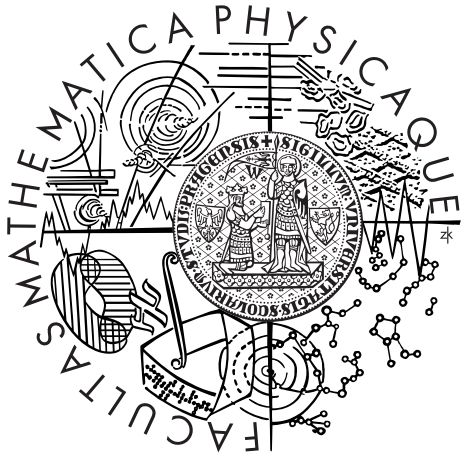
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Libuša Révészová

Jádrové odhady hustoty

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Ing. Marek Omelka, Ph.D.
Studijní program: Matematika
Studijní obor: Obecná matematika

Praha 2013

Rada by som poďakovala svojmu vedúcemu práce, Ing. Marekovi Omelkovi, Ph.D. za pravidelné konzultácie, dobré rady a pripomienky, ústretovosť a čas, ktorý mi venoval a svojej rodine a Tomášovi za podporu.

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Jádrové odhady hustoty

Autor: Libuša Révészová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Ing. Marek Omelka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Existují různé způsoby odhadu hustoty. Skupina metod, které odhadují hustotu jako funkci samotnou se nazývají neparametrické metody odhadu hustoty. Jednou z takových neparametrických metod je jádrový odhad. Tato práce se zabývá uvedením problematiky jádrových odhadů. Jako kritérium správnosti jádrového odhadu se bere střední čtvercová chyba MSE a středná integrovaná čtvercová chyba MISE odhadu hustoty. Na základě požadavku, aby tyto chyby byli co nejmenší se v práci popisují některé metody volby vyhlazovacího parametru. Ty jsou ilustrovány aplikací na data použitím statistického výpočetního prostředí *R*.

Klíčová slova: odhad hustoty, vyhladzovací parameter, chyba odhadu

Title: Kernel density estimation

Author: Libuša Révészová

Department: Department of Probability and Mathematical Statistics

Supervisor: Ing. Marek Omelka, Ph.D., Department of Probability and Mathematical Statistics

Abstract: There are various methods for estimating a density. A group of methods which estimate the density as a function are called nonparametric methods of density estimation. One of such methods is kernel density estimation. This thesis deals with introducing the issue of the kernel density estimation. As an error criteria for kernel density estimation we consider mean squared error MSE and mean integrated squared error MISE. Requiring these errors to be minimal, we describe some methods for choosing the smoothing parameter. These methods are illustrated by their application to data using software *R*.

Keywords: density estimation, smoothing parameter, error of an estimate

Názov práce:
Jadrové odhady hustoty

Autor:
Libuša Révészová

Katedra:
Katedra pravděpodobnosti a matematické statistiky

Vedúci bakalárskej práce:
Ing. Marek Omelka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Existujú rôzne spôsoby odhadu hustoty. Skupina metód, ktoré odhadujú hustotu ako funkciu samotnú sa nazývajú neparametrické metódy odhadu hustoty. Jednou z takýchto neparametrických metód je jadrový odhad. Táto práca sa zaoberá uvedením problematiky jadrových odhadov. Ako kritérium správnosti jadrového odhadu sa berie stredná štvorcová chyba MSE a stredná integrovaná štvorcová chyba MISE odhadu hustoty. Na základe požiadavky, aby tieto chyby boli čo najmenšie sa v práci popisujú niektoré metódy voľby vyhladzovacieho parametra. Tie sú ilustrované aplikáciou na dáta použitím štatistického výpočtového prostredia R.

