NEPARAMETRICKÉ ODHADY REGRESNÍCH KŘIVEK

Jaromír Antoch, MFF UK Praha

1. ÚVOD

Budiž $(X_i, Y_i)_{i=1}^n$ posloupnost pozorování náhodné dvojice (X,Y), kde X je náhodný vektor v R^p a Y je reálná náhodná veličina. Rozdělení P_{XY} vektoru (X,Y) budiž neznámé a položme si úkol zodpovědět pouze na základě znalostí posloupnosti $(X_i,Y_i)_{i=1}^n$ následující dvě otázky:

(A) je-li $G \subset R^p$, G neprázdná, jak se v průměru "mění" Y jako funkce X nabývá-li X hodnot v G;

(B) je-li (x,y) některé další pozorování vektoru (X,Y), kde x je známé a y neznámé, jak předpovědět hodnotu y.

Intuitivně můžeme úlohu vyřešit například takto (pro jednoduchost na okamžik předpokládejme, že $G = [0,1)$ a $p = 1$).

(ad A) Budiž $k_n \in N$ pevné přirozené číslo nezávisející na posloupnosti $(X_i,Y_i)_{i=1}^n$ a rozdělme interval $[0,1]$ na k_n disjunktních intervalů. Způsob, jakým se Y mění jako funkce X pak může být representována schodovitou funkcí, jež je v každém intervalu konstantní a rovna průměru těch Y_i, $i=1,\dots,n$, pro něž X_i padne do uvažovaného intervalu. Metoda tedy vpodstatě transformuje oblak našich dat v křivku (byť zatím dosti hrubou), která nám dává řešení úlohy (A).

(ad B) Budiž I některé (symetrické) okolí bodu x délky k_n, kde $k_n \in R^1$ je pevné reálné číslo nezávisející na posloupnosti $(X_i,Y_i)_{i=1}^n$. Potom predikce $\hat{y}$ může být rovna průměru těch Y_i, $i=1,\dots,n$, pro něž odpovídající X_i leží v I.

Otázky typu (A) a (B) patří mezi základní otázky matematické statistiky odedávna. Naznačená řešení se v té či oné formě intuitivně též používají odedávna. Teprve v posledních 25. letech však došlo k systematickému zkoumání daných problémů, které nejenže dalo odpověď i na řadu dalších podobných otázek, ale dalo především vzniknout velmi rozsáhlé třídě tzv. neparametrických odhadů regresních křivek, o jejichž vlastnostech bylo doposud publikováno několik set prací. Na tomto místě je třeba připomenout, že souběžně (a to velmi úzce) se rozvíjí studium vlastností neparametrických odhadů hustoty. Četné výsledky, důkazové postupy apod. lze velice často bez jakýchkoliv principiálních problémů použít ve směru druhém a naopak.

Cílem tohoto článku je podat základní informaci o vlastnostech některých nejtypičtějších zástupců této třídy. Pozornost je přitom soustředěna především na regresogram a jádrové odhady. Jsou uvedeny definice jednotlivých odhadů a zkoumány jejich vlastnosti (konzistence, vychýlení, rozptyl, asymptotické rozdělení atd.). Zároveň jsou diskutovány i praktické problémy s volbou optimálních odhadů a složitostí jejich výpočtu na počítači. Teorii ilustruje v posledním odstavci simulační experiment.

2. REGRESOGRAM

V tomto odstavci se budeme zabývat tzv. regresogramem, který je nejjednodušším typem neparametrického odhadu regresní křivky a je, de facto, analogií histogramu. Po definici tohoto odhadu budou shrnuty některé jeho základní vlastnosti a diskutovány výhody a nevýhody jeho použití.

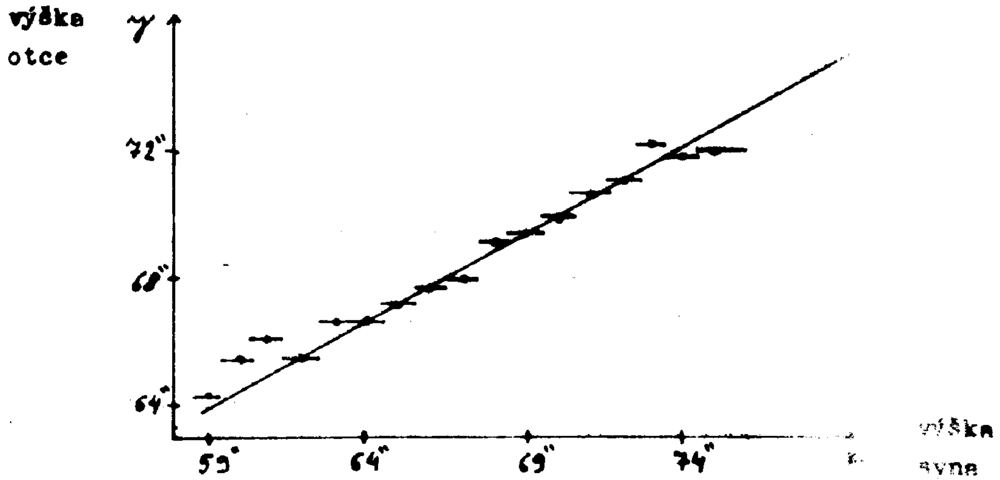
Definice 2.1. Nechť $(X_i,Y_i)_{i=1}^n$ je posloupnost nezávislých pozorování náhodného vektoru (X,Y) v prostoru $(R^{p+1}, \mathcal{B}^{p+1})$. Nechť $\mathcal{L}(X_i,Y_i) \sim \mathcal{L}(X,Y) \sim P_{XY}$ $\forall i$, $i=1,\dots,n$. Nechť $P_{XY} \ll \nu$, $p(x,y) = \frac{dP_{XY}}{d\nu}$ [kde ν je Lebesqueova míra na R^{p+1}] a nechť $\forall x \in R^p$ $f(x) = \int p(x,y)\,dy \neq 0$. Předpokládejme, že $\phi(x) = E(Y|X=x)$ existuje a uvažujme problém odhadu $\phi(x)$. Nechť $J_n = \{I_{n,m},\ m \in Z^p\}$ je některý rozklad prostoru R^p, $m=(m_1,\dots,m_p) \in Z^p$, $I_{n,m} = \prod_{k=1}^{p} [m_k h_n, (m_k+1) h_n)$, kde $\{h_n\}_{n=1}^{\infty}$ je některá posloupnost kladných čísel a Z značí množinu celých čísel v R^1. Nechť $\forall x \in R^p$ $I_n(x)$ označuje ten interval z J_n, který obsahuje bod x. Potom **regresogram** $\hat{\phi}_n(x)$ je definován vztahem

$$\hat{\phi}_n(x) = \begin{cases} \dfrac{1}{l_n(x)} \cdot \sum\limits_{\{i|X_i \in I_n(x)\}} Y_i & \text{jestliže } l_n(x) \neq 0, \\ 0 & \text{jinak,} \end{cases} \tag{2.1}$$

[illegible] $l_n(x)$ je počet těch X_i, $i=1,\dots,n$, jež leží v $I_n(x)$.

[illegible] 2.1 (a) V případě $p=1$ ihned vidíme, že se skutečně jedná o zřejmou analogii [illegible]togramu. Výsledkem je po částech konstantní funkce, která je v každém intervalu $I_n(\cdot)$ rozkladu J_n rovna průměru těch Y_i, $i=1,\dots,n$, pro něž $X_i \in I_n(\cdot)$. Vzhledem k tomu, že výsledky pro $p>1$ jsou velmi nepřehledné, budeme v dalším textu pro větší názornost většinu výsledků presentovat pouze pro případ $p=1$. Zájemce o obecný tvar odkážeme vždy na příslušnou literaturu.

(b) První známé použití této metody lze najít už v Pearsonově práci v Biometrice z roku 1903, v níž studoval závislost výšky syna na výšce otce. Na základě více než 1000 pozorování rozdělených do tříd po 1" zde Pearson "statisticky" dokazuje platnost "obecného zákona regrese" vysloveného Galtonem v roce 1889: "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 je to v této Pearsonově práci, kde je slovo regrese poprvé použito k označení "podmíněného očekávání", stejně jako poprvé je užito výrazu "regresní přímka". Pearsonovy výsledky jsou znázorněny na následujícím obrázku. Je zajímavé přitom všimnout, že Pearsonův regresogram přináší pro extrémní (ať již malé či velké) výšky otců více informace, než regresní přímka proložená daty.



Proložená regresní přímka : $y = 33.73 + 0.516x$

(c) První rigorózní definici regresogramu a studium jeho vlastností lze nalézt až u Tukey (1961). Můžeme zde přitom vysledovat vliv nastupující éry počítačů, které nabídly reálnou možnost aplikací neparametrických odhadů a podnítily zároveň zájem o zkoumání jejich vlastností. První opravdu důsledné studium vlastností regresogramu podal Bosq (1970).

(d) Zmiňme se zde ještě o jednom zobecnění regresogramu, navrženém a studovaném Bhattacharyou a Parthasarathym (1961). Idea spočívá (pro $p = 1$) v tom, rozdělit interval $[a,b]$ obsahující všechna X_i, $i=1,\dots,n$, na l_n disjunktních podintervalů, z nichž každý obsahuje právě k_n bodů X_i, $i=1,\dots,n$. $\phi(x)$ pak odhadneme funkcí po částech konstantní, která v každém podintervalu zvoleného dělení je rovna průměru těch Y_i, pro něž odpovídající X_i leží v daném podintervalu. Tento odhad je někdy nazýván <u>regresogram s náhodným krokem</u>. Jeho vlastnosti jsou blízké vlastnostem regresogramu a nebudeme se zde jimi blíže zabývat.

Vzhledem k tvrzení dokázaném v Bosq (1970) žádný z námi studovaných odhadů není nestranný. Lze však nalézt podmínky, za nichž studované odhady jsou asymptoticky nestranné. Pro regresogram je udává následující věta, z níž zároveň vyplývá, že jde o postačující podmínku pro konzistenci regresogramu podle kvadratického středu.

<u>Věta 2.1.</u> Nechť $\phi(.)$ je spojitá v bodě x a nechť dále

$$\lim_{n\to\infty} h_n = 0 \quad \text{a} \quad \lim_{n\to\infty} n\, h_n^p = +\infty. \tag{2.2}$$

Potom $\hat{\phi}_n(x)$ je asymptoticky nevychýlený odhad $\phi(x)$ a je konzistentní podle kvadratického středu.

<u>Důkaz:</u> Bosq (1970), Collomb (1977a).

