VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
ÚSTAV MATEMATIKY

FACULTY OF MECHANICAL ENGINEERING
INSTITUTE OF MATHEMATICS

STOCHASTICKÉ MODELOVÁNÍ DATOVÝCH SOUBORŮ

STOCHASTIC MODELING OF DATA SETS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — Bc. SVETOSLAV ORGONÍK
AUTHOR

VEDOUCÍ PRÁCE — doc. RNDr. ZDENĚK KARPÍŠEK, CSc.
SUPERVISOR

BRNO 2011

Vysoké učení technické v Brně, Fakulta strojního inženýrství

Ústav matematiky
Akademický rok: 2010/2011

ZADÁNÍ DIPLOMOVÉ PRÁCE

student(ka): Bc. Svetoslav Orgoník

který/která studuje v **magisterském navazujícím studijním programu**

obor: **Matematické inženýrství (3901T021)**

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma diplomové práce:

Stochastické modelování datových souborů

v anglickém jazyce:

Stochastic Modeling of Data Sets

Stručná charakteristika problematiky úkolu:

Studium efektivních metod jádrových odhadů rozdělení pravděpodobnosti z pozorovaných datových souborů pro kvantitativní znaky.

Cíle diplomové práce:

Popis, zhodnocení a implementace moderních statistických metod fitování rozdělení pravděpodobnosti pomocí jádrových odhadů vzhledem k možnostem jejich realizace na PC a aplikacím na konkrétních datových souborech.

Seznam odborné literatury:

1. Montgomery, D. C., Renger, G.: Probability and Statistics. New York: John Wiley & Sons, 1996.
2. Anděl, J.: Statistické metody. Praha: MATFYZPRESS, 2003.
3. Anděl, J.: Základy matematické statistiky. Praha: MATFYZPRESS, 2002.
4. Silverman, B. W.: Density Estimation for Statistics and Data Analysis. London: Chapman and Hall,1985.
5. Vajda, I.: Theory of Statistical Inference and Information. London: Kluwer Academic Press, 1989.
6. Scott, D.W.: Multivariate Density Estimation. Theory, Practice and Visualization. New York: Wiley, 1992.
7. Články a materiály z odborných časopisů, sborníků konferencí a Internetu dle pokynů vedoucího diplomové práce.

Vedoucí diplomové práce: doc. RNDr. Zdeněk Karpíšek, CSc.

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2010/2011.

V Brně, dne 19.11.2010

L.S.

prof. RNDr. Josef Šlapal, CSc. Ředitel ústavu	prof. RNDr. Miroslav Doupovec, CSc. Děkan fakulty

Abstrakt
Diplomová práca je zameraná na implementáciu moderných štatistických metód fitovania rozdelenia prevdepodobnosti pomocou jadrových odhadov vzhľadom k možnostiam ich realizácie na PC a aplikáciam na konkrétnych datových súboroch.

Diplomová práca je súčasťou riešenia projektu MŠMT Českej republiky čís. 1M06047 Centrum pro jakost a spolehlivost výroby.

Summary
Master's thesis is focused on implementing modern statistical methods for fitting propability distribution using kernel estimates with regard to the possibilities of their implementation on the PC and the application of specific data sets.

Master's thesis is a part of project from MSMT of the Czech Republic no. 1M06047 Center for Quality and Reliability of Production.

Klíčová slova
stochastické modelovanie, jadrové odhady

Keywords
stochastic modeling, kernel estimates

ORGONíK, S.*Stochastické modelování datových souborů*. Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2011. 76 s. Vedoucí doc. RNDr. Zdeněk Karpíšek, CSc.

Čestné prehlásenie

Prehlasujem, že som diplomovú prácu spracoval samostatne podľa pokynov vedúceho diplomovej práce a s použitím uvedenej literatúry.

V Brne, dňa 27. 5. 2011

Bc. Svetoslav Orgoník

Poďakovanie

Chcel by som poďakovať doc. RNDr. Z. Karpíškovi, CSc. za odborné vedenie a cenné rady, ktoré mi poskytol pri spracovaní diplomovej práce.

Bc. Svetoslav Orgoník

Obsah

1. Úvod

Diplomová práca je zameraná na popis, zhodnotenie a implementáciu moderných štatistických metód fitovania rozdelenia pravdepodobnosti pomocou jadrových odhadov vzhľadom k možnostiam ich realizácie na PC a aplikáciam na konkrétnych datových súboroch. V druhej kapitole sa zameriame na jednorozmerné odhady pre jednorozmerný náhodný výber. Začneme histogramom a naivným odhadom. Prejdeme k jadrovému odhadu, ktorý popíšeme aj s metódami hľadania vyhladzovacieho parametra. Túto kapitolu zakončíme metódou najbližšieho suseda, metódou premenlivého jadra a porovnanie odhadov trimodálnej hustoty. V tretej kapitole popíšeme jadrový odhad pre viacrozmerný náhodný výber. Vo štvrtej kapitole popíšeme naprogramované zdrojové makrá v Exceli, od histogramu až po dvojrozmerný jadrový odhad náhodného výberu.

Vlastné výsledky práce sú príklady v kapitolách 2 a 3, vrátane software pre naše výpočty vo 4. kapitole. Hlavnou časťou práce je druhá a štvrtá kapitola. V druhej kapitole vykreslíme jadrové odhady pre šesť metód odhadu hustoty pravdepodobnosti pre rôzne rozsahy náhodných výberov trimodálnej hustoty. Každá metóda sa bude robiť pre 9 náhodných výberov. K výpočtu použijeme vlastné naprogramované makrá v softwari Excel, ktoré budú podrobne popísané, včetne náhľadov vo štvrtej kapitole. Na všetky výpočty použijeme software Excel, včetne zobrazenia výsledkov až na dvojrozmerný prípad. V dvojrozmernom prípade zobrazíme napočítané odhady v Exceli pomocou Matlabu.

Teoretické partie používané v práci sme prevzali najmä z literatúry [1], [2], [3], [6] a [8]. Okrem tohto sú v kapitolách za teóriou vysvetľujúce príklady. Teória aj príklady sú doplnené názornými obrázkami.

2. Metódy odhadov pre jednorozmerný náhodný výber

V celej tejto kapitole budeme predpokladať, že máme náhodný výber z náhodnej veličiny X, tj. postupnosť nezávislých náhodných veličín $X_1, \ldots, X_n$ s rovnakým rozdelením, ktoré je spojité jednorozmerné a funkcia hustoty pravdepodobnosti, ktorú sa snažíme odhadnúť, je f. Je veľa praktických problémov, kde tieto predpoklady nie sú nutne oprávnené, ale napriek tomu poskytnú normálnu štruktúrú, o ktorej pojednávajú vlastnosti metód odhady hustoty.

V kapitole budeme uvažovať pozorovanú trimodálnu hustotu definovanú ako zmes Gaussových rozdelení $\frac{1}{3}N(3, 0.8) + \frac{1}{3}N(5, 0.5) + \frac{1}{3}N(7, 0.3)$.

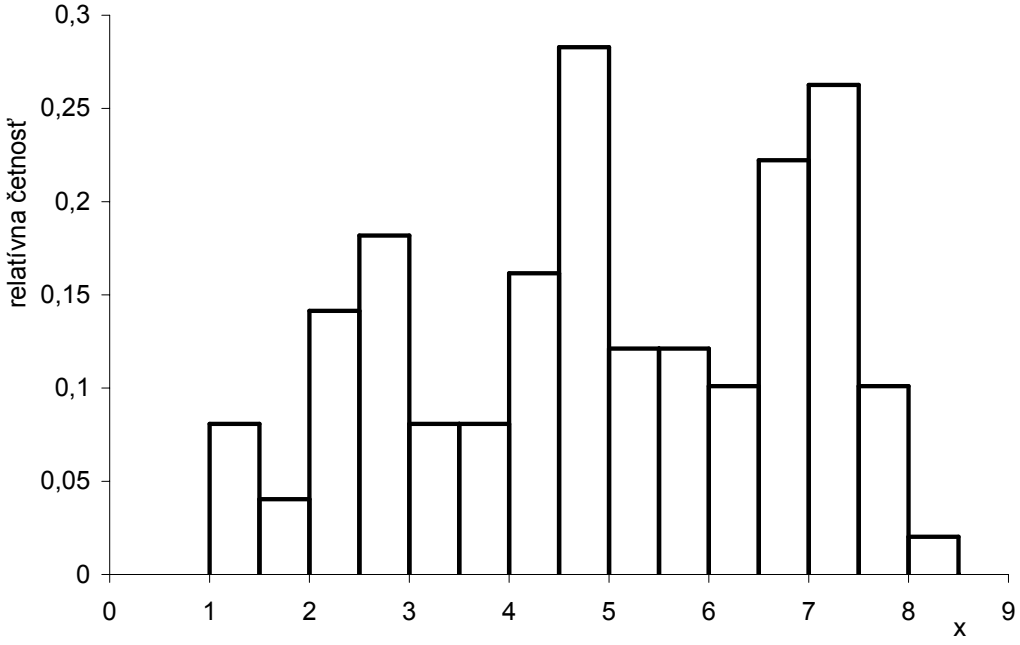
2.1. Histogram

Histogram je najstarší a najviac používaný odhad hustoty. Majme *počiatok* x_0 a *šírku triedy* h, definujeme *triedy* histogramu ako intervaly $[x_0 + mh, x_0 + (m + 1)h)$ pre kladné a záporné celé čísla m. Pre jednoznačnosť sú zvolené intervaly zľava uzavreté a zprava otvorené.

Histogram je potom definovaný

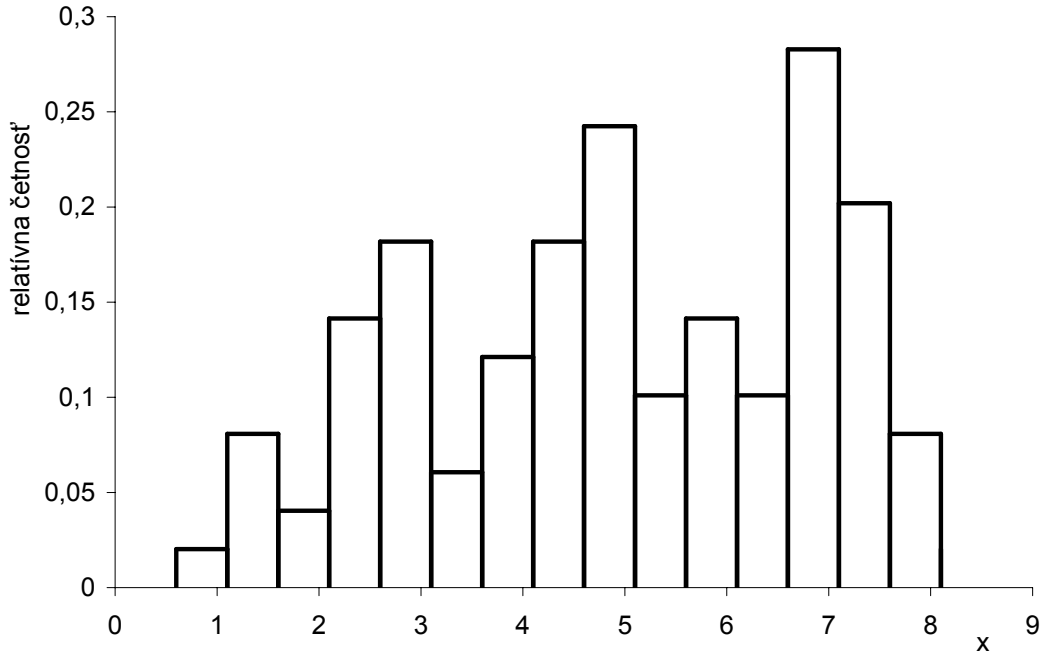
$$\hat{f}(x) = \frac{1}{nh}(\text{počet } X_i \text{ v rovnakej triede ako } x).$$

Poznamenajme ku konštrukcii histogramu, že musíme zvoliť počiatok a šírku triedy. To je voľba šírky triedy, ktorá primárne ovláda stupeň vyhladenia.

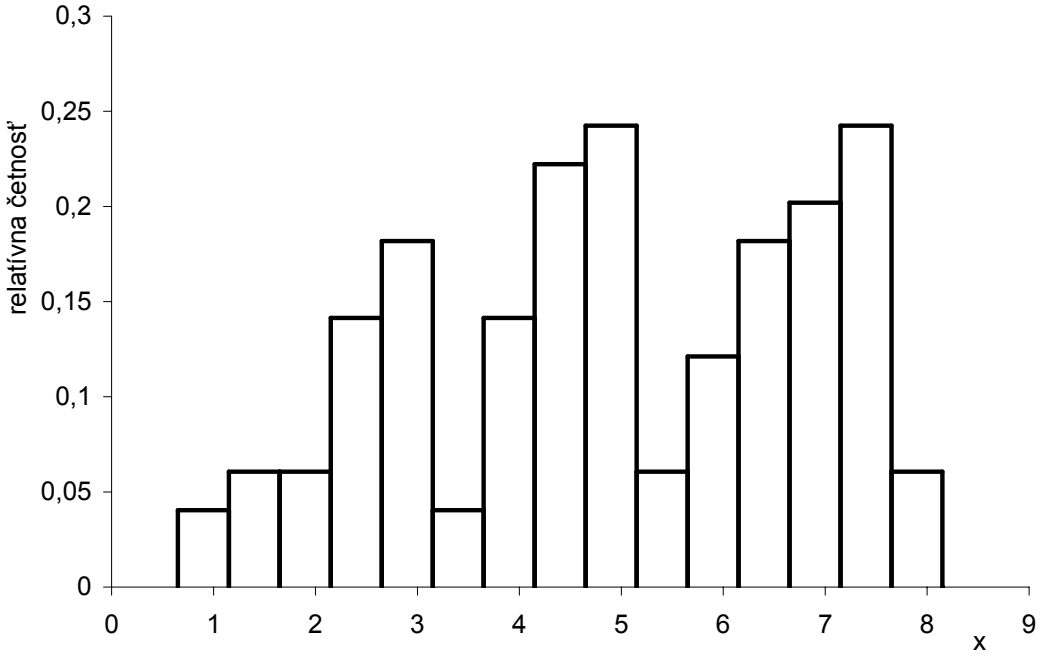


Obr. 2.1: Histogramy z trimodálnej hustoty, h=0.5, x_0=1.

Skeptici sa na odhad hustoty často pýtajú, prečo je niekedy nutné použiť metódy viac sofistikované než jednoduchý histogram. Dôvod pre tieto metódy a nevýhody histogramu



Obr. 2.2: Histogramy z trimodálnej hustoty, h=0.5, x_0=0.6



Obr. 2.3: Histogram z trimodálnej hustoty, h=0.5, x_0=0.65

závisí dosť podstatne na súvislosti. Rôzne matematické popisy presnosti môžu byť pomocou histogramu dosť podstatne zlepšené, a táto matematická nevýhoda sa premieňa na neúčinné použitie náhodného výberu, ak histogramy sú použité ako odhady hustôt v postupe ako zhluková analýza a neparametrická rozlišujúca analýza. Nespojitosť histogramov spôsobuje extrémny problém, ak sú potrebné derivácie odhadov. Keď sú potrebné odhady hustôt ako pomocná zložka ostatných metód, dôvod pre použitie alternatív k histogramom je celkom výrazný.

Histogramy sú samozrejme extrémne užitočné odhady hustôt pre prezentáciu a skúmnie náhodných výberov, špeciálne pre jednorozmerný prípad. Aj keď voľba počiatku v jednej dimenzie môže byť celkom úspešná, obrázky 2.1, 2.2 a 2.3 ukazujú histogramy trimodálnej hustoty s rovnakými triedami, ale rôznymi počiatkami. Hoci výsledok je rovnaký vo všetkých troch prípadoch, obzvlášť neštatistik môže získať iný dojem. Napríklad šírka vrcholu vpravo a delenie dvoch režimov.

Histogramy pre grafickú prezentáciu dvojrozmerných alebo trojrozmerných náhodných výberov majú niekoľko problémov. Napríklad nemôžeme jednoducho nakresliť obrys grafu reprezentujúci náhodný výber. V jednorozmernom prípade je závislosť odhadu na voľbe počiatku a šírke triedy. Nakoniec by mohlo byť stresujúce, že vo všetkých prípadoch, histogram stále vyžaduje voľbu stupňa vyhladenia.

Hoci histogram zostáva ako vynikajúci nástroj na prezentáciu náhodného výberu, je významný prinajmenšom vzhľadom k rôznym odhadom alternatívnych hustôt.

2.2. Naivný odhad

Z definície pravdepodobnosti hustoty, kde náhodná premenná X má hustotu f, potom

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P\left(x - h < X < x + h\right).$$

Pre dané h môžeme samozrejme odhadnúť $P\left(x - h < X < x + h\right)$ pomerom vzorky patriacej do intervalu $(x - h, x + h)$. Prirodzený odhad $\hat{f}$ hustoty je určený zvolením malého čísla h, potom

$$\hat{f}(x) = \tfrac{1}{2hn}[\text{ počet } X_1, \ldots, X_n \text{ patriacich do } (x - h, x + h)].$$

budeme volať naivný odhad.

Na vyjadrenie transparentnejšieho odhadu, definujeme váženú funkciu w

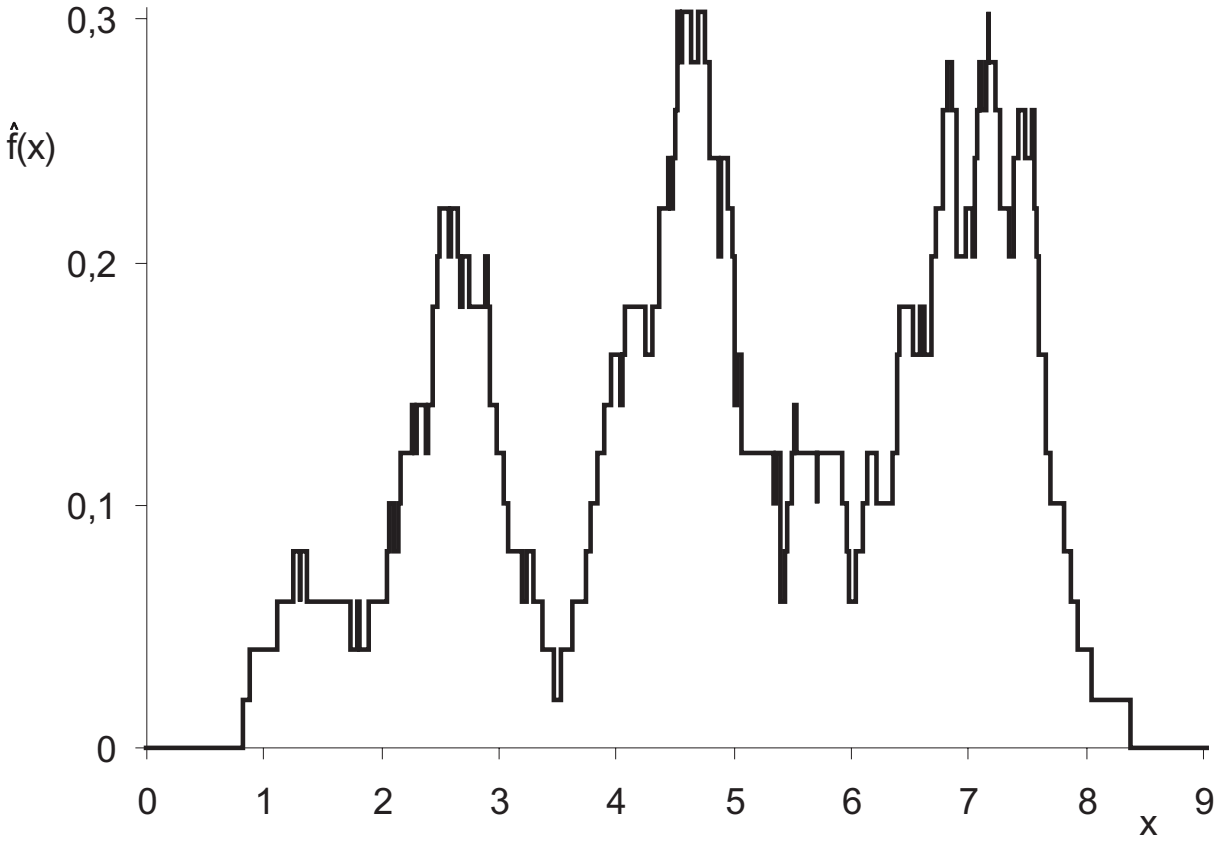
$$w(x) = \begin{cases} \frac{1}{2}, & |x| < 1 \\ \\ 0, & \text{inak.} \end{cases} \tag{2.1}$$

Potom vidíme, že naivný odhad môže byť zapísaný

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x - X_i}{h}\right).$$

Z funkcie 2.1 vidíme, že odhad je tvorený umiestnením "obdĺžnika" o šírke $2h$ a výške $(2nh)^{-1}$ na každé pozorovanie a sčítaním týchto obdĺžnikov dostaneme odhad. Vrátime sa k interpretácii dole, ale je poučné vidieť spojitosť s histogramom.

Uvažujme histogram vytvorený z náhodného výberu s použitím šírky triedy $2h$. Predpokladajme, že žiadne pozorovanie neleží presne na okraji triedy. Ak sa stane, že x je v strede jednej triedy histogramu, okamžite vyplýva z 2.1, že naivný odhad $\hat{f}(x)$ bude presne ordináta histogramu v x. Teda môžeme vidieť, že naivný odhad je pokus o vytvorenie histogramu, kde každá náhodná veličina náhodného výberu je stredom triedy, takže uvolnený histogram so špeciálnou voľbou umiestnenia triedy. Voľba šírky triedy je stále riadená parametrom h, ktorý ovláda stupeň, podľa ktorého je vyhladený náhodný výber k tvorbe odhadu.



Obr. 2.4: Naivný odhad z trimodálnej hustoty, h=0.25

Na používanie odhadov hustôt pre prezentácie nie je naivný odhad úplne vhodný. Vyplýva to z definície, že $\hat{f}$ nie je spojitá funkcia, ale má skoky v bodoch $X_i \pm h$ a všade inde má nulové derivácie. Toto dáva odhadom trochu strapatý charakter, ktorý nie je iba esteticky nevhodný, ale by mohol poskytnúť netrénovanému pozorovateľovi zavádzajúci dojem. Pre čiastočné prekonanie tejto prekážky, a pre iné technické dôvody, je záujem zvažovať zobecnenie naivného odhadu, ktoré je popísané v ďalšej časti.

Odhad hustoty používajúci naivný odhad má na obr. 2.4. "Stupňovitý" charakter odhadu, ktorý je zrejmý. Použité obdĺžniky na konštrukciu odhadu majú rovnakú šírku ako triedy v histogramoch na obrázkoch 2.1, 2.2 a 2.3.

2.3. Jadrový odhad

2.3.1. Úvod

V tejto kapitole sa budeme detailnejšie zaoberať základnými štatistickými vlastnosťami jadrového odhadu. Náš záujem pre jadrovú metódu nie je preto, že metóda je najlepšia na použitie za všetkých okolností, ale je tu niekoľko dôvodov pre prvé použitie. Metóda je široko aplikovateľná, špeciálne v jednorozmernom prípade, a jej chovanie je určite rozumné pred pokračovaním iných metód. Je to pravdepodobne metóda, ktorej vlastnosti sú najrozumnejšie. A pojednanie o týchto vlastnostiach vyzdvihuje problémy, ktoré majú ostatné metódy odhadu hustoty.

Zovšeobecníme si naivný odhad na prekonanie niektorých problémov pojednávaných vyššie. Váhovú funkciu w nahradíme za *jadrovú funkciu* K, ktorá spĺňa podmienku

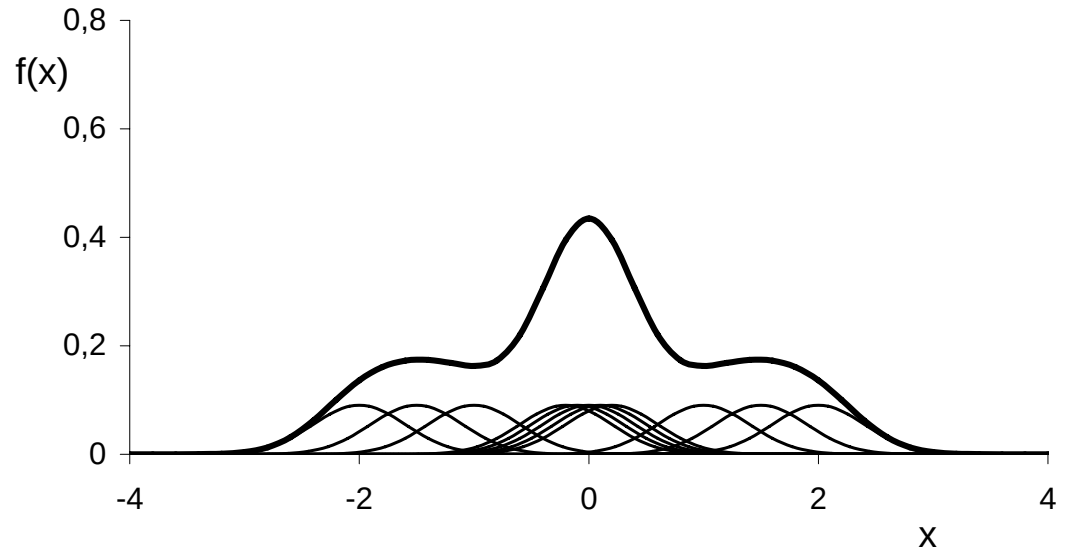
$$\int_{-\infty}^{\infty} K(x)dx = 1. \tag{2.2}$$

Obvykle, ale nie vždy, K bude symetrická funkcia hustoty pravdepodobnosti. Napríklad normálna hustota, alebo váhová funkcia w, ktorá je použitá v naivnom odhade. Analogicky pomocou definície naivného odhadu, je *jadrový odhad* s jadrom K definovaný

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{2.3}$$

kde h je *šírka okna*, tiež nazývaný *vyhladzovací parameter*. Matematickými vlastnosťami jadrového odhadu sa budeme zaoberať neskôr. Najprv intuitívne pojednáme s príkladmi, ktoré môže byť užitočné.

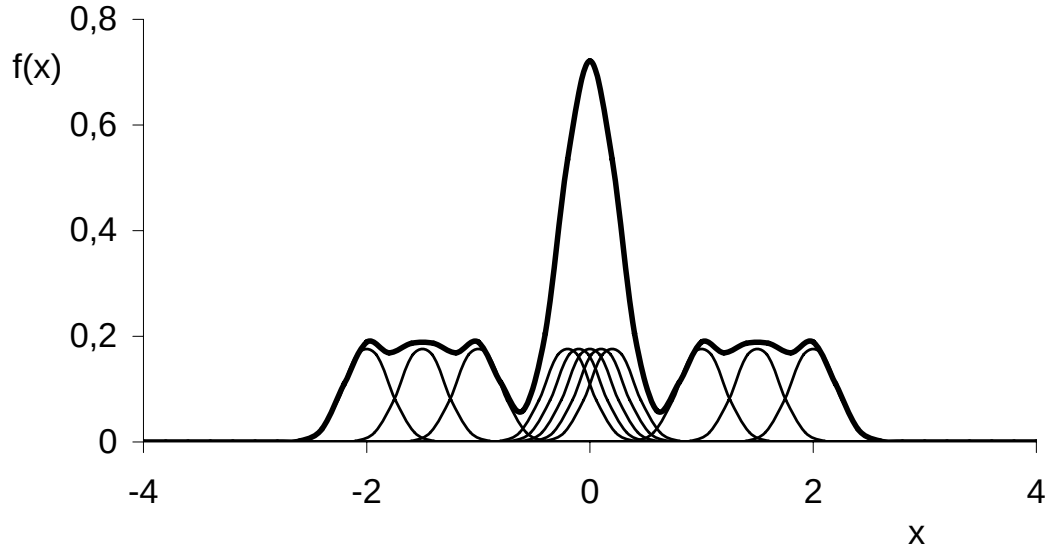
Rovnako ako naivný odhad, ktorý môžeme považovať za súčet vystredených "obdĺžnikov" na pozorovaniach, jadrový odhad je súčet "jadier" umiestnených na pozorovaniach. Jadrová funkcia K určuje tvar jadier, zatiaľ čo šírka okna h určuje ich šírku. Názorný príklad je na obr. 2.5 a obr. 2.6, kde jednotlivé jadrá $\frac{1}{nh}K\left(\frac{x-X_i}{h}\right)$ sa ukazujú rovnako ako odhad $\hat{f}$ vytvorený ich sčítaním. Môže to byť stresujúce, že na konštrukciu odhadu hustoty nie je vhodný taký malý rozsah náhodného výberu, ale tento náhodný výber s rozsahom 11 je použitý pre zrozumiteľnosť.



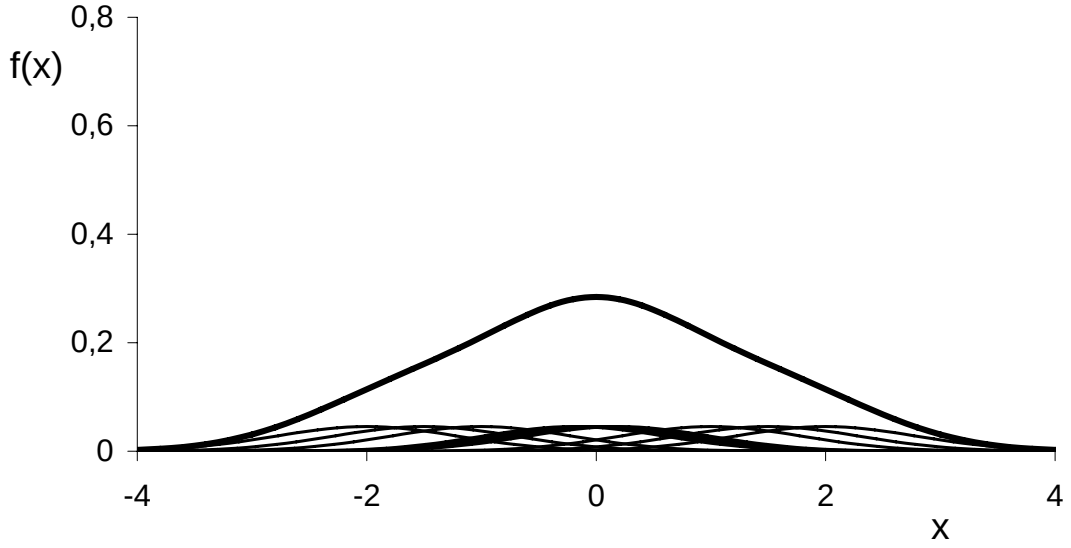
Obr. 2.5: Jadrový odhad ukazujúci jednotlivé jadrá. Šírka okna je 0.4.

Výsledok menenia šírky okna je znázornený na obr. 2.7. Keď h smeruje limitne k nule, je súčet dirakových delta funkcií na zložkách náhodného výberu, zatiaľ čo h narastá, všetky rušivé alebo iné detaily sú skryté.

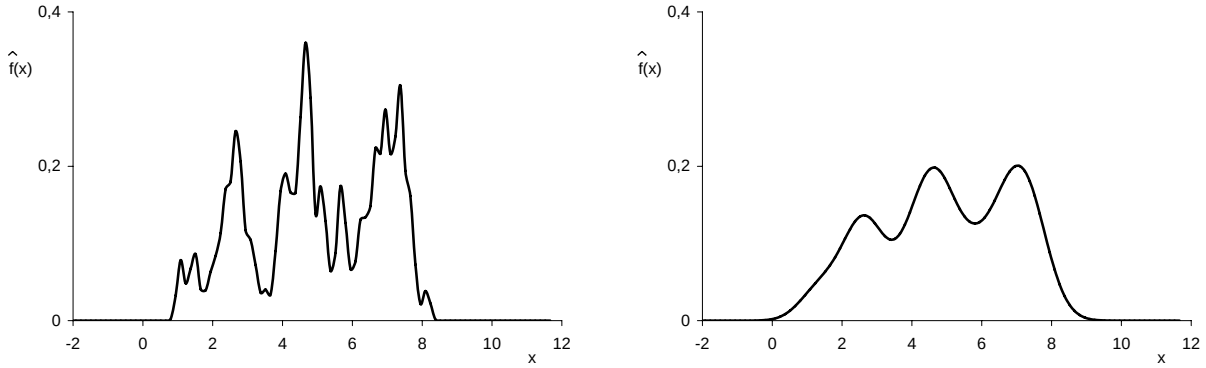
Ďalší názorný príklad menenia šírky okna je na obr. 2.8 a obr. 2.9 vľavo. Odhady tu boli vytvorené zo pseudonáhodného náhodného výberu s rozsahom 300, nakrelené z trimodálnej hustoty danej na obr. 2.9 vpravo. Na odhad boli použité normálne jadrá. Znova môžeme poznamenať, že ak je h zvolené príliš malé, potom sa jemná rušivá štruktúra stáva viditeľná. Zatiaľ ak je h príliš veľké, potom povaha trimodality rozdelenia je skrytá.



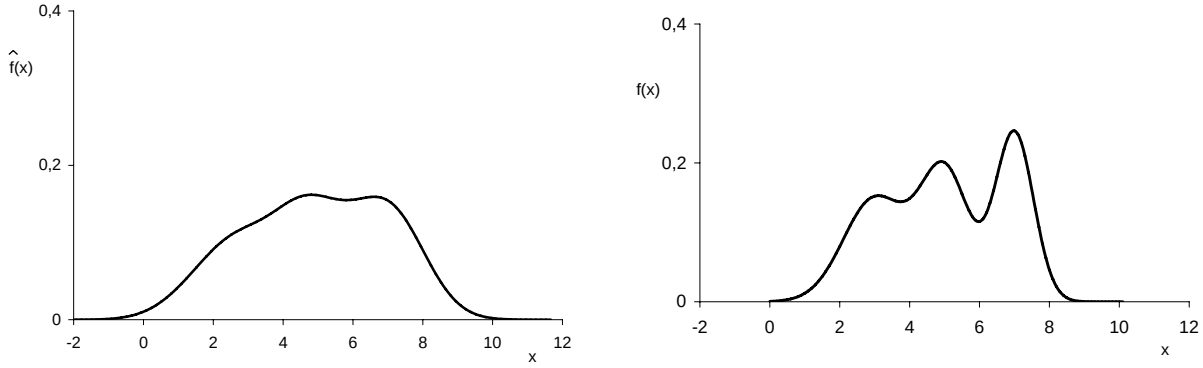
Obr. 2.6: Jadrový odhad ukazujúci jednotlivé jadrá. Šírka okna je 0.2.



Obr. 2.7: Jadrový odhad ukazujúci jednotlivé jadrá. Šírka okna je 0.8.

