty}^{\infty} E|f_n(x)-f(x)|^2 dx \\ \text{vychýlení} & \dots\dots & E\, f_n(x)-f(x) \text{ ap.} \end{array}$$

Uvedené míry samozřejmě nejsou jediné možné. O různých typech globálních měr je pojednáno např. v Rosenblatt (1979), o vlastnostech měr lokálních v přehledném článku Rosenblattové (1971).

Při odvozování asymptotických vlastností měr typu (2.8) se většinou používá následující postup, jenž byl poprvé použit Rosenblattem (1956) pro zkoumání vlastností odhadu tvaru (1.4), na němž i zde si hrubé odvození provedeme.

Nechť existují prvé tři derivace hustoty $f(y)$ v bodě x. Potom, použijeme-li rozvoj

(2.9) $$[F(x+h(n)) - F(x-h(n))] \approx 2h(n) f(x) + \frac{1}{3} \cdot f''(x) \cdot (h(n))^3 + o(h^4(n)),$$

kde $F(x)$ je distribuční funkce náhodné veličiny X, dostaneme

(2.10) $$\sigma^2(f_n^*(x)) = \frac{1}{4n(h(n))^2} \left[\{F(x+h(n))-F(x-h(n))\} + \{F(x+h(n))-F(x-h(n))\}^2\right] \sim \frac{f(x)}{2n\, h(n)} ,$$
(pokud $h(n) \to 0$ pro $n \to \infty$).

(2-11) $$E|f_n^*(x)-f(x)|^2 = \sigma^2(f_n^*(x)) + (E\, f_n^*(x)-f(x))^2 \sim \frac{f(x)}{2nh(n)} + \frac{(h(n))^4}{36} |f''(x)|^2 + o\left(\frac{1}{nh(n)} + (h(n))^4\right),$$

(opět pokud $h(n) \longrightarrow 0$ pro $n \to \infty$); a zintegrováním

(2.12) $$\int_{-\infty}^{\infty} E|f_n^*(x)-f(x)|^2 dx \sim \frac{1}{2nh(n)} + \frac{(h(n))^4}{36} \int_{-\infty}^{\infty} |f''(x)|^2 dx + o\left(\frac{1}{nh(n)} + (h(n))^4\right).$$

Vztahy (2.9)-(2.12) lze snadno překontrolovat pomocí základních aritmetických operací.

Vezmeme-li nyní v úvahu odhad (2.1) a $F_n(x)$ označíme empirickou distribuční funkci příslušnou k prvým n pozorováním, vidíme, že vztah (2.1) lze psát v ekvivalentní formě

$$(2.13) \qquad f_n(x) = \int_{-\infty}^{\infty} \frac{1}{h(n)} \cdot w\left(\frac{x-y}{h(n)}\right) \, d\, F_n(y).$$

Postupem analogickým předchozím (ale technicky náročnějším) pak lze získat následující aproximace (viz. např. Rosenblatt 1971, resp. 1979).

$$(2.14) \qquad \sigma^2[f_n(x)] = \frac{1}{n}\left[\frac{1}{h(n)} \int_{-\infty}^{\infty} w^2(x)\ f(x-h(n)u)du - \left(\int_{-\infty}^{\infty} w(u) \cdot f(x-h(n)u)du\right)^2\right] \sim$$

$$\sim \frac{f(x)}{n\, h(n)} \int_{-\infty}^{\infty} w^2(u)du \qquad \text{pokud} \quad f(x) > 0 \quad h(n) \to 0 \quad \text{pro} \quad n \to \infty$$

$$\sim \frac{h(n)}{n} \cdot f''(x) \int_{-\infty}^{\infty} w^2(u)du \qquad \text{pokud} \quad f(x)=0, \quad f''(x) \neq 0, \quad h(n) \to 0 \quad \text{pro} \quad n \to \infty \ ;$$

$$(2.15) \qquad E|f_n(x)-f(x)|^2 \sim \frac{f(x)}{n\, h(n)} \int_{-\infty}^{\infty} w^2(u)du + \frac{1}{4}(h(n))^4 (f''(x))^2 \left(\int_{-\infty}^{\infty} w(u)u^2 du\right)^2 + o\left(\frac{1}{nh(n)}\right) + (h(n))^4) ;$$

$$(2.16) \qquad \int_{-\infty}^{\infty} E|f_n(x)-f(x)|^2 dx \sim \frac{1}{nh(n)} \int_{-\infty}^{\infty} w^2(u)du + \frac{1}{4}(h(n))^4 \int_{-\infty}^{\infty} |f''(x)|^2 dx . \left(\int_{-\infty}^{\infty} w(u)u^2 du\right)^2 + o\left(\frac{1}{nh(n)} + (h(n))^4\right)$$