<u>Poznámka 2.2.</u> (a) Podmínka (2.2) intuitivně požaduje, aby s rostoucím n se délka intervalů $I_{n,m}$ sice zmenšovala, ale zároveň požaduje, aby nám zbývalo neustále dostatek pozorování v jednotlivých intervalech pro odhad. Uvažujeme-li posloupnost $\{h_n\}_{n=1}^{\infty}$ tvaru $c\, n^{-\gamma}$, (2.2) je zřejmě splněno pro $0 < \gamma < 1/p$.

(b) V Bosqově práci je též uvedena postačující podmínka, za níž

$$M_n = \sup_x |\hat{\phi}_n(x) - \phi(x)| \longrightarrow 0 \text{ s. j.}, \quad n \longrightarrow \infty. \tag{2.3}$$

Podívejme se nyní blíže, jak velké je vychýlení regresogramu pro $p = 1$.

<u>Věta 2.2.</u> Nechť $p = 1$ a podmínky věty 2.1. jsou splněny, nechť dále $\phi'(x)$ existuje. Potom

$$\hat{b}_n(x) = E\hat{\phi}_n(x) - \phi(x) = \tag{2.4}$$
$$= \phi'(x) \cdot h_n \cdot \left\{ \left(\left[\frac{x}{h_n} \right] + \tfrac{1}{2} \right) h_n - x \right\} + o(h_n).$$

<u>Důkaz.</u> Bosq (1970).

<u>Poznámka 2.3.</u> (a) Větu (2.3) lze samozřejmě zformulovat i pro $p > 1$, viz např. Collomb (1977a) či Collomb (1978).

(b) Vztah (2.4) potvrzuje naši intuitivní představu, že vychýlení podstatně závisí nejenom na délce intervalu $I_n(.)$ a poloze bodu x vzhledem ke středu intervalu, ale i fakt, že vychýlení výrazně vzroste s růstem strmosti odhadované křivky $\phi(x)$ a bude tím menší, čím plošší bude odhadovaná křivka.

Podívejme se nyní na rozptyl a střední kvadratickou chybu regresogramu.

<u>Věta 2.3.</u> Nechť jsou splněny podmínky věty 2.1. a nechť existuje podmíněný rozptyl Y vzhledem k X, tj.

(2.5) $v(.) = E((Y - E(Y|X = .))^2 | X = .)$,

který je spojitý v bodě x. Potom

(2.6) $$\hat{v}_n(x) = E(\hat{\phi}_n(x) - E\,\hat{\phi}_n(x))^2 = \frac{1}{n\,h_n P}\,\frac{v(x)}{f(x)} + o\left(\frac{1}{n\,h_n P}\right) .$$

Důkaz. Collomb (1977a).

Spojením výsledků vět (2.2) a (2.3) lze nyní okamžitě získat odhad střední kvadratické chyby regresogramu $\hat{g}_n(x)$ v bodě x, využitím známého vztahu

$$\hat{g}_n(x) = \hat{b}_n^2(x) + \hat{v}_n(x).$$

Větu opět vyslovíme pouze pro $p = 1$, i když ji lze zformulovat i pro $p > 1$.

Věta 2.4. Nechť jsou splněny podmínky vět 2.2. a 2.3. Je-li $p = 1$, potom

(2.7) $$\hat{g}_n(x) = E(\hat{\phi}_n(x) - \phi(x))^2 = \frac{1}{nh_n}\,\frac{v(x)}{f(x)} +$$

$$+ \phi'^2(x) \,.\, h_n^2 \,.\, \left\{\left(\left[\frac{x}{h_n}\right] + \frac{1}{2}\right) h_n - x\right\}^2 + o(h_n^2) + o\left(\frac{1}{n\,h_n}\right) .$$

Důkaz: Plyne bezprostředně z vět (2.2) a (2.3)

Poznámka 2.4. (a) Z (2.4) a (2.6) okamžitě vyplývá, že nelze vybrat posloupnost $\{h_n\}_{n=1}^{\infty}$ tak, aby současně zmenšovala jak rozptyl, tak vychýlení regresogramu. Z (2.7) lze dále snadno ukázat (pro $p=1$):

(2.8) jestliže $\lim_{n\to\infty} n\,h_n^3 = 0$, pak $\hat{g}_n(x) \sim \hat{v}_n(x)$, $n\to\infty$;
zatímco
jestliže $\lim_{n\to\infty} n\,h_n^3 = +\infty$, pak $\hat{g}_n(x) \sim \hat{b}_n^2(x)$, $n\to\infty$.

(b) Úvahy o volbě optimální posloupnosti $\{h_n\}_{n=1}^{\infty}$, tj. volbě optimální "délky okénka" rozkladu I_n, jsou velice podobné úvahám pro volbu optimální posloupnosti $\{h_n\}_{n=1}^{\infty}$ pro jádrové odhady. Vzhledem k tomu, že regresogram lze považovat především za odhad pomocný (a prvotní), nebudeme zde volbou optimální posloupnosti $\{h_n\}_{n=1}^{\infty}$ diskutovat.

(c) Ve všech předchozích případech jsme uvažovali pouze lokální chování regresogramu, nikoliv chování globální. Na toto by se však u odhadů tohoto typu nemělo zapomínat, neboť "parametrem", který odhadujeme, je vlastní regresní křivka, a název neparametrické metody zde tedy není úplně přesný. Pro posouzení "celkové kvality" odhadu jsou pak mnohem cennější míry tvaru

(2.10) $$\int_A |\hat{\phi}_n(x) - \phi(x)|\,\lambda(x)\,dx ,$$

(2.11) $$\int_A (\hat{\phi}_n(x) - \phi(x))^i\,\lambda(x)\,dx , \qquad i = 1,2,3,$$

či

(2.12) $$\max_{x\in A} |\hat{\phi}_n(x) - \phi(x)| \qquad \text{etc. ,}$$

kde A je nosič $\phi(x)$ a $\lambda(x)$, $x \in A$, je některá vhodná váhová funkce. Vzhledem k pomocnému charakteru regresogramu se však u této otázky nebudeme zastavovat.

Následující věta nám udává podmínky pro asymptotickou normalitu regresogramu, která nám umožňuje konstruovat (asymptotické) intervaly spolehlivosti pro $\phi(x)$. Větu opět vyslovíme pro $p = 1$.

Věta 2.4. Nechť $p = 1$ a podmínky věty 2.2. jsou splněny. Nechť dále

$$\lim_{n\to\infty} n\, h_n^3 = 0 \quad \text{a} \quad \lim_{n\to\infty} n\, h_n = +\infty . \tag{2.13}$$

Potom

$$\lim_{n\to\infty} P\left(\frac{\hat{\phi}_n(x) - \phi(x)}{\sqrt{\hat{W}_n(x)}} < t\right) = \Phi(t) , \tag{2.14}$$

kde $\Phi(t)$ je distribuční funkce $N(0,1)$ a

$$\hat{W}_n(x) = \frac{\sum_{i=1}^{n} (Y_i - \hat{\phi}_n(x))^2\, \mathbb{1}_{\{X_i \in I_n(x)\}}}{\left(\sum_{i=1}^{n} \mathbb{1}_{\{X_i \in I_n(x)\}}\right)^2} \tag{2.15}$$

Důkaz: Collomb (1978).

Důsledek: Pro každé $\alpha \in (0,1)$ je interval

$$\left[\hat{\phi}_n(x) - u_{1-\frac{\alpha}{2}} \cdot \sqrt{\hat{W}_n(x)} ,\quad \hat{\phi}_n(x) + u_{1-\frac{\alpha}{2}} \cdot \sqrt{\hat{W}_n(x)}\right] , \tag{2.16}$$

kde u_α je α-kvantil $N(0,1)$, asymptotickým dvojstranným intervalem spolehlivosti pro $\phi(x)$ s koeficientem spolehlivosti $1 - \alpha$.

Poznámka 2.5. (a) Tvar věty 2.4 pro $p > 1$ lze nalézt např. v práci Collomb (1978).

(b) Podmínka (2.13) je značně omezující na volbu posloupnosti $\{h_n\}_{n=1}^{\infty}$. Uvažujeme-li h_n tvaru $c\, n^{-\gamma}$, pak (2.13) je zřejmě splněna pokud $\frac{1}{3} < \gamma < 1$.

(c) Jsou-li splněny podmínky věty 2.4. na některém intervalu G, pak pomocí (2.16) můžeme sestrojit $\forall x \in G$ odpovídající interval spolehlivosti. Výsledkem pak bude pás složený z intervalů spolehlivosti.

POZOR: Nejedná se o pás spolehlivosti pro $\phi(x)$ v pravém slove smyslu.

(d) Za podmínek věty 2.4 dále platí, že

$$\hat{W}_n(x) \longrightarrow v(x) \quad \text{s. j.,} \quad n \to \infty . \tag{2.17}$$

2.1. VÝHODY A NEVÝHODY REGRESOGRAMU

Největší výhodou regresogramu je snadnost výpočtu spojená s nenáročností na paměť počítače a rychlostí, s níž získáme prvotní přehled o našich datech. Tuto vlastnost oceníme především tehdy, nemáme-li k disposici plotter pro vykreslení našich dat, jakož i v případě, kdy máme zpracovat stovky a tisíce neznámých dat.

Za tyto výhody však nutně platíme:

- získaná křivka je, na rozdíl od ostatních neparametrických odhadů, poměrně hrubá;
- metoda je vysoce nerobustní vzhledem k odlehlým pozorováním v Y;
- získaný odhad je pro většinu bodů dosti vychýlený (zvláště pro strmé úseky $\phi(x)$) atd.

Na druhé straně však neustále musíme mít na zřeteli ten fakt, že regresogram by měl být pouze jednou z prvních analýz našich dat a neočekávat od něj zázraky. V tomto okamžiku naopak vysokou nerobustnost oceníme, neboť příliš velké "kmitání", jakož i divoké a neočekávané skoky a oscilace získaného odhadu s největší pravděpodobností signalizují výskyt vážných chyb v našich datech spíše než cokoliv jiného.

3. JÁDROVÝ ODHAD

Nejznámější a nejlépe prostudovanou třídou neparametrických odhadů pro $\phi(x)$ tvoří tzv. jádrové (kernel) odhady, navržené nezávisle Nadarayou (1964) a Watsonem (1964). Dříve než přikročíme k jejich definici, zaveďme si nejprve pojem jádra.

Definice 3.1. Jádrem nazveme libovolnou funkci $K: (R^p, B^p) \to (R^1, B^1)$ takovou, že je symetrická $(K(x) = K(-x)\ \forall x \in R^p)$, nezáporná, ohraničená a

$$(3.1) \qquad \int_{R^p} K(x)\, dx = 1 \qquad \text{a} \qquad \lim_{|x| \to \infty} |x|^p \cdot K(x) = 0.$$

Poznámka 3.1. O některých nejpoužívanějších tvarech jádra, jejich vlastnostech a o výběru "optimálního" jádra bude pojednáno v odstavci 3.1.