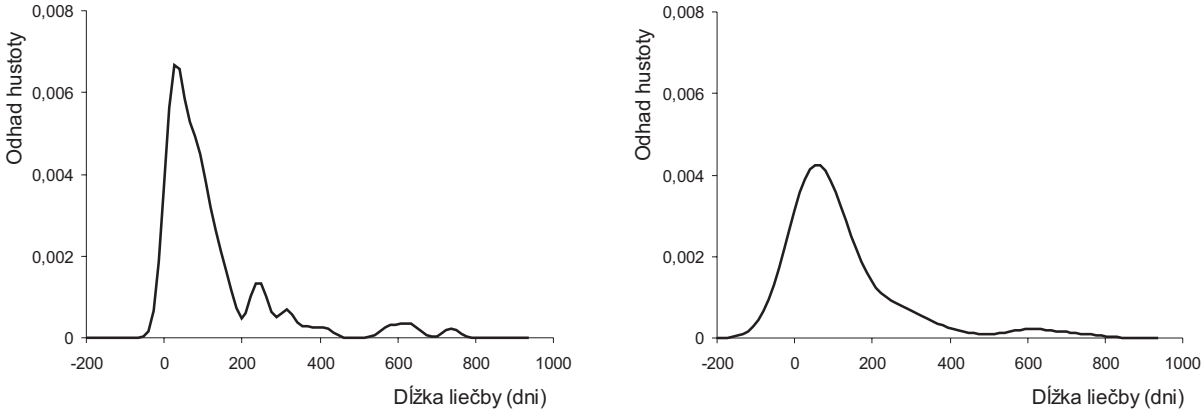


Obr. 2.8: Jadrové odhady náhodného výberu trimodálnej hustoty s rozsahom 99. Šírka okna obrázku vľavo je 0.1, a obrázku vpravo je 0.5.



Obr. 2.9: Vľavo je jadrový odhad náhodného výberu trimodálnej hustoty s rozsahom 99 so šírkou okna 0.9. Vpravo je skutočná trimodálna hustota, ktorú sme použili pri predchádzajúcich troch jadrových odhadoch.

Niektoré základné vlastnosti jadrových odhadov vyplývajú okamžite z definície. Pokiaľ je jadro K všade nezáporné a spĺňa podmienku 2.2, inými slovami funkcia hustoty pravdepodobnosti. Okamžite to vyplynie z definície, keď $\hat{f}$ bude hustota pravdepodobnosti. Okrem toho, $\hat{f}$ zdedí všetky vlastnosti, spojitosť a diferencovateľnosť jadra K. Takže ak napríklad K má normálnu funkciu hustoty, potom $\hat{f}$ bude hladká krivka s deriváciami všetkých rádov. To sú dôvody prečo niekedy použijeme jadrá, ktoré dávajú rovnako dobré i zlé hodnoty. Napriek tomu sa najviac používajú nezáporné jadrá .



Obr. 2.10: Jadrový odhad študijných údajov dĺžky liečenia. Vľavo je šírka okna 20. Vpravo je šírka okna 60.

V nasledujúcej tabuľke sú dĺžky liečby pacientov v dňoch.

1	13	22	31	39	56	76	90	111	134	228	311	640
1	14	25	32	39	62	79	91	112	144	231	314	737
1	14	27	34	40	63	82	92	119	147	235	322	
5	17	27	35	49	65	83	93	122	153	242	369	
7	18	30	36	49	65	84	93	123	163	256	415	
8	21	30	37	54	67	84	103	126	167	256	573	
8	21	31	38	56	75	84	103	129	175	257	609	

Nezávisle na histograme, je pravdepodobne jadrový odhad najbežnejšie používaný odhad a je určite najviac matematicky skúmaný. Ale má nevýhodu. Trpí nepatrným nedostatkom, keď je použitý náhodný výber s dlhým chvostom rozdelenia. Lebo šírka okna je pevná pozdĺž celého náhodného výberu. Na odhadoch chvostov sa javí tendencia k rušivému šumu. Ak sú odhady dosť vyhladené (máme na mysli dlhý chvost), podstatný detail hlavnej časti rozdelenia ostane skrytý. Príklad tohto chovania je daný bez ohľadu na fakt, že údaje o dĺžke liečenia sú samozrejme nezáporné, a odhadovanie hustoty liečenia je pozorovanie na $(-\infty, \infty)$. Na obrázku 2.10 vľavo je odhad so šírkou okna 20 zašumený po pravej strane na chvoste, zatiaľ čo na obrázku 2.10 vpravo je odhad so šírkou okna 60, ktorý ukazuje malý mod na chvoste a ešte zväčšuje šírku hlavného modu rozdelenia.

2.3.2. Miery odchýlky: stredná kvadratická chyba a stredná integrálna kvadratická chyba

Základná metodika teoretického zaobchádzania je pojednanie blízkosti odhadu $\hat{f}$ k skutočnej hustote f v rôznych zmysloch. Odhad $\hat{f}$ samozrejme závisí na údajoch práve tak, ako na jadre a šírke okna. Táto závislosť nebude obecne vyjadrená explicitne. Pre každé x, môže pokladať $\hat{f}$ za náhodnú premennú, kvôli závislosti na meraniach $X_1, \ldots, X_n$. Ľubovolné použitie pravdepodobnosti, rozptylu obsahujúceho $\hat{f}$ je s ohľadom na jeho náhodný výber rozdelenia ako štatistiky založené na týchto náhodných meraniach.

Okrem toho, kde je uvedená $\sum$ bude znamenať súčet od $i = 1$ do n, a $\int$ bude znamenať integrál cez $(-\infty, \infty)$.

Skúmali sa rôzne miery odchýlky odhadu hustoty $\hat{f}$ od skutočnej hustoty f. Keď uvážime odhad v jedom bode, prirodzená miera je *stredná kvadratická chyba* (označme si ju MSE), definovaná

$$\mathrm{MSE}_x\left(\hat{f}\right) = \mathrm{E}\left\{\hat{f}(x) - f(x)\right\}^2. \tag{2.4}$$

Pomocou základných vlastností strednej hodnoty a rozptylu,

$$\mathrm{MSE}_x\left(\hat{f}\right) = \left\{\mathrm{E}\hat{f}(x) - f(x)\right\}^2 + \mathrm{var}\hat{f}(x), \tag{2.5}$$

je súčet kvadrátu rozdielu strednej hodnoty $\hat{f}$ a f, a rozptylu na x. Vidíme, že vo veľa odvetví štatistiky, je porovnanie medzi chybou a rozptylom vo 2.5. Chyba môže byť zmen−šená rastúcim rozptylom, a naopak, regulovaná vyhladením.

Prvá a najširšie používaná miera s *globálnou* presnosťou $\hat{f}$ odhadu f je *stredná integrálna kvadratická chyba* (označme si ju MISE) definovaná

$$\mathrm{MISE}_x\left(\hat{f}\right) = \mathrm{E}\int\left\{\hat{f}(x) - f(x)\right\}^2 dx. \tag{2.6}$$

I keď sú iné globálne miery odchýlky, ktoré dávajú globálne dobrý odhad. MISE je zďaleka najtvárnejšia miera, a tak je dobre sa ňou zaoberať najskôr. Je užitočné poznamenať, že keď je integrand nezáporný, poradie integrácie a pravdepodobnosti v 2.6 môže byť prehodené na alternatívny tvar

$$\begin{aligned}\mathrm{MISE}_x\left(\hat{f}\right) &= \int \mathrm{E}\left\{\hat{f}(x) - f(x)\right\}^2 dx \\ &= \int \mathrm{MSE}_x\left(\hat{f}\right) dx\end{aligned} \tag{2.7}$$

$$= \int \left\{ \mathrm{E}\hat{f}\,(x) - f\,(x) \right\}^2 dx + \int \mathrm{var}\hat{f}\,(x)\,dx, \qquad (2.8)$$

kde MISE je súčet integrovaného kvadrátu rozdielu strednej hodnoty $\hat{f}$ a f a integrálneho rozptylu.

2.3.3. Približné vlastnosti

Mnoho teoretických prác odhadu hustoty sa zaoberá asymptotickými vlastnosťami rôznych metód vo veľa odvetviach matematickej štatistiky. Na konci tejto podkapitoly odvodíme približné vyjadrenie odchýlky a rozpytu odhadu, a použijeme ich na skúmanie, ako sa bude chovať stredná kvadratická chyba a stredná integrálna kvadratická chyba. V podkapitole 2.3.5 sa budeme odvolávať na rigorózne asymptotické výsledky, ktoré opravedlňujú tieto aproximácie, za podmienky vhodných predpokladov. Pre jednoduchosť budeme predpokladať v celej tejto rozprave, že:

$$\int K(t)dt = 1, \qquad \int tK(t)dt = 0, \qquad a \qquad \int t^2K(t)dt = k_2 \neq 0 \qquad (2.9)$$

a že,

neznáma hustota f má spojité derivácie všetkých požadovaných rádov (2.10)

Obvykle bude jadro K symetrická pravdepodobnostná funkcia hustoty, napríklad normálna hustota, a konštanta k_2 bude rozptyl rozdelenia s touto hustotou. Môže byť prekvapujúce, že na rozdiel od hustoty f, je jadro K ovládané uživateľom. A preto pre praktické použitie je potrebné uvažovať iba výsledky, ktoré platia pre konkrétne použitie jadra.

Odchýlka a rozptyl odhadu

Odchýlka odhadu $f(x)$ nezávisí priamo na rozsahu náhodného výberu, ale závisí na šírke okna h. Samozrejme, ak je h zvolené ako funkcia n, potom bude odchýlka odhadu (bias) závisieť nepriamo na n. Budeme písať

$$\begin{aligned} \mathrm{bias}_h(x) &= \mathrm{E}\hat{f}(x) - f(x) \\ &= \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)dy - f(x). \end{aligned} \qquad (2.11)$$

Odchýlka odhadu tiež závisí na jadre K, ale táto závislosť nebude vyjadrená explicitne. Použijeme 2.11 na získanie približného vyjadrenia odchýlky odhadu. Urobíme substitúciu $y = x - ht$ a použijeme predpoklad, že $\int K(t)dt = 1$, napíšeme

$$\begin{aligned} \mathrm{bias}_h(x) &= \int K(t)f(x-ht)dt - f(x) \\ &= \int K(t)\left(f(x-ht) - f(x)\right)dt. \end{aligned}$$

Taylorov rozvoj je

$$f(x-ht) = f(x) - htf'(x) + \frac{1}{2}h^2t^2f''(x) + \cdots$$

takže z predpokladov o K

$$\begin{aligned} \text{bias}_h(x) &= -hf'(x)\int tK(t)dt + \frac{1}{2}h^2f''(x)\int t^2K(t)dt + \cdots && (2.12)\\ &= \frac{1}{2}h^2f''(x)k_2 + \cdots && (2.13)\end{aligned}$$

Integrálna kvadratická chyba odhadu, potrebná v 2.8 pre strednú integrálnu kvadratickú chybu, je daná

$$\int \text{bias}_h(x)^2dx \approx \frac{1}{4}h^4h_2^2\int f''(x)^2dx. \quad (2.14)$$

Teraz vieme zmeniť rozptyl. Z [8] dostávame

$$\begin{aligned} \text{var}\hat{f}(x) &= \frac{1}{n}\int \frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(y)dy - \frac{1}{n}\left(f(x) + \text{bias}_h(x)\right)^2\\ &\approx \frac{1}{nh}\int f(x-ht)K(t)^2dt - \frac{1}{n}\left(f(x) + O(h^2)\right)^2, \end{aligned}$$

použitím substitúcie $y = x - ht$ v integráli, a aproximácie odchýlky odhadu 2.13. Predpokladajme, že h je malé a n je veľké, a rozvoj $f(x-ht)$ ako Taylorova rada, dostaneme

$$\begin{aligned} \text{var}\hat{f}(x) &\approx \frac{1}{nh}\int \left(f(x) - htf'(x) + \cdots\right)K(t)^2dt + O\left(\frac{1}{n}\right)\\ &= \frac{1}{nh}f(x)\int K(t)^2dt + O\left(\frac{1}{n}\right)\\ &\approx \frac{1}{nh}f(x)\int K(t)^2dt && (2.15)\end{aligned}$$

Pretože f je pravdepodobnostná funkcia hustoty, integrál 2.15 cez x dáva jednoduchú aproximáciu.

$$\int \text{var}\hat{f}(x)dx \approx \frac{1}{nh}\int K(t)^2dt \quad (2.16)$$

Predpokladajme, že chceme zvoliť h, aby bola stredná integrálna kvadratická chyba čo najmenšia. Porovnanie aproximácie 2.14 a 2.16 pre dva prvky strednej integrálnej kvadratickej chyby ukazuje jeden zo základných problémov odhadu hustoty. Ak sa pokúsime eliminovať odchýlku odhadu, použije sa veľmi malé h, a integrálny rozptyl bude veľký. Naopak, veľké h zníži náhodnú odchýlku určenú rozptylom, pridá do odhadu systematické chyby alebo odchýlku odhadu. Táto rozprava poskytuje matematické vysvetlenie chovania ilustrujúce na obrázkoch 2.8 a 2.9. Môže byť stresujúce, že pri použití ktorejkoľvek metódy odhadu hustoty, voľba vyhladzovacieho parametra naznačuje prijatie kompromisu medzi náhodou a systematickou chybou.

Ideálna šírka okna a jadro

Ideálnu hodnotu h dostaneme minimalizovaním približnej strednej integrálnej kvadratickej chyby

$$\frac{1}{4}h^4k_2^2\int f''(x)^2\,dx + n^{-1}h^{-1}\int K(t)^2\,dt, \tag{2.17}$$

je h_{opt}, kde

$$h_{opt} = k_2^{-\frac{2}{5}}\left(\int K(t)^2\,dt\right)^{\frac{1}{5}}\left(\int f''(x)^2\,dx\right)^{-\frac{1}{5}} n^{-\frac{1}{5}}. \tag{2.18}$$

Rovnica 2.18 pre optimálnu šírku okna je sklamaním, lebo ukazuje, že h_{opt} záleží na neznámej hustote, ktorú chceme odhadnúť. Napriek tomu boli získané niektoré užitočné závery. Za prvé, ideálna šírka okna bude konvergovať k nule s nárastom veľkosti vzorky, ale veľmi pomaly. Za druhé, keďže člen $\int f''^2$ meria v zmysle rýchlosť zmeny hustoty f. Z 2.18 môžeme vidieť, že menšie hodnoty h sú vhodné pre rýchlejšiu zmenu hustôt. Prirodzený prístup z 2.18 je voľba h k nejakej štandartnej hustote, ako je napríklad normálna hustota. Táto myšlienka bude prezkúmaná v podkapitole 2.3.4.

Substitúciou hodnoty h_{opt} z 2.18 späť do približnej strednej integrálnej kvadratickej chyby 2.17 ukazuje, že keď je h zvolený optimálne, približná hodnota strednej integrálnej kvadratickej chyby je

$$\frac{5}{4}C(K)\left(\int f''(x)^2dx\right)^{\frac{1}{5}} n^{-\frac{4}{5}} \tag{2.19}$$

kde konštanta $C(K)$ je daná

$$C(K) = K_2^{\frac{2}{5}}\left(\int K(t)^2dt\right)^{\frac{4}{5}}. \tag{2.20}$$

Predpis 2.19 ukazuje, že by sme mali vybrať jadro K s malou hodnotou $C(K)$. Ak zvolíme vyhladzovací parameter správne, teoreticky môžeme získať malú hodnotu strednej integrálnej kvadratickej chyby.

Zamerajme sa na jadrá, ktoré sú funkciami hustoty pravdepodobnosti. Tieto jadrá zaručia, že odhad $\hat{f}$ je všade nezáporný. Ak hodnota k_2 nie je rovná jednej, nahradíme jadro prepočítanou verziou $k_2^{\frac{1}{2}}K\left(k_2^{\frac{1}{2}}t\right)$. Toto nebude ovplyvňovať hodnotu $C(K)$.

Minimalizovaním $C(K)$ minimalizujeme $\int K(t)^2dt$ s obmedzeniami $\int K(t)dt = 1$ a $\int t^2K(t)dt = 1$. V inej súvislosti Hodges a Lehmann (1956) ukázal, že riešením je funkcia

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1-\frac{t^2}{5}\right), & -\sqrt{5} \le t \le \sqrt{5} \\ 0, & \text{inak.} \end{cases} \tag{2.21}$$

Označenie $K_e(t)$ je použité, lebo prvý ho použil v odhade hustoty Epanechnikov (1969), a tak sa často nazýva *Epanechnikove jadro.*

Môžeme posúdiť účinnosť hocijakého symetrického jadra K porovnaním s Epanechnikovým jadrom. Účinnosť jadra K je

$$\text{eff}(K) \quad = \quad \left(\frac{C(K_e)}{C(K)}\right)^{\frac{5}{4}} \tag{2.22}$$

$$= \frac{3}{5\sqrt{5}} \left(\int t^2 K(t) dt \right)^{-\frac{1}{2}} \left(\int K(t)^2 dt \right)^{-1}. \tag{2.23}$$

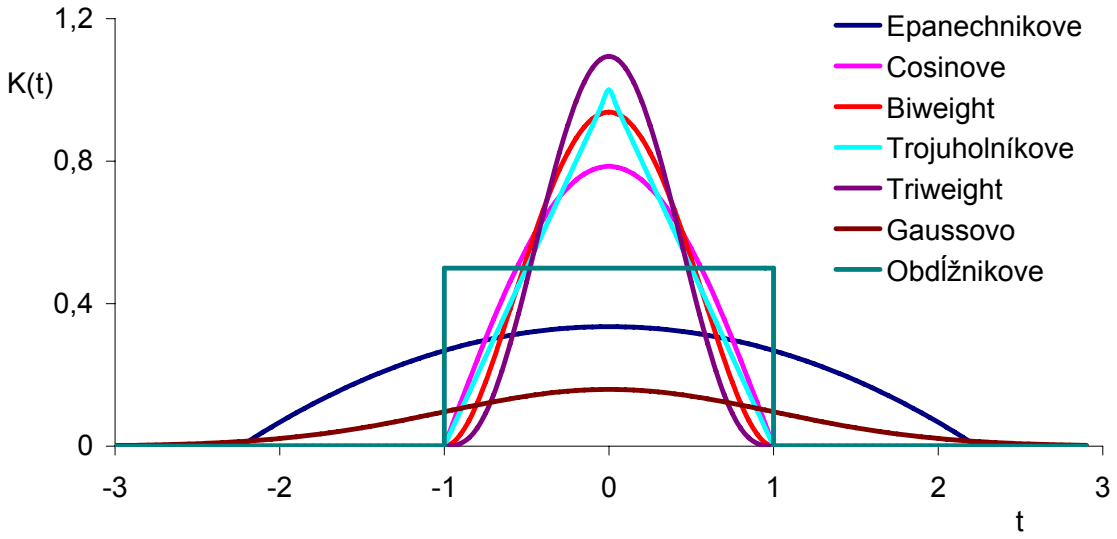
Dôvod pre mocninu $\frac{5}{4}$ v 2.22 je, aby pre veľké n bola stredná integrálna kvadratická chyba rovnaká keď použijeme rozsah náhodného výberu n a jadro K alebo keď použijeme rozsah náhodného výberu n *eff(K)* a jadro K_e. Niektoré jadrá a ich účinosti sú v tabuľke. Celkom je pozoruhodné, že získané účinnosti sú blízke jednej. Dokonca keď použijeme obdĺžnikové jadro, ktoré sa používa pri naivnom odhade, jeho účinnosť je skoro 0.93. Z tabuľky vidíme, že je veľmi malý rozdiel medzi rôznymi jadrami pre východziu strednú integrálnu kvadratickú chybu.

Jadro	K(t)		Účinnosť
Epanechnikove	$\frac{3}{4\sqrt{5}}\left(1-\frac{t^2}{5}\right)$	$\lvert t\rvert < \sqrt{5}$	1
	0	inak	
Cosinove	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}t\right)$	$\lvert t\rvert < 1$	$\left(\frac{48\sqrt{5}}{28\sqrt{\pi^2-8\pi^{2/5}}}\right) \approx 0.9995$
	0	inak	
Biweight	$\frac{15}{16}\left(1-t^2\right)^2$	$\lvert t\rvert < 1$	$\left(\frac{3087}{3125}\right)^{\frac{1}{2}} \approx 0.9939$
	0	inak	
Triweight	$\frac{35}{32}\left(1-t^2\right)^3$	$\lvert t\rvert < 1$	$\left(\frac{3861}{8750}\sqrt{5}\right) \approx 0.9867$
	0	inak	
Trojuholnikové	$1-\lvert t\rvert$	$\lvert t\rvert < 1$	$\left(\frac{243}{250}\right)^{\frac{1}{2}} \approx 0.9859$
	0	inak	
Gaussovo	$\frac{1}{\sqrt{2\pi}}e^{-\left(\frac{1}{2}\right)t^2}$		$\left(\frac{36\pi}{125}\right)^{\frac{1}{2}} \approx 0.9512$
Obdĺžnikové	$\frac{1}{2}$	$\lvert t\rvert < 1$	$\left(\frac{108}{125}\right)^{\frac{1}{2}} \approx 0.9295$
	0	inak	

2.3.4. Výber vyhladzovacieho parametra

Problém je, ako veľa vyhladiť, čo má rozhodujúci význam v odhade hustoty. Pred podrobnou diskusiou rôznych metód, má význam pozastaviť sa na nejaké poznámky všeobecnej povahy. Nikdy by nemalo byť zabudnuté, že vhodná voľba vyhladzovacieho parametra bude vždy ovplyvnená účelom, pre ktorý odhad hustoty je použitý. Ak účelom odhadu hustoty je posúdenie údajov na navrhnutie prijateľných modelov a hypotéz, potom pravdepodobne bude celkom samozrejme vhodná subjektívna voľba vyhladzovacieho parametra. Keď používame odhad hustoty na prezentáciu výsledkov, v prípade podhladenia, čitateľ môže ďalej vyhľadiť "od oka", ale nemôže jednoducho podhladiť.

Avšak, veľa aplikácií požaduje automatickú voľbu vyhladzovacieho parametra. Neskúsený uživateľ sa bude pravdepodobne cítiť šťastnejšie, ak je metóda plne automatická. Automatická voľba môže v každom prípade byť použitá ako štartovací bod pre nasledujúcu subjektívnu voľbu. Vedci prispievajú výsledkami a porovnávajú si ich, lebo chcú mať



Obr. 2.11: Jadrá

vzorovú metódu. Odhad hustoty je rutinne použitý na veľké rozsahy náhodných výberov, alebo časť veľkej procedúry, takže je potrebná automatická voľba.

V tomto pojednaní sme úmyselne použili slovo *automaticky* radšej ako *objektívne* pre metódy, ktoré nevyžadujú explicitné špecifikácie ovládacích parametrov. V pozadí procesu automatickej štatistickej procedúry leží vždy nebezpečie povzbudzujúce uživateľa nevenovať dosť uváženia na predchádzajúce predpoklady.

V nasledovných častiach budú popísané rôzne metódy na voľbu vyhladzovacieho parametra. Na riešenie tohto problému nie je doteraz všeobecne akceptovaný postup.

Subjektívna voľba

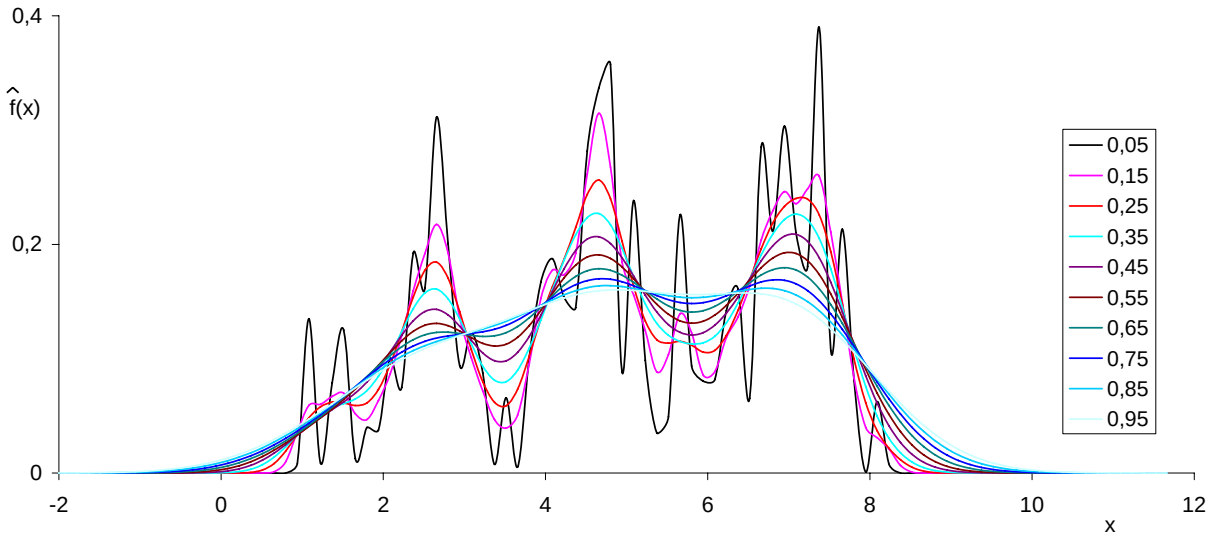
Prirodzená metóda voľby vyhladzovacieho parametra je nakreslenie niekoľkých kriviek a vybratie odhadu, ktorý je najviac podobný predchádzajúcej myšlienke o hustote. Pre veľa aplikácií bude tento prístup bezchybne vyhovujúci. Väčšie preniknutie do náhodného výberu môže dať postupné kreslenie niekoľkých kriviek vyhladenými rôznymi hodnotami, ako iba jednoduchá automatická krivka. Na obrázku 2.12 máme 10 odhadov trimodálnej hustoty s vyhladzovacími parametrami od 0.05 do 0.95.

Na obrázku je 2.13 je jadrový odhad trimodálnej hustoty pravdepodobnosti, kde $n = 99$. Gaussovo jadro je použité na odhad, vyhladzovací parameter je odhadnutý pomocou subjektívnej metódy a $h = 0.4$.

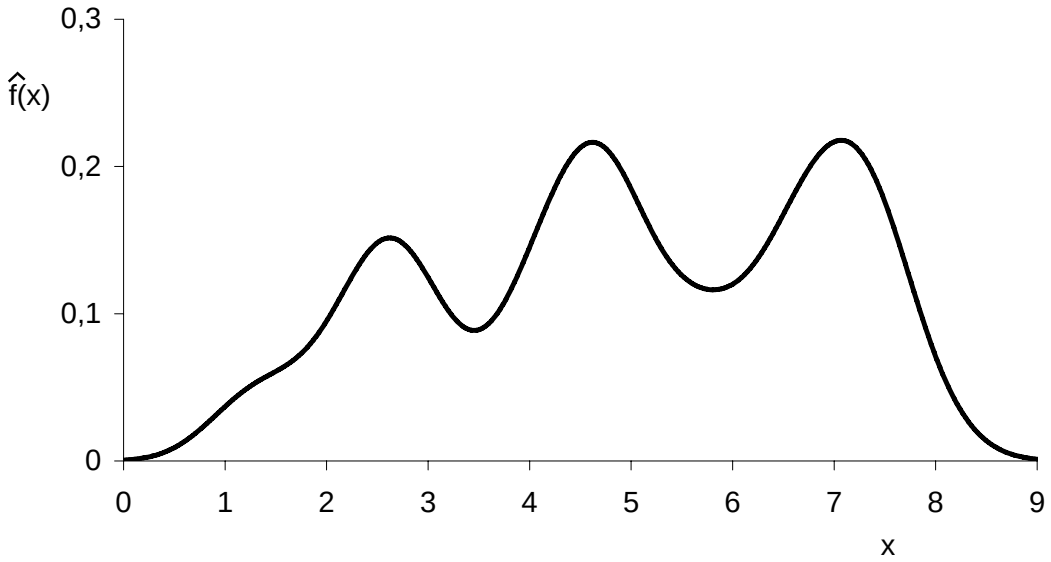
Normálne rozdelenie

Veľmi jednoduchý a prirodzený prístup je použitie štandardného rozdelenia na priradenie hodnoty členu $\int f''(x)^2 dx$ vo výraze 2.18 pre ideálnu šírku okna. Napríklad normálne rozdelenie s rozptylom σ^2 kde, člen ϕ má štandardnú normálnu hustotu,

$$\begin{aligned} \int f''(x)^2\,dx &= \frac{1}{\sigma^5}\int \phi''(x)^2\,dx \\ &= \frac{3}{8\sqrt{\pi}\sigma^5} \approx \frac{0.212}{\sigma^5}. \end{aligned} \tag{2.24}$$



Obr. 2.12: Odhady hustoty pre rôzne vyhladzovacie parametre



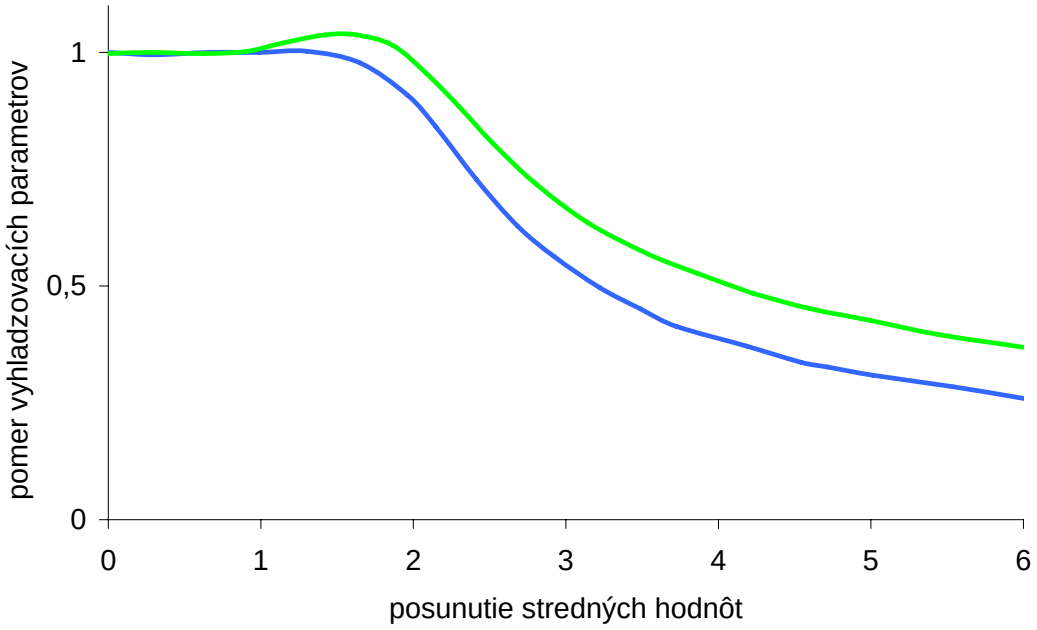
Obr. 2.13: Subjektívny odhad trimodálnej hustoty s vyhladzovacím parametrom 0.4.

Ak je použité Gaussovo jadro, potom šírku okna dostanieme z 2.18 a 2.24,

$$\begin{aligned} h_{opt} &= \frac{1}{\sqrt[10]{4\pi}} \frac{1}{\sqrt[5]{\frac{3}{8\sqrt{\pi}}}} \frac{\sigma}{\sqrt[5]{n}} \\ &= \frac{\sqrt[5]{\frac{4}{3}}\sigma}{\sqrt[5]{n}} \approx 1.06 \frac{\sigma}{\sqrt[5]{n}}. \end{aligned} \tag{2.25}$$

Voľba vyhladzovacieho parametra je rýchla, mali by sme odhadnúť σ z náhodného výberu a potom vložiť do 2.25. Pre odhad σ môže byť použitá smerodatná odchýlka alebo robustný odhad.

Keď má náhodný výber skutočne normálne rozdelenie, 2.25 bude fungovať dobre. Ale ak je náhodný výber multimodálny, môže trochu prehladzovať. A výsledok hodnoty



Obr. 2.14: Na obrázku je pomer asymptoticky optimálnej šírky okna 2.18 ku šírke okna danej subjektívnou voľbou. Zelená krivka: subjektívna voľba podľa smerodatnej odchýlky. Modrá krivka: subjektívna voľba podľa medzikvartilového rozpätia. Pomer je rátaný pre skutočnú hustotu, ktorá je zmes dvoch normálnych rozdelení, pre dané posunutie stredných hodnôt.

$(\int f''^2)^{\frac{1}{5}}$ je väčší vzhľadom k smerodatnej odchýlke. Tento výsledok je znázornený na obrázku 2.14, ktorý ukazuje pomer optimálneho vyhladzovaciho parametra 2.18 ku hodnote získanou použitím 2.25 ak je skutočné f rovné zmesi dvoch normálnych rozdelení, pre dané posunutie stredných hodnôt. Na obrázku vidíme, že pre posunutie v intervale (0,2) vzorec 2.25 počíta veľmi dobre. Samozrejme, keďže zmes hustôt je opäť normálna hustota na tomto intervale. Ale keď sa začne zmes podobať bimodálnej hustote, vzorec 2.25 začne prehladzovať čím ďalej viac, ku optimálnej voľbe vyhladzovacieho parametra.

Aby sme mohli skúmať citlivosť optimálnej šírky jadra na šikmosť a špicatosť na jednomodálnych rozdeleniach, odpovedajúce krivky ku obrázku 2.14, sú vypočítané pre lognormálne rozdelenie a t rozdelenia. Tieto sú na obrázkoch 2.15 a 2.16. Môžeme vidieť, že veľmi šikmé náhodné výbery, použitím 2.25 budú opäť prehladzovať, ale tento vzorec je pozoruhodne silný na špicatosť v t rozdeleniach. Lepší výsledok môžeme získať použitím robustného merania rozsahu. Vzorec 2.25 prepíšeme pomocou medzikvartilového rozpätia R normálneho rozdelenia (tj. polovica vzdialenosti kvartilov).