Prohlédneme-li si nyní vztahy (2.10)-(2.16) vidíme, že podstatnou se stala odpověď na otázku (IV). Uvažujme nejprve volbu posloupnosti $\{h(n)\}_{n=1}^{\infty}$. Požadujeme-li rychlost konvergence řádu $h(n) = k\, n^{-\alpha}$, $\alpha > 0$, k konst., ze vztahu (2.10)-(2.11) a (2.15)-(2.16) vidíme, že ve všech čtyřech případech je optimální volbou $\alpha = \frac{1}{5}$ a k to, jenž minimalizuje příslušný člen, jímž ten který výraz aproximujeme. V případě (2.11) je to např. ta konstanta, jež minimalizuje výraz

$$(2.17) \qquad \frac{f(y)}{2k_1} + \frac{k_1^4}{36} \cdot |f''(y)|^2 , \quad \text{tj.} \qquad k_1 = \left[\frac{9}{2} \frac{f(y)}{|f''(y)|^2}\right]^{1/5} .$$

V ostatních případech je tomu analogicky. Dostaneme přitom následující aproximace. Vztah (2.11) přejde na

$$(2.18) \qquad E|f_n^*(x) - f(x)|^2 \sim 0.24\ n^{-4/5}\ (f(x))^{4/5} |f''(x)|^{2/5} \ ;$$

vztah (2.12) na

$$(2.19) \qquad \int_{-\infty}^{\infty} E|f_n^*(x)-f(x)|^2 dx \sim 0.24\ n^{-4/5} \left[\int_{-\infty}^{\infty} |f''(x)|^2 dy\right]^{1/5} \ ;$$

vztah (2.15) na

$$(2.20) \qquad E|f_n(x)-f(x)|^2 \sim 1.52\ \left[f(x) \int_{-\infty}^{\infty} w^2(u)du\right]^{4/5} . \left|f''(x) \int_{-\infty}^{\infty} w(u)u^2 du\right|^{2/5} . n^{-4/5} + o(n^{-4/5}),\ n \to \infty \ ;$$

a vztah (2.16) na

$$(2.21) \qquad \int_{-\infty}^{\infty} E|f_n(x)-f(x)|^2 dx \sim 1.52\ \left[\int_{-\infty}^{\infty} w^2(u)du\right]^{4/5} . n^{-4/5} . \left[\int_{-\infty}^{\infty} w(u)u^2 du\right]^{2/5} . \left[\int_{-\infty}^{\infty} |f''(x)|^2 dx\right]^{1/5} + o(n^{-4/5}) \quad n \to \infty .$$

Pozn. (1) Na tomto místě je dobré si všimnout, že globální míra chování odhadu typu IMSE konverguje k nule řádově tak rychle jako míra lokální, totiž $o(n^{-4/5})$. Speciálně, tento řád konvergence pro K-odhady nezáleží na rychlosti poklesu chvostu odhadované hustoty v nekonečnu.

(2) Při hledání optimálních konstant $\{h(n)\}_{n=1}^{\infty}$ se dostaneme k jednomu ~~z~~ největších problémů celé teorie K-odhadů, neb volba α i k závisí na znalostech $f(x)$, $f''(x)$,... Toto je však předpoklad apriori nesmyslný uvědomíme-li si, že naším cílem je odhad $f(x)$. A není to jen při určení posloupnosti $\{h(n)\}_{n=1}^{\infty}$, ale i při odvození některých základních vlastností K-odhadů.

Tato potíž se nám obzvlášť výrazně projeví tehdy, jestliže odhady skutečně počítáme na základě simulovaných dat. Neznáme-li totiž (hypoteticky) rozdělení, z něhož výběr pochází, dostáváme výsledky výrazně horší než v případě, kdy na základě jeho znalosti můžeme optimálně volit $\{h(n)\}_{n=1}^{\infty}$ ... – což je ovšem tentýž bludný kruh jako výše.

Poslední hlavní úkol, jenž nám zbývá vyjasnit, je volba optimální váhové funkce. Tento úkol mj. řešil Epanechnikov (1969) a dosáhl některých velmi zajímavých výsledků.