Nyní již můžeme přistoupit k definici jádrových odhadů.

Definice 3.2. Nechť $(X_i, Y_i)_{i=1}^n$ je posloupnost nezávislých pozorování náhodného vektoru (X, Y) v prostoru (R^{p+1}, B^{p+1}). Nechť $\mathcal{L}(X_i, Y_i) \sim \mathcal{L}(X, Y) \sim P_{XY}$ $\forall i$, $i = 1, \dots, n$. Nechť $P_{XY} \ll \nu$, $p(x,y) = \frac{dP_{XY}}{d\nu}$ (kde ν je Lebesgueova míra na R^{p+1}) a nechť $\forall x \in R^p$ $f(x) = \int p(x,y)\, dy \neq 0$. Předpokládejme, že $\phi(x) = E(Y | X = x)$ existuje a uvažujme problém odhadu $\phi(x)$. Budiž $K(.)$ některé jádro a $\{h_n\}_{n=1}^{\infty}$ posloupnost nezáporných čísel.

Potom jádrový odhad $\tilde{\phi}_n(x)$ je definován $\forall x \in R^p$ vztahem

$$(3.2) \qquad \tilde{\phi}_n(x) = \begin{cases} \dfrac{\sum_{i=1}^{n} Y_i\, K\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)} & \text{jestliže } \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \neq 0 \\ 0 & \text{jinak.} \end{cases}$$

Poznámka 3.2. (a) Z tvaru (3.2) okamžitě vidíme, že zavedení jádra realizuje naši intuitivní snahu preferovat pro odhad ta pozorování, která jsou blíže bodu v němž odhadujeme. Úloha posloupnosti $\{h_n\}_{n=1}^{\infty}$ je přitom táž jako v případě regresogramu.

(b) Jádrový odhad je v podstatě vážený průměr pozorování Y_i, v němž váhy závisí pouze na vzdálenosti odpovídajících X_i od bodu, v němž odhadujeme. Z tohoto hlediska je zřejmé, že se stále jedná o odhad vysoce nerobustní vzhledem k odlehlým pozorováním v Y_i. Chceme-li získat odhad robustní, je třeba nahradit v definici (3.2) výběrový průměr vážený vzhledem k vzdálenosti X_i od bodu v němž odhadujeme některým odhadem robustním navíc vzhledem k hodnotám Y_i, např. α-useknutým průměrem či M-odhadem při současném zachování vah pomocí jádra K. Je však naprosto zřejmé, že časová náročnost na centrální jednotku počítače pro výpočet takovýchto odhadů enormě stoupne a zůstává otázkou, zda zvýšená "hladkost odhadů" vyrovná cenu, již musíme za výpočet zaplatit. Neparametrickými M-odhady se zabýval např. Härdle (1983).

(c) Zatímco regresogram je odhad spíše popisný , poskytující nám především hrubý odhad $\phi(x)$ a neumožňující téměř žádnou přesnou predikci, u jádrového odhadu tomu bývá naopak. Kromě mnohem detailnějšího popisu odhadované regresní křivky zpravidla slouží právě pro predikci (resp. může sloužit). Je-li $\phi(x)$ spojitá (resp. diferencovatelná řádu $g, g \in N$), regresogram $\hat{\phi}_n(x)$ je nevhodný odhad, neb není spojitý (ani diferencovatelný), zatímco zvolíme-li spojité jádro (resp. diferencovatelné řádu g), jádrový odhad bude spojitý, byť ne vždy ideálně hladký (resp. diferencovatelný řádu g).

Jádrový odhad, podobně jako regresogram, je odhad vychýlený. Následující věta udává postačující podmínky pro asymptotickou nevychýlenost $\tilde{\phi}_n(x)$, které jsou zároveň postačující pro konzistenci jádrového odhadu podle kvadratické středu.

<u>Věta 3.1.</u> Nechť $\phi(.)$ a $f(.)$ jsou spojité v bodě x, $f(x) \neq 0$ a nechť dále

(3.3) $$\lim_{n \to \infty} h_n = 0 \qquad \text{a} \qquad \lim_{n \to \infty} n\, h_n^p = +\infty .$$

Potom $\tilde{\phi}_n(x)$ je asymptoticky nevychýlený odhad $\phi(x)$ a je konzistentní podle kvadratického středu.

<u>Důkaz:</u> Bosq (1970), Collomb (1977a).

<u>Poznámka 3.3.</u> (a) Podmínka (3.3) je pro posloupnost h_n tvaru $cn^{-\gamma}$ zřejmě splněna pro $0 < \gamma < 1/p$.

(b) V práci Devroye a Wagner (1979) lze najít obecné podmínky, za nich pro jádrový odhad $\tilde{\phi}_n(x)$ a $g \geq 1$ platí

(3.4) $$E(\tilde{\phi}_n(x) - \phi(x))^g \longrightarrow 0, \qquad n \longrightarrow \infty .$$

Označme si

(3.5) $$M_n = \sup_{x \in G} \left| \tilde{\phi}_n(x) - \phi(x) \right| ,$$

kde G je některá ohraničená množina v R^p a podívejme se, za jakých podmínek $M_n \longrightarrow 0$.

<u>Věta 3.2</u> Nechť $\phi(x)$ je spojitá pro $\forall x \in G$.

Jestliže

(3.6) $$\lim_{n \to \infty} h_n = 0 \qquad \text{a} \qquad \lim_{n \to \infty} \frac{n\, h_n^{2p}}{\log n} = +\infty ,$$

potom

(3.7) $$M_n \longrightarrow 0 \qquad \text{s.j.,} \quad n \longrightarrow \infty .$$

<u>Důkaz:</u> Devroye (1979).

Podívejme se nyní blíže , jak velké je vychýlení jádrového odhadu. Pro přehlednost se omezíme pouze na případ $p = 1$.

<u>Věta 3.3.</u> Nechť $p = 1$ a podmínky věty 3.1. jsou splněny. Nechť dále buď

(a) ϕ (resp. f) má derivace 2. (resp. 1.) řádu v bodě x v případě, je-li nosič jádra ohraničený;

nebo

(b) ϕ (resp. f) má ohraničené derivace 3. (resp. 2.) řádu v bodě x a $\int |z|^5 K(z)\, dz < +\infty$, není-li nosič jádra ohraničený.

Potom

$$(3.8) \qquad \widetilde{b}_n(x) = E\,\widetilde{\phi}_n(x) - \phi(x) = h_n^2\left(\frac{\phi''(x)}{2} + \phi'(x)\,.\,\frac{f'(x)}{f(x)}\right)\int z^2K(z)\,dz + o(h_n^2).$$

Důkaz: Collomb (1977a).

Poznámka 3.4. (a) V citované Collombově práci lze nalézt i obecný výsledek pro $p > 1$.
(b) Z tvaru (3.8) je zřetelně vidět, jakým způsobem vychýlení závisí na tvaru odhadované regresní křivky $\phi(x)$ a lokálním chování hustoty $f(x)$. Mj. lze říci, že jádrový odhad podhodnocuje $\phi(x)$ v okolí lokálního maxima konkávních funkcí a nadhodnocuje $\phi(x)$ v okolí lokálního minima konvexních funkcí. Vychýlenost přitom bude tím menší, čím plošší je odhadovaná $\phi(x)$. Viz též obrázky ilustrující numerický příklad v odstavci 5.

Podívejme se nyní na rozptyl a střední kvadratickou chybu jádrových odhadů.

Věta 3.4. Nechť jsou splněny podmínky věty 3.1. a nechť dále existuje podmíněný rozptyl Y vzhledem k X, tj.

$$(3.9) \qquad v(.) = E((Y - E(Y|X = .))^2 \mid X = .),$$

který je spojitý v bodě x. Potom

$$(3.10) \qquad \widetilde{v}_n(x) = \frac{1}{n\,h^p}\,.\,\frac{v(x)}{f(x)}\int K^2(z)\,dz + o\left(\frac{1}{n\,h_n^p}\right).$$

Důkaz: Collomb (1977a).

Spojením vět 3.3. a 3.4. okamžitě dostaneme odhad střední kvadratické chyby jádrového odhadu $\widetilde{\phi}(.)$ v bodě x (vzhledem k větě 3.3. pouze pro $p = 1$).

Věta 3.5. Nechť $p = 1$ a podmínky vět 3.3. a 3.4 jsou splněny. Potom

$$(3.11) \qquad \widetilde{g}_n(x) = E(\widetilde{\phi}_n(x) - \widetilde{\phi}(x))^2 = \frac{1}{n\,h_n}\,.\,\frac{v(x)}{f(x)}\int K^2(z)\,dz + h_n^4\left(\frac{\phi''(x)}{2} + \phi'(x)\,.\,\frac{f'(x)}{f(x)}\right)^2\left(\int z^2K(z)\,dz\right)^2 + o\left(\frac{1}{n\,h_n}\right) + o(h_n^4).$$

Důkaz: Plyne bezprostředně z vět 3.3. a 3.4.

Poznámka 3.5. (a) Z (3.8) a (3.10) okamžitě vyplývá, že nelze vybrat posloupnost $\{h_n\}_{n=1}^{\infty}$ tak, aby současně zmenšovala jak rozptyl, tak vychýlení jádrového odhadu. Z (3.11) lze dále snadno ukázat (pro $p = 1$):

$$(3.12) \qquad \text{jestliže} \quad \lim_{n\to\infty} n\,h_n^5 = 0, \quad \text{pak} \quad \widetilde{g}_n(x) \sim \widetilde{v}_n(x), \quad n \to \infty;$$

zatímco

$$(3.13) \qquad \text{jestliže} \quad \lim_{n\to\infty} n\,h_n^5 = +\infty, \quad \text{pak} \quad \widetilde{g}_n(x) \sim \widetilde{b}_n^2(x), \quad n \to \infty.$$

Srovnej s poznámkou 2.4. (a).

(b) Nechť $\lambda(x)$ je některá vhodná váhová funkce a jsou splněny určité podmínky na funkce $\phi''(x)$ a $f'(x)$, umožňující integraci ve výrazu (3.11). Podívejme se na globální míru chování jádrového odhadu $\widetilde{\phi}_n(x)$ typu integrované střední kvadratické chyby. Lze ukázat, že

(3.14) $$\int_{R^1} E(\tilde{\phi}_n(x) - \phi(x))^2 \lambda(x)\, dx = \frac{A}{n h_n} + B h_n^4 + o(h_n^4) + o(\frac{1}{n h_n}), \quad n \to \infty ;$$

odtud plyne, že

(3.15) $$\min_{\{h_n\}>0} \int_{R^1} E(\tilde{\phi}_n(x) - \phi(x))^2 \lambda(x)\, dx \sim C n^{-4/5}, \quad n \to \infty ,$$

kde A, B a C jsou kladné konstanty nezávislé na n. Srovnej též s výrazy (3.25) a (3.26).