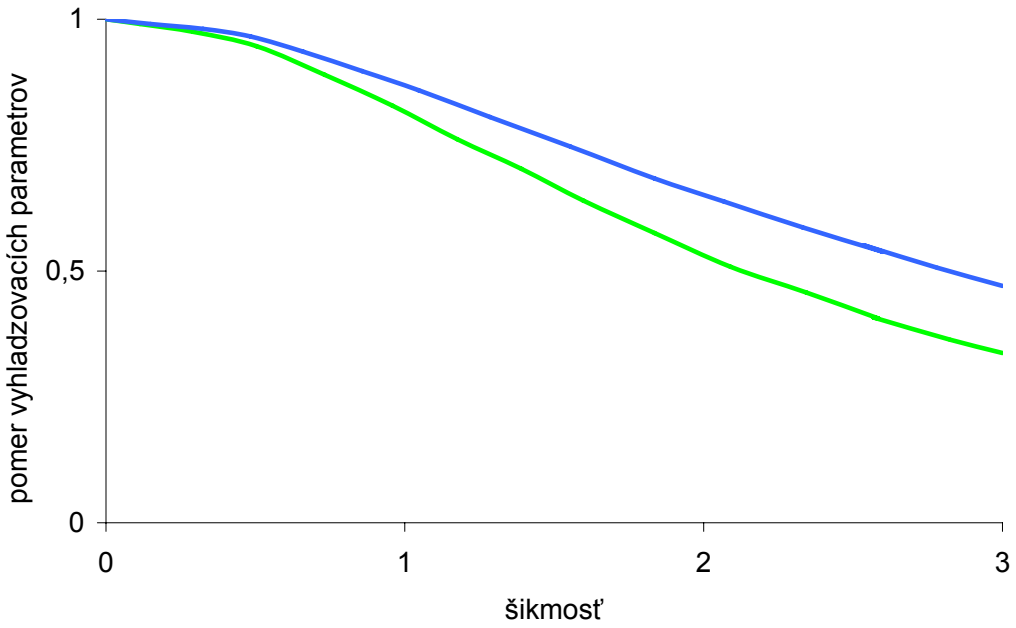
$$h_{opt} = 0.79Rn^{-\frac{1}{5}}. \tag{2.26}$$

Použitím 2.26 na dlhé chvosty a šikmé rozdelenia je oveľa lepšia modrá krivka na obrázkoch 2.15 a 2.16. Bohužiaľ pri použití 2.26 na bimodálnej hustote, prehladzuje dokonca ešte viacej. Najlepšie je použiť adaptivný odhad rozsahu

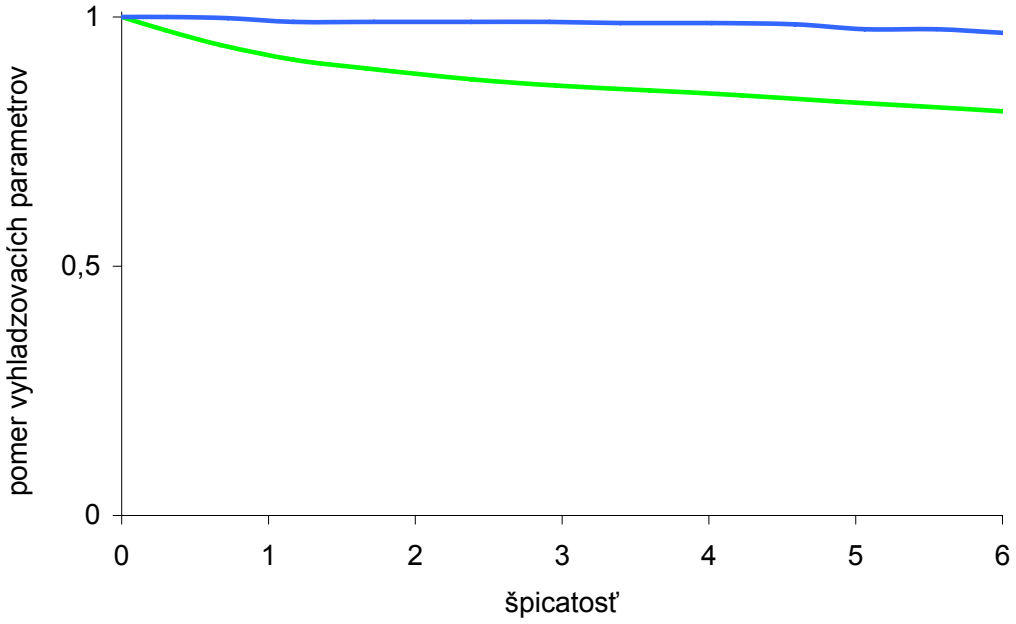
$$A = \min\left(\sigma, \frac{R}{1.34}\right) \tag{2.27}$$

namiesto σ v 2.25. Toto si dobre poradí s unimodálnymi hustotami, a celkom aj s bimodálnymi hustotami. Napríklad pre Gaussovo jadro

$$h = 0.9An^{-\frac{1}{5}}. \tag{2.28}$$



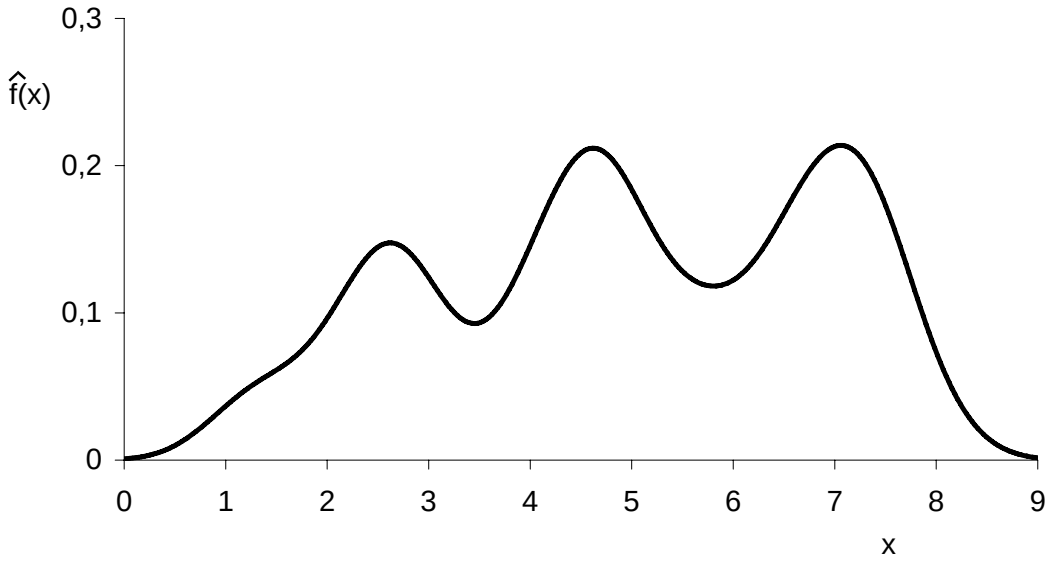
Obr. 2.15: Na obrázku je pomer asymptoticky optimálnej šírky okna ku šírke okna danej subjektívnou voľbou, ako na obrázku 2.14, pre lognormálne rozdelenie s danou šikmosťou.



Obr. 2.16: Pomer asymptoticky optimálnej šírky okna ku šírke okna danej subjektívnou voľbou, ako na obrázku 2.14, pre t rozdelenie s danými koeficientami špicatosti.

bude stredná integrálna kvadratická chyba do 10% optima pre všetky t rozdelenia, pre lognormálne rozdelenia so šikmosťou do 1.8, a pre zmes normálnych rozdelení s posunutím do troch smerodatných odchýliek. Podľa simulácií, ktoré urobil Silverman, i mimo týchto bodov bude zrejmá šikmosť a bimodalita pomocou 2.28, aj keď bude hustota trochu prehladená. Teda voľba 2.28 pre vyhladzovací parameter bude veľmi dobrá pre široký rozsah hustôt a jednoduchá na výpočet. Pre veľa účelov to bude určite primeraná voľba šírky okna, a pre iných to bude dobrý štartovací bod pre následné jemné doladenie. Na

obrázku je 2.17 je jadrový odhad trimodálnej hustoty pravdepodobnosti, kde $n = 99$. Gaussovo jadro je použité a vyhladzovací parameter je $h = 0.423$. Predpokladáme, že mody majú normálne rozdelenie.



Obr. 2.17: Odhad trimodálnej hustoty pomocou normálneho rozdelenia s vyhladzovacím parametrom 0.423.

Vzájomná kontrola pomocou najmenších kvadrátov

Vzájomná kontrola pomocou najmenších kvadrátov je plne automatická metóda (auto) pre voľbu vyhladzovacieho parametra. Táto metóda bola formulovaná pred pár rokmi, ale je postavená na extrémne jednoduchej myšlienke. Metódu navrhol Rudemo (1982) a Bowman (1984).

Majme odhad $\hat{f}$ a hustotu f, integrálnu kvadratickú chybu môžeme napísať

$$\int \left(\hat{f} - f\right)^2 = \int \hat{f}^2 - 2\int \hat{f}f + \int f^2. \tag{2.29}$$

Teraz výraz 2.29 nezávisí iba na $\hat{f}$, a tak ideálna voľba šírky okna (v zmysle minimalizovať integrálknu kvadratickú chybu) sa bude zhodovať voľbe, ktorá minimalizuje veličinu R definovanú

$$R\left(\hat{f}\right) = \int \hat{f}^2 - 2\int \hat{f}f. \tag{2.30}$$

Základný princíp vzájomnej kontroly pomocou najmenších kvadrátov je vytvorenie odhadu $R\left(\hat{f}\right)$ z náhodného výberu, a potom minimalizovať odhad cez h, aby sme dostali šírku okna. Člen $\int \hat{f}^2$ môžeme vytvoriť z $\hat{f}$. Definujme $\hat{f}_{-i}$ ako hustotu odhadu vytvorenú z celého náhodného výberu *okrem* X_i, teda

$$\hat{f}_{-i}(x) = \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{x - X_j}{h}\right). \tag{2.31}$$

Teraz definujme

$$M_0(h) = \int \hat{f}^2 - \frac{2}{n}\sum_i \hat{f}_{-i}(X_i)\,. \tag{2.32}$$

Hodnota M_0 závisí iba od náhodného výberu (hoci to nie je veľmi vhodný tvar pre jednoduché počítanie). Myšlienka vzájomná kontrola pomocou najmenších kvadrátov je minimalizovať hodnotu M_0 cez h. Ďalej budeme diskutovať, prečo od tejto procedúry môžeme očakávať dobré výsledky a tiež výpočetne jednoduchšiu aproximáciu M_0.

Aby sme mohli pochopiť, prečo minimalizovať M_0, je rozumný spôsob ako postupovať. Predpokladajme očakávanú hodnotu $M_0(h)$. Súčet výrazu 2.31 má pravdepodobnosť

$$\begin{aligned} \mathrm{E}\frac{1}{n}\sum_i \hat{f}_{-i}(X_i) &= \mathrm{E}\hat{f}_{-n}(X_n) \\ &= \mathrm{E}\int \hat{f}_{-n}(x)f(x)dx = \mathrm{E}\int \hat{f}(x)f(x)dx \end{aligned} \tag{2.33}$$

lebo $\mathrm{E}\left(\hat{f}\right)$ závisí iba na jadre a šírke okna, nie na rozsahu náhodného výberu. Vložením 2.33 späť do definície $M_0(h)$ ukazuje, že $\mathrm{E}M_0(h) = \mathrm{E}R\left(\hat{f}\right)$. Vyplýva to z 2.29, že $M_0(h) + \int f^2$ je pre všetky h nestranný odhad strednej integrálnej kvadratickej chyby vzhľadom k tomu, že člen $\int f^2$ je ten istý ako pre všetky h. Minimalizovanie $\mathrm{E}M_0(h)$ znamená minimalizovanie strednej integrálnej kvadratickej chyby. Predpokladajme, že minimum M_0 je blízko minimu $\mathrm{E}M_0$, teda môžeme dúfať, že minimum M_0 dáva dobrú voľbu vyhladzovacieho parametra.

Vyjadrenie hodnoty M_0 v tvare, ktorý je vhodný pre výpočet, najprv definujeme K^2 ako konvolúciu jadra so sebou. Napríklad ak K je normálne Gaussovo jadro, potom K^2 bude Gaussova hustota s rozptylom 2. Teraz predpokladajme, že K je symetrické a máme substitúciu $u = \frac{x}{h}$

$$\begin{aligned} \int \hat{f}(x)^2 dx &= \int \sum_i \frac{1}{nh}K\left(\frac{x - X_i}{h}\right) \times \sum_j \frac{1}{nh}K\left(\frac{x - X_j}{h}\right) dx \\ &= \frac{1}{n^2 h}\sum_i\sum_j \int K\left(\frac{X_i}{h} - u\right) K\left(u - \frac{X_j}{h}\right) du \\ &= \frac{1}{n^2 h}\sum_i\sum_j K^2\left(\frac{X_i - X_j}{h}\right). \end{aligned} \tag{2.34}$$

Tiež

$$\begin{aligned} \frac{1}{n}\sum \hat{f}_{-i}(X_i) &= \frac{1}{n}\sum_i \frac{1}{n-1}\sum_{j\neq i}\frac{1}{h}K\left(\frac{X_i - X_j}{h}\right) \\ &= \frac{1}{n(n-1)}\sum_i\sum_j \frac{1}{h}K\left(\frac{X_i - X_j}{h}\right) - \frac{1}{(n-1)h}K(0) \end{aligned} \tag{2.35}$$

Na nájdenie $M_0(h)$, výrazy 2.34 a 2.35 môžeme dosadiť do 2.32. Veľmi blízko súvisiaca hodnota funkcie $M_1(h)$, stále jednoduchšie zrátateľná, je daná vymenením zložky $\frac{1}{n-1}$ v 2.35 za jednoduchší $\frac{1}{n}$, a potom substitúciou do 2.32 dostanieme

$$M_1(h) = \frac{1}{n^2 h} \sum_i \sum_j K^* \left(\frac{X_i - X_j}{h} \right) + \frac{2}{nh} K(0) \tag{2.36}$$

kde funkcia K^* je definovaná

$$K^*(t) = K^2(t) - 2K(t). \tag{2.37}$$

Použitie vzťahu 2.36 zaberie počítaču veľa času, potrebuje $\frac{1}{2}n(n-1)$ výpočtov funkcie K^*, ktoré sú potrebné pre každú funkciu $M_1(h)$, lebo sú potrebné k minimalizácii cez h. Výpočty sa môžu jednoducho vymknúť spod kontroly.

Pre všetky, avšak veľmi malé rozsahy náhodných výberov je priame použitie výpočtu podľa definície jadrového odhadu veľmi neefektívna. Pre oveľa rýchlejší výpočet si všimnime, že jadrový odhad je konvolúcia náhodného výberu s jadrom a použitie Fourierovej transformácie na výpočet konvolúcie. Použitie rýchlej Fourierovej transformácie umožňuje nájdenie priamej a inverznej Fourierovej transformácie veľmi rýchlo. Tento algoritmus vyvinul [9].

Majme funkciu g, označme $\tilde{g}$ jej Fourierovu transformáciu

$$\tilde{g}(s) = \frac{1}{\sqrt{2\pi}} \int e^{ist} g(t) dt.$$

Definujme $u(s)$ ako Fourierovu transformáciu náhodného výberu,

$$u(s) = \frac{1}{\sqrt{2\pi} n} \sum_{j=1}^{n} e^{isX_j}.$$

Nech $\tilde{f}_n(s)$ je Fourierova transformácia jadra odhadu hustoty

$$\tilde{f}_n(s) = \sqrt{2\pi} \tilde{K}(hs) u(s), \tag{2.38}$$

čo je konvolúcia pre Fourierove transformácie. Použili sme vlastnosť Fourierovej transformácie, pomocou jadra $\frac{1}{h} K\left(\frac{t}{h}\right)$ je $\tilde{K}(hs)$. Vzorec 2.37 je mimoriadne vhodný pre použitie, keď K je Gaussovo jadro. V tomto prípade Fourierova transformácia K môže byť explicitne nahradená

$$\tilde{f}_n(s) = e^{-\frac{1}{2}h^2 s^2} u(s). \tag{2.39}$$

Základná idea algoritmu, ktorá bude ujasnená v tejto sekcii, je použitie rýchlej Fourierovej transformácie a nájdením funkcie u a tiež inverznej $\tilde{f}_n$ na nájdenie odhadu hustoty $\tilde{f}$.

Hľadanie hodnoty $M_1(h)$ pomocou vzájomnej kontroly najmenších kvadrátov z Fourierovej transformácie je priamočiare. Definujme

$$\begin{aligned} v(s) &= \frac{1}{\sqrt{2\pi} n^2} \sum \sum e^{is(X_j - X_k)} \\ &= \sqrt{2\pi} \left| u(s) \right|^2 . \end{aligned} \tag{2.40}$$

Fourierova transformácia funkcie K^* definovanej v (2.38) je

$$\begin{aligned} \tilde{K^*}(s) &= \tilde{K^2}(s) - 2\tilde{K}(s) \\ &= \sqrt{2\pi} \tilde{K}(s)^2 - 2\tilde{K}(s) \qquad (2.41) \\ &= \frac{1}{\sqrt{2\pi}} \left(e^{-s^2} - 2e^{-\frac{1}{2}s^2} \right) \qquad (2.42) \end{aligned}$$

v špeciálnom prípade Gaussovho jadra. Definujme funkciu

$$\psi(t) = \frac{1}{n^2} \sum_i \sum_j \frac{1}{h} K^* \left(\frac{X_i - X_j}{h} - t \right). \tag{2.43}$$

Potom je hodnota $M_1(h)$ pomocou vzájomnej kontroly najmenších kvadrátov

$$M_1(h) = \psi(0) + \frac{2}{nh} K(0). \tag{2.44}$$

Teraz

$$\begin{aligned} \tilde{\psi}(s) &= \sqrt{2\pi} \tilde{K}^*(hs) v(s) \\ &= 2\pi \tilde{K}^*(hs) \left| u(s) \right|^2 \end{aligned}$$

a tak

$$\begin{aligned} \psi(0) &= \frac{1}{\sqrt{2\pi}} \int \tilde{\psi}(s) ds \\ &= \sqrt{2\pi} \int \tilde{K}^*(hs) \left| u(s) \right|^2 ds \\ &= \int \left(e^{-h^2 s^2} - 2e^{-\frac{1}{2} h^2 s^2} \right) \left| u(s) \right|^2 ds \end{aligned} \tag{2.45}$$

ak je použité Gausovo jadro. Substitúciou 2.45 do 2.44 dostanieme hodnotu $M_1(h)$ pomocou vzájomnej kontroly najmenších kvadrátov. Poznamenajme, že nie je dokonca potrebné invertovať transformáciu na nájdenie hodnoty $M_1(h)$. Keďže rýchla Fourierova transformácia dáva diskrétnu Fourierovu transformáciu postupnosti skôr ako Fourierova transformácia funkcie, je potrebné urobiť nepatrné úpravy v postupe. Predpokladajme interval $[a, b]$, v ktorom leží náhodný výber. Tento interval by mal byť zvolený dosť veľký. Položíme $a < \min(X_i) - 3h$ a $b > \max(X_i) + 3h$, tento interval je postačujúci pre použitie Gaussovho jadra.

Zvoľme $M = 2^r$ pre nejaké celé číslo r. Odhad hustoty bude nájdený na M bodoch na intervale $[a, b]$, a zvolením $r = 7$ alebo 8 dostaneme výborné výsledky. Definujme

$$\begin{aligned} \delta &= \frac{b-a}{M} \\ t_k &= a + k\delta \qquad \text{pre } k = 0, 1, \ldots, M-1. \end{aligned}$$

Diskretizujme náhodný výber nasledovne. Ak náhodná veličina náhodného výberu X leží v intervale $[t_k, t_{k+1}]$, rozdelí sa na váhu $\frac{1}{n\delta^2}(t_{k+1} - X)$ v t_k a váhu $\frac{1}{n\delta^2}(X - t_k)$ v t_{k+1}. Tieto váhy sa spočítajú cez všetky body X_i, a dostaneme postupnosť (ξ_k). Teraz, pre $-\frac{1}{2}M \leq l \leq \frac{1}{2}M$, definujme Y_l ako diskrétnu Fourierovu transformáciu

$$Y_l = \frac{1}{M} \sum_{k=0}^{M-1} \xi_k e^{\frac{i2\pi kl}{M}}$$

ktorá môže byť nájdená pomocou Fourierovej transformácie.

Definujme

$$s_l = \frac{2\pi l}{b-a}$$

a na chvíľu predpokladajme, že $a = 0$. Potom použitím definície váh ξ_k,

$$\begin{aligned} Y_l &= \frac{1}{M}\sum_k \xi_k e^{it_k s_l} \\ &\approx \frac{1}{nM\delta}\sum_j e^{is_l X_j} \qquad (2.46) \\ &= \frac{\sqrt{2\pi}}{b-a} u\left(s_l\right). \qquad (2.47) \end{aligned}$$

Aproximácia 2.47 sa bude zhoršovať s rastúcim $|s_l|$, odkedy ďalší krok v algoritme vynásobený všetkými Y_l pre veľké $|l|$ bude veľmi malý faktor, v praxi toto nenastane.

Definujme postupnosť ζ_l^* pomocou

$$\zeta_l^* = e^{-\frac{1}{2}h^2 s_l^2} Y_l \qquad (2.48)$$

a nech ζ_k je inverzná diskrétna Fourierova transformácia ζ_l^* . Potom

$$\begin{aligned} \zeta_k &= \sum_{l=-\frac{M}{2}}^{\frac{M}{2}} e^{-\frac{2\pi i k l}{M}} \zeta_l^* \\ &\approx \sum_l e^{\frac{-is_l t_k \sqrt{2\pi}}{b-a}} e^{-\frac{1}{2}h^2 s_l^2} u(s_l) \\ &\approx \frac{1}{\sqrt{2\pi}} \int e^{-ist_k} e^{-\frac{1}{2}h^2 s^2} u(s) ds \qquad (2.49) \\ &= \hat{f}\left(t_k\right) \end{aligned}$$

keďže (2.49) je inverzná Fourierova transformácia $\hat{f}$ ako odvodená v 2.38. Prípad obecného a je trochu komplikovanejší, ale konečný výsledok (2.48) je úplne rovnaký.

Teda odhad hustoty môžeme nájsť na delení t_k pomocou nasledovného algoritmu:

1. Diskretizujte na nájdenie váhovej postupnosti ξ_k.
2. Rýchla Fourierova transformácia na nájdenie postupnosti Y_l.
3. Použi 2.48 k nájdeniu postupnosti ζ_l^*.
4. Inverzná Fourierova transformácia na nájdenie postupnosti $\hat{f}\left(t_k\right)$.
5. Ak sú potrebné odhady s inými šírkami okien pre rovnaký náhodný výber, opakujte iba kroky 3 a 4.

Použitím faktu, že (Y_l) je Fourierova transformácia reálnej postupnosti, dosiahneme značné úspory: pamäťové a skrátený času výpočtu.

Zaoberajme sa teraz otázkou hľadania hodnoty $M_1(h)$ pomocou vzájomnej kontroly najmenších kvadrátov. Aproximujeme integrál 2.45 sumou, a substitúciou 2.47,

$$\begin{aligned} \psi(0) &= (b-a)\sum_{l=-\frac{M}{2}}^{\frac{M}{2}} \left(e^{-h^2 s_l^2} - 2e^{-\frac{1}{2}h^2 s_l^2}\right)|Y_l|^2 \\ &= -1 + 2(b-a)\sum_{l=1}^{\frac{M}{2}} \left(e^{-h^2 s_l^2} - 2e^{-\frac{1}{2}h^2 s_l^2}\right)|Y_l|^2 \qquad (2.50) \end{aligned}$$

keďže $Y_0 = \frac{1}{M}\sum \xi_k = \frac{1}{M\delta} = \frac{1}{b-a}$ a $|Y_l| = |Y_{-l}|$ pre všetky l. Substitúciou (2.50) späť do (2.44) dostávame

$$\frac{1}{2}\left(1 + M_1(h)\right) = (b-a)\sum_{l=1}^{\frac{M}{2}} \left(e^{-h^2 s_l^2} - 2e^{-\frac{1}{2}h^2 s_l^2}\right)|Y_l|^2 + \frac{1}{nh\sqrt{2\pi}}. \tag{2.51}$$

Toto kritérium sa jednoducho nájde pre rozsah hodnoty h. Pre hodnoty, ktoré budú zaujímavé, exponenciálny člen sa rýchlo stáva zanedbateľný, a súčet aktuálnych výpočtov bude oveľa menší ež $\frac{1}{2}M$ členov. Pre hľadanie minima M_1 by sme mali začať hľadať h na intervale

$$\frac{1}{4\sqrt[5]{n}}\sigma < h < \frac{3}{2\sqrt[5]{n}}\sigma \tag{2.52}$$

a potom rozšíriť interval, ak minimum leží na hranách intervalu. Konzervatívna minimalizačná stratégia hľadania h je kvazi newtonovské približovanie.

Metóda testovanie grafu

Metóda testovanie grafu je vyvinutá [9], je úplne rozdielna od predchádzajúcich metód. Prináša odhady, ktoré sú rovnomerne blízko ku skutočnej hustote. Majme konečný interval, to je trochu silnejšia požiadavka, než malá integrálna chyba, potom konvergencia $\sup\left|\hat{f} - f\right|$ k nule je dostatočná, ale nie nutná pre konvergenciu $\int\left(\hat{f} - f\right)^2$ k nule.

Princíp tejto metódy je veta, uvedená a dokázaná [9], ktorá dáva nasledovný výsledok. Predpokladajme symetrické, dvakrát diferencovateľné jadro K splňujúce isté podmienky regularity, a že $\int x^2 K(x)dx$ je nenulové. Predpokladajme tiež, že neznáma hustota f má rovnomerne spojité a ohraničené druhé derivácie. Teraz predpokladajme, že h je zvolené ako funkcia n, na zabezpečenie najrýchlejšej možnej konvergencie $\sup\left|\hat{f} - f\right|$ k nule. Inak povedané, h je zvolené minimalizovaním maximálnej chyby v odhade hustoty. Potom použitie rovnakej voľby šírky okna bude prípad, že keď $n \to \infty$,

$$\frac{\sup\left|\hat{f}'' - \mathrm{E}\hat{f}''\right|}{\sup\left|\mathrm{E}\hat{f}''\right|} \to k \tag{2.53}$$

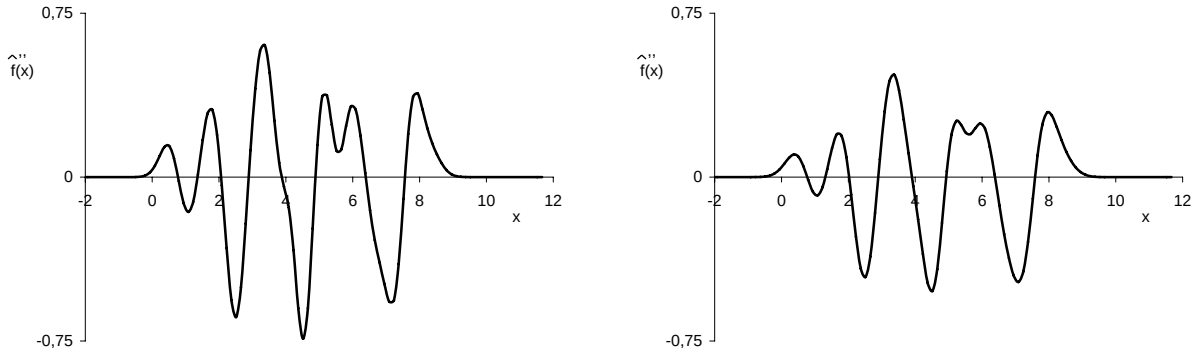
kde konštanta k závisí iba na jadre, a je daná

$$k = \frac{1}{2}\int\left|x^2 K(x)dx\right|\left(\frac{\int\left(K''\right)^2 dx}{\int K^2 dx}\right)^{\frac{1}{2}}$$

Keď je K Gaussovo jadro, potom konštanta k je približne 0.4.

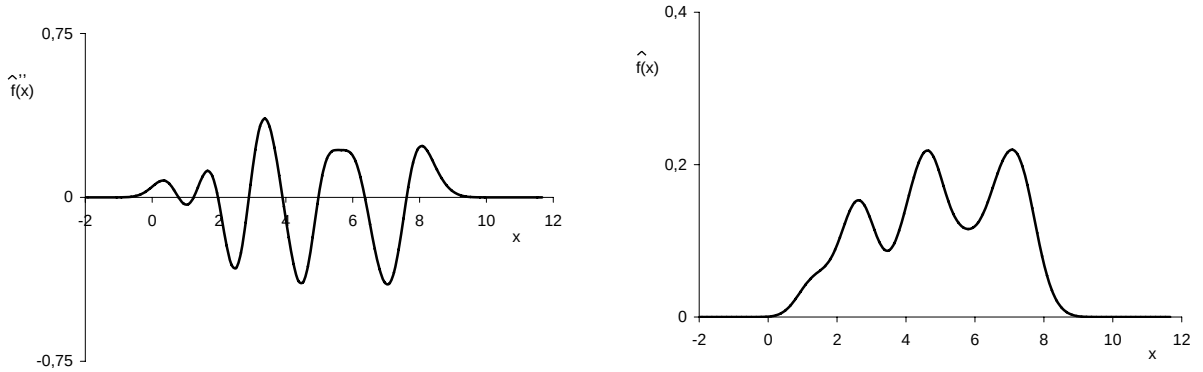
Člen $\hat{f}'' - \mathrm{E}\hat{f}''$ v čitateli 2.53 predstavuje náhodný šum krivky $\hat{f}''$, zatiaľ čo menovateľ $\mathrm{E}\hat{f}''$ je trend tejto krivky. Takže 2.53 môžeme opäť napísať, že pre dobrý odhad hustoty, velkosť šumu $\hat{f}''$ bude asi polovica maximálne hodnoty trendu krivky. Pre rozumné veľké rozsahy náhodného výberu sa bude javiť šum ako prudké zmeny krivky $\hat{f}''$.

Pri tejto metóde voľby vyhladzovacieho parametra postupujeme nasledovne. Nakreslíme "testovacie grafy" druhých derivácií $\hat{f}$ pre rôzne hodnoty h. Na základe malej rozpravy hore, ideálny testovací graf by mal mať prudké zmeny, ktoré sú celkom zretelné, ale neskrývajú úplne systematické striedanie. Vyberieme šírku okna, ktorého výnosy testovacieho grafu vyhovujú tomuto princípu, a použijeme túto šírku okna pre odhad hustoty.



Obr. 2.18: Testové grafy pre náhodný výber s rozsahom 99 z trimodálnej hustoty. Šírka okna obrázku vľavo je 0.33, a obrázku vpravo je 0.39.

Príklad aplikácie princípu testovaného grafu je na obrázkoch 2.18 a 2.19 vľavo. Môžeme vidieť, že šírka okna narastá cez celkom úzky rozsah, testovací graf sa "láme" a sú viditeľné prudké lokálne zmeny. Ak šírka okna je nižšia ako 0.33, test graf sa stáva veľmi rušivý. Keď je šírka okna nad 0.45, tak test graf má hladké krivky s malým, alebo neviditeľným náhodným šumom. Zistíme, že šírka okna je asi 0.39, čo je vhodný odhad f, odpovedajúci na obrázku 2.19 vpravo.



Obr. 2.19: Vľavo je test graf pre náhodný výber s rozsahom 99 z trimodálnej hustoty so šírkou okna 0.45. Vpravo je odhad trimodálnej hustoty so šírkou okna 0.39.

2.3.5. Asymptotické vlastnosti

O asymptotických vlastnostiach odhadu hustoty a obzvlášť jadrového odhadu je veľa literatúry. Uvedieme si zopár asymptotických výsledkov. Mnoho techník, ktoré sa používajú k overeniu diskutovaných výsledkov sú extrémne chytré a nepokúsime sa ich tu popísať.

Obvyklý asymptotický systém, v ktorom vety o jadrovom odhade hustoty sú dokázané, je predpokladať, že jadro K a neznáma hustota f sú pevné a splňujú dané regulárne podmienky. Zvažované odhady hustoty sú konštruované z prvých n meraní v nezávislom identickom rozdelení postupnosti $X_1, X_2, \ldots$ vytiahnuté z f. Predpokladá sa, že šírka okna h závisí do istej miery na rozsahu náhodného výberu n. Omedzujúce výsledky sú potom získané chovaním odhadu keď n ide do nekonečna. K tomu, aby sme explicitne urobili závislosť na n, musíme v tomto odstavci šírku okna označiť h_n.

Konzistentné výsledky

Veľká pozornosť bola venovaná podmienkam, za ktorých je jadrový odhad v rôznych zmysloch, konzistentný odhad skutočnej hustoty. Podmienky pre konzistenciu sú prekvapivo ľahké, hoci rýchlosť s ktorou odhad hustoty konverguje ku skutočnej hodnote môže byť veľmi pomalá.

Konzistentný odhad f v jednom bode x bola skúmaná [8]. Jeho predpoklad na jadro K bolo, že K bola ohraničená borelovská funkcia splňujúca

$$\int |K(t)|\, dt < \infty \qquad \text{a} \qquad \int K(t)dt = 1 \tag{2.54}$$

a

$$|tK(t)| \to 0 \qquad \text{pre} \qquad |t| \to \infty. \tag{2.55}$$

Tieto vlastnosti splňuje skoro každé jadro.

Predpokladalo sa, že šírka okna h_n splňuje

$$h_n \to 0 \qquad \text{a} \qquad nh_n \to \infty \qquad \text{pre} \qquad n \to \infty, \tag{2.56}$$

za týchto podmienok bolo ukázané, že pokiaľ f je spojitá v x,

$$\hat{f}(x) \to f(x) \text{ pravdepodobne keď } n \to \infty.$$

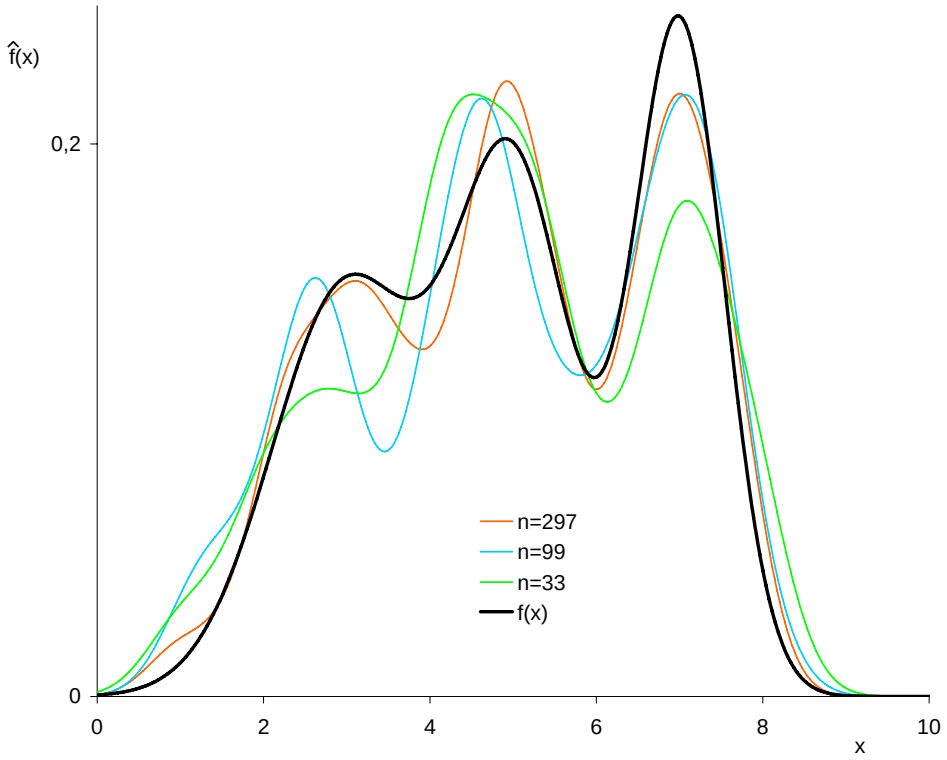
Vlastnosti 2.56 sú typické v tejto požadovanej konzistencii. Vyplýva z nich, že šírka okna sa musí zmenšovať s narastajúcim rozsahom náhodného výberu. Nesmie rýchlo konvergovať k nule ako $\frac{1}{n}$. Teda očakávaný rozsah náhodného výberu patriaceho do intervalu $x \pm h_n$ musí konvergovať do nekonečna, avšak pomaly, keď n konverguje k nekonečnu.