Kľúčové slová:
odhad hustoty, vyhladzovací parameter, chyba odhadu

Obsah

Kapitola 1

Úvod

„Odhad hustoty zažil za posledných 20 rokov širokú explóziu záujmu. Aplikuje sa vo veľa rôznych oblastiach, ktoré zahŕňajú archeológiu, bankovníctvo, klimatológiu, ekonómiu, genetiku, hydrológiu a fyziológiu.“[3]. V týchto a ďalších oblastiach je potrebné interpretovať dáta získané pozorovaniami a exprimentami, skúmať ich vnútornú štruktúru a podobne. V tejto práci sa budeme venovať neparametrickým metódam odhadovania hustoty. To znamená, že nebudeme o dátach vopred predpokladať, z akého rozdelenia pochádzajú a odhadovať parametre tohto rozdelenia, no budeme sa snažiť odhadnúť hustotu ako funkciu samotnú. V rámci tohto prístupu k odhadu hustôt existujú viaceré metódy. My sa budeme venovať hlavne jadrovým odhadom, ktorých problematika je obsiahla, no skúsime stručne zhrnúť a popísať základné poznatky z odporučenej literatúry.

V kapitole 2 popíšeme základnú a široko používanú metódu konštrukcie histogramov. Je to obľúbená metóda a prináša užitočný náhľad do štruktúry dát. Má však isté nedostatky a ako výrazne vylepšenie tejto metódy si zadefinujeme jadrový odhad hustoty. V tejto kapitole vychádzame z knihy [1], konkrétne z kapitol 1 a 2.

V tretej kapitole popíšeme základné metódy, pomocou ktorých sa overuje správnosť jadrového odhadu. Budeme tu rozlišovať chybu odhadu v jednom bode a tiež chybu odhadu ako odchýlku dvoch funkcií v L_2. Zistíme, že na jadrový odhad má veľký vplyv voľba vyhladzovacieho parametra. Získané výsledky budeme potrebovať aproximovať, aby sme dostali transparentnejšie vyjadrenie chýb v závislosti na vyhladzovacom parametri a vedeli tak lepšie interpretovať jeho úlohu v jadrových odhadoch. Táto kapitola je založená na literatúre [1], kapitola 2 a [2], kapitola 3.

Posledná kapitola 4 uvádza niektoré možnosti voľby vyhladzovacieho parametra, v závisloti na tom, na čo odhad hustoty budeme potrebovať. Predstavujú sa tam dve, v praxi často používané skupiny metód, pomocou ktorých sa volí vyhladzovací parameter. Výsledky sa aplikujú na dáta pomocou štatistického softvéru *R*. Script z programu *R* je možné nájsť v prílohe na CD. Táto časť práce je založená na [2], kapitola 3, [1], kapitola 3 a [3], časť 3.

Kapitola 2

Neparametrické metódy odhadu hustoty

2.1 Histogram

Nech $X_1, X_2, ..., X_n$ sú rovnako rozdelené nezávislé náhodné veličiny, pochádzajúce z rozdelenia s hustotou $f(x)$. Túto hustotu chceme odhadnúť.

Veličiny $X_1, X_2, ..., X_n$ teda tvoria náhodný výber z rozdelenia, o ktorom vopred nebudeme nič predpokladať. Na odhad hustoty v takomto prípade budeme potrebovať neparametrickú metódu. Najbežnejšou neparametrickou metódou odhadu hustoty je konštrukcia histogramu. Najprv si teda odhad hustoty pomocou histogramu definujeme.

Definícia 2.1. Nech $X_1, ..., X_n$ je náhodný výber z rozdelenia s hustotou $f(x)$. Nech $a \in \mathbb{R}$, $h > 0$, $h \in \mathbb{R}$ a nech D je také delenie reálnej osi $\mathbb{R}$ na intervaly, že $D = \bigcup_{j\in\mathbb{Z}} \big(a + jh, a + (j + 1)h\big]$. Zvoľme $x \in \mathbb{R}$, potom $\exists\ k \in \mathbb{Z}$ tak, aby $x \in \big(a + kh, a + (k + 1)h\big]$.
Odhad hustoty f v bode x pomocou histogramu definujeme ako

$$\hat{f}_{HIST}(x; h) = \frac{P}{nh}, \tag{2.1}$$

kde $P = \sum_{i=1}^{n} \mathbb{1}\{X_i \in \big(a + kh, a + (k + 1)h\big]\}$ a n je rozsah výberu.