Mějme k dispozici odhady splňující podmínky. Definice 2.1 a nechť navíc váhová funkce $w(u)$ splňuje tyto normující podmínky:

$$\text{(c)} \qquad w(u) = w(-u) \qquad , \qquad u \in R_1$$

$$\text{(2.22)} \qquad \text{(d)} \qquad \int_{-\infty}^{\infty} w(u)u^2 \, du = 1 \; .$$

Zvolíme-li za kritérium optimalizace IMSE, pak ze vztahu (2.21) vidíme, že pro určení optimální váhové funkce při pevných u, k a $\{h(n)\}$ a splnění (2.22) stačí najézt váhovou funkci minimalizující

$$\text{(2.23)} \qquad \int_{-\infty}^{\infty} w^2(y) \, dy .$$

Epanechnikov řešil tuto úlohu metodami variačního počtu a ukázal, že existuje právě jedno řešení tvaru

$$\text{(2.24)} \qquad w_o(y) = \frac{3}{4\sqrt{5}} - \frac{3y^2}{20\sqrt{5}} \qquad |y| \leq 5$$

$$= 0 \qquad \text{jinak,}$$

a tato optimální váhová funkce nezáleží ani na původní hustotě, ani velikost výběru či, v případě j-rozměrné hustoty, na rozměru prostoru.

Ukažme si několik typických váhových funkcí. Typy $w_o - w_5$ navrhl Epanechnikov, $w_6 - w_{11}$ lze nalézt v práci Parzenově (1962).

Je zajímavé si povšimnout, že zatím co Epanechnikov navrhoval typy s ohledem k minimalizaci M, tj. ohledem k optimální váhové funkci, Parzen je patrně volil intuitivně.

Označme $L_i = \int_{-\infty}^{\infty} w^2(y) \, dy$ a $M_i = L_i / L_o$, $i = 0, \dots, 11$.

TABULKA 1.

i	w(y)	obor y	L	M
0	$\frac{3}{4\sqrt{5}} - \frac{3y^2}{20\sqrt{5}}$	$\lvert y \rvert \leq \sqrt{5}$	$\frac{3}{5\sqrt{5}}$	1
	0	$\lvert y \rvert > \sqrt{5}$		
1	$\frac{1}{4}\sqrt{\pi^2 - 8} \, . \cos \frac{\sqrt{\pi^2 - 8}}{2}$	$\lvert y \rvert \leq \frac{\pi}{\sqrt{\pi^2 - 8}}$		
	0	$\lvert y \rvert > \frac{\pi}{\sqrt{\pi^2 - 8}}$	$\frac{\pi}{\sqrt{\pi^2 - 8}}$	1.001

i	w(y)	obor y	L	M
2	$\frac{1}{\sqrt{6}} - \frac{\lvert y\rvert}{6}$ 0	$\lvert y\rvert \leq \sqrt{6}$ $\lvert y\rvert > \sqrt{6}$	$\frac{\sqrt{6}}{9}$	1.015
3	$\frac{1}{\sqrt{2\pi}}\, e^{-\frac{y^2}{2}}$	$y \in R_1$	$\frac{1}{2\sqrt{\pi}}$	1.051
4	$\frac{1}{2\sqrt{3}}$ 0	$\lvert y\rvert \leq \sqrt{3}$ $\lvert y\rvert > \sqrt{3}$	$\frac{1}{2\sqrt{3}}$	1.077
5	$\frac{1}{\sqrt{2}}\, e^{-\sqrt{2}\, y}$ 0	$y \geq 0$ $y < 0$	$\frac{1}{\sqrt{8}}$	1.320
6	$\frac{1}{2}\, e^{-\lvert y\rvert}$	$y \in R^1$	$\frac{1}{2}$	1.863
7	$\frac{1}{2}$ 0	$\lvert y\rvert \leq 1$ $\lvert y\rvert > 1$	$\frac{1}{2}$	1.863
8	$1 - \lvert y\rvert$ 0	$\lvert y\rvert \leq 1$ $\lvert y\rvert > 1$	$\frac{3}{2}$	2.485
9	$\frac{1}{\pi} \cdot \frac{1}{1+y^2}$	$y \in R^1$	$\frac{1}{\pi}$	1.186
10	$\frac{1}{2\pi}\left[\frac{2}{y}\sin\frac{y}{2}\right]^2$ 0	$0 \leq u \leq 2\pi$ $u < 0;\ u > 2\pi$	$\frac{1}{3\pi}$	0.395
11	$\frac{4}{3} - 8y^2 + 8\lvert y\rvert^3$ $\frac{8}{3}(1-\lvert y\rvert)^3$ 0	$\lvert y\rvert < \frac{1}{2}$ $\frac{1}{2} \leq \lvert y\rvert \leq 1$ $\lvert y\rvert > 1$	0.96	3.578

Máme-li takto shrnuty některé výsledky pro jednorozměrné K-odhady, naskýtá se přirozená otázka, jak je tomu v případě K-rozměrných hustot. Tuto otázku řešil Epanečnikov (1969) pro následující model.