Uvedieme si príklad predchádzajúceho tvrdenia, kde hľadáme jadrové odhady trimodálnej hustoty pravdepodobnosti pomocou štyroch metód, pre rozsahy náhodných výberov od 33 do 8019. Gaussove jadrá boli použité v nasledovných príkladoch. Medzivýsledky sú v nasledovnej tabuľke.

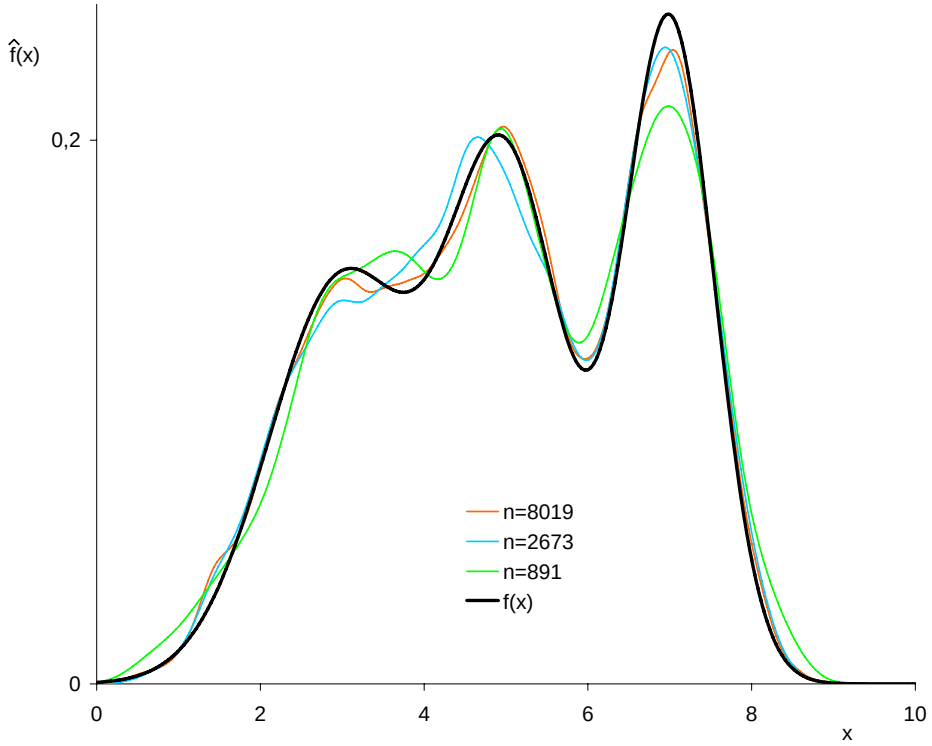
n	h_n	Metóda	nh_n	MISE
33	0.44	subjektívna voľba	14.5	5.03E−7
	0.463	normálne rozdelenie	15.3	5.16E−7
	0.81	auto	26.7	8.01E−7
	0.4	test graf	13.2	4.84E−7
99	0.4	subjektívna voľba	39.6	2.21E−7
	0.423	normálne rozdelenie	41.8	2.12E−7
	0.331	auto	32.8	2.90E−7
	0.39	test graf	38.6	2.27E−7
297	0.32	subjektívna voľba	95	8.64E−8
	0.353	normálne rozdelenie	105	9.43E−8
	0.279	auto	82.7	8.77E−8
	0.3	test graf	89.1	8.55E−8

891	0.25	subjektívna voľba	223	7.91E−8
	0.268	normálne rozdelenie	239	8.55E−8
	0.248	auto	221	7.85E−8
	0.235	test graf	209	7.49E−8
2673	0.2	subjektívna voľba	535	3.02E−8
	0.213	normálne rozdelenie	569	3.25E−8
	0.181	auto	483	2.77E−8
	0.18	test graf	481	2.76E−8
8019	0.15	subjektívna voľba	1203	1.59E−8
	0.171	normálne rozdelenie	1370	1.85E−8
	0.133	auto	1066	1.49E−8
	0.14	test graf	1123	1.52E−8

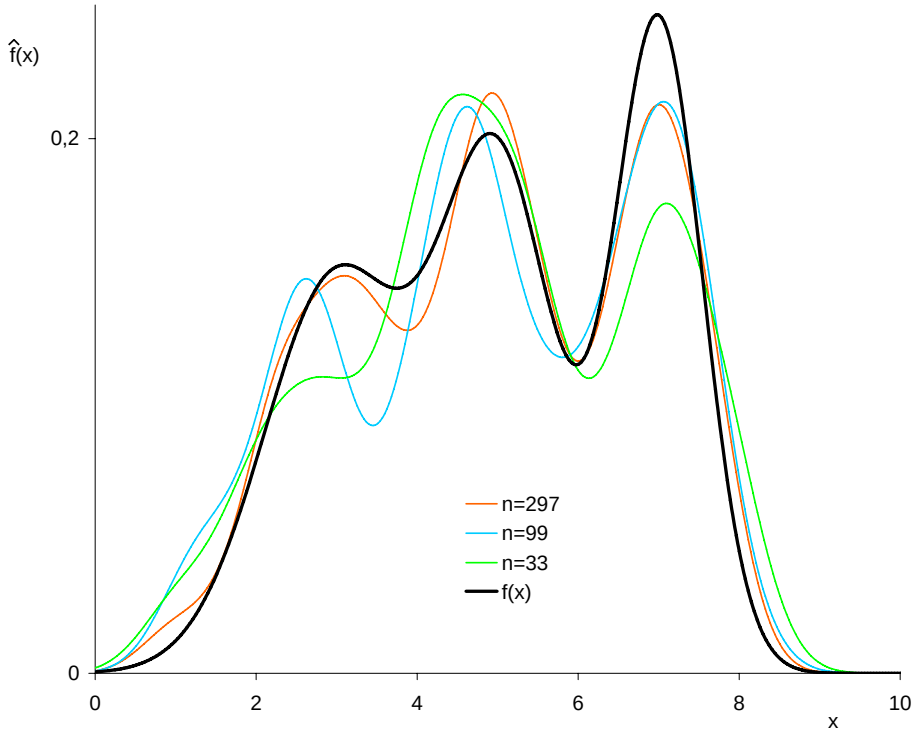
Na obrázkoch 2.20, 2.21, 2.22, 2.23, 2.24, 2.25, 2.26 a 2.27 vidíme, že jadrové odhady hustoty pravdepodobnosti, kde vyhladzovacie parametre sú určené pomocou rôznych metód sa s rastúcim rozsahom náhodného výberu približujú skutočnej hustote $f(x)$. To isté je vidieť z obrázku 2.30, kde tiež s rastúcim rozsahom náhodného výberu stredná integrálna kvadratická chyba klesá. Na obrázku 2.28 vidíme, že s rastúcim rozsahom náhodného výberu klesá vyhladzovací parameter h_n. A na obrázku 2.29 vidíme, že s rastúcim rozsahom náhodného výberu narastá nh_n.



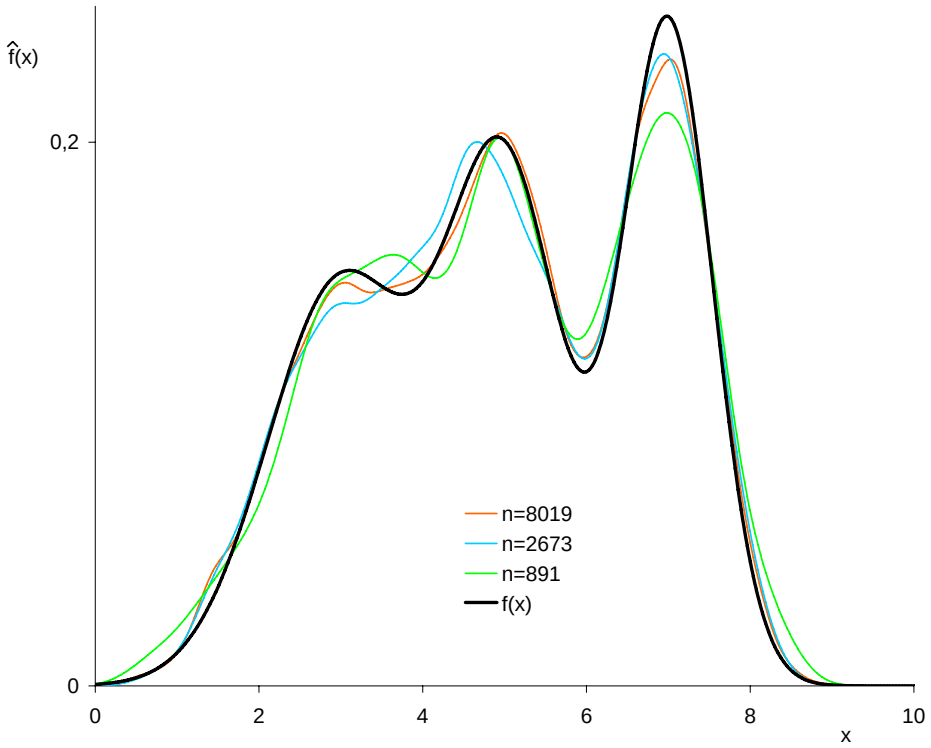
Obr. 2.20: Jadrové odhady hustoty, kde vyhladzovacie parametre sú určené subjektívnou voľbou pre rôzne n, a skutočná trimodálna hustota $f(x)$.



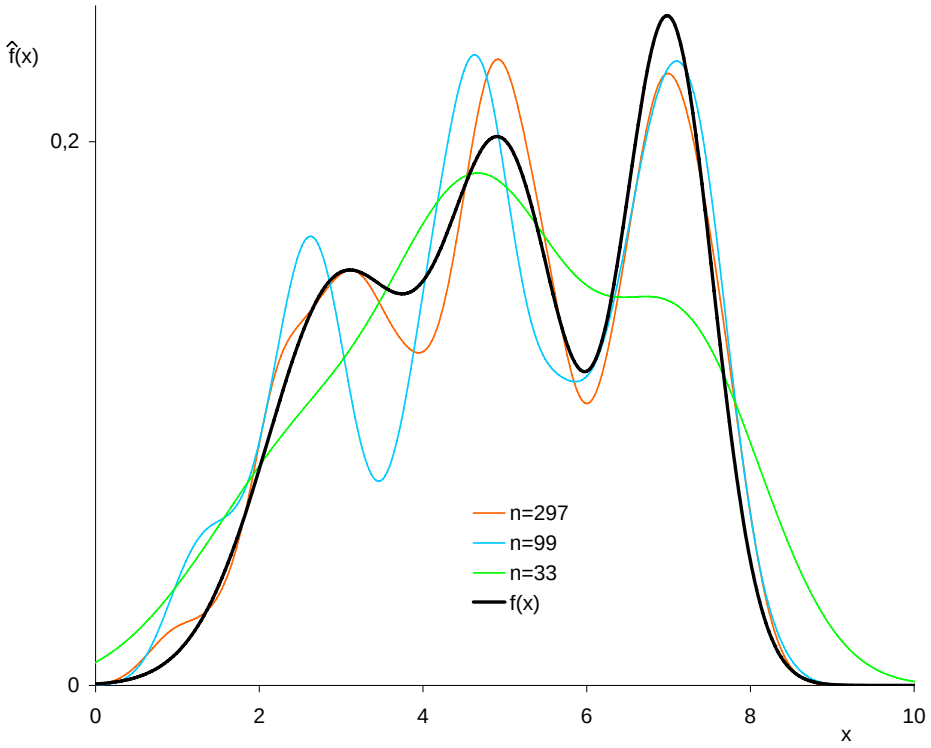
Obr. 2.21: Jadrové odhady hustoty, kde vyhladzovacie parametre sú určené subjektívnou voľbou pre rôzne n, a skutočná trimodálna hustota $f(x)$.



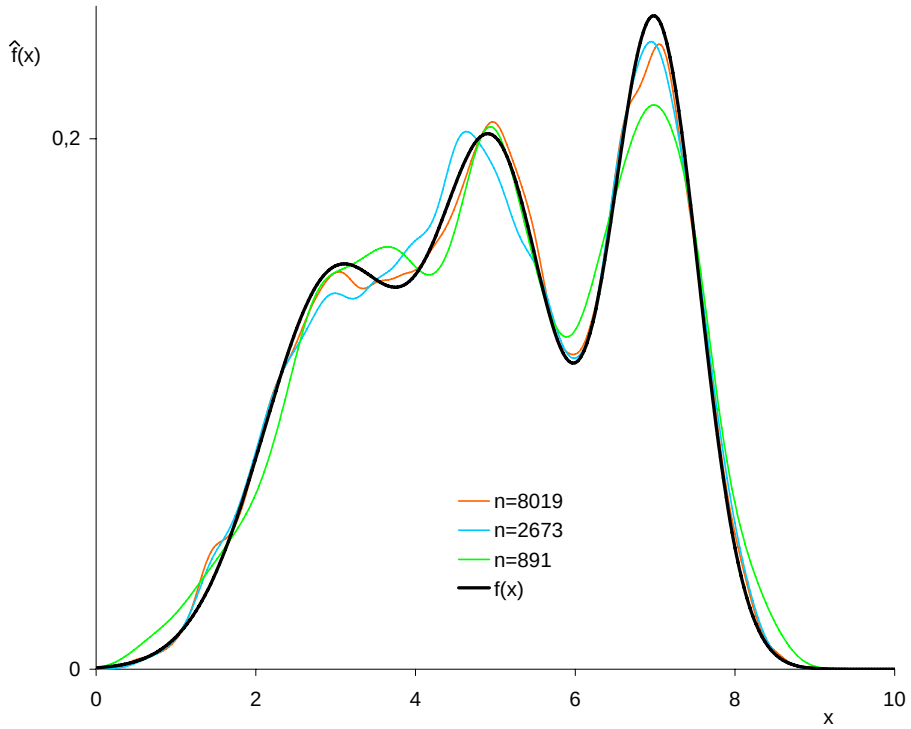
Obr. 2.22: Jadrové odhady hustoty, kde vyhladzovacie parametre sú určené pomocou normálnych rozdelení pre rôzne n, a skutočná trimodálna hustota $f(x)$.



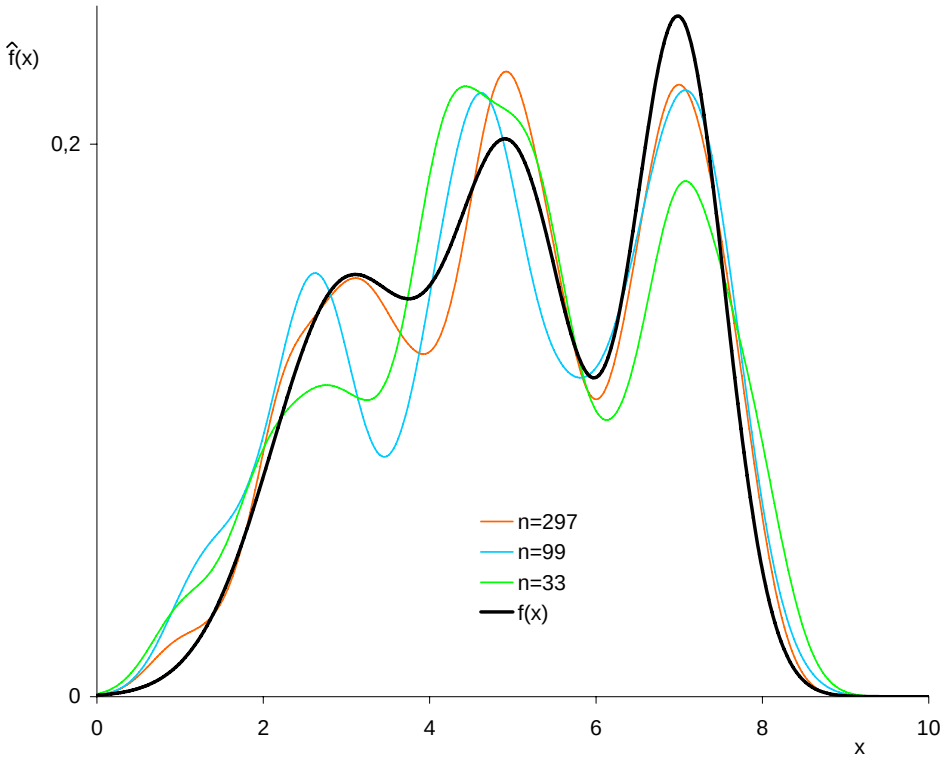
Obr. 2.23: Jadrové odhady hustoty, kde vyhladzovacie parametre sú určené pomocou normálnych rozdelení pre rôzne n, a skutočná trimodálna hustota $f(x)$.



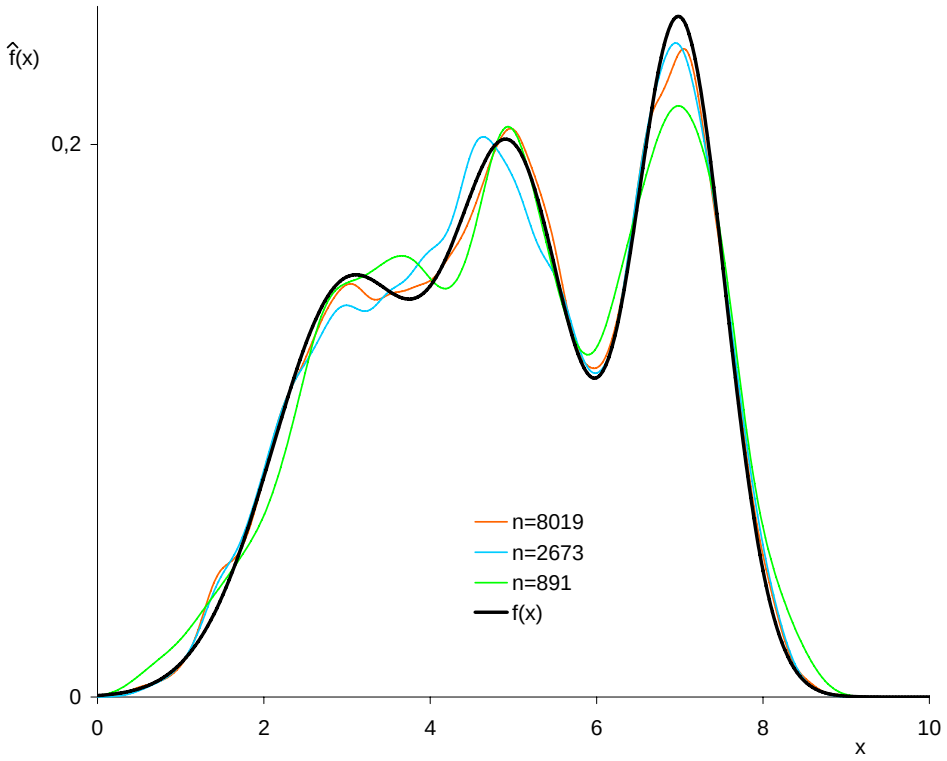
Obr. 2.24: Jadrové odhady hustoty, kde vyhladzovacie parametre sú určené metódou vzájomnej kontroly pomocou najmenších kvadrátov pre rôzne n, a skutočná trimodálna hustota $f(x)$.



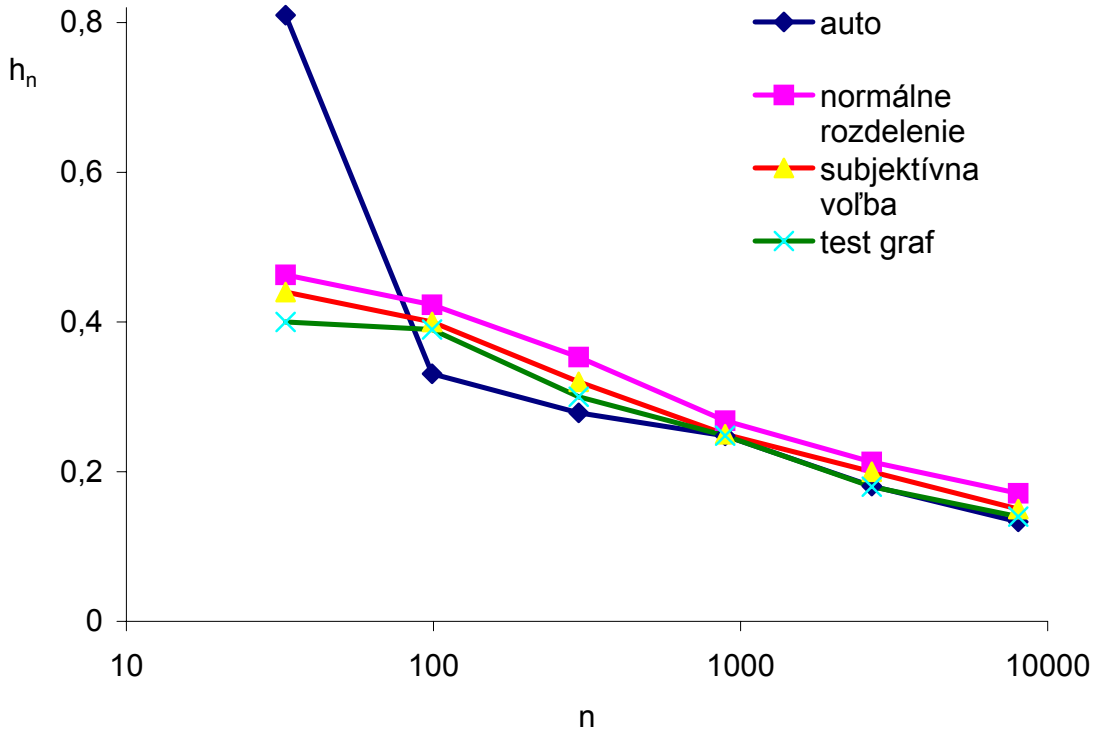
Obr. 2.25: Jadrové odhady hustoty, kde vyhladzovacie parametre sú určené metódou vzájomnej kontroly pomocou najmenších kvadrátov pre rôzne n, a skutočná trimodálna hustota $f(x)$.



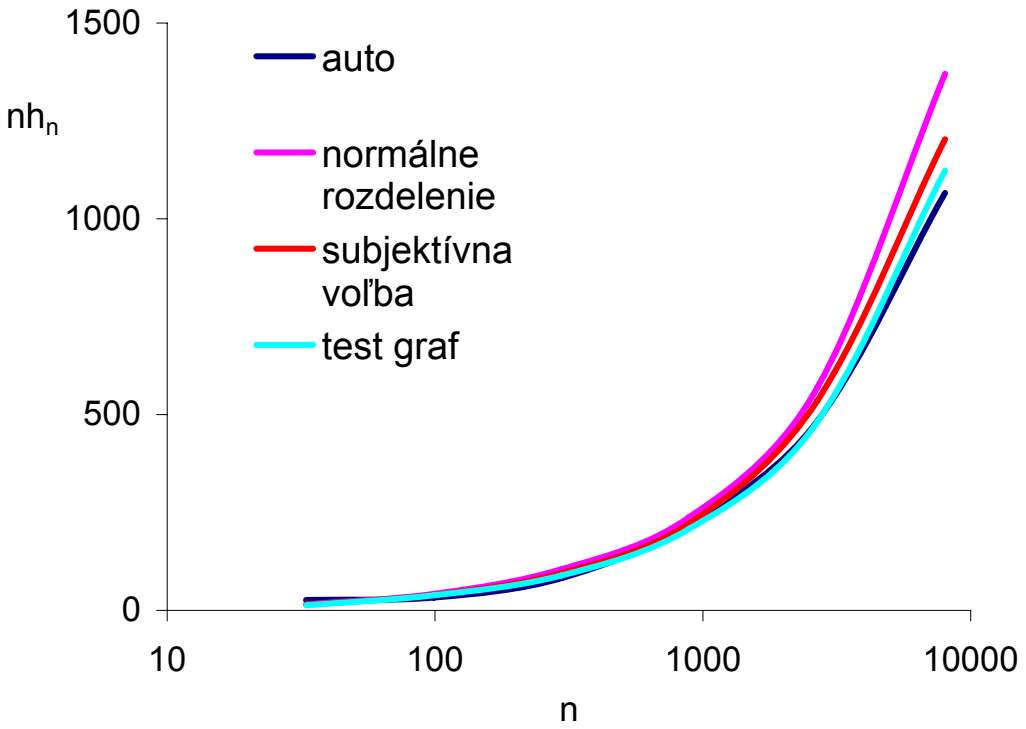
Obr. 2.26: Jadrové odhady hustoty, kde vyhladzovacie parametre sú určené metódou test graf pre rôzne n, a skutočná trimodálna hustota $f(x)$.



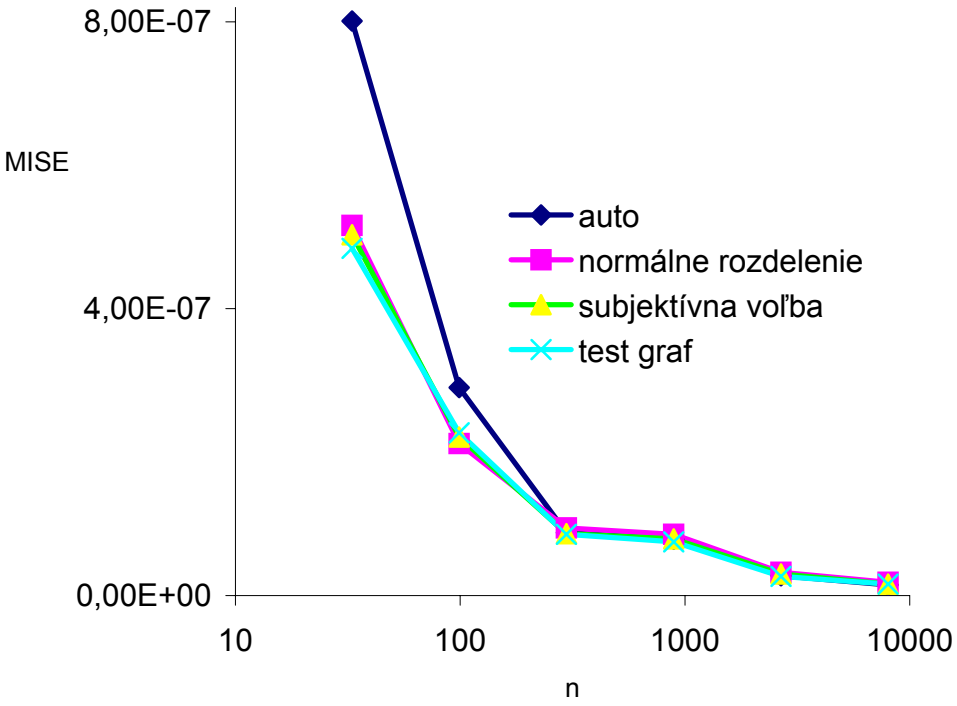
Obr. 2.27: Jadrové odhady hustoty, kde vyhladzovacie parametre sú určené metódou test graf pre rôzne n, a skutočná trimodálna hustota $f(x)$.



Obr. 2.28: Závislosť h_n na n pre rôzne metódy odhadu hustoty pravdepodobnosti trimodálnej hustoty.



Obr. 2.29: Závislosť nh_n na n pre rôzne metódy odhadu hustoty pravdepodobnosti trimodálnej hustoty.



Obr. 2.30: Závislosť MISE na n pre rôzne metódy odhadu hustoty pravdepodobnosti trimodálnej hustoty.

Akonáhle sa vzdialime od konzistencii v jednom bode, je potrebné určiť, ktorým smerom odhad krivky $\hat{f}$ aproximuje skutočnú hustotu f. Rovnomerná konzistencia je, že

pravdepodobne konvergencia $\sup\left|\hat{f}(x) - f(x)\right|$ ide do nuly. Toto sa domnieva niekoľko autorov [8].

Predpokladajme, že jadro K je ohraničené, má ohraničené zmeny (BV funkcia) a vyhovuje 2.54, a že množina nespojitostí ma Lebesgueovu mieru nula. Opäť týmto podmienkam vyhovuje prakticky každé jadro. Predpokladajme, že

$$f \text{ je rovnomerne spojitá na } (-\infty, \infty) \tag{2.57}$$

a že šírka okna h_n splňuje

$$h_n \to 0 \qquad a \qquad \frac{nh_n}{\log n} \to \infty \qquad \text{pre} \qquad n \to \infty \tag{2.58}$$

[8] ukázal podľa dômyselného komplikovaného argumentu, že potom s pravdepodobnosťou 1 nastane tento prípad

$$\sup_x \left|\hat{f}(x) - f(x)\right| \to 0 \qquad \text{pre} \qquad n \to \infty \tag{2.59}$$

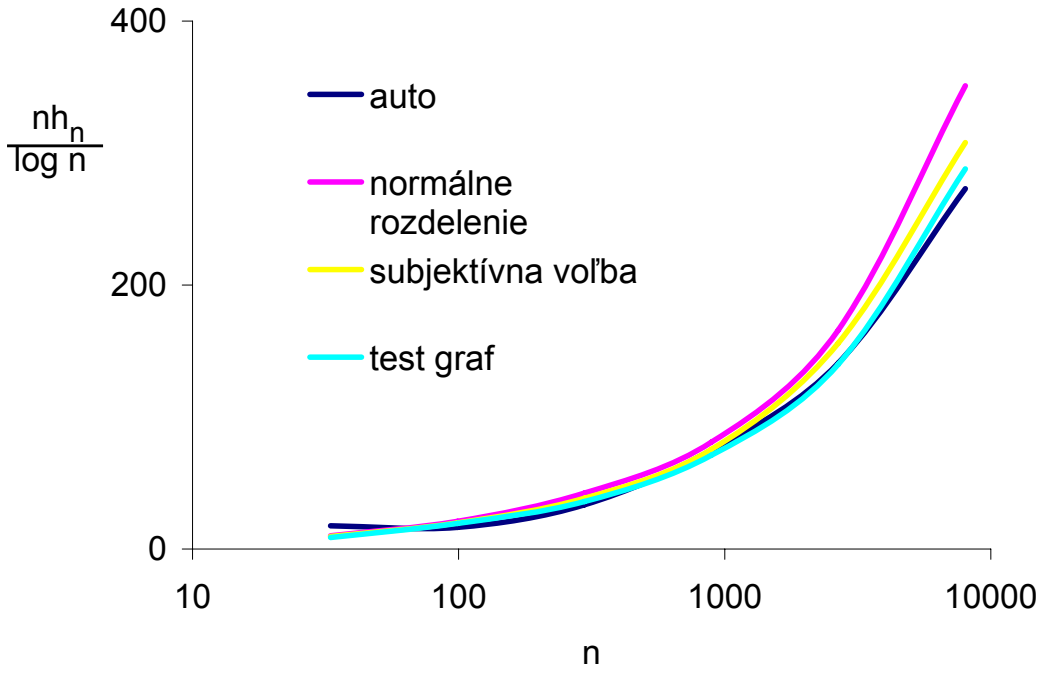
a okrem toho, že podmienky (2.58) a (2.59) sú nutné práve tak ako rovnomerná konzistencia. Podmienky (2.59) sú iba veľmi málo silnejšie ako (2.56) potrebné pre bodovú konzistenciu.

Uvedieme si príklad predchádzajúceho tvrdenia, kde hľadáme jadrové odhady trimodálnej hustoty pravdepodobnosti pomocou štyroch metód, pre rozsahy náhodných výberov od 33 do 8019. Medzivýsledky sú v nasledovnej tabuľke.

n	h_n	Metóda	$\frac{nh_n}{\log n}$	$\sup_x \left\|\hat{f}(x) - f(x)\right\|$
33	0.44	subjektívna voľba	9.56	7.16E−2
	0.463	normálne rozdelenie	10.1	7.46E−2
	0.81	auto	17.6	1.05E−1
	0.4	test graf	8.69	6.63E−2
99	0.4	subjektívna voľba	19.8	5.97E−2
	0.423	normálne rozdelenie	21	5.58E−2
	0.331	auto	16.4	7.30E−2
	0.39	test graf	19.4	6.15E−2
297	0.32	subjektívna voľba	38.4	2.85E−2
	0.353	normálne rozdelenie	42.4	3.39E−2
	0.279	auto	33.5	2.84E−2
	0.3	test graf	36	2.51E−2
891	0.25	subjektívna voľba	75.5	3.38E−2
	0.268	normálne rozdelenie	81	3.56E−2
	0.248	auto	74.9	3.36E−2
	0.235	test graf	71	3.24E−2

2673	0.2	subjektívna voľba	156	1.74E−2
	0.213	normálne rozdelenie	166	1.73E−2
	0.181	auto	141	1.77E−2
	0.18	test graf	140	1.78E−2
8019	0.15	subjektívna voľba	308	1.73E−2
	0.171	normálne rozdelenie	351	1.77E−2
	0.133	auto	273	1.74E−2
	0.14	test graf	288	1.73E−2

Na obrázku 2.31 vidíme, že s rastúcim rozsahom náhodného výberu narastá $\frac{nh_n}{\log n}$. Na obrázku 2.32 vidíme, že s rastúcim rozsahom náhodného výberu klesá $\sup_x \left|\hat{f}(x) - f(x)\right|$.

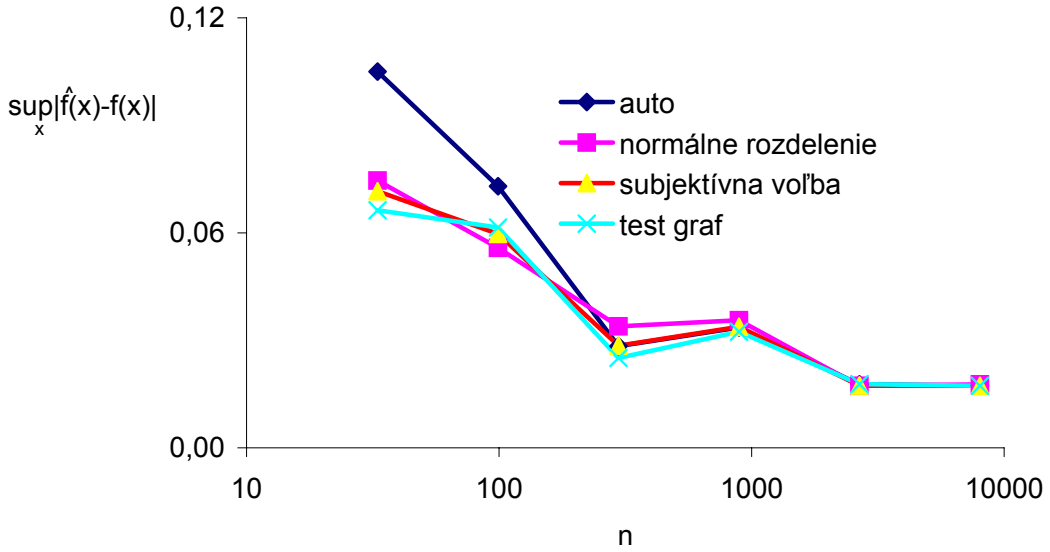


Obr. 2.31: Závislosť $\frac{nh_n}{\log n}$ na n pre rôzne metódy odhadu hustoty pravdepodobnosti trimodálnej hustoty.