Táto metóda má široké využitie a je často používaná. Skúsime ju v stručnosti popísať na nasledujúcom príklade.

Pomocou štatistického softvéru R si použitím príkazu *rnorm* vygenerujeme náhodný výber z normovaného normálneho rozdelenia. Jeho rozsah bude n=50.

V tabuľke 2.1 je reprezentovaný náhodný výber $X_1, ..., X_{50}$. Teraz si zvolíme reálne číslo h, ktoré bude predstavovať dĺžku intervalu a reálnu os rozdelíme na intervaly s touto dĺžkou. Parameter h sa zvykne nazývať vyhladzovací parameter. Nech v našom prípade $h = 0.5$ a bod $a = 0$.

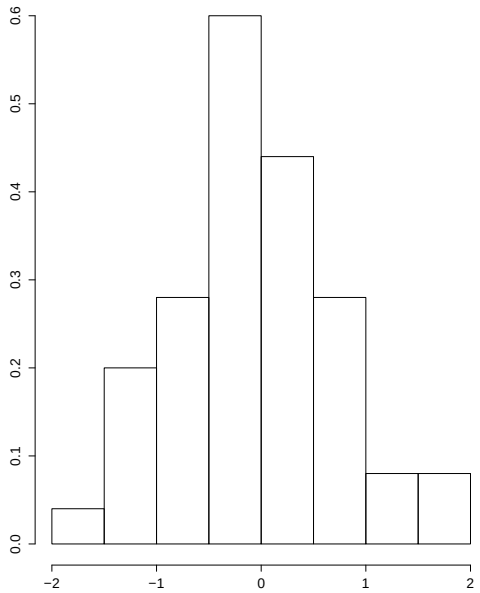
Tabuľka 2.1: Náhodný výber z $N(0,1)$

-1.38507062	0.03832318	-0.76303016	0.21230614	1.42553797
0.74447982	0.70022940	-0.22935461	0.19709386	1.20715377
0.31833673	-1.42379885	-0.40509086	0.99538657	0.95881779
-0.15096960	-1.22306879	-0.86882429	-1.04248536	-1.10363778
0.44418506	-0.20495061	1.67563243	-0.13132225	-0.19988298
0.05491242	-0.68216549	-0.72770415	-0.86190429	-0.03752311
-1.63142324	0.17716660	-0.01250080	-0.39431713	0.35156293
0.87876756	0.20465408	-0.88738071	-0.47721606	-0.26774095
1.58585916	0.04690059	0.35649678	-0.12138001	0.91808790
-0.03609184	-0.98114749	-0.43425983	-0.0674843	0.98189457

Prakticky to znamená, že si zvolíme bod x, čím máme zároveň zvolený interval dĺžky h, do ktorého toto x patrí. Spočítame, koľko pozorovaní je umiestnených v tom istom intervale ako daný bod, čím získame číslo P. Vydelením čisla P rozsahom výberu n a dĺžkou intervalov h dostávame výšku stĺpca histogramu nad bodom x.

Nech napríklad je $x = 0.6$, potom $x \in (0.5, 1]$. Z vyššie uvedeného výberu v tabuľke 2.1 spočítame, koľko hodnôt patrí do tohoto intervalu, teda $P = 7$. Takže máme $\hat{f}_{HIST}(0.6; 0.5) = \frac{7}{50 \cdot 0.5} = 0.28$. Týmto spôsobom skonštruujeme celý histogram, znázornený na obrázku 2.1:

Obrázok 2.1: Histogram pre náhodný výber z N(0,1) s parametrom h=0.5

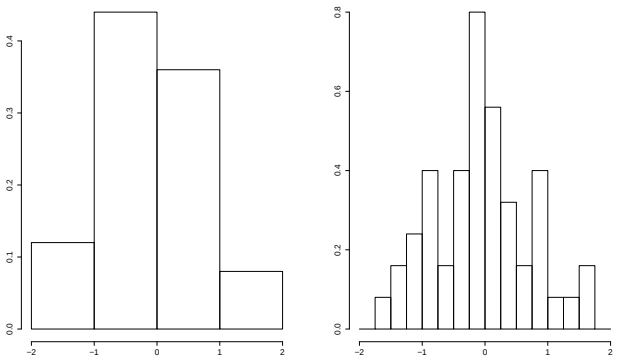


Parameter h sme v predchádzajúcom prípade zvolili ľubovoľne. Z definície (2.1) je však vidieť, že pri meniacej sa dĺžke intervalu sa bude meniť aj tvar histogramu. Keď interval zmenšíme, t.j. máme nové $h^* < h$, tak pre číslo P^*,

označujúce nový počet pozorovaní, ktoré padnú do intervalu spolu s x bude platiť $P^* \leq P$. Čiže to, aký vzťah bude medzi $\hat{f}_{HIST}(x,h)$ a $\hat{f}_{HIST}(x,h^*)$, t.j. výškou stĺpca nad bodom x, záleží na rozmiestnení pozorovaní vzhľadom k meniacej sa dĺžke intervalu h . Výška stĺpca nad bodom x sa nezmení len v tom prípade, keď sa pri zmene parametra h zachová pomer počtu pozorovaní v danom intervale k jeho dĺžke. Inak sa bude meniť aj výška stĺpcov nad jednotlivými bodmi aj ich šírka zmenšovaním h a výsledný histogram bude viac „kostrbatý“. Pri zväčšení h sa intervaly rozšíria a histogram bude vyzerať „uhladenejšie“.

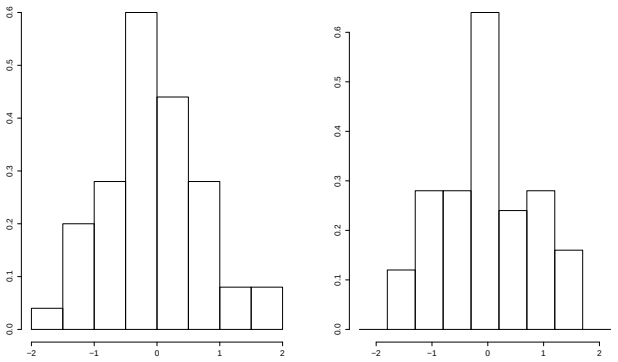
Na nasledujúcom obrázku 2.2 je vidieť, ako sa pôvodný histogram zmení zväčšením h na $h = 1$ a jeho zmenšením na $h = 0.25$, pre ten istý náhodný výber.

Obrázok 2.2: Vľavo je histogram pre $h = 1$, vpravo $h = 0.25$.



Histogram teda závisí na tom, akú dĺžku intervalov zvolíme. Pri konštrukcii histogramov sa ale navyše musíme rozhodnúť, do ktorého bodu chceme delenie reálnej osi umiestniť, t.j. ako zvoliť bod a. Aj pri zachovaní jednej dĺžky intervalu h sa tvar histogramu v závislosti na jeho umiestnení mení. Na obrázku 2.3 vidíme, ako sa pôvodný histogram pre $h = 0.5$ mení s meniacim sa bodom umiestnenia delenia, čo je spôsobené tým, že sa mení počet P pozorovaní patriaiacich do intervalu spolu s bodom x.

Obrázok 2.3: Na histograme vľavo je $a = 0$, vpravo $a = 0.2$.



Existujú však iné neparametrické metódy, ktoré problém s voľbou umiestnenia intervalu nezdieľajú. Naviac, oproti histogramu už disponujú vlastnosťami

ako napríklad spojitosť, hladkosť a podobne. Jednou z takých metód je metóda jadrových odhadov hustoty.