Nechť $X_1,\dots,X_n$ je n nezávislých realizací k-rozměrné náhodné veličiny $X(x_1,\dots,x_k)$, tj.

(2.25) $$X_i = X(x_1^{(i)},\dots,x_k^{(i)}), \qquad i=1,\dots,n.$$

Definujme mnohorozměrnou empirickou hustotu $f_n(x_1,\dots,x_k)$ ve tvaru

(2.26) $$f_n^{(6)}(x_1,\dots,x_k) = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{k}\frac{1}{h_l(n)}\, w_l\left[\frac{x_l - x_l^i}{h_l(n)}\right] ,$$

kde váhová funkce $w_l(y)$ splňuje $\forall \ell$

$$0 \le w_l(y) < c < +\infty \quad , \quad y \in R_1 ;$$
$$w_l(y) = w_l(-y) \quad , \quad y \in R_1 ;$$
(2.27)
$$\int_{-\infty}^{\infty} w_l(y)dy = 1 ;$$
$$\int_{-\infty}^{\infty} w_l(y)y^2dy = 1;$$
$$\int_{-\infty}^{\infty} w_l(y)y^m < +\infty \qquad \text{pro některé } m \ge 2;$$

a nechť $\forall_l, \{h_l(n)\} \longrightarrow 0$ pro $n \to \infty$.

Ve své práci autor ukázal některé základní vlastnosti odhadů tvaru (2.26) a soustředil se především na určení optimální volby $h_l(n)$, $w(y)$ vzhledem k minimalizaci IMSE. Výsledky, které obdržel, jsou analogické těm, získaným pro náš jednorozměrný model a tyto vesměs vyplývají jako zřejmý důsledek.

Posledním typem odhadu, jenž je opět velmi úzce spojen s odhady z Definice 2.1, a o němž se krátce zmíníme, budou tak zvané k-nejbližší sousední odhady. Tyto jsou obvykle definovány následovně : nechť $X_1,\dots,X_n$ jsou nezávislé kopie náhodné veličiny X s hustotou $f(x)$, $w(u)$ váhová funkce splňující podmínky Definice 2.1 a $R_n = R_n(x)$ označíme vzdálenost mezi x a k-tým nejbližším sousedem k x mezi $X_1,\dots,X_n$. Potom definujme odhad $f(x)$ předpisem

$$f_n^{(7)}(x) = \frac{1}{n\,R_n} \sum_{j=1}^{n} w\left(\frac{x-X_j}{R_n}\right), \qquad x \in R_1 .$$

Analogie s (2.1) je evidentní. Studiem těchto odhadů se potvrdila, dle očekávání, značná podobnost s výsledky pro odhady (2.1) (alespoň asymptoticky). Např. je-li $f(x)$ ohraničená, spojitě diferencovatelná v okolí x a $f(x) > 0$, pak MSE klesá k nule řádově nejrychleji $n^{-4/5}$ ap. Obě metody by si nyní jistě zasloužily některá numerická srovnání pro porovnání jejich skutečné síly.

3. Aplikace

Na závěr svého vystoupení v Podkosti jsem byl dotázán : "Dobře tedy, ale k čemu to všechno vlastně je dobré?" Pokusím se tedy krátce na tuto otázku odpovědět.

Pomineme-li odpovědi typu:

a) je to zajímavé samo o sobě;

b) vždyť je to právě hustota s níž tolik pracujeme a měli bychom snad o ní znát co nejvíce;

c) pohled na histogram nám řekne mnohem víc o datech než pohled na empirickou distribuční funkci ... ,

lze argumentovat i následujícím.

Metody, původně rozvíjené pro odhady hustoty byly s úspěchem a značným užitkem použity např. v časových řadách při odhadu spektrální hustoty ap. (viz např. Rosenblatt 1971). Dále, stále více prací je věnováno užitím výše uvedených myšlenek v regresní analýze. Použití metod K-odhadů regresní funkce rozpracovali např. Gasser-Müller (viz [9]), užití metod k-tého nejbližšího souseda pro odhad regresní funkce podrobně rozpracoval Stone ...