(c) Jedna z možností, jak volit "optimální" posloupnost $\{h_n\}_{n=1}^{\infty}$ vychází přímo z minimalizace $\tilde{g}_n(x)$. Blíže viz. odstavec 3.1.

Následující věta nám udává podmínky pro asymptotickou normalitu jádrového odhadu, která nám zároveň umožňuje konstruovat (asymptotické) intervaly spolehlivosti pro $\phi(x)$. Větu opět vyslovíme pro $p = 1$.

Věta 3.6. Nechť $p = 1$ a podmínky věty 3.3. jsou splněny. Nechť dále

(3.16) $$\lim_{n \to \infty} n h_n = +\infty \quad \text{a} \quad \lim_{n \to \infty} n h_n^5 = 0.$$

Potom

(3.17) $$\lim_{n \to \infty} P\left(\frac{\tilde{\phi}_n(x) - \phi(x)}{\sqrt{\tilde{W}_n(x)}} < t\right) = \Phi(t),$$

kde $\Phi(t)$ je distribuční funkce $N(0,1)$ a

(3.18) $$\tilde{W}_n(x) = \frac{\sum_{i=1}^{n} (Y_i - \tilde{\phi}_n(x))^2 K(\frac{x - X_i}{h_n})}{(\sum_{i=1}^{n} K(\frac{x - X_i}{h_n}))^2} \int K^2(z)\, dz.$$

Důkaz: Collomb (1977b).

Důsledek. Pro každé $\alpha \in (0,1)$ je interval

(3.19) $$\left[\tilde{\phi}_n(x) - u_{1-\frac{\alpha}{2}} \sqrt{\tilde{W}_n(x)},\ \tilde{\phi}_n(x) + u_{1-\frac{\alpha}{2}} \sqrt{\tilde{W}_n(x)} \right],$$

kde u_α je α-kvantil $N(0,1)$, asymptotickým dvojstranným intervalem spolehlivosti pro $\phi(x)$ s koeficientem spolehlivosti $1 - \alpha$.

Poznámka 3.6. (a) Tvar věty 3.6. pro $p > 1$ s podrobným důkazem lze nalézt v Collomb (1977b).

(b) Podmínka (3.16) je značně omezující na volbu posloupnosti $\{h_n\}_{n=1}^{\infty}$. Uvažujeme-li h_n tvaru $c\, n^{-\gamma}$, pak (3.16) je zřejmě splněn pokud $1/5 < \gamma < 1$.

(c) Jsou-li splněny podmínky věty 3.6. na některém intervalu G, pak pomocí (3.19) můžeme sestrojit $\forall x \in G$ odpovídající interval spolehlivosti. Výsledkem pak bude pás složený z intervalů spolehlivosti.

POZOR: Opět se nejedná o pás spolehlivosti pro $\phi(x)$ v pravém slova smyslu.

(d) Za podmínek věty 3.6. lze dále mj. ukázat, že

(3.20) $$\frac{\widetilde{W}_n(x)}{v_n(x)} \longrightarrow 1 \quad \text{s. j.}, \quad n \longrightarrow \infty ,$$

(3.21) $$\frac{n\, h_n^p}{\int K^2(z)\,dz}\ \widetilde{W}_n(x) \longrightarrow \frac{v(x)}{f(x)} \quad \text{s.j.}, \quad n \longrightarrow \infty .$$

3.1. VOLBA OPTIMÁLNÍHO JÁDRA A POSLOUPNOSTI $\{h_n\}_{n=1}^{\infty}$ (případ $p = 1$).

Jak zřetelně vyplývá z předchozího textu, jednou z nejdůležitějších otázek zůstává volba optimálního jádra a posloupnosti $\{h_n\}_{n=1}^{\infty}$. Podívejme se nyní proto na tuto otázku podrobněji.

Klasický přístup k této otázce spočívá ve volbě toho jádra, které minimalizuje střední kvadratickou chybu odhadu. Označme si K_1 třídu těch jader dle definice 3.1., pro která navíc $\int z^2 K(z)\,dz = 1$ a žádná dvě různá jádra z K_1 nelze odvodit jedno z druhého pouhou změnou měřítka na osách. Hledejme nyní ve třídě K_1 to jádro $K^*(z)$, jež minimalizuje střední kvadratickou chybu odpovídajícího odhadu. Tento úkol lze poměrně snadno vyřešit pomocí věty 3.5. a následujícího lemmatu, dokázaného ve spojitosti s neparametrickými odhady hustoty.

<u>Lemma 3.1.</u> Problém

$$\min_{K \in K_1} \int K^2(z)\,dz$$

má řešení $K^*(z)$ definované vztahem

(3.22) $$K^*(z) = \begin{cases} 0 & |z| \geq \sqrt{5} \\ \dfrac{3}{4\sqrt{5}} - \dfrac{3z^2}{20\sqrt{5}} & |z| < \sqrt{5} \end{cases}$$

<u>Důkaz:</u> Epanechnikov (1969).

Vrátíme-li se nyní zpět k (3.11), vidíme, že odhad s jádrem $K^*(z)$ definovaným vztahem (3.22) skutečně minimalizuje asymptoticky $\widetilde{g}_n(x)$ a jádro $K^*(z)$ je tudíž v daném smyslu optimální. Daný výsledek je prakticky důležitý, neboť $K^*(z)$ nijak nezávisí na P_{XY}.

Podívejme se nyní na to, jak se zvětší rozptyl odhadu, užijeme-li místo K^* jiné jádro $K \in K_1$. V tabulce 3.1., která nám mj. udává přehled některých nejčastěji doporučovaných jader, najdeme hodnotu konstanty c $\left[c = \int K^2(z)\,dz \Big/ \int K^{*2}(z)\,dz \right]$ udávající, kolikrát je větší rozptyl jádrového odhadu, nahradíme-li optimální Epanechnikovovo jádro K^* jádrem K. Z výsledků tabulky 3.1. můžeme konstatovat, že rozdíly mezi jádry nejsou příliš výrazné. Zvláště příjemně to působí u odhadu typu klouzavného okénka, který je v praxi běžně používán.

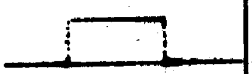
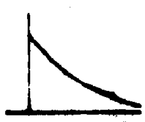
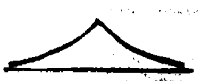
Z praktického hlediska můžeme říci, že pro prvotní ohledání dat spravidla stačí odhad typu klouzavého okénka, který lze na uspořádaných datech realizovat velice rychle. Užitím odhadu s optimálním jádrem Epanechnikovova typu mnoho na přesnosti nezískáme (viz též tab. 3.1), zato výrazně prodloužíme čas potřebný k výpočtu. Ideálním řešením by byl pro úlohy tohoto typu maticový koprocesor, který však zpravidla k dispozici nemáme. Připomeňme, že uspořádání dat je nezbytně potřebné pouze pro ty druhy jádrových odhadů, které užívají jádro s konečným nosičem. Vzhledem k existenci rychlých třídících algoritmů se nám "cena" za uspořádání rychle vrátí.

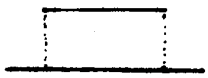
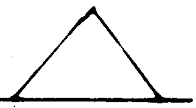
Společnou nevýhodou všech jádrových odhadů s jádrem majícím konečný nosič je to, že špatně odhadují tvar regresní křivky na krajích jejího nosiče. Z tohoto hlediska lepší výsledky dají odhady s jádrem majícím nekonečný nosič (např. hustotu Cauchyho rozdělení, N(0,1) etc.), která jsou schopny mnohem přesněji zvládnout tvar počátku (konce) odhadované křivky (zvláště pokud $\phi(x)$ je zde podstatně různá od nuly. Pro menší úlohy (řádově do 10^3 pozorování) lze doporučit použít přímo klouzavého okénka jádrový odhad s jádrem tvaru hustota Cauchyho rozdělení či jiné jádro, jež má za nosič R^1 a lze jej rychle počítat. Odpadne uspořádání, kraje regresní křivky budou odhadnuty a čas potřebný pro výpočet zůstane poměrně rozumný (nebudeme-li ovšem odhadovat příliš hustě).

Označme

(3.23)
$$c = \frac{\int K^2(z)\,dz}{\int K^{*2}(z)\,dz} .$$

__Tabulka 3.1.__

i	K(y)	obor y	Tvar jádra	c
0	$\frac{3}{4\sqrt{5}} - \frac{3y^2}{20\sqrt{5}}$ 0	$\lvert y\rvert \le \sqrt{5}$ $\lvert y\rvert > \sqrt{5}$		1
1	$\frac{1}{4}\sqrt{\pi^2-8}\cdot\cos\frac{\sqrt{\pi^2-8}\,y}{2}$ 0	$\lvert y\rvert \le \frac{\pi}{\sqrt{\pi^2-8}}$ $\lvert y\rvert > \frac{\pi}{\sqrt{\pi^2-8}}$		1.001
2	$\frac{1}{\sqrt{6}} - \frac{\lvert y\rvert}{6}$ 0	$\lvert y\rvert \le \sqrt{6}$ $\lvert y\rvert > \sqrt{6}$		1.015
3	$\frac{1}{\sqrt{2\pi}}\, e^{-\frac{y^2}{2}}$	$y \in R_1$		1.051
4	$\frac{1}{2\sqrt{3}}$ 0	$\lvert y\rvert \le \sqrt{3}$ $\lvert y\rvert > \sqrt{3}$		1.077
5	$\frac{1}{\sqrt{2}}\, e^{-\sqrt{2}\,y}$ 0	$y > 0$ $y < 0$		1.320
6	$\frac{1}{2}\, e^{-\lvert y\rvert}$	$y \in R^1$		1.863

7	$\frac{1}{2}$ 0	$\lvert y\rvert \leq 1$ $\lvert y\rvert > 1$		1.772
8	$1 - \lvert y\rvert$ 0	$\lvert y\rvert \leq 1$ $\lvert y\rvert > 1$		2.362
9	$\frac{1}{\pi} \cdot \frac{1}{1+y^2}$	$y \in R^1$		1.186

O jádrech vhodných pro případ $p > 1$ se lze dočíst v pracech Elkins (1968) a Epanechnikov (1969). Výsledky, ač původně pro odhady hustoty, jsou snadno přenositelné pro jádrové odhady regresní křivky.