[8] pojednáva o trochu slabšom druhu konzistencie. Opäť používajú komplikované technické argumenty k získaniu väčšiny významných výsledkov. Predpokladajme, že K je nezáporná borelovská funkcia, ktorej integrál je 1. Pomocou žiadnych predpokladov na neznámu hustotu f ukázali, že podmienky (2.56) sú nutné a dostatočné pre konvergenciu

$$\int \left|\hat{f}(x) - f(x)\right| dx \to 0 \qquad \text{s pravdepodobnosťou 1 pre} \qquad n \to \infty. \tag{2.60}$$

Uvediemi si príklad predchádzajúceho tvrdenia, kde hľadáme jadrové odhady trimodálnej hustoty pravdepodobnosti pomocou štyroch metód, pre rozsahy náhodných výberov od 33 do 8019. Medzivýsledky sú v nasledovnej tabuľke.

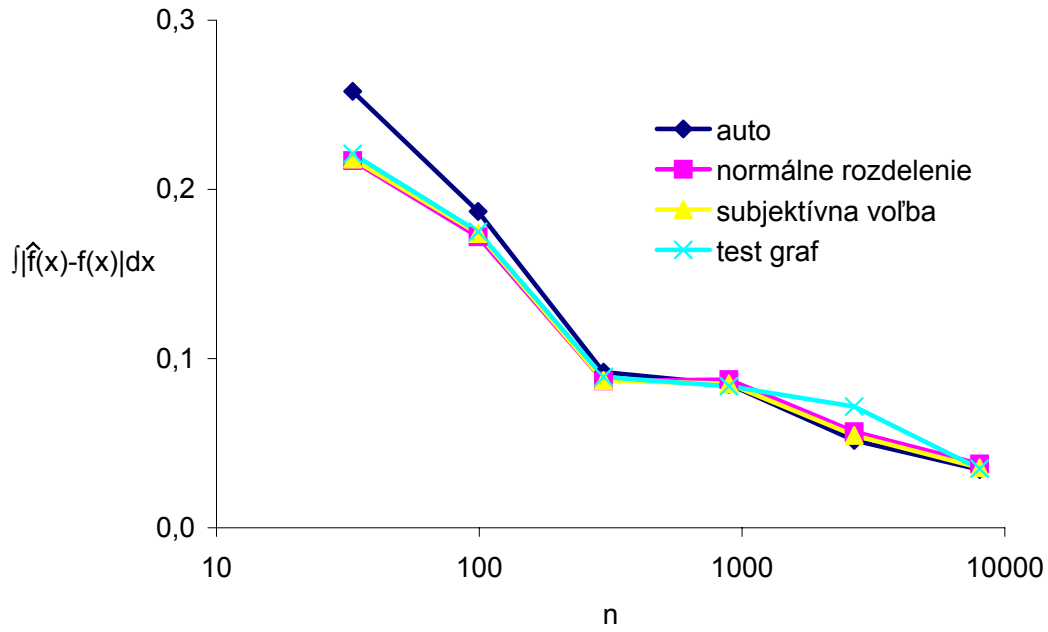


Obr. 2.32: Závislosť $\sup_x \left|\hat{f}(x) - f(x)\right|$ na n pre rôzne metódy odhadu hustoty pravdepodobnosti trimodálnej hustoty.

n	h_n	Metóda	nh_n	$\int \left\|\hat{f}(x) - f(x)\right\| dx$
33	0.44	subjektívna voľba	14.5	2.18E−1
	0.463	normálne rozdelenie	15.3	2.17E−1
	0.81	auto	26.7	2.58E−1
	0.4	test graf	13.2	2.21E−1
99	0.4	subjektívna voľba	39.6	1.74E−1
	0.423	normálne rozdelenie	41.8	1.72E−1
	0.331	auto	32.8	1.87E−1
	0.39	test graf	38.6	1.75E−1
297	0.32	subjektívna voľba	95	8.71E−2
	0.353	normálne rozdelenie	105	8.68E−2
	0.279	auto	82.7	9.20E−2
	0.3	test graf	89.1	8.91E−2
891	0.25	subjektívna voľba	223	8.51E−2
	0.268	normálne rozdelenie	239	8.76E−2
	0.248	auto	221	8.48E−2
	0.235	test graf	209	8.37E−2
2673	0.2	subjektívna voľba	535	5.46E−2
	0.213	normálne rozdelenie	569	5.68E−2
	0.181	auto	483	5.17E−2
	0.18	test graf	481	5.17E−2
8019	0.15	subjektívna voľba	1203	3.55E−2

0.171	normálne rozdelenie	1370	3.77E−2
0.133	auto	1066	3.44E−2
0.14	test graf	1123	3.49E−2

Na obrázku 2.33 vidíme, že s rastúcim rozsahom náhodného výberu klesá $\int \left|\hat{f}(x) - f(x)\right| dx$.



Obr. 2.33: Závislosť $\int \left|\hat{f}(x) - f(x)\right| dx$ na n pre rôzne metódy odhadu hustoty pravdepodobnosti trimodálnej hustoty

Rýchlosť konvergencie

Veľmi slabé podmienky (2.56) a (2.58) za ktorých odhad je uvedený, konzistencia môže dobre nalákať neopatrného na falošný pocit bezpečia, doporučením, že dobré odhady hustoty môžu byť získané zo širokého rozsahu hodnôt šírky okna h. Pozorovaná citlivosť odhadov na šírke okna môže byť zmierená so sekciou Konzistentné výsledky tým, že zvažuje rýchlosť v ktorej odhad $\hat{f}$ konverguje ku skutočnej hustote f. Aproximácia strednej integrálnej kvadratickej chyby rozvinutá v 2.58 je príklad dávajúci rýchlosť konvergencie $\hat{f}$ k f.

Predpokladajme aproximáciu (2.17) strednej integrálnej kvadratickej chyby. Pripomeňme, že (2.19) ukazuje, že ak je h zvolené optimálne, za vhodných regulárnych podmienok je aproximácia strednej integrálnej kvadratickej chyby pôjde k nule s rýchlosťou $n^{-\frac{4}{5}}$. Teraz predpokladajme, že namiesto voľby optimálneho h, zvolíme $h = n^{-\frac{1}{2}}$, a význam, ktorý by mohol byť sugestovaný podľa konzistentných vlastností (2.56). Substitúcia tejto hodnoty späť do (2.17) ukazuje, že stredná integrálna kvadratická chyba potom ide k nule s rýchlosťou $n^{-\frac{1}{2}}$.

Výsledky môžu byť tiež získané pre rýchlosť konvergencie $\hat{f}$ ku f rôznymi inými spôsobmi ako strednou integrálnou kvadratickou chybou. Napríklad Silverman (1978a) získal

exaktne rýchlosti so sup $\left|\hat{f}(x) - \mathrm{E}\hat{f}(x)\right|$ a sup $\left|\mathrm{E}\hat{f}(x) - f(x)\right|$ konvergenciou k nule. Tieto výsledky boli použité ako základ teórie metódy test grafu na voľbu šírky okna.

Avšak asymptotické vety sú užitočné, ak s nimi zaobchádzame opatrne. Napríklad ich môžeme použiť ako štartovací bod na modelovanie štúdie, alebo modelovanie základnej procedúry, a môžu nám pomôcť dať intuitívne cítenie, ako dobre sa budú chovať metódy praxi.

2.4. Metóda najbližšieho suseda

Metóda odhadu najbližšieho suseda predstavuje pokus prispôsobiť stupeň vyhladenia k "lokálnej" hustote náhodného výberu. Stupeň vyhladenia je ovládaný celým číslom k, ktoré je vybrané podstatne menšie ako rozsah náhodného výberu, typicky $k \approx \sqrt{n}$. Definujme vzdialenosť $d(x, y)$ medzi dvoma bodmi $|x-y|$, a pre každé t definujme vzdialenosti

$$d_1 t \leq d_1 t \leq \ldots \leq d_n t,$$

usporiadané vo vzostupnom poradí, od t k náhodným veličinám náhodného výberu.

Odhad hustoty $k-$tého najbližšieho suseda je definovaný

$$\hat{f} = \frac{k}{2nd_k(t)}. \tag{2.61}$$

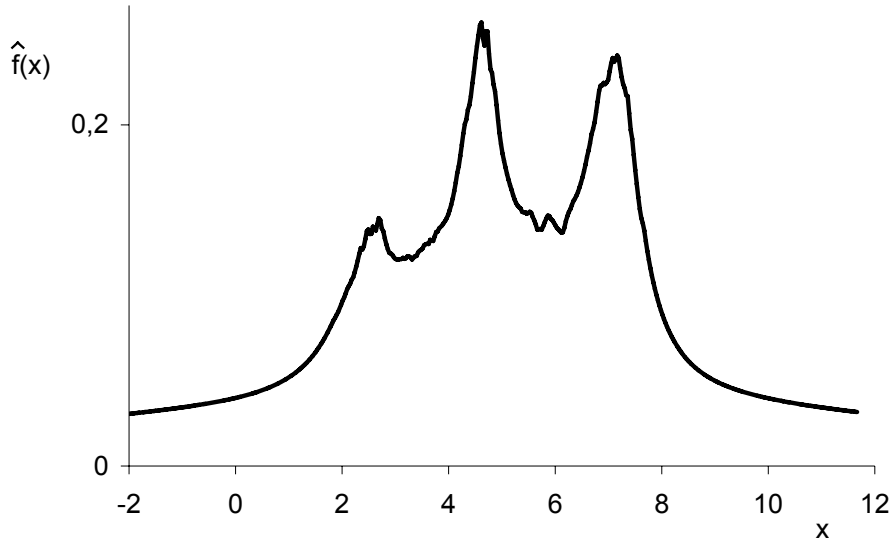
Aby sme pochopili definíciu, predpokladajme, že hustota v t je $f(t)$. Potom z rozsahu náhodného výberu n budeme očakávať, že asi $2rnf(t)$ pozorovaní spadá do intervalu $[t - -r, t+r]$ pre každé $r > 0$. Všimnime si pojednanie v naivnom odhade. Pretože z definície, presne k pozorovaní spadne do intervalu $[t-d_k, t+d_k]$. Odhad hustoty v t môžeme dostať položením

$$k = 2d_k(t)n\hat{f}(t),$$

a z tohoto môžeme vyjadriť definíciu odhadu hustoty $k-$tého najbližšieho suseda.

Zatiaľ čo naivný odhad je založený na počte meraní spadajúcich do vystredených obdĺžnikov s pevnou šírkou k náhodnej veličine náhodného výberu, odhad najbližšieho suseda je nepriamo úmerný k šírke obdĺžnika potrebného obsiahnuť daný rozsahom náhodného výberu. V prípade chvosta rozdelenia, vzdialenosť $d_k(t)$ bude väčšia ako hlavná časť rozdelenia, a problém podhladenia chvosta by mal byť zmenšený.

Podobne ako naivný odhad, s ktorým súvisí, odhad najbližšieho suseda definovaný 2.61 nie je hladká krivka. Funkcia $d_k(t)$ je spojitá, ale jej derivácie budú mať nespojitosť v každom bode $\frac{1}{2}\left(X_{(j)} + X_{(j+k)}\right)$, kde $X_{(j)}$ sú usporiadané štatistiky náhodného výberu. Okamžite vyplýva z týchto poznámok a z definície, že $\hat{f}(t)$ bude kladné a spojité všade, ale bude mať nespojité derivácie vo všetkých rovnakých bodoch ako $d_k(t)$. Na rozdiel od jadrového odhadu, odhad najbližšieho suseda nebude hustota pravdepodobnosti, lebo nebude integrovať k jednotke. Pre t menšie ako najmenšia náhodná veličina náhodného výberu položíme $d_k(t) = X_{(k)} - t$ a pre $t > X_{(n)}$ položíme $d_k(t) = t - X_{(n-k+1)}$. Substitúciou do 2.61 dostaneme, že $\int_{-\infty}^{\infty} \hat{f}(t)dt$ je nekonečno, a že chvost $\hat{f}$ zaniká pomerom $\frac{1}{t}$, inými slovami extrémne pomaly. Teda odhad najbližšieho suseda je nevhodný, ak je požadovaný odhad celej hustoty. Na obrázku 2.34 je odhad metódou najbližšieho suseda trimodálnej hustoty. Vážne chvosty a nespojitosti v deriváciách sú zrejmé.



Obr. 2.34: Odhad metódou najbližšieho suseda trimodálnej hustoty s Gaussovým jadrom $k = 15$.

Je možné zobecniť odhad najbližšieho suseda pridaním jadrového odhadu. Majme $K(x)$ jadrovú funkciu integrujúcu do jednotky. Potom odhad zovšeobecneného $k-$teho najbližšieho suseda je definovaný

$$\hat{f}(t) = \frac{1}{nd_k(t)} \sum_{i=1}^{n} K\left(\frac{t - X_i}{d_k(t)}\right) \tag{2.62}$$

Okamžite môžeme vidieť, že $\hat{f}(t)$ je presne jadrový odhad určený v t so šírkou okna $d_k(t)$. Teda celkový stupeň vyhladenia je riadený voľbou celého čísla k, ale použitá šírka okna v každej náhodnej veličine náhodného výberu závisí na hustote pozorovaní v blízkosti tejto náhodnej veličiny náhodného výberu.

Obyčajný odhad $k-$teho najbližšieho suseda je špeciálnym prípadom 2.62, keď K je obdĺžnikove jadro w z 2.1. Teda vzťah medzi 2.62 a 2.61 je rovnaký ako jadrový odhad s naivným odhadom. Ale derivácia zobecneného odhadu najbližšieho suseda bude nespojitá vo všetkých bodoch, kde funkcia $d_k(t)$ má nespojité derivácie. Presná integrabilita a vlastnosti chvosta budú zavisieť na presnom tvare jadra, a nebude tu pojednávaná.

2.5. Metóda premenlivého jadra

Metóda premenlivého jadra trochu súvisí s postupom metódy najbližšieho suseda. Je to ďalšia metóda, ktorá prispôsobuje vyhladenie lokálnej hustote náhodného výberu. Odhad je spravený podobne ako klasický jadrový odhad, ale parameter miery "jadro" je umiestnený na náhodných veličinách náhodného výberu a je požadovaná zmena od jednej náhodnej veličiny náhodného výberu, do inej náhodnej veličiny náhodného výberu.

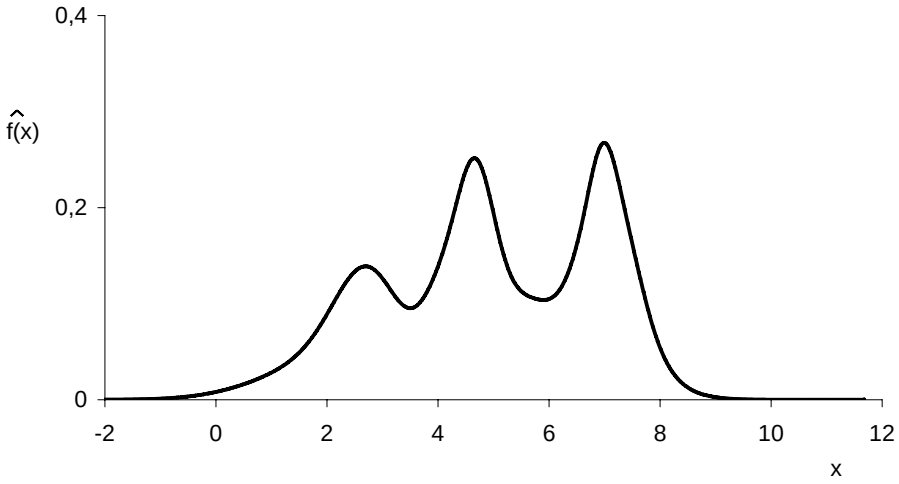
Nech K je jadrová funkcia a k je kladné celé číslo. Definujme $d_{j,k}$ ako vzdialenosť z X_j ku $k-$tej najbližšej náhodnej veličine v množine obsahujúcu ďalších $n - 1$ náhodných veličín náhodného výberu. Potom *odhad premenlivým jadrom* s vyhladzovacím parametrom h je definovaný

$$\hat{f}(t) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{hd_{j,k}} K\left(\frac{t - X_j}{hd_{j,k}}\right).$$

Šírka okna jadra umiestnená v náhodnej veličine náhodného výberu X_j je úmerná s $d_{j,k}$. Kde sú náhodné veličiny náhodného výberu riedke, budú mať priradené plochšie jadrá. Pre každé pevné k bude celkový stupeň vyhladenia závisieť na parametri h. Voľba k určuje, ako bude citlivá voľba šírky okna na lokálny detail.

Niektoré porovnania odhadu s premenlivým jadrom so všeobecným odhadom metódou najbližšieho suseda (2.4) môžu byť poučné. V (2.4) je použitá šírka okna na konštrukciu odhadu v t závislá na vzdialenostiach z t k náhodným veličinám náhodného výberu. V (2.5) šírky okna nezávisia na bode t, ktorého hustota sa odhaduje, a závisí iba na vzdialenostiach medzi náhodnými veličinami náhodného výberu.

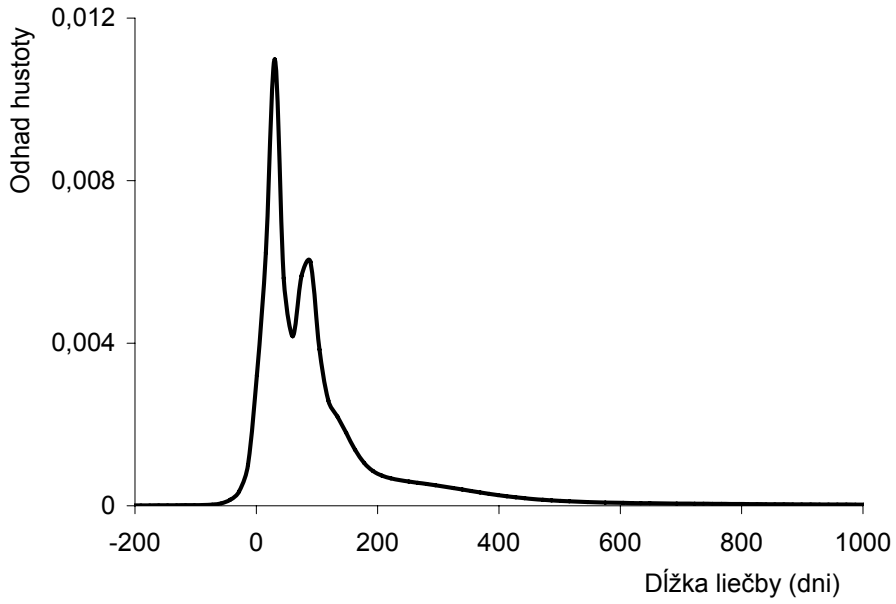
Na rozdiel od všeobecného odhadu metódou najbližšieho suseda, odhad s premenlivým jadrom bude sám pravdepodobnostná funkcia hustoty, pokiaľ je K jadro. To je okamžitý následok definície. Okrem toho, ako u obyčajného jadrového odhadu, odhad zdedí všetky lokálne vlastnosti vyhladenia jadra. Na obrázku 2.36 je použitá metóda na získanie odhadu dĺžky liečenia pacienta. Šum na chvoste krivky je eliminovaný, ale je zaujímavé si všimnúť, že metóda odkrýva štruktúru v hlavnej časti rozdelenia, ktorá nie je skutočne viditeľná dokonca v podhladenej krivke na obrázku 2.10. Na obrázku 2.35 je odhad trimodálnej hustoty gaussovým premenlivým jadrom.



Obr. 2.35: Odhad trimodálnej hustoty gaussovým premenlivým jadrom $k = 30, h = 0.33$.

2.6. Porovnanie jadrových odhadov trimodálnej hustoty

Uvedieme si porovnanie jadrových odhadov trimodálnej hustoty pre rôzne jadrové odhady pravdepodobnosti pre rozsahy náhodných výberov od 33 do 2673. Gaussovo jadro je použité vo všetkých odhadoch. Každá metóda sa robila pre 9 náhodných výberov s rovnakým rozsahom. Rozsah náhodného výberu metódy najbližšieho suseda a premenlivého jadra sú z dôvodu náročnosti na výpočet od 33 do 297. Ku každému odhadu sme určili strednú



Obr. 2.36: Odhad dĺžky liečenia pacienta s premenlivým gaussovým jadrom $k = 8, h = 5$.

integrálnu kvadratickú chybu. V nasledujúcej tabuľke sú odhady s metódami hľadania vyhladzovacieho parametra pre rozsah náhodného výberu 33.

n	Metóda	h_n	MISE	Metóda	h_n	MISE
33	subjektívna	0.44	2.52E−6	normálne	0.463	2.59E−6
	volba	0.45	5.65E−6	rozdelenie	0.493	5.44E−6
		0.5	9E−7		0.533	1.05E−6
		0.45	4.63E−6		0.425	6.16E−6
		0.36	1.54E−6		0.462	1.27E−6
		0.36	4.01E−6		0.465	2.05E−6
		0.45	6.31E−6		0.4	7.05E−6
		0.4	7.53E−6		0.541	5.69E−6
		0.5	2.89E−6		0.422	2.95E−6
33	auto	0.81	4.02E−6	test graf	0.4	2.43E−6
		0.82	5.14E−6		0.4	6E−6
		0.493	8.57E−7		0.425	7,82E−7
		0.244	1.51E−5		0.55	3.41E−6
		0.637	1.72E−6		0.41	1.33E−6
		0.216	1.36E−5		0.5	1.77E−6
		0.716	4.47E−6		0.425	6.66E−6
		0.343	9.09E−6		0.485	6.19E−6
		0.892	3.98E−6		0.4	3.04E−6

V nasledujúcej tabuľke sú odhady metódami najbližšieho suseda a premenlivého jadra pre rozsah náhodného výberu 33.

n	Metóda	k	MISE	Metóda	k	h_n	MISE
33	najbližší	8	3.42E−6	premenlivé	7	0.5	3.02E−6
	sused	6	5.97E−6	jadro	9	0.5	8.09E−6
		6	1.65E−6		9	0.5	3.06E−6
		7	2.39E−5		7	0.6	3.97E−5
		6	1.74E−6		8	0.5	4.04E−6
		7	1.08E−5		8	0.6	2.23E−5
		8	5.3E−6		6	0.7	1.21E−5
		8	6.99E−6		10	0.7	8.72E−6
		7	3.03E−6		7	0.5	2.49E−6

V nasledujúcej tabuľke sú odhady s metódami hľadania vyhladzovacieho parametra pre rozsah náhodného výberu 99.

n	Metóda	h_n	MISE	Metóda	h_n	MISE
99	subjektívna	0.4	1.11E−6	normálne	0.423	1.06E−6
	volba	0.37	1.05E−6	rozdelenie	0.386	1.06E−6
		0.39	1.91E−6		0.391	1.92E−6
		0.41	7.31E−7		0.405	7.23E−7
		0.46	2.08E−6		0.449	2.05E−6
		0.38	4.48E−7		0.407	4.82E−7
		0.42	6.56E−7		0.394	6.07E−7
		0.43	9.54E−7		0.425	9.51E−7
		0.37	8.87E−7		0.385	9.35E−7
99	auto	0.331	1.45E−6	test graf	0.39	1.14E−6
		0.453	1.19E−6		0.375	1.05E−6
		0.459	2.11E−6		0.38	1.9E−6
		0.229	1.75E−6		0.38	7.01E−7
		0.203	3.49E−6		0.36	1.97E−6
		0.307	5.9E−7		0.385	4.51E−7
		0.343	6.27E−7		0.425	6.69E−7
		0.247	1.93E−6		0.41	9.47E−7
		0.475	1.31E−6		0.355	8.44E−7

V nasledujúcej tabuľke sú odhady metódami najbližšieho suseda a premenlivého jadra pre rozsah náhodného výberu 99.

n	Metóda	k	MISE	Metóda	k	h_n	MISE
99	najbližší	15	1.04E−6	premenlivé	20	0.8	1.57E−6
	sused	13	1.32E−6	jadro	18	0.65	2.03E−6
		15	1.83E−6		19	0.7	1.76E−6
		15	1.05E−6		21	0.8	7.17E−7
		14	1.49E−6		18	0.7	1.25E−6
		15	5.4E−7		19	0.7	1.23E−6
		16	6.92E−7		17	0.7	2.01E−6
		15	1.26E−6		20	0.7	1.14E−6
		11	7.88E−7		12	0.8	1.72E−6

V nasledujúcej tabuľke sú odhady s metódami hľadania vyhladzovacieho parametra pre rozsah náhodného výberu 297.

n	Metóda	h_n	MISE	Metóda	h_n	MISE
297	subjektívna	0.32	4.34E−7	normálne	0.353	4.73E−7
	volba	0.35	8.66E−7	rozdelenie	0.308	9E−7
		0.28	1.32E−6		0.328	1.32E−6
		0.32	7.46E−7		0.319	7.46E−7
		0.3	4.35E−7		0.320	4.75E−7
		0.32	7.02E−7		0.341	7.57E−7
		0.35	4.94E−7		0.329	4.7E−7
		0.3	7.89E−7		0.319	8,36E−7
		0.3	4.24E−7		0.335	4,62E−7
297	auto	0.279	4.4E−7	test graf	0.3	4.29E−7
		0.212	1.42E−6		0.35	8.66E−7
		0.307	1.31E−6		0.335	1.33E−6
		0.208	1.21E−6		0.33	7.47E−7
		0.344	5.33E−7		0.3	4.35E−7
		0.323	7.09E−7		0.31	6.79E−7
		0.26	5.27E−7		0.33	4.71E−7
		0.389	1.05E−6		0.29	7.67E−7
		0.271	4.33E−7		0.33	4.54E−7

V nasledujúcej tabuľke sú odhady metódami najbližšieho suseda a premenlivého jadra pre rozsah náhodného výberu 297.

n	Metóda	k	MISE	Metóda	k	h_n	MISE
297	najbližší	26	5.91E−7	premenlivé	65	0.75	4.86E−7
	sused	30	1.4E−6	jadro	60	0.6	6.75E−7
		30	1.53E−6		54	0.65	9.41E−7
		29	1.24E−6		50	0.7	2.18E−6
		29	4.05E−7		40	0.8	2.87E−6
		30	5.03E−7		35	0.75	1.02E−6
		29	6.54E−7		35	0.75	1.84E−6
		29	6.98E−7		50	0.6	3.16E−7
		31	5.04E−7		60	0.6	7.4E−7

V nasledujúcej tabuľke sú odhady s metódami hľadania vyhladzovacieho parametra pre rozsah náhodného výberu 2673.

n	Metóda	h_n	MISE	Metóda	h_n	MISE
2673	subjektívna	0.2	1.51E−7	normálne	0.213	1.63E−7
	volba	0.2	1.09E−7	rozdelenie	0.215	1.09E−7
		0.19	1.39E−7		0.213	1.57E−7
		0.21	1.71E−7		0.214	1.74E−7
		0.2	1.01E−7		0.212	1.16E−7
		0.2	6.22E−8		0.219	5.35E−8
		0.19	2.6E−7		0.207	2.85E−7
		0.2	1.37E−7		0.212	1.53E−7
		0.21	1.53E−7		0.211	1.53E−7
2673	auto	0.181	1.39E−7	test graf	0.18	1.39E−7
		0.138	1.74E−7		0.205	1.08E−7
		0.141	1.53E−7		0.195	1.42E−7
		0.176	1.58E−7		0.19	1.6E−7
		0.134	1.54E−7		0.2	1.1E−7
		0.165	1.02E−7		0.2	6.22E−8
		0.176	2.44E−7		0.195	2.67E−7
		0.164	1.05E−7		0.2	1.37E−7
		0.146	1.75E−7		0.215	1.56E−7

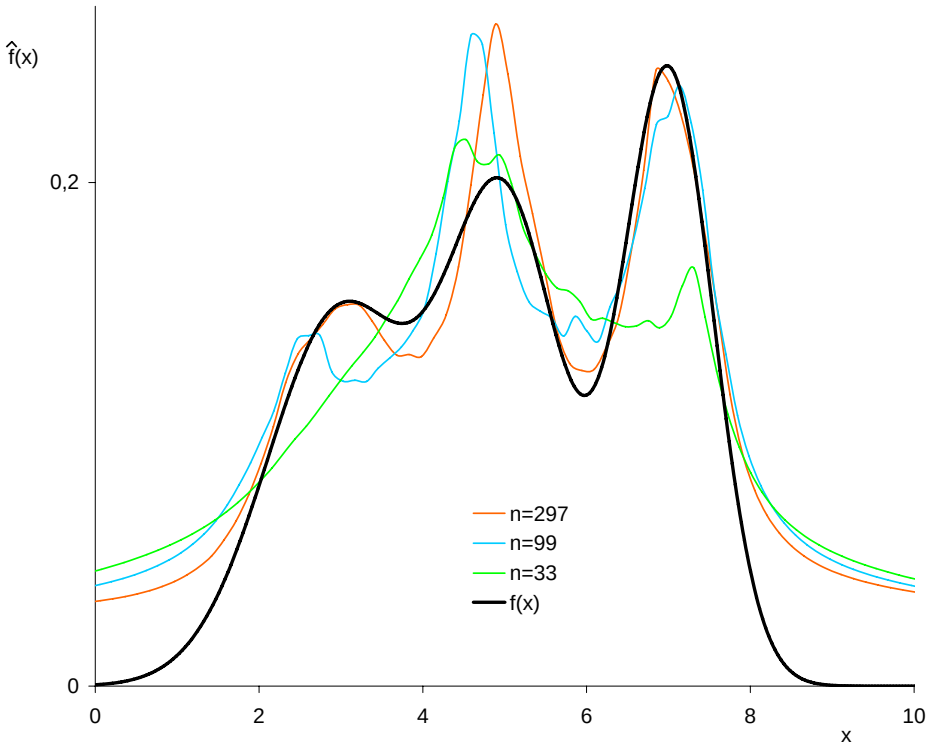
V nasledovnej tabuľke sú stredné hodnoty a rozptyly vyhladzovacích parametrov a stredných integrálnych kvadratických chýb pre metódy odhadu s hľadaním vyhladzovacieho parametra.

n	Metóda	E (h_n)	D (h_n)	E (MISE)	D (MISE)
33	subjektívna volba	0.434	2.4E−3	4E−6	4.44E−12
	normálne rozdelenie	0.467	2.09E−3	3.81E−6	4.62E−12
	auto	0.575	5.98E−2	6.44E−6	2.27E−11
	test graf	0.444	2.64E−3	3.51E−6	4.44E−12
99	subjektívna volba	0.403	8E−4	1.09E−6	2.72E−13
	normálne rozdelenie	0.407	4.11E−4	1.09E−6	2.66E−13
	auto	0.39	9.56E−3	1.61E−6	6.89E−13
	test graf	0.384	4.36E−4	1.07E−6	2.50E−13
297	subjektívna volba	0.316	4.91E−4	6.9E−7	7.56E−14
	normálne rozdelenie	0.328	1.63E−4	7.15E−7	7.31E−14
	auto	0.288	3.14E−3	8.48E−7	1.41E−13
	test graf	0.319	3.58E−4	6.86E−7	7.63E−14
2673	subjektívna volba	0.2	4.44E−5	1.43E−7	2.68E−15
	normálne rozdelenie	0.213	9.21E−6	1.52E−7	3.46E−15
	auto	0.158	2.97E−4	1.56E−7	1.58E−15
	test graf	0.198	8.4E−5	1.42E−7	2.75E−15

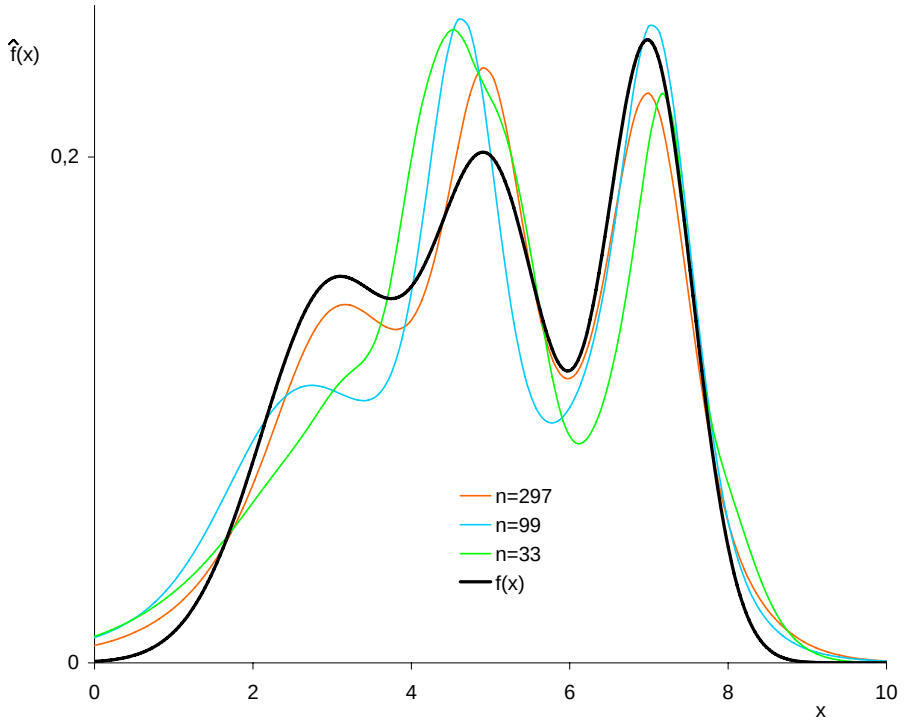
V nasledovnej tabuľke sú stredné hodnoty a rozptyly vyhladzovacích parametrov a stredných integrálnych kvadratických chýb pre metódy odhadu premenlivého jadra najbližšieho suseda.

n	Metóda	E(k)	E (h_n)	D(k)	D (h_n)	E (MISE)	D (MISE)
33	najbli. sused	7		0.667		6.98E−6	4.32E−11
	premen. jadro	7.89	0.567	1.43	6.67E−3	1.15E−5	1.35E−10
99	najbli. sused	14.3		2		1.11E−6	1.49E−13
	premen. jadro	18.2	0.728	6.17	2.84E−3	1.49E−6	1.71E−13
297	najbli. sused	29.2		1.73		8.36E−7	1.65E−13
	premen. jadro	49.9	0.689	110	5.43E−3	1.23E−6	6.7E−13

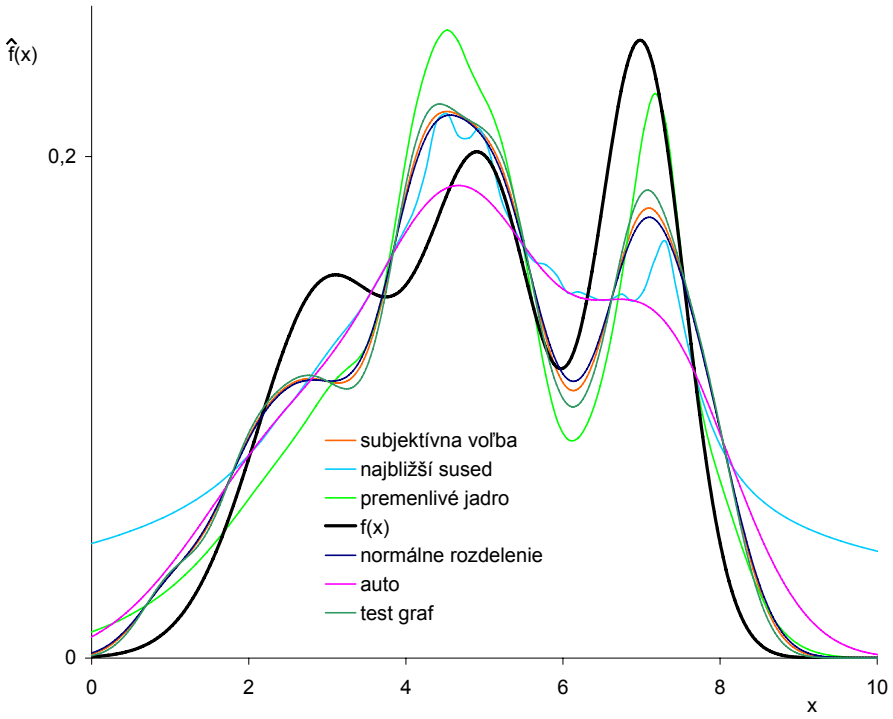
Na obrázkoch 2.37, 2.38, 2.39, 2.40 a 2.41 vidíme, že jadrové odhady hustoty pravdepodobnosti sa s rastúcim rozsahom náhodného výberu približujú skutočnej hustote f(x). Na obrázkoch 2.39, 2.40 a 2.41 vidíme porovnanie odhadov jadrových metód pre konkrétne rozsahy náhodných výberov. Na obrázku 2.42 vidíme, že so stúpajúcim rozsahom náhodného výberu klesá rozptyl vyhladzovacích parametrov. Na obrázku 2.43 vidíme, že so stúpajúcim rozsahom náhodného výberu klesá stredná hodnota stredných integrálnych kvadratických chýb. Na obrázku 2.44 je závislosť rozptylu stredných integrálnych kvadratických chýb na rozsahu náhodného výberu.