2.2 Definícia jadrových odhadov

Definícia 2.2. Nech $X_1, ..., X_n$ je náhodný výber z jednorozmerného spojitého rozdelenia s neznámou hustotu f. Potom *jadrový odhad* hustoty f definujeme ako

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big),$$

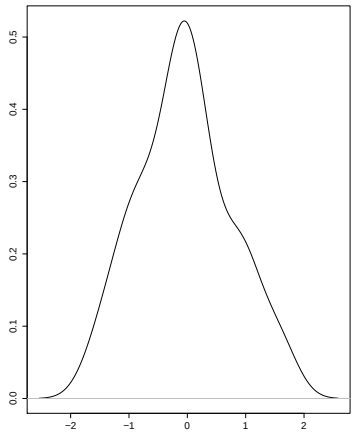
kde n je rozsah výberu, K je spojitá funkcia spĺňajúca $\int K(x)dx = 1$ nazývaná *jadro* a h je parameter, ktorý sa nazýva *vyhladzovacie okno*.
Po zavedení $K_h(u) = h^{-1}K(u/h)$ upravíme výraz pre tento odhad do tvaru

$$\hat{f}_h(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i)}{n}. \tag{2.2}$$

Podľa definície je teda $\hat{f}_h(x)$ náhodnou veličinou - závisí na náhodnom výbere $X_1, ..., X_n$.

Pre náhodný výber z tabuľky 2.1 teraz pomocou príkazu *density* v programe R vytvoríme jadrový odhad, obrázok 2.5. Použitá jadrová funkcia je hustota normovaného normálneho rozdelenia. Vyhladzovací parameter je v tomto prípade vypočítaný automaticky, metódou, ktorá je v R prednastavená pre príkaz *density*. Je to metóda nazývaná „Silverman's rule of thumb", ktorá bude bližšie popísaná v kaptiole 4. V tomto prípade vyšlo $h = 0.3029$.

Obrázok 2.4: Jadrový odhad hustoty pre náhodný výber z N(0,1) s h=0.3029 a $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$



Za jadrovú funkciu $K(x)$ sa najčastejšie volí nejaká hustota, napríklad hustota normálneho rozdelenia. Voľba jadra má určitý vplyv na výsledný tvar odhadu, no pri zväčšujúcom sa rozsahu výberu sa tento vplyv výrazne znižuje. Ilustrovať to môžme na nasledujúcich obrázkoch. Opäť si vygenerujeme náhodné výbery z normovaného normálneho rozdelenia, jeden s malým rozsahom $n = 5$, druhý s rozsahom $n = 100$. Pre oba výbery vytvoríme jadrový odhad, jeden s jadrom

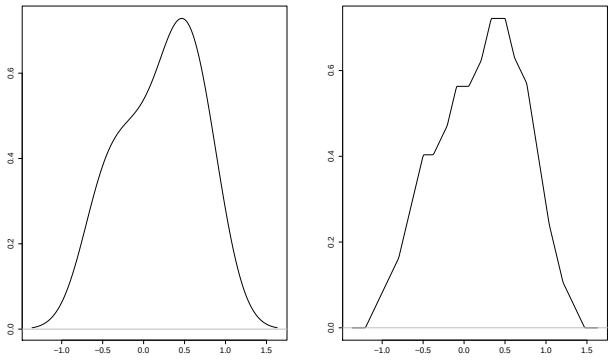
$$K(u) = (1 - |u|)\mathbb{1}(|u| \leq 1)$$

(v R pod názvom „triangular“) a druhý s jadrom

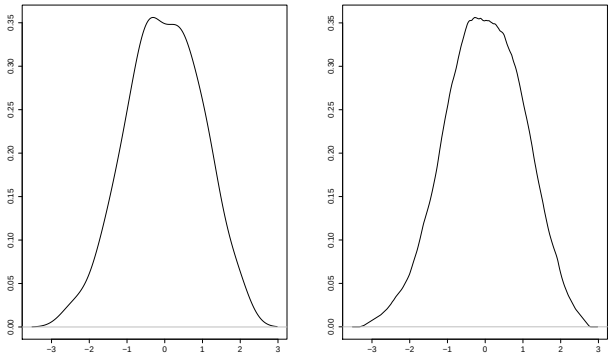
$$K(v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}}.$$

Na obrázku 2.5 vidíme, že pre malý rozsah výberu sa tvar odhadu líši pre rôzne jadrové funkcie. Avšak, na ďalšom obrázku 2.6 už rozdiel takmer nie je vidieť. V [1], kapitole 2.7 je ukázané, že skutočne voľba jadrovej funkcie na odhad hustoty nemá až taký významný vplyv, na rozdiel od výberu vyhladzovacieho parametra, ktorý má v jadrových odhadoch kľúčovú úlohu. Preto sa ďalej budeme zaoberať vplyvom voľby vyhladzovacieho parametra na jadrové odhady.