Nyní již můžeme přistoupit k hledání "optimální" posloupnosti $\{h_n\}_{n=1}^{\infty}$. Zvolíme-li střední kvadratickou chybu za kritérium kvality odhadu, pak s přihlédnutím k (3.11) a (3.22) můžeme říci, že odhad $\tilde{\phi}_n^*(x)$ takový, že $K = K^*$ a

(3.24) $$h_n = h_n^* = \left(\frac{3}{20\sqrt{5}}\right)^{1/5} \left\{ \frac{v(x)}{f(x).\left(\frac{\phi''(x)}{2} + \phi'(x).\frac{f'(x)}{f(x)}\right)^2} \right\}^{1/5} . n^{-1/5}, \quad n \in N,$$

je nejlepším odhadem ve třídě jádrových odhadů definovaných vztahem (3.2). Střední kvadratická chyba tohoto odhadu pak je

(3.25) $$\tilde{\varepsilon}_n^{*}(x) \sim \left[5^{1/5} . \left(\frac{3}{4\sqrt{5}}\right)^{4/5} \left(\frac{v(x)}{f(x)}\right)^{4/5} \left(\frac{\phi''(x)}{2} + \phi'(x) . \frac{f'(x)}{f(x)}\right)^{2/5} \right] n^{-4/5} .$$

Z tohoto výrazu plyne, že

(3.26) $$\min_{\{h_n\} \in R^1} E(\tilde{\phi}_n(x) - \phi(x))^2 \approx M(x) n^{-4/5}, \qquad n \to \infty,$$

kde $M(x)$ je kladná konstanta nezávisející na n.

Všechny asymptotické výsledky, které jsme zde uvažovali, nedovolují odpovědět na velmi důležitou praktickou otázku: "Jak pro pevné n vybrat optimální h?". Intuice nabízí zvolit takové h, pro které $\tilde{\phi}_n(x)$ opticky "co nejlépe" prokládá mrak dat $(X_i, Y_i)_{i=1}^n$. Nejdoporučovanějším postupem se zdá být metoda tzv. krosvalidace (cross validation). V podstatě jde o to zvolit h takové, jež by nám minimalizovalo empirickou kvadratickou chybu (bylo by však možné zvolit i některou jinou míru, např. empirickou absolutní chybu etc.).

Přesněji řečeno, je třeba nalézt takové $\hat{h}$, pro něž

(3.27) $$S_n(\hat{h}) = \min_{h > 0} S_n(h),$$

kde

$$S_n(h) = \sum_{j=1}^{n} (Y_j - \tilde{\varphi}_{n-1}(h, X_j))^2. \tag{3.28}$$

Zde $\tilde{\varphi}_{n-1}(h, X_j)$ značí jádrový odhad, počítaný podle (3.2) z (n-1) pozorování (X_i, Y_i), $i = 1, \ldots, n$, $i \neq j$ a s daným h v bodě $x = X_j$. Jádrový odhad, počítaný s h dle (3.27) bývá zván <u>krosvalidační jádrový odhad.</u>

Výše popsaná metoda, která je intuitivně jasná a prakticky důležitá, však s sebou přináší řadu problémů teoretických, blíže viz. např. Wang (1983), Hall (1984), Härdle a Marron (1984) či Collomb et all.(1985), jakož i numerických.

4. NĚKTERÉ DALŠÍ NEPARAMETRICKÉ ODHADY

V tomto odstavci se pro ilustraci krátce zmíníme o některých dalších typech neparametrických odhadů, aniž bychom se podrobně zabývali jejich vlastnostmi. Tyto si zájemce může nalézt v citované literatuře.

4.1. NEPARAMETRICKÝ ODHAD UŽÍVAJÍCÍ ORTOGONÁLNÍ FUNKCE

Tato metoda je analogií metody navržené Čencovem (1962) pro odhad hustoty, využívající některé ortogonální base v $L_2(\nu)$, kde ν je Lebesgueova míra. Příkladem takovéto base (pro $p=1$) je např. posloupnost Hermitových funkcí

$$e_i(x) = (2^i \cdot i! \sqrt{\pi}) \cdot H_i(x) \cdot \exp(-x^2/2), \tag{4.1}$$

kde

$$H_i(x) = \exp(x^2) \cdot \frac{d^i \exp(-x^2)}{d x^i}, \qquad i = 0, 1, \ldots; \tag{4.2}$$

nebo posloupnost trigonometrických funkcí (pro X nabývající hodnot v $[-\pi, \pi]$)

$$e_0(x) = \frac{1}{\sqrt{2\pi}}, \quad e_{2i-1}(x) = \frac{1}{\sqrt{2\pi}} \cos(ix) \quad \text{a} \tag{4.3}$$

$$e_{2i}(x) = \frac{1}{\sqrt{2\pi}} \sin(ix), \qquad i = 1, 2, \ldots$$

<u>Definice 4.1.</u> Nechť $(X_i, Y_i)_{i=1}^{n}$ je posloupnost nezávislých pozorování náhodného vektoru (X, Y) v prostoru (R^{p+1}, B^{p+1}). Nechť $\mathcal{L}(X_i, Y_i) \sim \mathcal{L}(X, Y) \sim P_{XY}$ $\forall i$, $i = 1, \ldots, n$. Nechť $P_{XY} \ll \nu$, $p(x, y) = \frac{dP_{XY}}{d\nu}$ (kde ν je Lebesgueova míra na R^{p+1}) a nechť $\forall x \in R^p$ $f(x) = \int p(x, y)\, dy \neq 0$. Předpokládejme, že $\varphi(x) = E(Y | X = x)$ existuje a uvažujme problém odhadu $\varphi(x)$. Nechť $\{e_j\}_{j=0}^{\infty}$ je některá ortogonální base $L_2(\nu)$. Potom odhad Čencovova typu je definován vztahem

$$\varphi_n^{o}(x) = \frac{\sum_{j=0}^{n} K_{nj}\, e_j(x)}{\sum_{j=0}^{n} D_{nj}\, e_j(x)} \tag{4.4}$$

kde

$$K_{nj} = \frac{1}{n} \sum_{i=1}^{n} Y_i \cdot e_j(X_i) \quad \text{a} \quad D_{nj} = \frac{1}{n} \sum_{i=1}^{n} e_j(X_i). \tag{4.5}$$

Vlastnosti odhadu $\overset{o}{\phi}_n(x)$ mohou být poměrně jednoduše vyvozeny z výsledků práce Bleues a Bosq (1978), zabývající se odhady podobného typu pro hustoty. Vzhledem k složitosti a časové náročnosti tohoto odhadu lze jen těžko počítat s tím, že se tato metoda prosadí výrazněji do praxe.

4.2. METODY UŽÍVAJÍCÍ SPLINOVÉ FUNKCE

Citace na práce zabývající se těmito typy neparametrických odhadů lze nalézt v Collomb (1985).

4.3. NEPARAMETRICKÉ SEKVENČNÍ ODHADY

Uvažujme tutéž situaci jako v odstavci 4.1. Potom, následujíc Ahmada a Lina (1976), sekvenční neparametrický odhad může být definován rekurzivním vztahem

(4.6) $$\hat{\phi}_n(x) = \frac{N_n(x)}{D_n(x)} = \frac{N_{n-1}(x) + Y_n Z_n}{D_{n-1}(x) + Z_n} ,$$

kde

(4.7) $$N_o(x) = D_o(x) = 0 \qquad \text{a} \qquad Z_n(x) = \frac{K((x - X_n)/h_n)}{h_n^p} ,$$

K je některé jádro a $\{h_n\}_{n=1}^{\infty}$ je posloupnost kladných čísel taková, že

(4.8) $$\lim_{n\to\infty} h_n = 0 \qquad \text{a} \qquad \lim_{n\to\infty} n\, h_n^p = +\infty .$$

Někteří autoři, např. Devroye a Wagner (1980), studují odhad (4.6), v němž ale podmínky (4.7) a (4.8) jsou nahrazeny "přirozenějšími" podmínkami

(4.9) $$N_o(x) = D_o(x) = 0 \qquad \text{a} \qquad Z_n(x) = K((x - X_n)/h_n),$$

(4.10) $$\lim_{n\to\infty} h_n = 0 \qquad \text{a} \qquad \sum_{n=1}^{\infty} h_n = +\infty .$$

Jiný přístup navrhl Révész ve svých pracech z let (1973) a (1977), v nichž studoval sekvenční neparametrický odhad Robbins-Monroova typu definovaný vztahem (pro $p = 1$)

(4.11) $$\bar{\phi}_n(x) = \bar{\phi}_{n-1}(x) + (Y_n - \bar{\phi}_{n-1}(x)) \cdot \frac{K((x - X_n)/h_n)}{n\, h_n} ,$$

kde

(4.12) $$\bar{\phi}_o(x) = 0 \qquad \text{a} \qquad h_n = n^{-\gamma}, \qquad 0 < \gamma < 1.$$

Citace na další práce zabývající se sekvenčními neparametrickými odhady lze naléz v práci Collomb (1985).

4.4. ODHAD TYPU k_n NEJBLIŽŠÍCH SOUSEDŮ (k - NN ODHAD)

Uvažujme opět tutéž situaci jako v odstavci 4.1. Potom k-NN odhad je definován, viz. Bhattacharya a Parthasaraty (1961), vztahem

(4.13) $$\tilde{\tilde{\phi}}_n(x) = \frac{1}{k_n} \sum_{j \in I_n(x)} Y_j , \qquad \forall x \in R^p ,$$

kde

(4.14) $\quad I_n(x) = \left\{ i \mid X_i \text{ je jedním z } k_n \text{ nejbližších pozorování vzhledem k } x \right\}.$

Od posloupnosti $\{k_n\}_{n=1}^{\infty}$ požadujeme, aby

(4.15) $$\lim_{n\to\infty} k_n = +\infty \quad \text{a} \quad \lim_{n\to\infty} \frac{k_n}{n} = 0.$$

V literatuře můžeme nalézt několik zobecnění odhadu $\tilde{\tilde{\varphi}}_n(x)$. Jedno z nich navrhl a podrobně studoval Royal (1976). Jeho odhad je definován vztahem

(4.16) $$\tilde{\tilde{\varphi}}_n^{1}(x) = \sum_{i=1}^{n} Y_i\, w_{ni}\,,$$

kde

(4.17) $$w_{ni} \geq 0,\ i = 1,\ldots,n\,, \quad \sum_{i=1}^{n} w_{ni} = 1\,,$$

$w_{ni} = v_{nj}$, kde j je pořadí $|x - X_j|$ v množině $\left\{|x - X_j|,\ j = 1,\ldots,n\right\}$. Posloupnost $\{v_{nj}\}_{j=1}^{n}$ musí přitom splňovat $\forall n \in N$ podmínku

(4.18) $$v_{n1} \geq \ldots \geq v_{nn} \quad \text{a} \quad \lim_{n\to\infty} \max_{1\leq j\leq n} v_{nj} = 0.$$

Jedná se tedy s váženou versí odhadu $\tilde{\tilde{\varphi}}_n(x)$, kde váhy u pozorování klesají spolu se stoupajícím pořadím vzdálenosti X_i od bodu x v rámci $I_n(x)$. Zvolíme-li

(4.19) $$v_{ni} = \frac{1}{k_n} \qquad i = 1,\ldots,k_n\,,$$
$$= 0 \qquad i = k_n + 1,\ldots,n\,,$$

potom zřejmě odhady $\tilde{\tilde{\varphi}}_n(x)$ a $\tilde{\tilde{\varphi}}_n^{1}(x)$ jsou totožné.