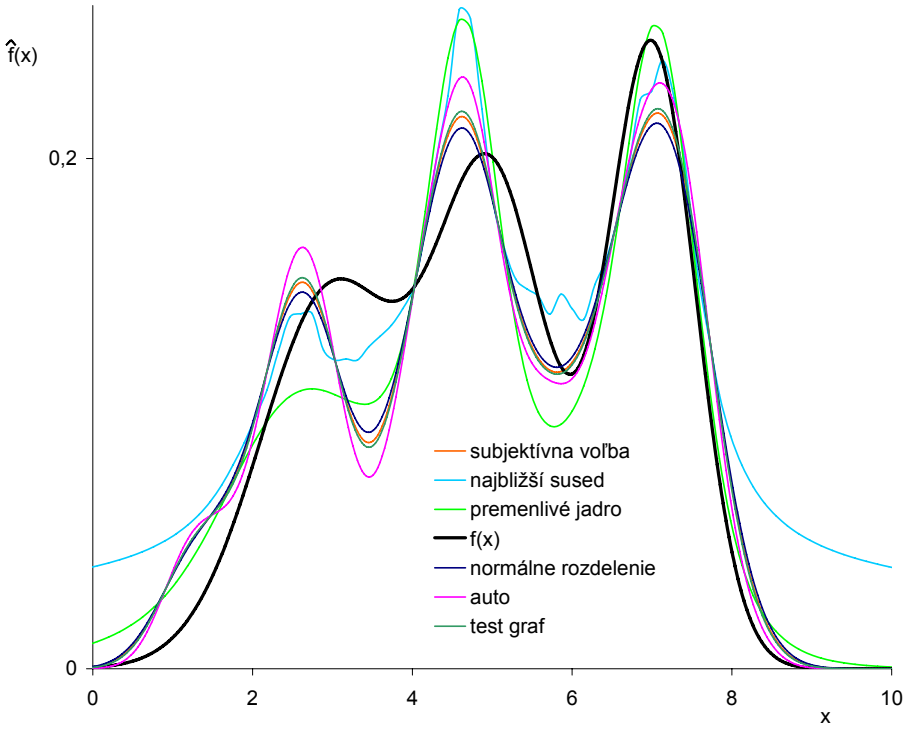
Obr. 2.37: Odhady hustoty pravdepodobnosti metódou najbližšieho suseda prvých náhodných výberov pre rôzne rozsahy náhodných výberov n a skutočná trimodálna hustota $f(x)$.



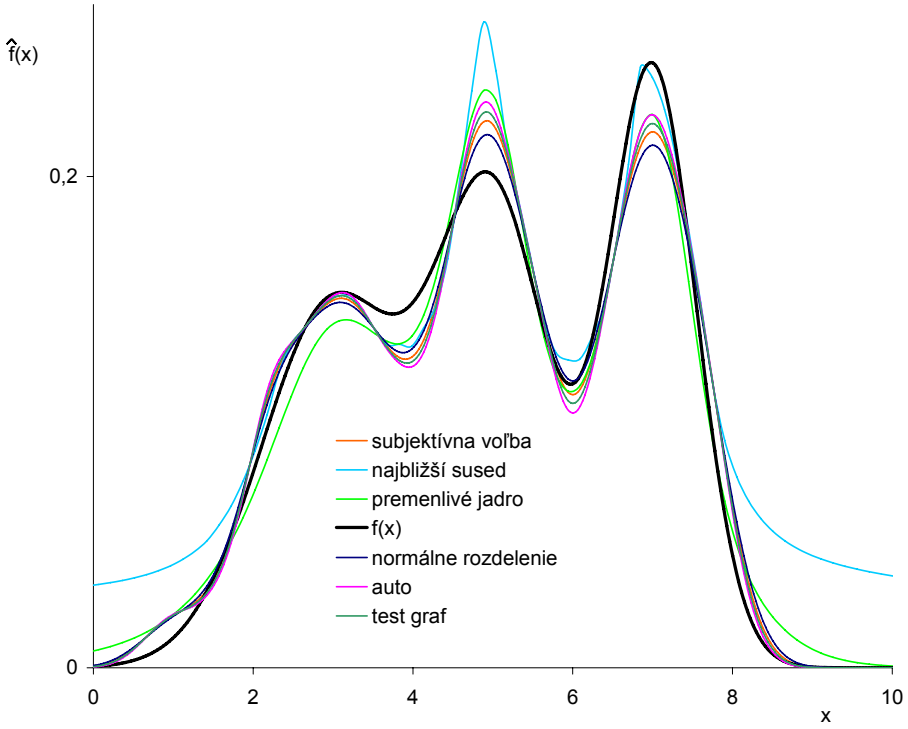
Obr. 2.38: Odhady hustoty pravdepodobnosti metódou premenlivého jadra prvých náhodných výberov pre rôzne rozsahy výberov výberov n a skutočná trimodálna hustota $f(x)$.



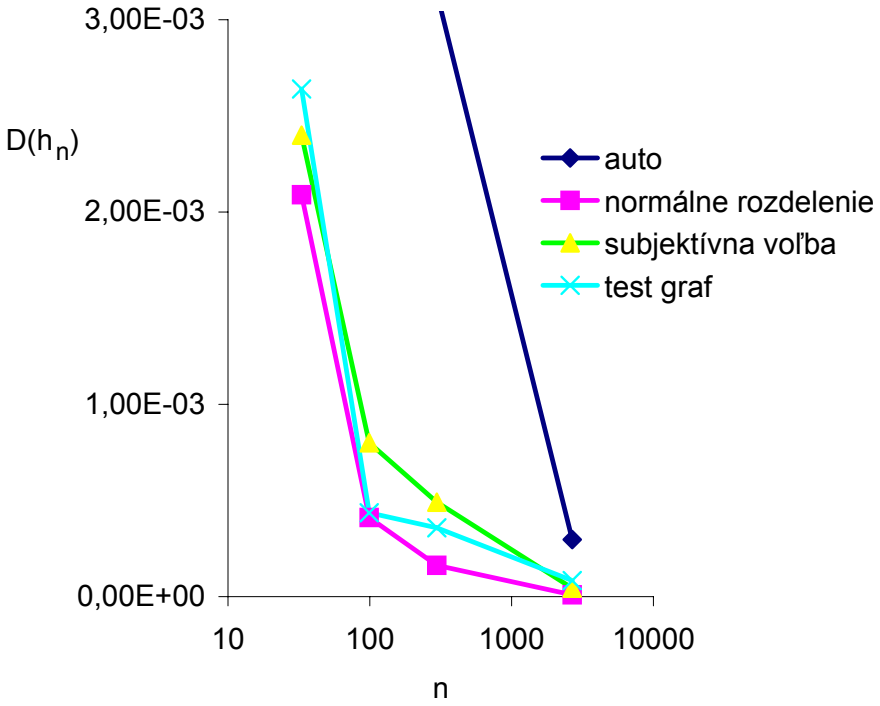
Obr. 2.39: Odhady hustoty pravdepodobnosti s rozsahom náhodného výberu $n = 33$ prvých náhodných výberov pre rôzne metódy odhadu a skutočná trimodálna hustota $f(x)$.



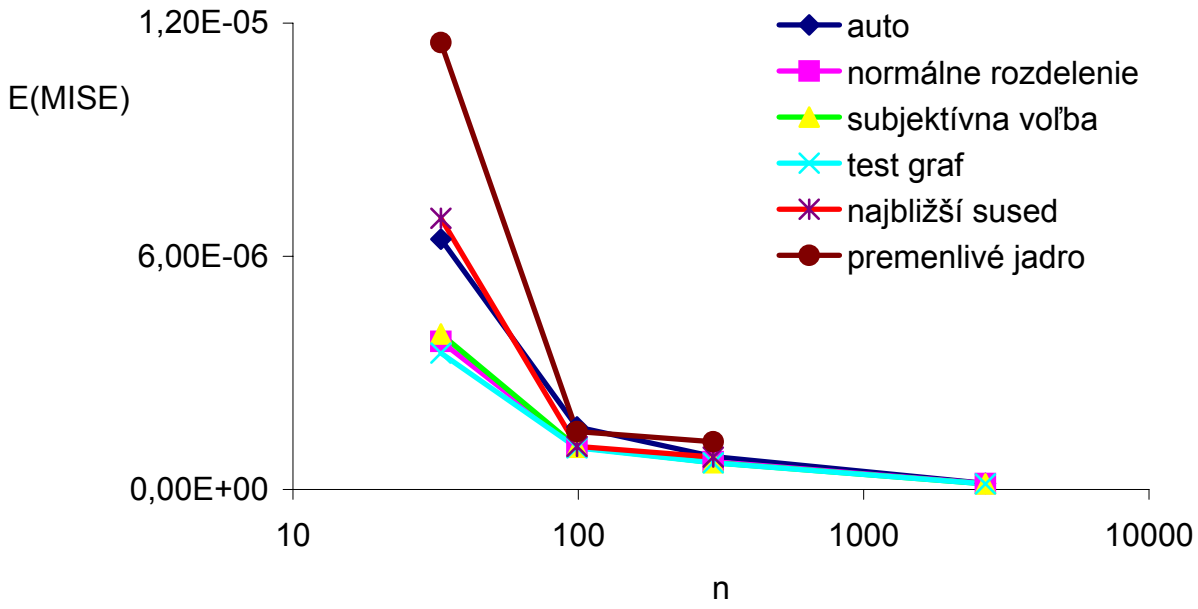
Obr. 2.40: Odhady hustoty pravdepodobnosti s rozsahom náhodného výberu $n = 99$ prvých náhodných výberov pre rôzne metódy odhadu a skutočná trimodálna hustota $f(x)$.



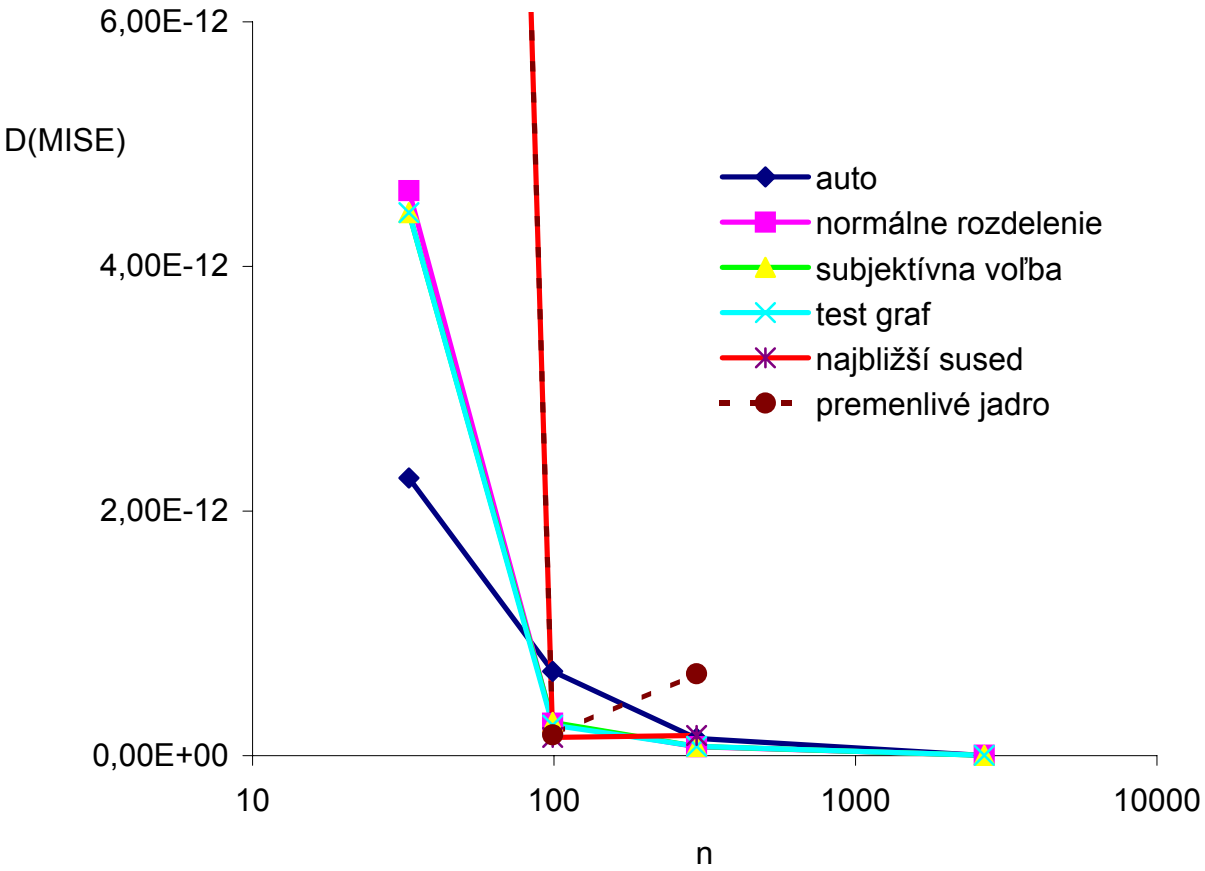
Obr. 2.41: Odhady hustoty pravdepodobnosti s rozsahom náhodného výberu $n = 297$ prvých náhodných výberov pre rôzne metódy odhadu a skutočná trimodálna hustota $f(x)$.



Obr. 2.42: Závislosť $\mathrm{D}\,(h_n)$ na n pre rôzne metódy odhadu hustoty pravdepodobnosti trimodálnej hustoty.



Obr. 2.43: Závislosť E(MISE) na rozsahu náhodného výberu n pre rôzne metódy odhadu hustoty pravdepodobnosti trimodálnej hustoty.



Obr. 2.44: Závislosť D(MISE) na rozsahu náhodného výberu n pre rôzne metódy odhadu hustoty pravdepodobnosti trimodálnej hustoty.

3. Metódy odhadov pre viacrozmerný náhodný výber

Doteraz sme sa zaujímali o odhady hustoty pravdepodobnosti jednorozmerného náhodného výberu. Veľa, ak nie najviac dôležitých aplikácií odhadu hustoty vyžaduje analýzu viacrozmerného náhodného výberu, ktorý pojednáme v tejto kapitole. Opäť bude kladená veľká pozornosť jadrovému odhadu. To neznamená že jadrový odhad je jediný, alebo najlepší odhad hustoty pravdepodobnosti viacrozmerného náhodného výberu.

Pre viacrozmerný náhodný výber je rozlíšenie medzi rôznymi aplikáciami odhadu hustoty dôležitejšie, ako pre jednorozmerný náhodný výber. Jednoduché je pochopiť nakreslený tvar alebo perspektívny pohľad na dvojrozmernú funkciu hustoty. Ale prezentované problémy sú nepravdepodobné, že hustota bude použiteľná pre výskumný účel vo viacej ako dvoch dimenziach. V prípade nutnosti by mohol skúsený užívateľ s prístupom k sofistikovanému grafickému zariadeniu preskúmať a získať užitočný pohľad trojrozmernej funkcie pravdepodobnosti. Napríklad Scott a Thompson (1983) dokonca uvažovali prezentáciu štvor a päť rozmernej hustoty pravdepodobnosti. Na druhej strane, ak cieľ nesmeruje k hustote funkcie, ale miesto toho sa použije ako gradient v nejakej štatistickej metóde, potom vzhľad prezentácie sa stáva menej dôležitý, a môže byť užitočné a nevyhnutné odhadnúť hustoty viacrozmerného náhodného výberu.

3.1. Jadrový odhad pre viacrozmerný náhodný výber

V tejto časti bude predstavená jadrová metóda pre viacrozmerný náhodný výber. V celej tejto časťi budeme tučné písmo používať pre náhodné vektory náhodného výberu d rozmerného priestoru. Budeme predpokladať, že $\mathbf{X}_1, \dots \mathbf{X}_n$ je viacrozmerný náhodný výber s rovnakou hustotou, ktorú chceme odhadnúť.

3.1.1. Definícia jadrového odhadu pre viacrozmerný náhodný výber

Jadrový odhad ako súčet "jadier" vystredených na pozorovaniach si jednoducho zobecníme na viacrozmerný prípad. Jadrový odhad pre viacrozmerný náhodný výber s jadrom K a šírkou okna h je definovaný

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right). \tag{3.1}$$

Jadrová funkcia $K(\mathbf{x})$ je funkcia definovaná pre d rozmerné $\mathbf{x}$ splňujúca

$$\int_{R^d} K(\mathbf{x}) d\mathbf{x} = 1. \tag{3.2}$$

Obvykle bude K radiálne symetrická jednomodálna pravdepodobnostná funkcia hustoty, napríklad štandardná viacrozmerná normálna funkcia hustoty je

$$K(\mathbf{x}) = (2\pi)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right). \tag{3.3}$$

Epanechnikove jadro je iné viacrozmerné možné jadro

$$K_e(\mathbf{x}) = \begin{cases} \frac{1}{2c_d}(d+2)\left(1 - \mathbf{x}^T\mathbf{x}\right) & \text{ak} \quad \mathbf{x}^{\mathrm{T}}\mathbf{x} < 1 \\ 0 & \text{inak} \end{cases} \tag{3.4}$$

kde c_d je objem jednotkovej d rozmernej sféry: $c_1 = 2$, $c_2 = \pi$, $c_3 = \frac{4\pi}{3}$ atď. Ďalšie užitočné jadrá pre dvojrozmerný náhodný výber sú

$$K_2(\mathbf{x}) = \begin{cases} \frac{3}{\pi}\left(1 - \mathbf{x}^T\mathbf{x}\right)^2 & \text{ak} \quad \mathbf{x}^{\mathrm{T}}\mathbf{x} < 1 \\ 0 & \text{inak} \end{cases} \tag{3.5}$$

a

$$K_3(\mathbf{x}) = \begin{cases} \frac{4}{\pi}\left(1 - \mathbf{x}^T\mathbf{x}\right)^3 & \text{ak} \quad \mathbf{x}^{\mathrm{T}}\mathbf{x} < 1 \\ 0 & \text{inak.} \end{cases} \tag{3.6}$$

Výhody týchto dvoch jadier oproti jadru (3.4) sú že jadrá, a z nich plynúce hustoty odhady majú väčšiu diferencovateľnosť. Naviac môžu byť rátané oveľa rýchlejšie ako normálne jadro (3.3).

3.2. Voľba jadra a šírky okna

V kapitole 2 sme sa zaoberali voľbou jadra a šírkou okna, ktorú môžeme rozšíriť s vhodnými modifikáciami na viacrozmerný prípad. Technické detaily rôznych potrebných zmien sú uvedené v tejto sekcii.

3.2.1. Vlastnosti náhodného výberu

Ako v sekcii 2.3.3, približné vyjadrenie pre odchýlku odhadu a rozptyl odhadov môže byť odvodené. Tieto môžeme použiť s vhodnou voľbou jadra a šírky okna. Predpokladajme, že jadro K je radiálna symetrická funkcia hustoty pravdepodobnosti a že neznáma hustota f má ohraničené a spojité druhé derivácie.

Definujme konštanty α a β ako

$$\alpha = \int t_1^2 K(\mathbf{t}) d\mathbf{t}$$

a

$$\beta = \int K(\mathbf{t})^2 d\mathbf{t}. \tag{3.7}$$

V podstate rovnakými manipuláciami ako predtým, použitím viacrozmernej formy Taylorovej vety dostaneme aproximácie

$$\mathrm{bias}_h(\mathbf{x}) \approx \frac{1}{2} h^2 \alpha \nabla^2 f(\mathbf{x}) \tag{3.8}$$

a

$$\mathrm{var}\hat{f}(\mathbf{x}) \frac{1}{nh^d} \beta f(\mathbf{x}). \tag{3.9}$$

Ich spojením ako v sekcii 2.3.3 dáva aproximáciu strednej integrálnej kvadratickej chyby

$$\frac{1}{4}h^4\alpha^2\int\left\{\nabla^2 f(\mathbf{x})\right\}^2 d\mathbf{x}+\frac{1}{nh^d}\beta. \tag{3.10}$$

Preto aproximovaná optimálna šírka okna, v zmysle minimalizovania strednej integrálnej kvadratickej chyby, je daná

$$h_{opt}^{d+4}=d\beta\alpha^{-2}\left\{\int\left(\nabla^2 f\right)^2\right\}^{-1}n^{-1}. \tag{3.11}$$

Výsledky, ktoré môžu byť nakreslené z tejto diskusie sú paralélne veľmi blízke v sekcii 2.3.3 pre jednorozmerný prípad. Aproximovaná optimálna šírka okna (3.11) konverguje k nule s narastajúcim n, ale veľmi pomaly, s rýchlosťou $n^{-\frac{1}{d+4}}$. Okrem toho aproximácia hodnoty h závisí na neznámej hustote, ktorú chceme odhadnúť.

Hodnota h_{opt} môže byť späť dosadená do (3.10) a dostaneme aproximovanú minimálnu možnú strednú integrálnu kvadratickú chybu, a ako v sekcii 2.3.3 vyberieme jadro. Epanechnikove jadro (3.4) je optimálne spoločne s nezápornými jadrami v zmysle minimalizovanian najmenšej dosiahnuteľnej strednej integrálnej kvadratickej chyby, ale práve ako predtým, ostatné jadrá definované v sekcii 3.1.1 môžu dosiahnuť veľmi podobné stredné integrálne kvadratické chyby.

3.2.2. Voľba šírky okna vzhľadom k normálnemu rozdeleniu

Prvý krok pred voľbou vyhladzovacieho parametra, je použitie výrazu (3.11) k stanoveniu vhodnej hodnoty šírky okna, keď f je normálna hustota, ako viacrozmerná normálna. Ak ϕ je jednotková d rozmerná normálna hustota, potom môže byť ukázané, že

$$\int\left(\nabla^2\phi\right)^2=\left(2\sqrt{\pi}\right)^{-d}\left(\frac{d}{2}+\frac{d^2}{4}\right). \tag{3.12}$$

Hodnota daná (3.12) môže byť späť vložená do (3.11), aby sme dostali optimálnu šírku okna pre vyhladenie normálneho rozdelenia náhodného výberu s jednotkovým rozptylom. Šírka okna je potom daná

$$h_{opt}=A(K)n^{-\frac{1}{d+4}}. \tag{3.13}$$

$A(K)$ je v nasledovnej tabuľke.

Jadro	Dimenzia	$A(K)$
Viacrozmerné normálne K	2	1
ako v rovnici (3.3)	d	$\left(\frac{4}{d+2}\right)^{\frac{1}{d+4}}$
K_e ako v rovnici (3.4)	2	2.4
	3	2.49
	d	$\left(\frac{8}{c_d}(d+4)\left(2\sqrt{\pi}\right)^d\right)^{\frac{1}{d+4}}$
K_2 ako v rovnici (3.5)	2	2.78
K_3 ako v rovnici (3.6)	2	3.12

Konštanta

$$A(K) = \left[d\beta\alpha^{-2} \left\{ \int \left(\nabla^2 \phi \right)^2 \right\}^{-1} \right]^{\frac{1}{d+4}} \tag{3.14}$$

závisí na jadre.

Uvedieme si príklad dvojrozmerného jadrového odhadu. Hustotu si zvolíme zmes troch normálnych dvojrozmenrých rozdelení $\frac{1}{3}N\left((0,0),\begin{pmatrix}1&0\\0&1\end{pmatrix}\right) + \frac{1}{3}N\left((4,0),\begin{pmatrix}1&0\\0&1\end{pmatrix}\right) + \frac{1}{3}N\left((2,3.464),\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$. Gaussovo jadro (3.3) je použité na tento odhad. Rozsah náhodného výberu je 150.
Na obrázkoch:

- (3.1), (3.3) a (3.5) sú jadrové odhady trimodálnej hustoty
- (3.2), (3.4) a (3.6) sú ich vrstevnice
- (3.1) máme podhladený jadrový odhad, kde $h = 0.3$
- (3.2) máme vrstevnicu podhladeného jadrového odhadu, kde $h = 0.3$
- (3.3) máme jadrový odhad, kde $h = 0.8$
- (3.4) máme vrstevnicu podhladeného jadrového odhadu, kde $h = 0.8$
- (3.5) máme prehladený jadrový odhad, kde $h = 1.6$
- (3.6) máme vrstevnicu podhladeného jadrového odhadu, kde $h = 1.6$

3.2.3. Viac sofistikované metódy voľby šírky okna

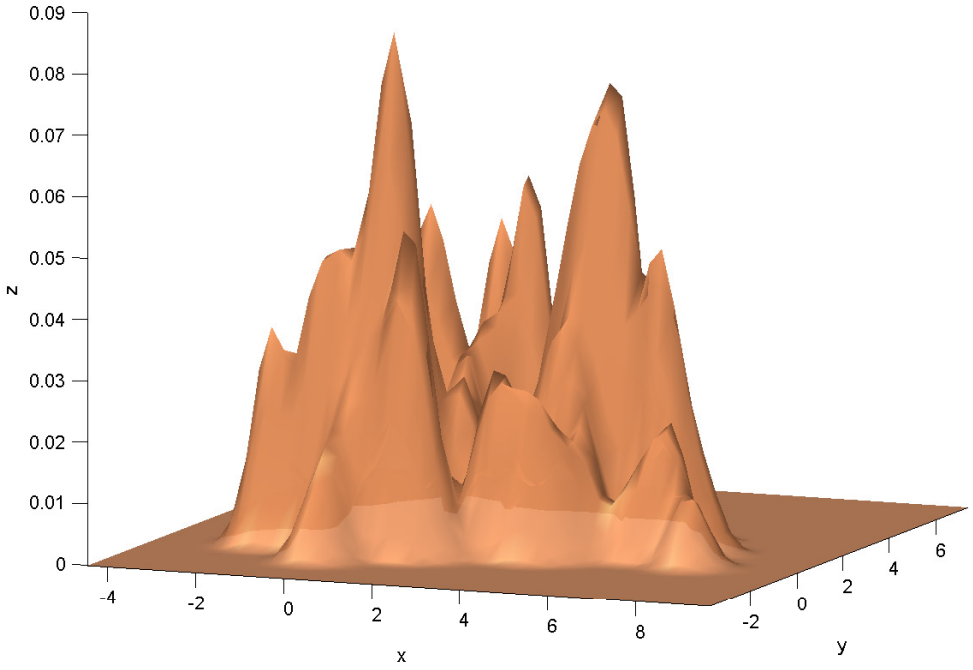
Metódu vzájomnej kontroly najmenších kvadrátov môžeme previesť na viacrozmerný prípad bez podstatných úprav. Hodnota $M_1(h)$ z (2.36) bude jednotková d rozmerná normálna hustota, potom môže byť ukázané, že

$$M_1(h) = \frac{1}{n^2 h^d} \sum_i \sum_j K^* \left\{ \frac{\mathbf{X}_i - \mathbf{X}_j}{h} \right\} + \frac{2}{nh^d} K(\mathbf{0}). \tag{3.15}$$

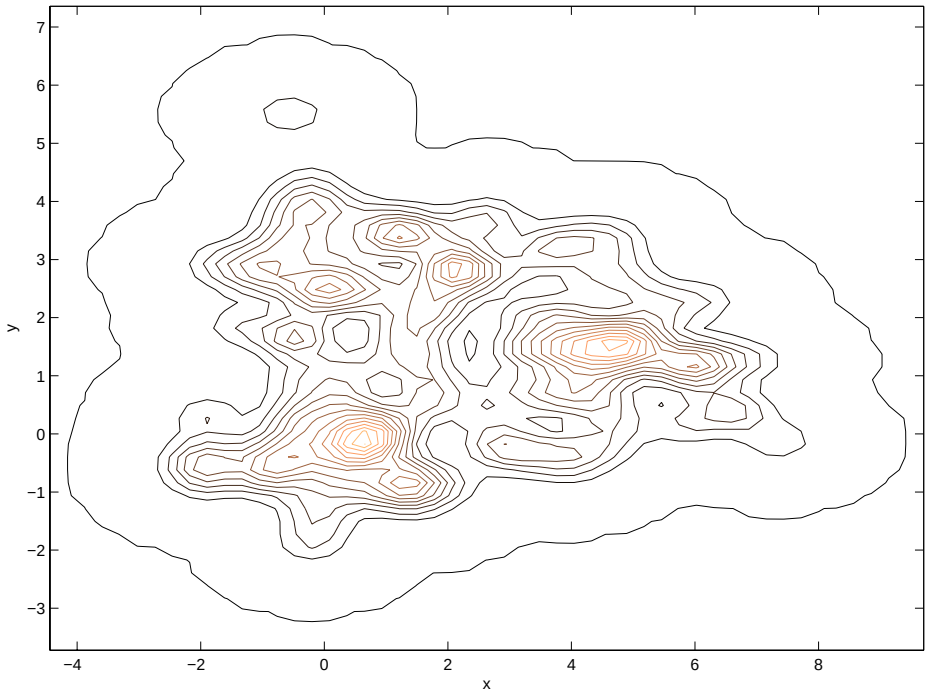
Je zaujímavé poznamenať, že čas potrebný na výpočet $M_1(h)$ z (3.15) závisí na rozmere d, na čase výpočtu kvadrátu rozdielov $(\mathbf{X}_i - \mathbf{X}_j)^T$.

Metóda test graf v podsekcii 2.3.4 môže byť rozšírená, hoci nie je pravdepodobne praktická vo viacej ako v dvoch rozmeroch. Ako dvojrozmerná metóda potrebuje uvažovať "testovaciu plochu" dávajúcu nakreslený $\nabla^2 \hat{f}$. Toto je časovo náročné. Testovacia plocha je citlivejšia na malé zmeny h.

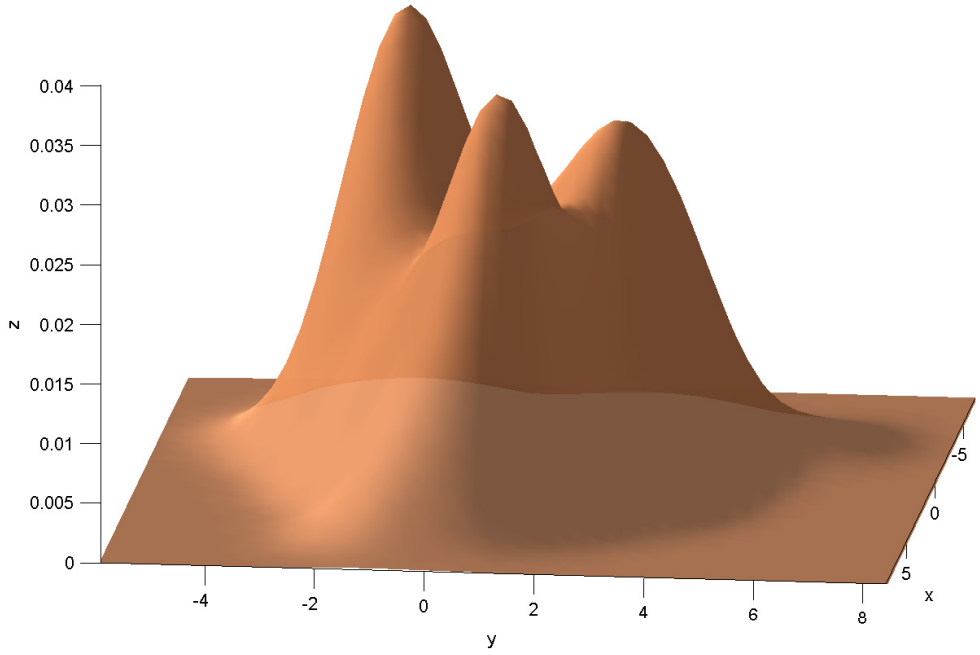
Dokonca pri použití rýchleho počítača je dôležité dávať pozor vo výpočtoch odhadu viacrozmernej hustoty. Obzvlášť keď rozsah náhodného výberu je veľký, použitie nevhodného algoritmu môže viesť k veľmi dlhému výpočtu.