Protože každá z výše uvedených oblastí by si sama o sobě vyžádala samostatné shrnutí, ukáži využití myšlenek spojených s odvozením K-odhadů na následujícím jednoduchém příkladu.

Nechť $X_1,\dots,X_n$ jsou nezávislé kopie téže náhodné veličiny X s distribuční funkcí $F(x)$, inverzní distribuční funkcí $F^{-1}(y)$ a hustotou $f(x)$. Označme $X_{(1)},\dots,X_{(n)}$ odpovídající pořádkové statistiky. Tak jako je empirická distribuční funkce $F_n(x)$ přirozeným odhadem $F(x)$, u $F^{-1}(y)$ tímto přirozeným odhadem bývá kvantilová funkce

$$(3.1)\qquad Q_n(y) = X_{(k)}\,, \quad \frac{k-1}{n} < y \le \frac{k}{n}\,, \quad k=1,\dots,n,$$
$$Q_n(0) = X_{(1)}, \quad Q_n(y) = 0 \quad \text{pro} \quad y \notin [0,1] \;.$$

O vlastnostech $Q_n(y)$ pojednává řada prací, stejně jako o možnostech charakterizace náhodných veličin pomocí $F^{-1}(t)$ [viz např. (1) a Parzenův článek v (9)].

Sestrojujeme-li konfidenční interval pro $F^{-1}(y)$ [odhadujeme-li pomocí $Q_n(y)$] , narazíme na potřebu odhadnout veličinu $G(t) = \frac{1}{f(F^{-1}(t))}$. S nutností odhadnout $G(t)$ se však nesetkáváme pouze zde, ale i v některých partiích neparametrických metod ap. V minulosti byly sice navrženy některé odhady $G(t)$, bohužel vesměs dosti složité a neohrabané.

Uvědomíme-li si, že $\frac{d}{dy} F^{-1}(y) = \frac{1}{f(F^{-1}(y))} = G(y)$ [jsou-li splněny vhodné podmínky regularity] , proč nepoužít analogie s (1.4) a za odhad $G(t)$ nevzít

$$(3.2)\qquad G(y) = \frac{Q_n(y+a_n) - Q_n(y-a_n)}{2\,a_n}\;,$$

kde $\{a(n)\}_{n=1}^{\infty}$ je některá posloupnost kladných konstant a jenž je přirozenou analogií odhadu (1.4). Rozšíření odhadu (3.2) na třídu odhadů

$$(3.3)\qquad G_w(y) = \frac{1}{a_n} \int_0^1 w\left(\frac{y-u}{a_n}\right)\, d\, Q_n(u),$$

kde $w(u)$ je některá hustota na (0,1), je potom již zřejmé z odvození K-odhadů. Zkoumání vlastností odhadů (3.3) lze provést analogicky.

<u>L i t e r a t u r a :</u> Jak již bylo jednou zmíněno, podrobný soupis literatury lze nalézt v práci Wertzově. Proto jsou zde citovány pouze některé klíčové práce.

(1) Csörgő-Révész (1981) : Strong Approximations in Probability and Statistics. Academia Press, New York.

(2) Епанечников В.А.(1969): Непараметрическая оценка многомерной плотности вероятностей. Теория вероятностей и её применения,Том XVI. ,156-161.

(3) Parzen E.(1962) : On estimation of a probability density function and mode. AMS,33, 1065-1076.

(4) Rosenblatt M.(1956) : Remarks on some nonparametric estimates of a density function. AMS,27, 832-837.

(5) Rosenblatt M.(1971) : Curve estimation. AMS,42, 1815-1842.

(6) Rosenblatt M.(1979) : Global measures of deviation for kernel and nearest neighbor density estimates. Lecture Notes 757, 181-190, Springer-Verlag.

(7) Stone C.J.(1977): Consistent nonparametric regression. With discussion. AS 5,595-645.

(8) Wertz W., Scheiner B.(1979): Statistical density estimation : a Bibliography. International Statistical Review, 47, 155-175.

(9) Lecture Notes in Mathematics, vol.757, Smoothing Techniques for Curve Estimation, Proceedings Heidelberg 1979, Edited by Th.Gasser and M.Rosenblatt, Springer-Verlag.

Adresa: J.Antoch, MFF UK, KPMS, Sokolovská 83, 186 00 Praha 8-Karlín, tel. 2317683