Jiné "zobecnění" odhadu $\tilde{\tilde{\varphi}}_n(x)$ studoval např. Mack (1982). Jeho odhad je definován vztahem

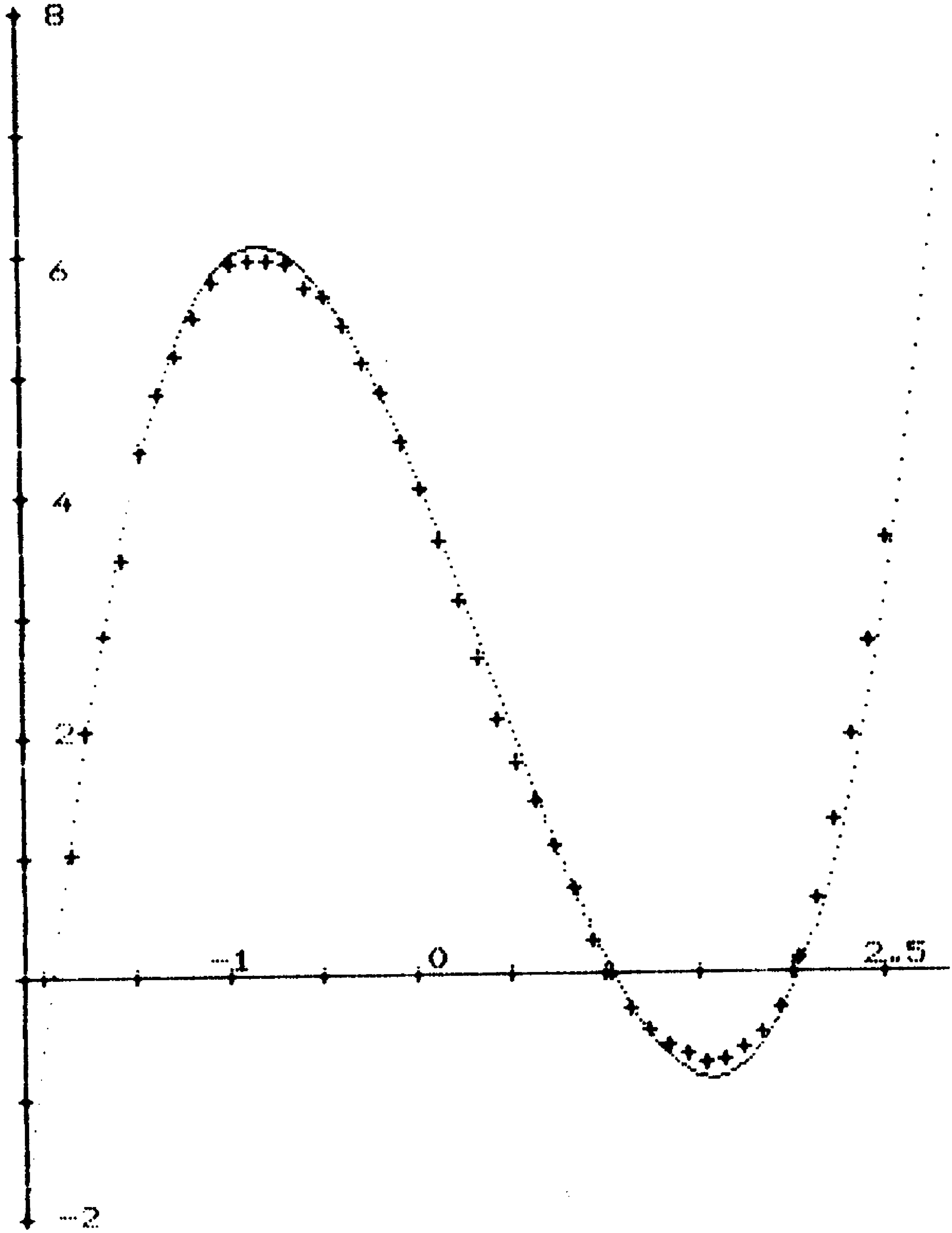
(4.20) $$\tilde{\tilde{\varphi}}_n^{2}(x) = \frac{\sum_{i=1}^{n} Y_i \cdot K\left(\frac{x - X_i}{H_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{H_n}\right)}\,, \qquad x \in R^p,$$

kde K je některé jádro definované v odstavci 3. a H_n je euklidovská vzdálenost mezi bodem x a k_n-tým nejbližším sousedem mezi $\{X_i,\ i = 1,\ldots,n\}$.

Vlastnosti k-NN odhadů a jádrových odhadů si jsou, dle očekávání, dosti podobné, důkazy vycházejí z velmi si blízkých předpokladů a postupují analogicky. Řády konvergence pro vychýlení, rozptyl (a tudíž i střední kvadratickou chybu) jsou stejné. Z výpočetního hlediska lze říci, že na předem uspořádaných datech je výpočet k-NN odhadů výrazně rychlejší než výpočet jádrových odhadů. Podrobný soupis prací zabývajících se k-NN viz Collomb (1985).

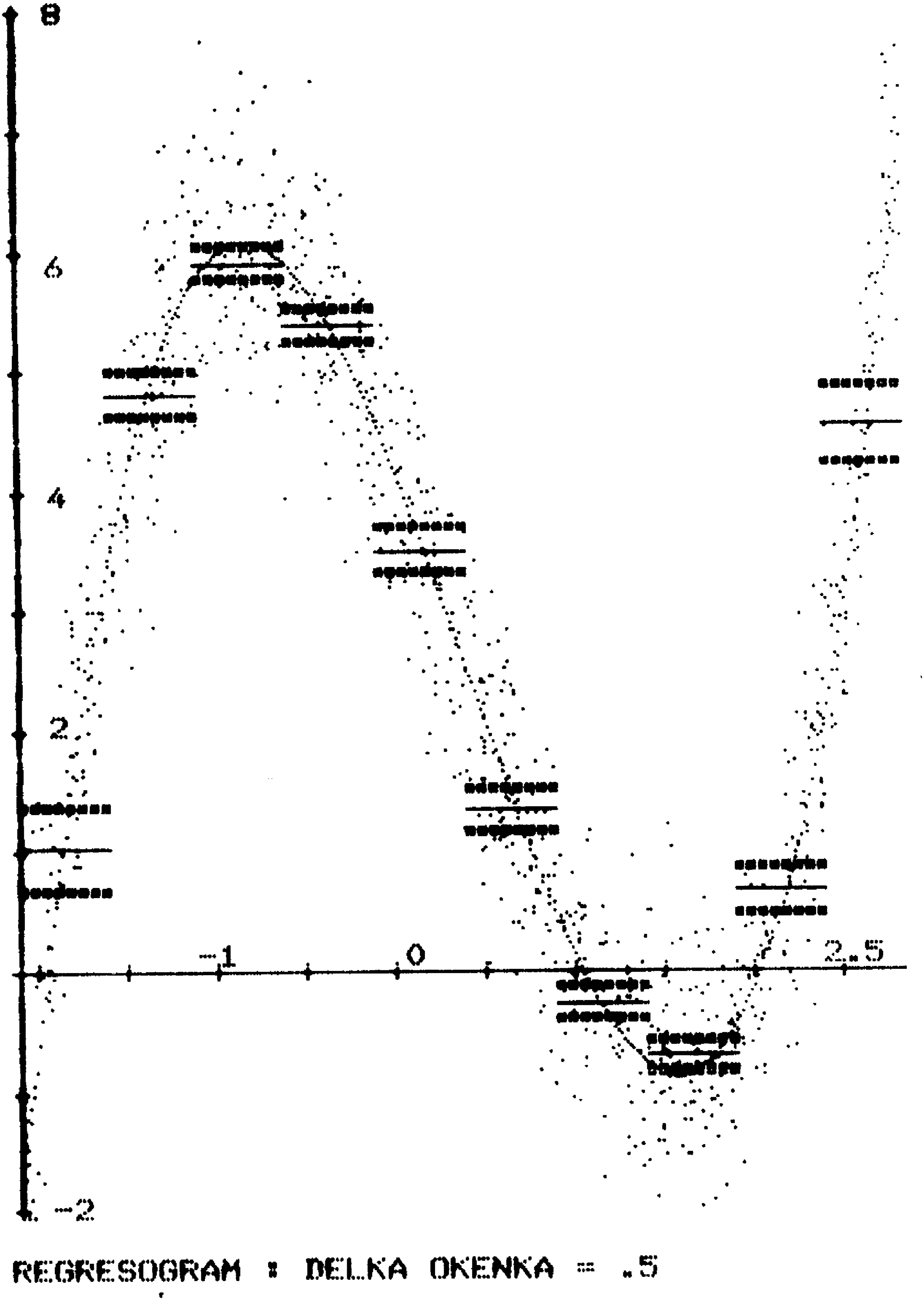
5. PŘÍKLAD

Ilustrujme nyní chování uvažovaných odhadů na následujícím simulačním experimentu. Pozorování Y_i, $i = 1,\ldots, 1000$, byla generována podle vztahu