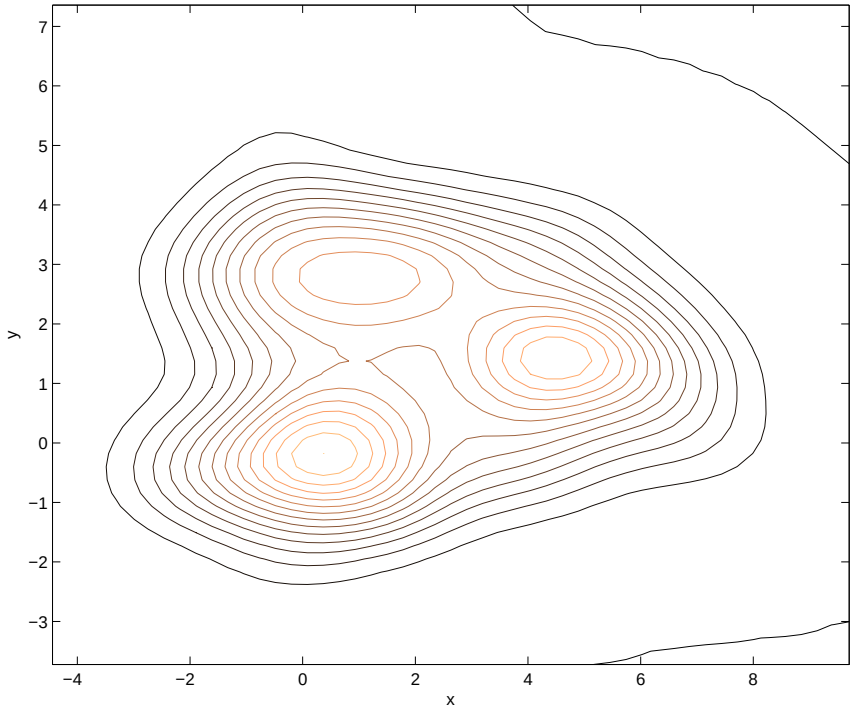
Obr. 3.1: Jadrový odhad trimodálnej hustoty, $h = 0.3$.



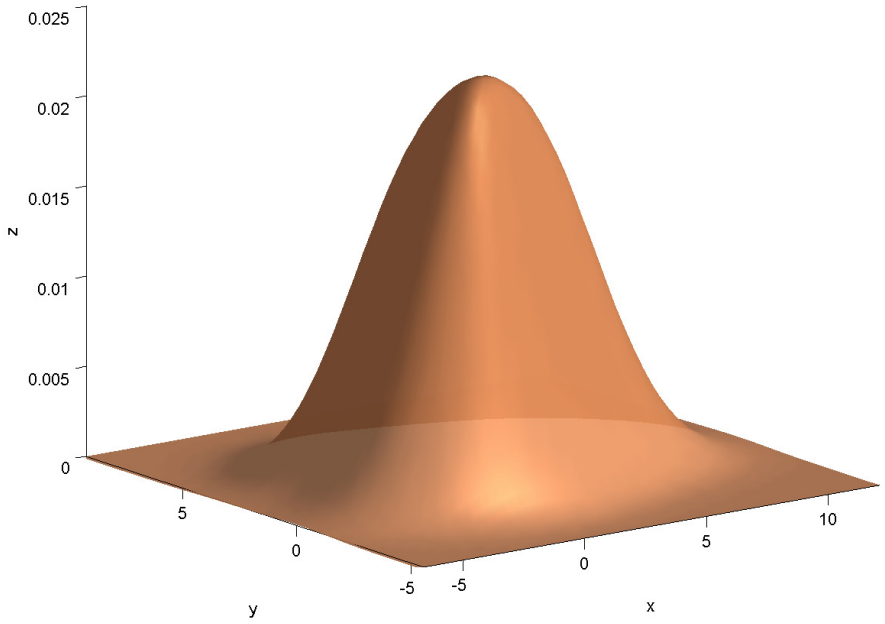
Obr. 3.2: Vrstevnice jadrového odhadu trimodálnej hustoty, $h = 0.3$.



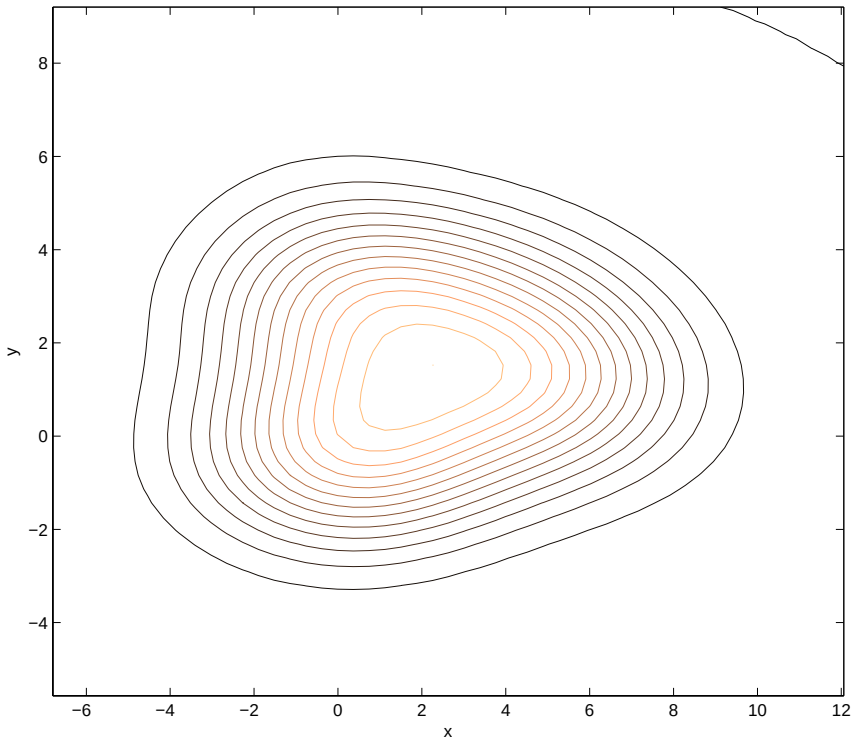
Obr. 3.3: Jadrový odhad trimodálnej hustoty, $h = 0.8$.



Obr. 3.4: Vrstevnice jadrového odhadu trimodálnej hustoty, $h = 0.8$.



Obr. 3.5: Jadrový odhad trimodálnej hustoty, $h = 1.6$.



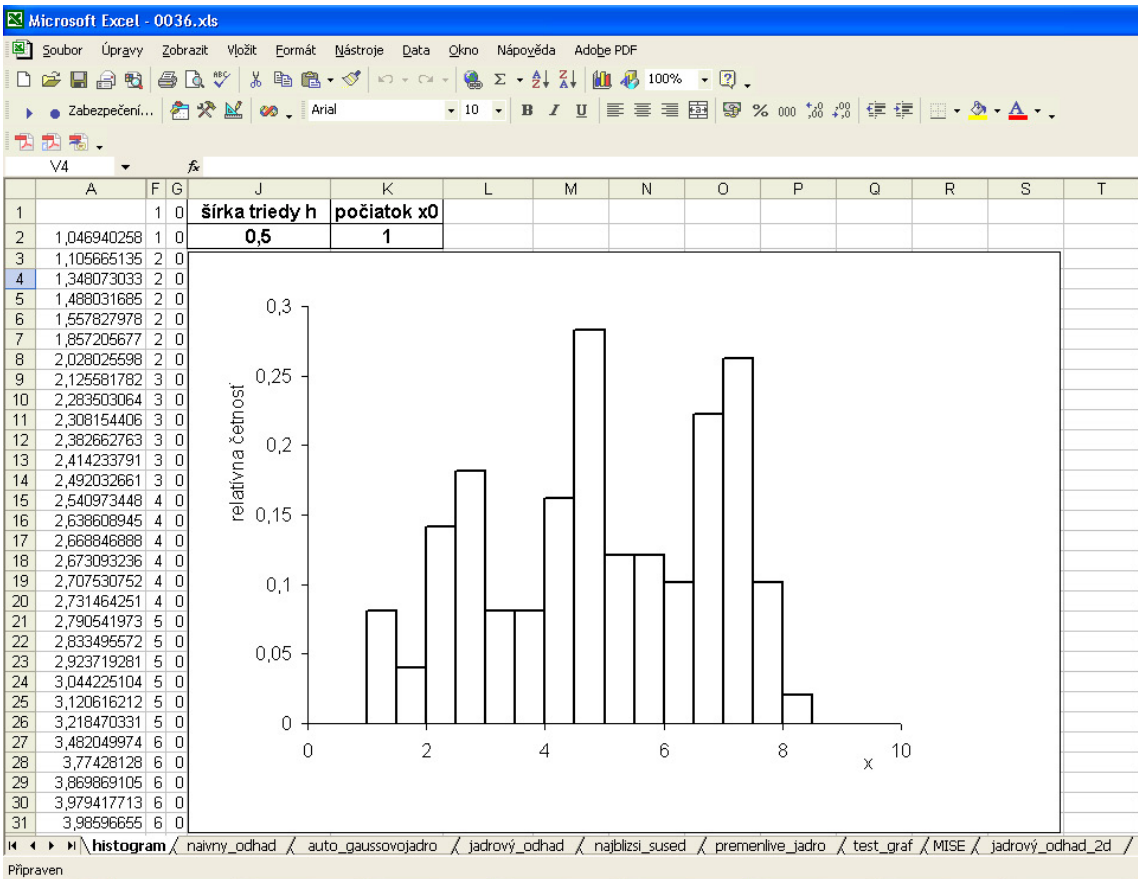
Obr. 3.6: Vrstevnice jadrového odhadu trimodálnej hustoty, $h = 1.6$.

4. Makrá v Exceli a ich popis

V tejto kapitole uvedieme implementáciu jadrových metód odhadu rozdelenia do makier v Exceli. Metódy sú rozdelené do listov v zošite Exceli. Náhodný výber sa zadáva do stĺpca A, okrem prvej bunky $A1$. Makrá majú rovnaké názvy ako metódy. V test grafe, subjektívnej voľbe vyhladzovacieho parametra v jadrovom odhade, metóde najbližšieho suseda a metóde premenlivého jadra je možné vybrať jedno z päť jadier. A to z gaussovho, epanechnikovho, biweight, trojuholníkovho a obdĺžnikovho jadra. Percento výpočtu sa ukazuje okrem histogramu a výpočtu vyhladzovacieho parametra metódou vzájomnej kontroly pomocou najmenších kvadrátov a normálnym rozdelením. Prvý stĺpec pre graf bude definičný obor $\hat{f}$ a druhý bude $\hat{f}$. Makrá dokážu pracovať s maximálnym rosahom náhodného výberu 65534. Bližší popis je uvedený pri každom makre.

4.1. Histogram

Do bunky $J2$ sa zadáva šírka triedy h a do bunky $K2$ sa zadáva počiatok x_0. Graf vykreslíme zo stĺpcov F a G. Na obrázku 4.1 vidíme makro histogram. Nasleduje makro histogram.



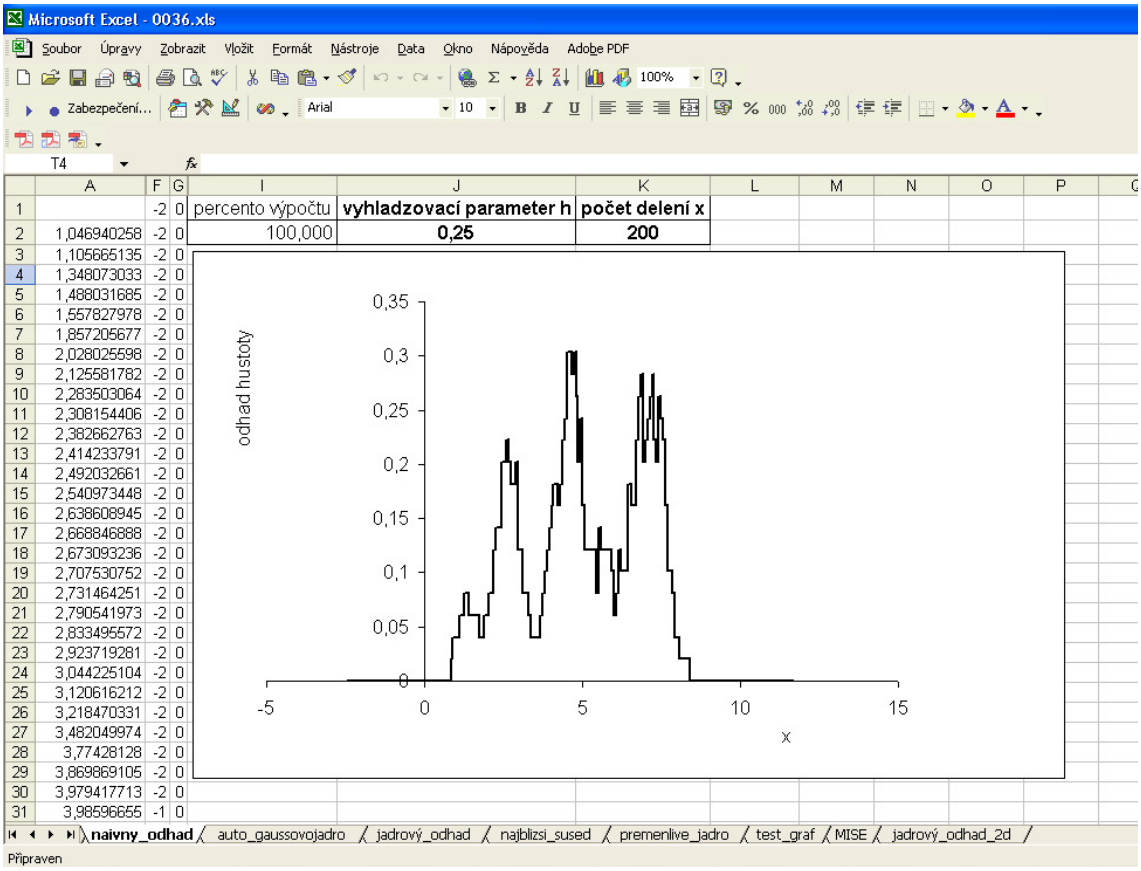
Obr. 4.1: Makro histogram.

```
Sub histogram()
Cells(1, 2) = "=COUNT(R[1]C[-1]:R[65535]C[-1])"
```

```
n = (Cells(1, 2))   'rozsah vyberu
h = (Cells(2, 10))  'šírka triedy
x0 = (Cells(2, 11)) 'počiatok

Range("B2:G65536").Select
Selection.ClearContents
Range("A2:A65536").Select
Selection.Sort Key1:=Range("A2"), Order1:=xlAscending, Header:=xlGuess, _
OrderCustom:=1, MatchCase:=False, Orientation:=xlTopToBottom, _
DataOption1:=xlSortNormal
Max = Cells(n + 1, 1)
Min = Cells(2, 1)
stlpcov = (Max - Min) / h + 1
  For m = 0 To stlpcov
   For i = 2 To n + 1
      Cells(i, 2) = ((x0 + m * h + h / 2) - Cells(i, 1)) / h
       'Histogram Jadro
         If Abs(Cells(i, 2)) < 1 / 2 Then
            Cells(i, 2) = 1
         ElseIf (Cells(i, 2)) = -1 / 2 Then
            Cells(i, 2) = 1
         Else
            Cells(i, 2) = 0
         End If
        Cells(m + 2, 3) = Cells(m + 2, 3) + Cells(i, 2) / h
    Next i
        Cells(m + 2, 5) = Cells(m + 2, 3) / n
        Cells(m + 2, 4) = x0 + m * h
  Next m
     Cells(1, 6) = Cells(2, 4)
     Cells(2, 6) = Cells(2, 4)
For i = 1 To stlpcov
     Cells(i * 3, 6) = Cells(i + 2, 4)
     Cells(i * 3 + 1, 6) = Cells(i + 2, 4)
     Cells(i * 3 + 2, 6) = Cells(i + 2, 4)
Next i
For i = 1 To 2 * stlpcov
     Cells((i - 1) * 3 + 1, 7) = 0
     Cells((i - 1) * 3 + 2, 7) = Cells(i + 1, 5)
     Cells((i - 1) * 3 + 3, 7) = Cells(i + 1, 5)
Next i
     Cells(6, 6) = Cells(4, 4)
     Cells(7, 6) = Cells(4, 4)
     Cells(7, 7) = 0
Range("B1:E65536").Select
Selection.ClearContents
End Sub
```

4.2. Naivný odhad

Do bunky $J2$ sa zadáva vyhladzovací parameter h a do bunky $K2$ sa zadáva ekvidistantné delenie definičného oboru odhadu. Graf vykreslíme zo stĺpcov F a G. Na obrázku 4.2 vidíme makro naivny_odhad. Nasleduje makro naivny_odhad.



Obr. 4.2: Makro naivny_odhad.

```
Sub naivny_odhad()
Cells(1, 2) = "=COUNT(R[1]C[-1]:R[65535]C[-1])"
n = Cells(1, 2)      'rozsah výberu
h = Cells(2, 10)     'vyhladzovací parameter
nn = Cells(2, 11)    'počet delení x

Range("A2:A65536").Select
Selection.Sort Key1:=Range("A2"), Order1:=xlAscending, Header:=xlGuess, _
OrderCustom:=1, MatchCase:=False, Orientation:=xlTopToBottom, _
DataOption1:=xlSortNormal
Range("B2:G65535").Select
Selection.ClearContents
sirka = (Cells(n + 1, 1) - Cells(2, 1)) * 2
pociatok = Cells(2, 1) - (sirka - (Cells(n + 1, 1) - Cells(2, 1))) / 2
mm = sirka / (nn)
  For m = 0 To nn
    For i = 2 To n + 1
```

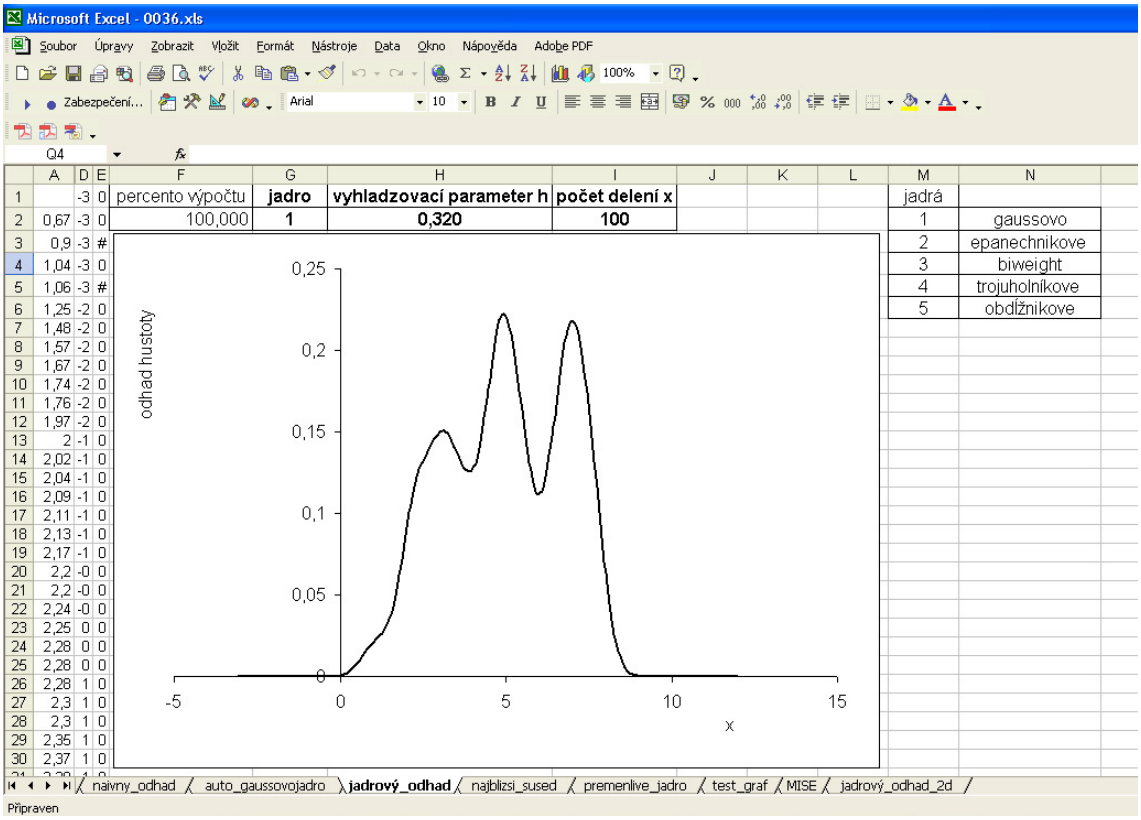
```
      Cells(i, 2) = ((pociatok + m * mm) - Cells(i, 1)) / h
         If Abs(Cells(i, 2)) < 1 Then
            Cells(i, 2) = 1 / 2
         Else
            Cells(i, 2) = 0
         End If
            Cells(m + 2, 3) = Cells(m + 2, 3) + Cells(i, 2) / h
     Next i
        Cells(m + 2, 5) = Cells(m + 2, 3) / n
        Cells(m + 2, 4) = pociatok + m * mm
        Cells(2, 9) = (m + 1) / (nn + 1) * 100
  Next m
    For i = 2 To nn + 2
        Cells((i - 2) * 2 + 1, 6) = Cells(i, 4)
        Cells((i - 2) * 2 + 2, 6) = Cells(i, 4)
    Next i
       Cells(1, 7) = Cells(2, 5)
    For i = 1 To nn
          Cells(2 * i, 7) = Cells(i + 1, 5)
          Cells(2 * i + 1, 7) = Cells(i + 1, 5)
    Next i
Range("B1:E65535").Select
Selection.ClearContents
End Sub
```

4.3. Jadrový odhad pre jednorozmerný náhodný výber

4.3.1. Subjektívna voľba

Do bunky $G2$ sa zadáva jadro. Do bunky $H2$ sa zadáva vyhladzovací parameter h a do bunky $I2$ sa zadáva ekvidistantné delenie definičného oboru odhadu. Graf vykreslíme zo stĺpcov D a E. Na obrázku 4.3 vidíme makro jadrovy_odhad. Nasleduje makro jadrovy_odhad.

```
Sub jadrovy_odhad()
Cells(1, 2) = "=COUNT(R[1]C[-1]:R[65535]C[-1])"
n = Cells(1, 2)     'rozsah výberu
h = Cells(2, 8)     'vyhladzovaci parameter
nn = Cells(2, 9)    'počet delení x

Range("A2:A65536").Select
Selection.Sort Key1:=Range("A2"), Order1:=xlAscending, Header:=xlGuess, _
OrderCustom:=1, MatchCase:=False, Orientation:=xlTopToBottom, _
DataOption1:=xlSortNormal
Range("B2:E65536").Select
```



Obr. 4.3: Makro jadrovy_odhad.

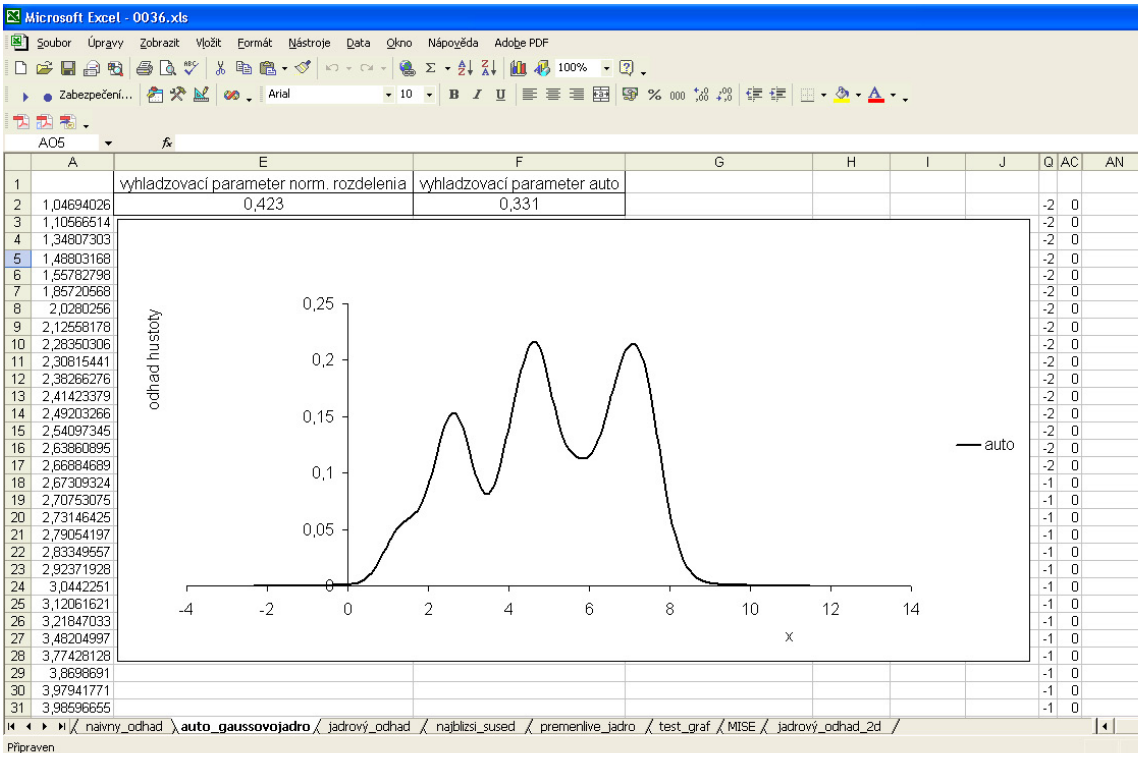
```
Selection.ClearContents
Pi = 3.14159265358979
sirka = (Cells(n + 1, 1) - Cells(2, 1)) * 2
pociatok = Cells(2, 1) - (sirka - (Cells(n + 1, 1) - Cells(2, 1))) / 2
mm = sirka / (nn)
   For m = 0 To nn
   For i = 2 To n + 1
      Cells(i, 2) = ((pociatok + m * mm) - Cells(i, 1)) / h
         If Cells(2, 7) = 1 Then              'gaussovo jadro
             Cells(i, 2) = 1 / (2 * Pi) ^ 0.5 * Exp(-0.5 * (Cells(i, 2)) ^ 2)
         ElseIf Cells(2, 7) = 2 Then          'epanechnikove jadro
            If Abs(Cells(i, 2)) < 5 ^ 0.5 Then
               Cells(i, 2) = 3 / 4 * (1 - 1 / 5 * Cells(i, 2) ^ 2) / 5 ^ 0.5
            Else
               Cells(i, 2) = 0
            End If
         ElseIf Cells(2, 7) = 3 Then          'biweight jadro
            If Abs(Cells(i, 2)) < 1 Then
                Cells(i, 2) = 15 / 16 * (1 - Cells(i, 2) ^ 2) ^ 2
            Else
                Cells(i, 2) = 0
            End If
        ElseIf Cells(2, 7) = 4 Then           'trojuholníkove jadro
            If Abs(Cells(i, 2)) < 1 Then
```

```
            Cells(i, 2) = 1 - Abs(Cells(i, 2))
         Else
            Cells(i, 2) = 0
         End If
      ElseIf Cells(2, 7) = 5 Then            'obdĺžnikove jadro
         If Abs(Cells(i, 2)) < 1 Then
            Cells(i, 2) = 0.5
         Else
            Cells(i, 2) = 0
         End If
      End If
      Cells(m + 2, 3) = Cells(m + 2, 3) + Cells(i, 2)
   Next i
       Cells(m + 2, 5) = Cells(m + 2, 3) / (h * n)
       Cells(m + 2, 4) = pociatok + m * mm
       Cells(2, 6) = (m + 1) / (nn + 1) * 100
  Next m
Cells(1, 4) = Cells(2, 4)
Cells(1, 5) = 0
Cells(nn + 3, 4) = Cells(nn + 2, 4)
Cells(nn + 3, 5) = 0
Cells(nn + 3, 4) = Cells(nn + 2, 4)
Cells(nn + 3, 5) = Cells(nn + 2, 5)
Range("B1:C65536").Select
Selection.ClearContents
End Sub
```

4.3.2. Auto a pomocou normálneho rozdelenia

Zadáva sa iba náhodný výber. Graf vykreslíme zo stĺpcov Q a AC. V bunke $E2$ dostaneme vyhladzovací parameter pomocou normálneho rozdelenia. V bunke $F2$ dostaneme vyhladzovací parameter pomocou vzájomnej kontroly najmenších kvadrátov. Na výpočet vyhlazovacieho parametra pomocou normálneho rozdelenia je použitý adaptivný odhad rozsahu. V auto metóde na hľadanie minima hodnoty M_1 sme použili Newtonovu metódu. Newtonova metóda môže uviaznuť v lokálnom minime. Keďže sa počiatočný bod hľadá náhodne v doporučenom intervale, je dobré spustiť toto makro niekoľko krát a porovnať hodnoty $F2$. Na obrázku 4.4 vidíme makro auto_gaussovojadro, je to jadrový odhad pomocou vzájomnej kontroly najmenších kvadrátov (auto). Nasleduje makro auto_gaussovojadro.

```
Sub auto_gaussovojadro()
Range("A2:A65536").Select
Selection.Sort Key1:=Range("A2"), Order1:=xlAscending, Header:=xlGuess, _
OrderCustom:=1, MatchCase:=False, Orientation:=xlTopToBottom, _
DataOption1:=xlSortNormal
Range("B2:AB65536").Select
Selection.ClearContents
```



Obr. 4.4: Makro auto_gaussovojadro, jadrový odhad pomocou vzájomnej kontroly najmenších kvadrátov (auto).

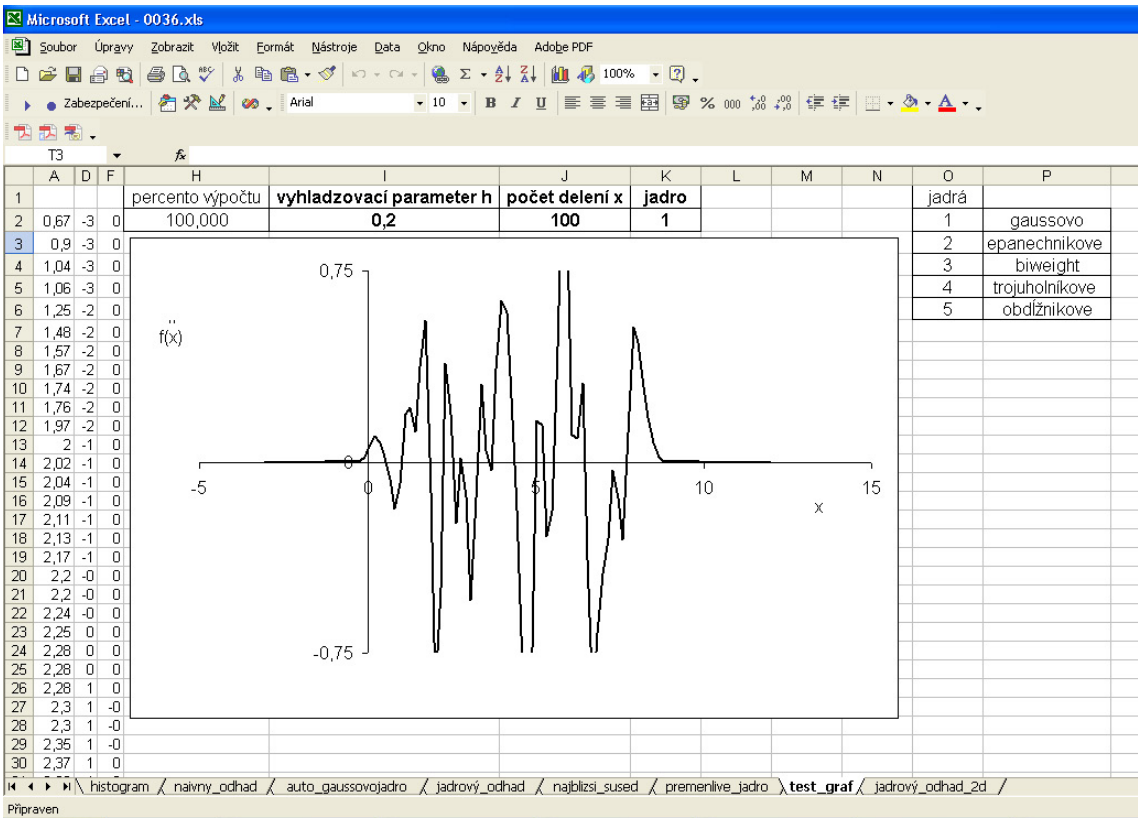
```
Cells(2, 4) = "=COUNT(RC[-3]:R[65534]C[-3])"
n = Cells(2, 4)
Pi = 3.14159265358979
Cells(2, 3) = "=QUARTILE(RC[-2]:R[65534]C[-2],3)"
Cells(3, 3) = "=QUARTILE(RC[-2]:R[65533]C[-2],1)"
Cells(2, 2) = "=StDevP(RC[-1]:R[65534]C[-1])"
 If Cells(2, 2) < ((Cells(2, 3) - Cells(3, 3)) / 2 / 1.34) Then
   Cells(2, 5) = 0.9 * Cells(2, 2) * n ^ (-1 / 5)
   Else: Cells(2, 5) = 0.9 * ((Cells(2, 3) - Cells(3, 3))/2/1.34)*n^(-1/5)
 End If
smerod_odchylka = Cells(2, 2)
h_spodne = smerod_odchylka / (4 * n ^ (1 / 5))
h_vrchne = 3 * smerod_odchylka / (2 * n ^ (1 / 5))
 B = Cells(n + 1, 1) + 3 * h_vrchne
 A = Cells(2, 1) - 3 * h_vrchne
 R = 8
 m = 2 ^ R
 delta = (B - A) / m
 For i = 1 To m
       Cells(i + 1, 17) = A + (i - 1) * delta
 Next i
 Range("R2:R257").Select
 Selection.Formula = 0
 For j = 2 To m
```

```
    For i = 2 To n + 1
        If (Cells(j + 1, 17) - Cells(i, 1)) <= delta Then If (Cells(j + 1, 17)
         - Cells(i, 1)) > 0 Then Cells(j, 18) = Cells(j, 18) + 1 / (n * delta
          * delta) * (Cells(j + 1, 17) - Cells(i, 1))
        If (Cells(j + 1, 17) - Cells(i, 1)) <= delta Then If (Cells(j + 1, 17)
         - Cells(i, 1)) > 0 Then Cells(j + 1, 18) = Cells(j + 1, 18) + 1 / (n
         * delta * delta) * (Cells(i, 1) - Cells(j, 17))
    Next i
 Next j
  Application.Run "ATPVBAEN.XLA!Fourier", ActiveSheet.Range("$R$2:$R$257"), _
        ActiveSheet.Range("$T$2:$T$257"), True, False
h1 = (h_spodne + h_vrchne) / 2
krok = 0.000001
delta = 1
Do While delta >= 0.00000001
    For hs_spod = 1 To 3
    hs = h1 - ((hs_spod - 1) * krok)
       For i = 1 To m
           Cells(i + 1, 19) = 2 * Pi * i / (B - A)
           Cells(i + 1, 22) = Exp(-1 / 2 * hs ^ 2 * Cells(i + 1, 19) ^ 2)
           Cells(i + 1, 23) = "=COMPLEX(RC[-1],0)"
           Cells(i + 1, 21) = "=IMPRODUCT(RC[2],RC[-1])"
           Cells(i + 1, 27) = "=IMABS(RC[-7])"
           Cells(i + 1, 28) = (Exp(-hs ^ 2 * Cells(i + 1, 19) ^ 2)
            - 2 * Exp(-1 / 2 * hs ^ 2 * Cells(i + 1, 19) ^ 2))
            * Cells(i + 1, 27) ^ 2
        Next i
        Cells(2, 30) = "=SUM(RC[-2] :R[129]C[-2])"
        Cells(2, 31) = 2 * (B - A) * Cells(2, 30) + 2 /(n * hs *(2*Pi)^0.5)-1
        Cells(1 + hs_spod, 33) = Cells(2, 31)
  diff1 = (Cells(2, 33) - Cells(3, 33)) / krok
  diff2 = (Cells(2, 33) - 2 * Cells(3, 33) + Cells(4, 33)) / (krok * krok)
 hopt = h1 - diff1 / diff2
 delta = Abs(h1 - hopt)
 Cells(8, 12) = delta
 h1 = hopt
Cells(2, 39) = "=RANDBETWEEN(1,1000)"
If h1 > h_vrchne Then h1 = Cells(2, 39) / 1000 * h_vrchne
If h1 < 0 Then h1 = Cells(2, 39) / 1000 * h_vrchne
Cells(2, 6) = h1
Loop
  Application.Run "ATPVBAEN.XLA!Fourier", ActiveSheet.Range("U2:U257"), _
        ActiveSheet.Range("X2:X257"), False, False
  For i = 1 To m
      Cells(i + 1, 25) = "=IMREAL(RC[-1])"
  Next i
 Cells(2, 35) = "=MIN(RC[-10]:R[255]C[-10])"
```

```
 mmin = Cells(2, 35)
  For i = 1 To m
      Cells(i + 1, 29) = (Cells(i + 1, 25) - mmin) * 2
  Next i
Range("B1:D65536").Select
Selection.ClearContents
Range("L8:L8").Select
Selection.ClearContents
Range("R1:AB257").Select
Selection.ClearContents
Range("AD1:AM257").Select
Selection.ClearContents
End Sub
```

4.3.3. Test graf

Do bunky $K2$ sa zadáva jadro. Do bunky $I2$ sa zadáva vyhladzovací parameter h a do bunky $J2$ sa zadáva ekvidistantné delenie definičného oboru odhadu. Graf vykreslíme zo stĺpcov D a F. Na obrázku 4.5 vidíme makro test_graf. Nasleduje makro test_graf.