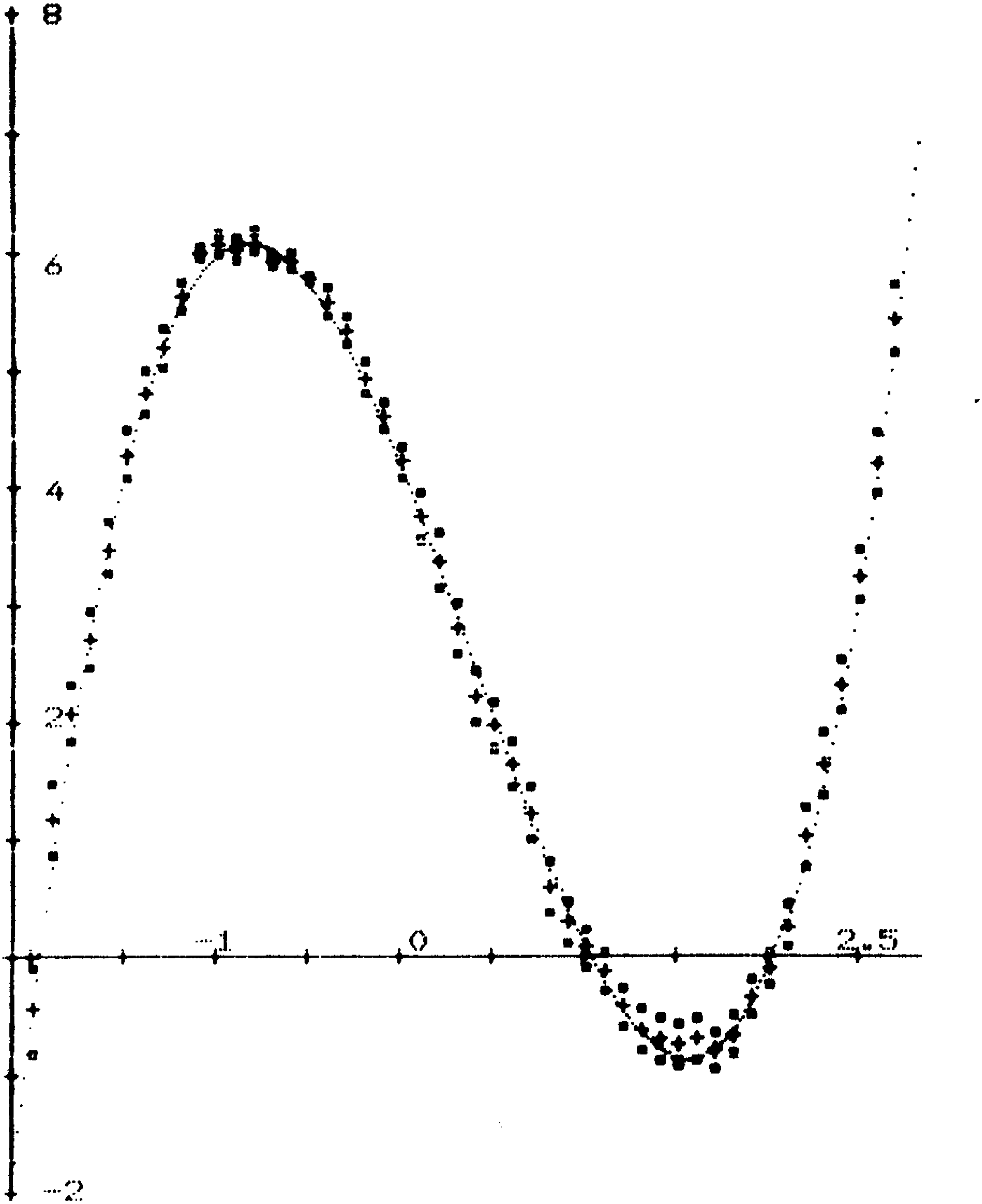


ODHAD TYPU KLOUZAVEHO OKENKA DELKY H = .5

Obr. č. 1.

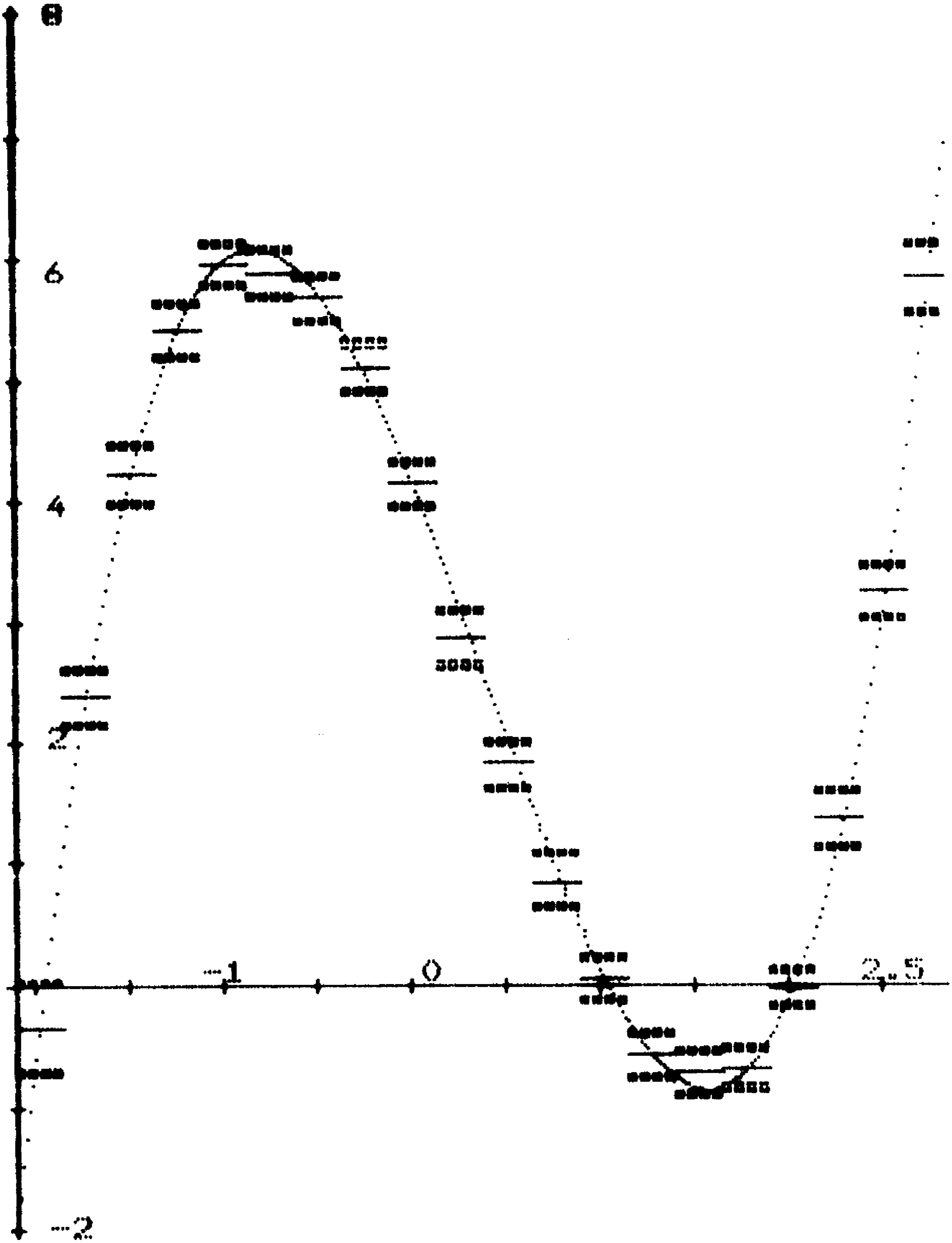


Obr. č. 2.



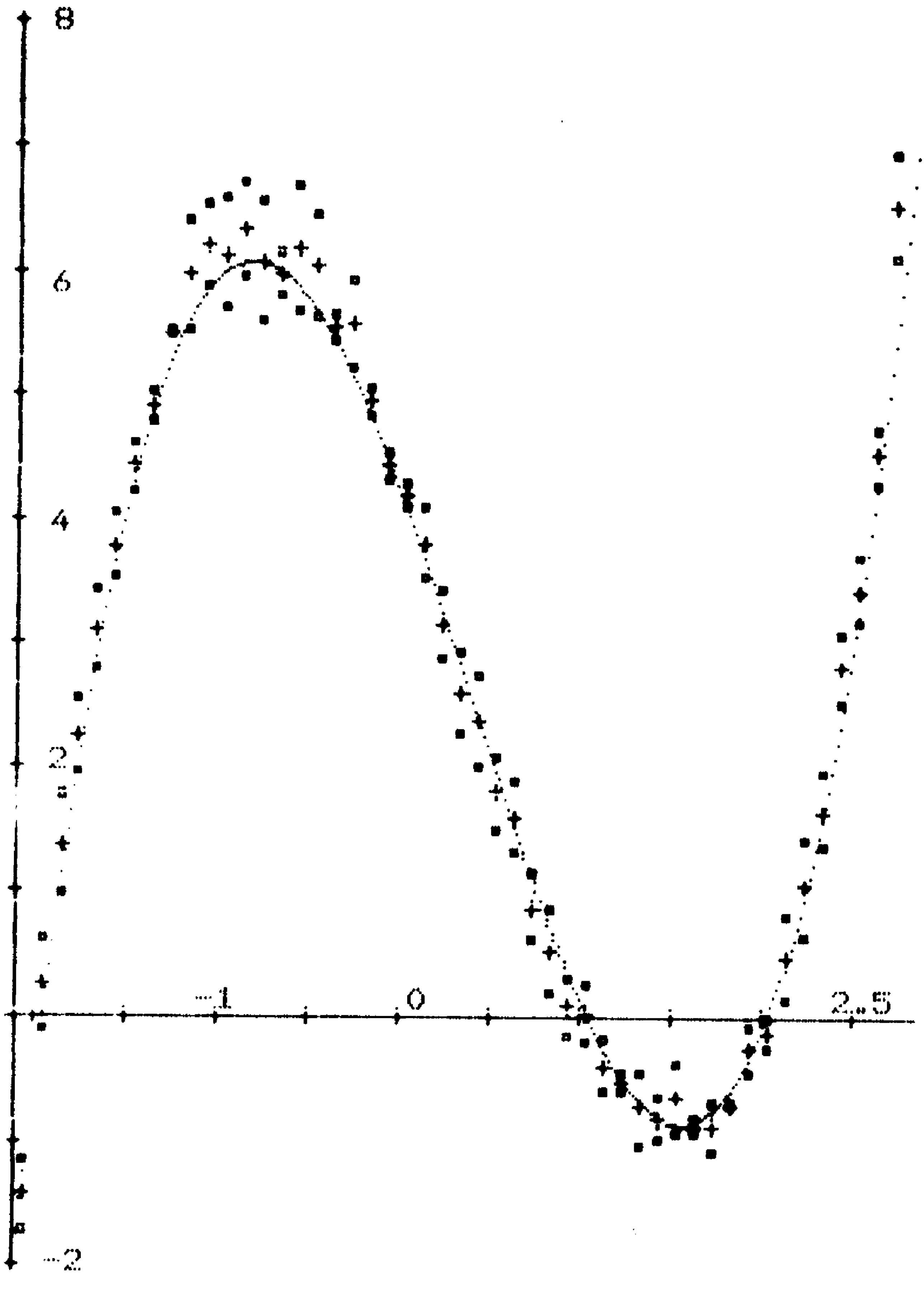
ODHAD TYPU KLOUZAVEHO OKENKA DELKY H = .25

Obr. č. 3.



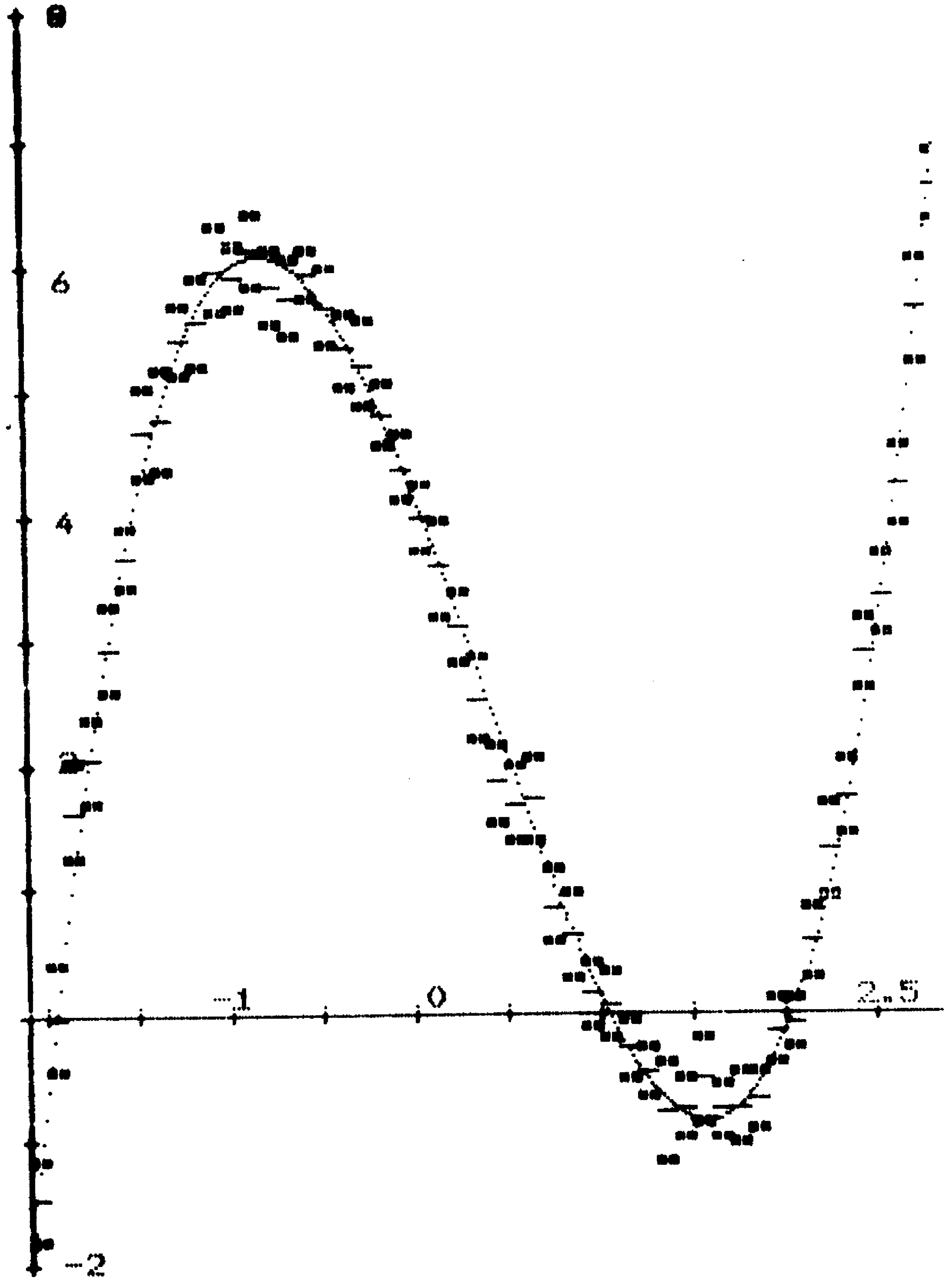
REGRESOGRAM : DELKA OKENKA = .25

Obr. č. 4.



ODHAD TYPU KLOUZAVEHO OKENKA DELKY H = .1

Obr. č. 5.



REGRESOGRAM : DELKA OKENKA = .1

Obr. č. 6.

$$(5.1) \qquad Y_i = X_i^3 - X_i^2 - 4\,X_i + 4 + \varepsilon_i \quad , \qquad i = 1,\ldots,1000,$$

kde $X_1,\ldots,X_{1000}$ tvoří náhodný výběr z $R[-2.15, 2.85]$ a $\varepsilon_1,\ldots,\varepsilon_{1000}$ tvoří náhodný výběr z $N(0, 0.7^2)$ [neuvažujeme zde tedy žádná odlehlá pozorování]. Pro takto získaná pozorování Y_i, $i = 1,\ldots,1000$, byl zpětně odhadován tvar původní regresní křivky pomocí šesti odhadů, spec.

(a) regresogramu s délkou okénka d_1 (d_1 = 0.50; 0.25; 0.10);

(b) jádrového odhadu tvaru klouzavého okénka pevné délky d_2 (d_2 = 0.50; 0.25; 0.10).

Obrázky 2,4 a 6 nám ukazují výsledky pro regresogram (s různou délkou okénka d_1). Obrázky 1,3,5 nám ukazují výsledky pro jádrový odhad tvaru klouzavého okénka délky d_2, odhadovaného v intervalu $[x_{min}, x_{max}]$ s krokem 0.05; hodnoty x_{min}, x_{max} pro dané d_2 udává následující tabulka:

d_2	0.50	0.25	0.10
x_{min}	-1.85	-1.975	-2.05
x_{max}	+2.55	+2.675	+2.75

Na všech obrázcích můžeme vidět jak jednotlivá pozorování (značená tečkami), tak tvar původní křivky (tečkovaná čára) a odhad této křivky (úsečky v případě regresogramu, křížky v případě klouzavého okénka). Znakem □ jsou značeny 5%-ní intervaly spolehlivosti pro $\phi(x)$. Celkově lze říci, že shoda teorie s praxí je v případě tohoto simulačního experimentu velmi dobrá.

6. LITERATURA

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

Antoch J. (1982). Neparametrické odhady hustoty. Sborník z zimní školy JČSMF ROBUST 82, 1-9.

Bhattacharya P.K. a Partasarathy K.R. (1961). Some limit theorems in regression theory. Sankhyā A 23, 91-102.

Bleuez J. a Bosq D. (1978). Etude d'une classe d'estimateurs non parametriques de la densité. Ann. Inst. Henri Poincaré 14, 479-498.

Bosq D. (1970). Contribution à la théorie de l'estimation fonctionelle. Publications de l'ISUP, vol. XIX, fasc. 2 et 3.

Collomb G. (1977a). Quelgues proprietés de la méthode du noyau pour l'estimation non-paramétrique de la régression en un point fixé. Compte Rendus Acad. Sci. Paris 285, A, 289-292.

Collomb G. (1977b). Estimation non-paramétrique de la régression par la méthode du noyau : propriété de convergence asymptotiquement normale indépendante. Annales Scientifiques de l'Université de Clermont, 24-46.

Collomb G. (1978). Estimation non-parametrique de la regression. Preprint 07-78, LPS, Univ. P. Sabatier, Toulouse.

Collomb G. (1985). Nonparametric regression: An up-to-date bibliography. Statistics 16, 309-324.

Collomb G. et all. (1985). Weak pointwise consistency of the cross validatory window estimate in non parametric regression estimation. CMUC.

Devroye L.P. (1979). The uniform convergence of the Nadaraya-Watson regression function estimate.Can. J. Statist. 6,179-191.

Devroye L.P. a Wagner T.J. (1979). Distribution-free consitency results in nonparametric discrimination and regression function estimation. Ann. Statist. 8, 231-239.

Devroye L.P. a Wagner T.J. (1980). On the L_1- convergence of kernel regression function estimators with applications in discrimination. Zeitschrift für Wahr. und verw. Gebiete 51, 15 - 25.

Elkins T.A. (1968). Cubical and spherical estimation of multivariate probability density. J.A. S.A. 63, 1495 - 1513.

Hall P. (1984). Asymptotic properties of integrated square error and crossvalidation for kernel estimation of a regression function. Zeitschrift für Wahrscheinlichkeitstheorie u. verw. Gebiete 67, 175 - 196.

Härdle W. (1983). Asymptotic maximal deviation of M-smoothers. Preprint, University of Heidelberg, FRG.

Härdle W. and Marron J.S. (1984). Optimal bandwidth selection in non-parametric regression function estimation. Preprint, University of Heidelberg, FRG.

Mack Y.P. (1982). Local properties of k-NN regression estimates. To appear in SIAM.

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

Révész P. (1973). Robbins-Monro procedure in a Hilbert space and its application in the theory of learning processes I. Stud. Scien. Math. Hungarica 8, 391-398.

Révész (1977). How to apply the method of stochastic approximation in the non-parametric estimation of a regression function. Math. Oper. und Stochastik, Ser. Statistics, 8, 119 - 126.

Révész P. (1978). On the non-parametric estimation of the regression function. Preprint, Dept. of Mathematics. Carlethon Univ., Ottawa.

Royall R.M. (1976). A class of non-parametric estimators of smooth regression function. PhD Dissertation, Stanford University.

Stone C.J. (1977). Consistent non-parametric regression. Annals of Statistics 5, 595-645.

Tukey J.W. (1961). Curves as parameters, and touch estimation. Proc. 4th Berkeley Symp., 681 - 694.

Watson G.S. (1964). Smooth regression analysis. Sankhyā A 26, 359 - 372.

Wong W.W. (1983). On the consistency of cross validation in kernel nonparametric regression. Annals of Statistics 11, 1136 - 1141.

Эпанечников В.А.(1969). Непараметрическая оценка многомерной плотности вероятностей. Теория вероятностей и ее применения. Том 15.,156-131.

Надарая А.Е.(1983). Непараметрические оценки плотности вероятностей и кривых регрессии. Издательство Тбилисского Университета.

Надарая А.Е.(1964). Непараметрическая оценка регрессии. Теория вероятностей и ее применения. Том 9,141-142.

Ченцов Н.Н.(1962). Оценка неизвестной плотности вероятностей. Сов. математика,3,1559-63.