Obr. 4.5: Makro test_graf.

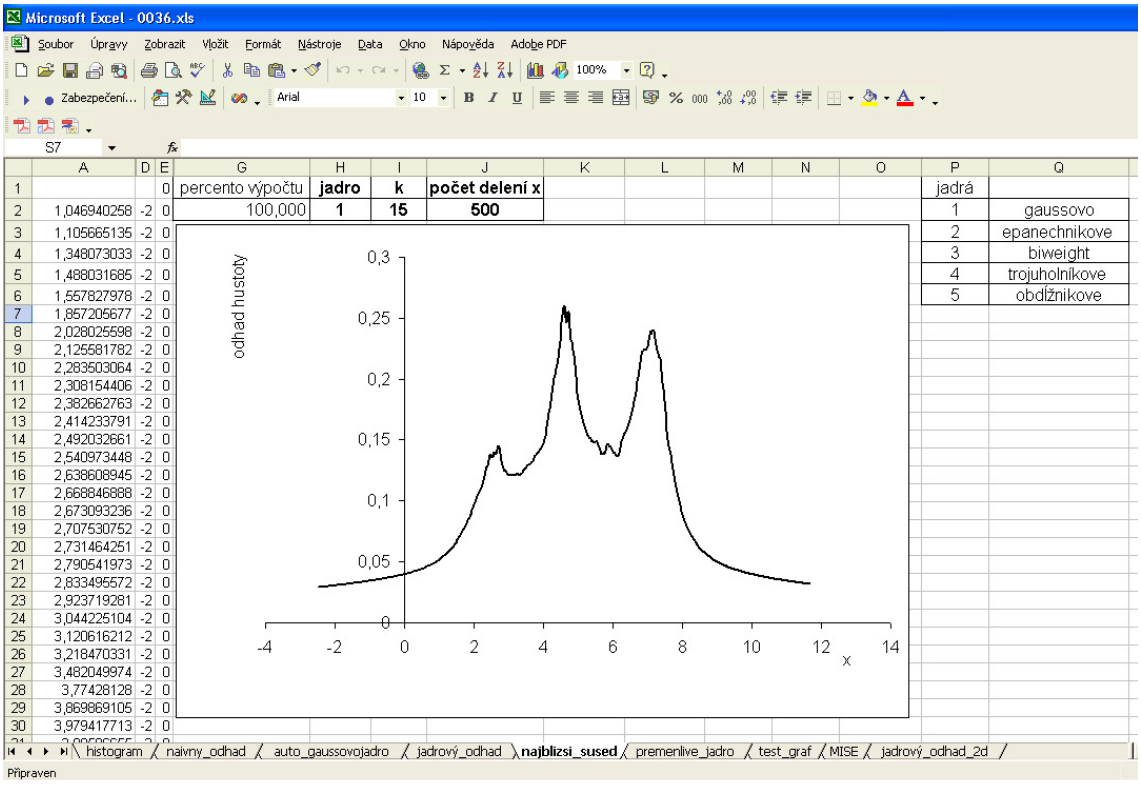
```
Sub test_graf()
Cells(1, 2) = "=COUNT(R[1]C[-1]:R[65535]C[-1])"
n = Cells(1, 2)     'rozsah výberu
h = Cells(2, 9)     'vyhladzovaci parameter
```

```
nn = Cells(2, 10)  'počet delení x

Range("A2:A65536").Select
Selection.Sort Key1:=Range("A2"), Order1:=xlAscending, Header:=xlGuess, _
OrderCustom:=1, MatchCase:=False, Orientation:=xlTopToBottom, _
DataOption1:=xlSortNormal
Range("B2:F65536").Select
Selection.ClearContents
Pi = 3.14159265358979
sirka = (Cells(n + 1, 1) - Cells(2, 1)) * 2
pociatok = Cells(2, 1) - (sirka - (Cells(n + 1, 1) - Cells(2, 1))) / 2
mm = sirka / (nn)
   For m = 0 To nn
   For i = 2 To n + 1
      Cells(i, 2) = ((pociatok + m * mm) - Cells(i, 1)) / h
            If Cells(2, 11) = 1 Then                'gaussovo jadro
            Cells(i, 2) = 1 / (2 * Pi) ^ 0.5 * Exp(-0.5 * (Cells(i, 2))^2)
        ElseIf Cells(2, 11) = 2 Then                'epanechnikove jadro
           If Abs(Cells(i, 2)) < 5 ^ 0.5 Then
              Cells(i, 2) = 3 / 4 * (1 - 1 / 5 * Cells(i, 2) ^ 2) / 5 ^0.5
           Else
              Cells(i, 2) = 0
           End If
        ElseIf Cells(2, 11) = 3 Then                'biweight jadro
           If Abs(Cells(i, 2)) < 1 Then
               Cells(i, 2) = 15 / 16 * (1 - Cells(i, 2) ^ 2) ^ 2
           Else
               Cells(i, 2) = 0
           End If
         ElseIf Cells(2, 11) = 4 Then               'trojuholníkove jadro
           If Abs(Cells(i, 2)) < 1 Then
              Cells(i, 2) = 1 - Abs(Cells(i, 2))
           Else
              Cells(i, 2) = 0
           End If
       ElseIf Cells(2, 11) = 5 Then                 'obdĺžnikove jadro
           If Abs(Cells(i, 2)) < 1 Then
              Cells(i, 2) = 0.5
           Else
              Cells(i, 2) = 0
           End If
       End If
       Cells(m + 2, 3) = Cells(m + 2, 3) + Cells(i, 2)
    Next i
        Cells(m + 2, 5) = Cells(m + 2, 3) / (h * n)
        Cells(m + 2, 4) = pociatok + m * mm
        Cells(2, 8) = (m + 1) / (nn + 1) * 100
```

```
  Next m
For i = 2 To n - 1                          'druha derivácia dopredná i spätná
  Cells(i, 6) = (Cells(i + 2, 5) - 2 * Cells(i + 1, 5) + Cells(i, 5))/(mm*mm)
Next i
For i = n To n + 1
  Cells(i, 6) = (Cells(i, 5) - 2 * Cells(i - 1, 5) + Cells(i - 2, 5))/(mm*mm)
Next i
Cells(nn + 2, 4) = Cells(nn + 2, 4)
Cells(nn + 2, 6) = Cells(nn + 2, 6)
Range("B1:C65536").Select
Selection.ClearContents
Range("E2:E65536").Select
Selection.ClearContents
End Sub
```

4.4. Metóda najbližšieho suseda

Do bunky $H2$ sa zadáva jadro. Do bunky $I2$ sa zadáva k a do bunky $J2$ sa zadáva ekvidistantné delenie definičného oboru odhadu. Graf vykreslíme zo stĺpcov D a E. Na obrázku 4.6 vidíme makro sajblizsi_suded. Nasleduje makro najblizsi_suded.



Obr. 4.6: Makro najblizsi_suded.

```
Sub najblizsi_suded()
Cells(1, 2) = "=COUNT(R[1]C[-1]:R[65535]C[-1])"
n = Cells(1, 2)              'rozsah výberu
```

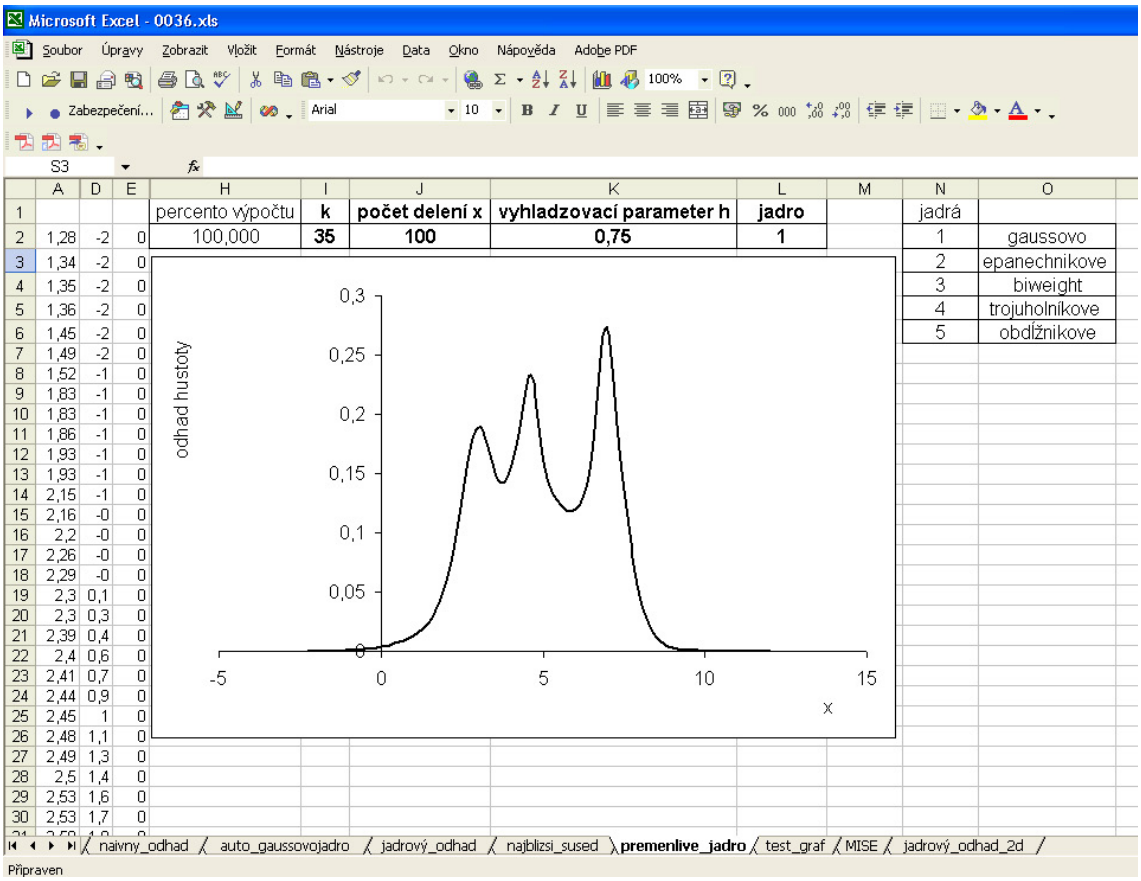
```
kk = Cells(2, 9)            'k
nn = Cells(2, 10)           'počet deleni x

Range("A2:A65536").Select
Selection.Sort Key1:=Range("A2"), Order1:=xlAscending, Header:=xlGuess, _
OrderCustom:=1, MatchCase:=False, Orientation:=xlTopToBottom, _
DataOption1:=xlSortNormal
Range("B2:E65536").Select
Selection.ClearContents
Range("F2:F65536").Select
Selection.Formula = 100000
Pi = 3.14159265358979
sirka = (Cells(n + 1, 1) - Cells(2, 1)) * 2
pociatok = Cells(2, 1) - (sirka - (Cells(n + 1, 1) - Cells(2, 1))) / 2
mm = sirka / (nn)
dkminus = 10000
pravy = 0
lavy = 0
mmm = 0
   For m = 0 To nn
      Xj = pociatok + m * mm
        For k = 2 To n + 1
           Cells(k, 6) = Abs(Xj - Cells(k, 1))
        Next k
        Cells(kk + 2, 6) = 100000
        Range("F2:F65536").Select
Selection.Sort Key1:=Range("F2"), Order1:=xlAscending, Header:=xlGuess, _
        OrderCustom:=1, MatchCase:=False, Orientation:=xlTopToBottom, _
        DataOption1:=xlSortNormal
     If Xj < Cells(2, 1) Then
           dk = Abs(Cells(1 + kk, 1) - Xj)
     End If
     If Xj > Cells(n + 1, 1) Then
           dk = Abs(Xj - Cells(n + 1 - kk + 1, 1))
     Else
           dk = Cells(1 + kk, 6)
     End If
   For i = 2 To n + 1
        Cells(i, 2) = (Xj - Cells(i, 1)) / dk
         If Cells(2, 8) = 1 Then                          'gausovo jadro
             Cells(i, 2) = 1 / (2 * Pi) ^ 0.5 * Exp(-0.5*(Cells(i,2))^2)
         ElseIf Cells(2, 8) = 2 Then                'epanechnikove Jadro
            If Abs(Cells(i, 2)) < 5 ^ 0.5 Then
               Cells(i, 2) = 3 / 4 * (1 - 1 / 5 * Cells(i, 2) ^ 2)/5^0.5
            Else
               Cells(i, 2) = 0
            End If
```

```
         ElseIf Cells(2, 8) = 3 Then                    'biweight jadro
            If Abs(Cells(i, 2)) < 1 Then
                Cells(i, 2) = 15 / 16 * (1 - Cells(i, 2) ^ 2) ^ 2
            Else
                Cells(i, 2) = 0
            End If
        ElseIf Cells(2, 8) = 4 Then                'trojuholníkove jadro
            If Abs(Cells(i, 2)) < 1 Then
               Cells(i, 2) = 1 - Abs(Cells(i, 2))
            Else
               Cells(i, 2) = 0
            End If
       ElseIf Cells(2, 8) = 5 Then                   'obdlznikove jadro
            If Abs(Cells(i, 2)) < 1 Then
               Cells(i, 2) = 0.5
            Else
               Cells(i, 2) = 0
            End If
       End If
         Cells(m + 2, 3) = Cells(m + 2, 3) + Cells(i, 2)
   Next i
       Cells(m + 2, 5) = Cells(m + 2, 3) / (n * dk)
       Cells(m + 2, 4) = Xj
       Cells(2, 7) = (m + 1) / (nn + 1) * 100
       If Xj < Cells(2, 1) Then
          If m > 1 Then
             If Cells(m + 2, 5) < Cells(m + 1, 5) Then
                Cells(m + 1, 5) = 0
                Cells(2, 5) = 0
                lavy = 1
                mmmm = m
             End If
          End If
       End If
       If Xj > Cells(n + 1, 1) Then
          If m < nn - 1 Then
             If Cells(m + 2, 5) > Cells(m + 1, 5) Then
                Cells(m + 1, 5) = 0
                Cells(nn + 1, 5) = 0
                Cells(nn, 5) = 0
                mmm = m
                pravy = 1
             End If
          End If
       End If
       dkminus = dk
 Next m
```

```
If lavy = 1 Then
    For i = 1 To mmmm
       Cells(i + 1, 5) = 0
    Next i
End If
If pravy = 1 Then
    For i = mmm To nn + 1
       Cells(i + 1, 5) = 0
    Next i
End If
Range("B1:C65536").Select
Selection.ClearContents
Range("F1:F65536").Select
Selection.ClearContents
End Sub
```

4.5. Metóda premenlivého jadra

Do bunky $L2$ sa zadáva jadro a do bunky $K2$ sa zadáva vyhladzovací parameter. Do bunky $I2$ sa zadáva k a do bunky $J2$ sa zadáva ekvidistantné delenie definičného oboru odhadu. Graf vykreslíme zo stĺpcov D a E. Na obrázku 4.7 vidíme makro premenlive_jadro. Nasleduje makro premenlive_jadro.



Obr. 4.7: Makro premenlive_jadro.

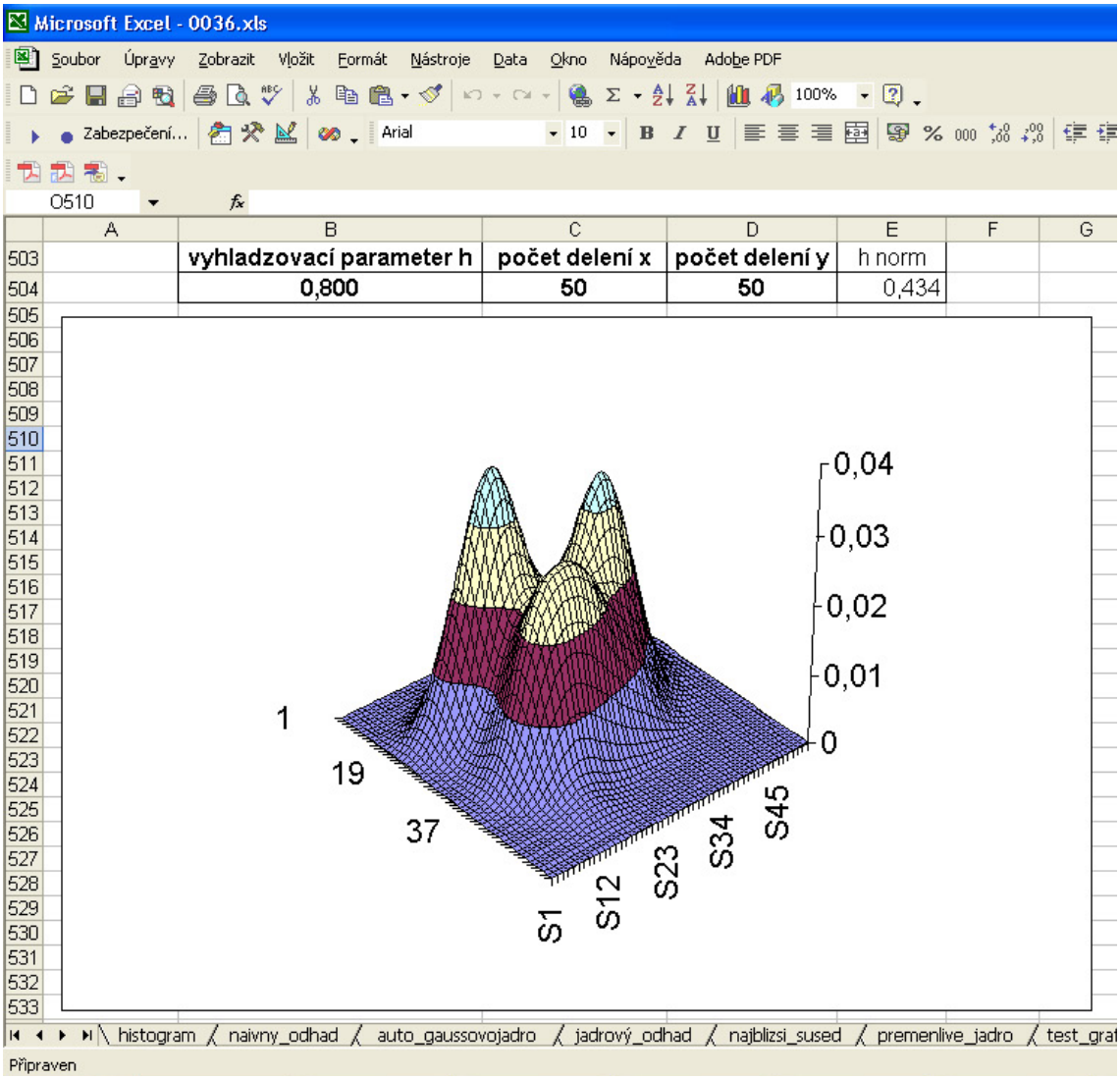
```
Sub premenlive_jadro()
Cells(1, 2) = "=COUNT(R[1]C[-1]:R[65535]C[-1])"
n = Cells(1, 2)          'rozsah výberu
kk = Cells(2, 9)         'k
nn = Cells(2, 10)        'počet delení x
h = Cells(2, 11)         'vyhladzovací parameter

Range("A2:A65536").Select
Selection.Sort Key1:=Range("A2"), Order1:=xlAscending, Header:=xlGuess, _
OrderCustom:=1, MatchCase:=False, Orientation:=xlTopToBottom, _
DataOption1:=xlSortNormal
Range("B2:E65536").Select
Selection.ClearContents
Range("F2:F65536").Select
Selection.Formula = 100000
Pi = 3.14159265358979
sirka = (Cells(n + 1, 1) - Cells(2, 1)) * 2
pociatok = Cells(2, 1) - (sirka - (Cells(n + 1, 1) - Cells(2, 1))) / 2
mm = sirka / (nn)
  For m = 0 To nn
    For i = 2 To n + 1
     Xj = pociatok + m * mm
        For k = 2 To n + 1
           Cells(k, 6) = Abs(Cells(i, 1) - Cells(k, 1))
           If Cells(k, 6) = 0 Then
              Cells(k, 6) = 100000
           End If
        Next k
        Cells(kk + 2, 6) = 100000
        Range("F2:F65536").Select
Selection.Sort Key1:=Range("F2"), Order1:=xlAscending, Header:=xlGuess, _
        OrderCustom:=1, MatchCase:=False, Orientation:=xlTopToBottom, _
        DataOption1:=xlSortNormal
        Djk = Cells(1 + kk, 6)
        Cells(i, 2) = (Xj - Cells(i, 1)) / (h * Djk)
         If Cells(2, 12) = 1 Then                          'gaussovo jadro
             Cells(i, 2) = 1 / (2 * Pi) ^ 0.5 * Exp(-0.5 * (Cells(i,2))^2)
         ElseIf Cells(2, 12) = 2 Then                  'epanechnikove jadro
            If Abs(Cells(i, 2)) < 5 ^ 0.5 Then
               Cells(i, 2) = 3 / 4 * (1 - 1 / 5 * Cells(i, 2) ^ 2) / 5^0.5
            Else
               Cells(i, 2) = 0
            End If
         ElseIf Cells(2, 12) = 3 Then                      'biweight jadro
            If Abs(Cells(i, 2)) < 1 Then
                Cells(i, 2) = 15 / 16 * (1 - Cells(i, 2) ^ 2) ^ 2
            Else
```

```
                Cells(i, 2) = 0
            End If
         ElseIf Cells(2, 12) = 4 Then                'trojuholníkove jadro
            If Abs(Cells(i, 2)) < 1 Then
               Cells(i, 2) = 1 - Abs(Cells(i, 2))
            Else
               Cells(i, 2) = 0
            End If
        ElseIf Cells(2, 12) = 5 Then                   'obdĺžnikove jadro
            If Abs(Cells(i, 2)) < 1 Then
               Cells(i, 2) = 0.5
            Else
               Cells(i, 2) = 0
            End If
        End If
          Cells(m + 2, 3) = Cells(m + 2, 3) + Cells(i, 2) / (h * Djk)
    Next i
        Cells(m + 2, 5) = Cells(m + 2, 3) / n
        Cells(m + 2, 4) = Xj
        Cells(2, 8) = (m + 1) / (nn + 1) * 100
  Next m
    Cells(nn + 2, 4) = Cells(nn + 2, 4)
    Cells(nn + 2, 5) = Cells(nn + 2, 5)
Range("B1:C65536").Select
Selection.ClearContents
Range("F1:F65536").Select
Selection.ClearContents
End Sub
```

4.6. Jadrový odhad pre dvojrozmerný náhodný výber

Do bunky $B504$ sa zadáva vyhladzovací parameter. Do bunky $C504$ a $D504$ sa zadáva ekvidistantné delenie definičného oboru odhadu. V bunke $E504$ dostaneme doporučený štartovací vyhladzovací parameter pomocou normálneho rozdelenia. Graf vykreslíme z nasledovných oblastí: stĺpce G a H a oblasť $J2 : BH52$ pre hodnotu $C504 = D504 = 50$. Maximálny rozsah náhodného výberu je 500. V bunke $A1$ sa uvádza percento výpočtu. Na obrázku 4.8 vidíme makro jadr_2dodhadgauss. Nasleduje makro jadr_2dodhadgauss.

```
Sub jadr_2d_odhadgauss()
Cells(1, 2) = "=COUNT(R[1]C[-1]:R[501]C[-1])"
n = Cells(1, 2)      'rozsah výberu
h = Cells(504, 2)    'vyhladzovaci parameter
nn = Cells(504, 3)   'pocet deleni x
ny = Cells(504, 4)   'pocet deleni y

Cells(504, 5) = Cells(1, 2) ^ (-1 / 6)
Range("C2:EE501").Select
```



Obr. 4.8: Makro jadr_2dodhadgauss.

```
Selection.ClearContents
Pi = 3.14159265358979
Range("A2:B501").Select
Selection.Sort Key1:=Range("A2"), Order1:=xlAscending, Header:=xlGuess, _
OrderCustom:=1, MatchCase:=False, Orientation:=xlTopToBottom, _
DataOption1:=xlSortNormal
Cells(1, 3) = "=MAX(R[1]C[-1]:R[500]C[-1])"
Cells(1, 4) = "=MIN(R[1]C[-2]:R[500]C[-2])"
xsirka = (Cells(n + 1, 1) - Cells(2, 1)) * 1.5
xpociatok = Cells(2, 1) - (xsirka - (Cells(n + 1, 1) - Cells(2, 1))) / 2
xmm = xsirka / nn
ysirka = (Cells(1, 3) - Cells(1, 4)) * 1.5
ypociatok = Cells(1, 4) - (ysirka - (Cells(1, 3) - Cells(1, 4))) / 2
ymm = ysirka / ny
 For ym = 0 To ny
   For xm = 0 To nn
```

```
   For i = 2 To n + 1
      Cells(i, 3) = ((xpociatok + xm * xmm) - Cells(i, 1)) / h
      Cells(i, 4) = ((ypociatok + ym * ymm) - Cells(i, 2)) / h
              ' Gausovo jadro
     Cells(i, 5) = 1 / (2 * Pi) * Exp(-0.5 * ((Cells(i, 3)) * (Cells(i, 3))
                  + (Cells(i, 4)) * (Cells(i, 4))))
        Cells(xm + 2, 6) = Cells(xm + 2, 6) + Cells(i, 5)
    Next i
        Cells(xm + 2, 10 + ym) = Cells(xm + 2, 6) / (h * h * n)
        Cells(xm + 2, 7) = xpociatok + xm * xmm
        Cells(1, 1) = (ym + 1) / (nn + 1) * 100
  Next xm
Range("C2:F501").Select
Selection.ClearContents
Cells(ym + 2, 8) = ypociatok + ym * ymm
Next ym
End Sub
```

5. Záver

Cieľom práce bol popis, zhodnotenie a implementácia moderných štatistických metód fitovania rozdelenia pravdepodobnosti pomocou jadrových odhadov vzhľadom k možnostiam ich realizácie na PC a aplikáciam na konkrétnych datových súboroch. Zamerali sme sa na jednorozmerné odhady: jadrový odhad, metóda najbližšieho suseda, premenlivé jadro a viacrozmerné jadrové odhady. Na tieto odhady sme urobili makrá v Exceli, vrátane naivného odhadu, histogramu a automatickej metódy hľadania vyhlazovacieho parametra jadrového odhadu pomocou vzájomnej kontroly najmenších kvadrátov.

Časová náročnosť výpočtu odhadov narastá s rastúcim rozsahom náhodného výberu a zjemňovaním ekvidistantného delenia definičného oboru odhadu. V príkladoch sme rozsahy náhodných výberov a ekvidistantné delenia definičných oborov odhadov volili tak, aby najnáročnejší odhad trval okolo jednej hodiny. Najviac náročný odhad na výpočet je odhad s premenlivým jadrom. Odhady s premenlivým jadrom majú najväčšie rozsahy náhodných výberov 297. Preto interval na ktorom sme porovnávali odhady trimodálnej hustoty, sme ekvidistantným delením rozdelili na 100 intervalov. V príkladoch asymptotických vlastností sme tento interval rozdelili na 500 intervalov. Keďže makro ukazuje percento výpočtu, vieme podľa neho odhadnúť, ako dlho bude trvať výpočet. Uvádzame pokusné zlepšenia pre tento odhad. Nerobili by sa usporiadania celého náhodného výberu, ale len okolia k náhodnej veličiny X_j. Toto bolo vyskúšané, ale nedošlo k poznateľnému zrýchleniu. Nerobilo by sa usporiadanie náhodného výberu pre každú náhodnú veličinu X_j, ale len pre prvú, a potom by sa drobnými úpravami zistil $d_{j,k}$. Tento postup dokonca spomalil výpočet. Uvádzame možné zlepšenia pre tento odhad. Pri dolaďovaní vhodného vyhladzovacieho parametra by sa neprehľadával celý interval, na ktorom hľadáme odhad hustoty, ale len malá časť intervalu. Toto platí i pre metódu najbližšieho suseda, test graf, subjektívna voľba jadrového odhadu, naivný odhad a jadrový odhad pre dvojrozmerný náhodný výber. Akurát v poslednom odhade sa interval zmení na kartézsky súčin intervalov. V prípade veľkého rozsahu náhodného výberu, veľkej jemnosti a hlavne použitia Gaussoveho jadra by sa dal jadrový odhad značne zrýchliť tým, že sa vhodný vyhladzovací parameter natvrdo zadá do makra auto_gaussovojadro, a vyradí sa z činnosti hľadanie minimálnej hodnoty M_1. V prípade odhadu pomocou vzájomnej kontroly najmenších kvadrátov sa dá trochu meniť rýchlosť odhadu na úkor presnosti, poprípade nenájdenia vyhladzovacieho parametra. V tejto metóde je použitá Newtonova metóda hľadania minima. Keď nájde vhodný štartovací bod, konverguje k minimu veľmi rýchlo. Táto metóda je naprogramovaná v makre auto_gaussovojadro, presnosť je nastavená na $1E-8$.

Jednoduchými úpravami makier by sa dali naprogramovať odhady pre ďalšie viacrozmerné náhodné výbery (trojrozmerné atď.). Ekvidistantné delenie intervalu definičného oboru odhadu, na ktorom hľadáme odhad, by záviselo na zmene odhadu. Aj keď je jadrový odhad hustoty pravdepodobnosti najrozšírenejší, nie je to univerzálny odhad. Napriek tomu veľmi dobre odhaduje tvar (asymetriu a modalitu) hustoty pravdepodobnosti. Mimo literatúry citovanej v texte sme preštudovali veľa článkov, najmä [4], [7], [9] a [10].

Literatúra

[1] Anděl, J.: Statistické metody. Praha: MATFYZPRESS, 2003.

[2] Anděl, J.: Základy matematické statistiky. Praha: MATFYZPRESS, 2002.

[3] Montgomery, D. C., Renger, G.: Probability and Statistics. New York: John Wiley and Sons, 1996.

[4] Müller, D. W., Sawitzky, G.: Excess mass estimates and tests for multimodality. J. Amer. Statist. Assoc. 86, 738−746.

[5] Neradová, V.: Progresivní metody odhadů neznámých rozdělení pravděpodobnosti. Diplomová práca−vedúci Karpíšek, Z. Ústav matematiky FSI VUT v Brně, Brno 2007.

[6] Scott, D.W.: Multivariate Density Estimation. Theory, Practice and Visualization. New York: Wiley, 1992.

[7] Seather, S. J., Jones, M. C.: A reliable data−based bandwidth selection method for kernel density estimation, Journal of the Royal Statistical Society, Series B, 53, 683−690.

[8] Silverman, B. W.: Density Estimation for Statistics and Data Analysis. London: Chapman and Hall,1985.

[9] Silverman, B. W.: Using kernel density estimates to investigate multimodality. J. Roy. Statist. Soc. Ser. B 43, 97−99.

[10] Simonoff, J. S.: Three Sides of Smoothing: Categorical Data Smoothing, Nonparametric Regression, and Density Estimation. International Statistical Review, Vol. 66, No. 2. (Aug.,1998), pp. 137−156.

[11] Vajda, I.: Theory of Statistical Inference and Information. London: Kluwer Academic Press, 1989.

[12] URL:<http://en.wikipedia.org/wiki/Kernel_(statistics).html > [cit. 2011 − 05 − 11]

[13] URL:<http://fedc.wiwi.hu−berlin.de/xplore/ebooks/html/csa/node145.html> [cit. 2011 − 05 − 11]

[14] URL:<http://nedwww.ipac.caltech.edu/level5/March02/Silverman/Silver_contents.html> [cit. 2011 − 05 − 11]

6. Zoznam použitých skratiek a symbolov

n	rozsah náhodného výberu
X_i	$i-$té pozorovanie náhodnej veličiny X
f	hustota pravdepodobnosti
$N(\mu, \sum)$	viacrozmerné Gaussovo rozdelenie pravdepodobnosti s kovariančnou maticou $\sum$
$\hat{f}$	odhad hustoty pravdepodobnosti
h	vyhladzovací parameter
K	jadro
MSE	stredná kvadratická chyba
MISE	stredná integrálna kvadratická chyba
bias	odchýlka odhadu
var, D	rozptyl
eff	účinnosť (efektívnosť) jadra
$\int$	integrál od $-\infty$ do ∞
$\sum$	suma od 1 do n
E	stredná hodnota
sup	suprémum
$X_{(j)}$	$j-$tá usporiadaná štatistika
∇	operátor nabla