Obrázok 2.5: Jadrové odhady pre náhodný výber z $N(0, 1)$ s rozsahom $n = 5$. Vľavo použité jadro „gaussian“ $K(v)$, vpravo jadro „triangular“ $K(u)$.



Obrázok 2.6: Jadrové odhady pre náhodný výber z $N(0, 1)$ s rozsahom $n = 100$. Vľavo použité jadro „gaussian“ $K(v)$, vpravo jadro „triangular“ $K(u)$.

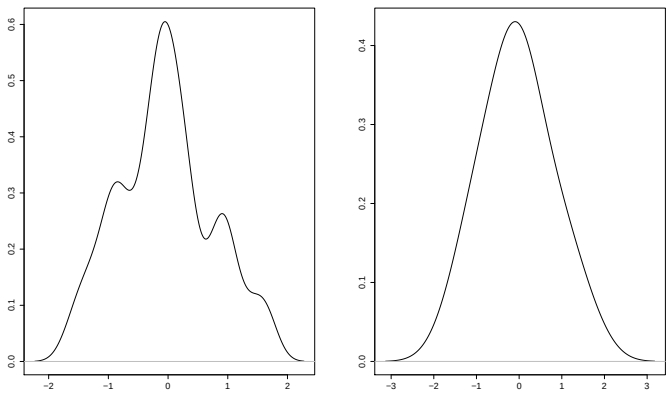


Kapitola 3

Určovanie chyby odhadu

Podobne ako sa pri histogramoch menením vyhladzovacieho parametra menil jeho tvar, aj pri jadrových odhadoch bude tvar odhadu hustoty závisieť od vyhladzovacieho parametra h. Preto bude dôležité vedieť si zvoliť parameter h správne v tom zmysle, aby odhad hustoty splnil naše kritéria na jeho presnosť. Na nasledujúcich obrázkoch je vidieť, ako veľmi na voľbe vyhladzovacieho parametra pri jadrových odhadoch záleží. Keď ho zmenšíme, vidíme že výsledný odhad viac „odráža“ skutočné rozloženie dát, v literatúre sa takýto odhad nazýva „undersmoothed“ alebo „podhladený“, naopak zas pri väčšom vyhladzovacom okne je odhad „oversmoothed“, t.j. „nadhladený“

Obrázok 3.1: Jadrové odhady výber uvedený v tabuľke 2.1 pri $h = 0.2$ (vľavo) a pri $h = 0.5$ (vpravo)



Majme jadrový odhad $\hat{f}_h$ pre náhodný výber $X_1, ..., X_n$. Na to, aby sme zistili či tento odhad spĺňa naše očakávania na jeho presnosť respektíve správnosť si musíme špecifikovať, ako si jeho presnosť resp. správnosť budeme predstavovať. Záleží na tom, akým spôsobom budeme chcieť tento odhad využiť, či nám pôjde o odhad samotný alebo ho potrebujeme ako medzikrok k ďalšej práci a podobne.

V našom prípade sa budeme snažiť o to, aby sa minimalizovala vzdialenosť odhadovanej funkcie hustoty od teoretickej. Odchýlku funkcie odhadu $\hat{f}_h$ od skutočnej funkcie f môžme určovať buď v danom bode $x \in \mathbb{R}$ alebo ako L_2 vzdialenosť týchto dvoch funkcií.

3.1 Stredná štvorcová chyba MSE

Existuje viacero spôsobov, ako určiť odchylku dvoch funkcií v nejakom konkrétnom bode. Majme daný bod $x \in \mathbb{R}$. Pre nás najvýhodnejší spôsob je takzvaná stredná štvorcová chyba MSE (mean squared error). Jej výhodnosť spočíva napríklad v tom, že sa dá rozložiť na súčet rozptylu a štvorca vychýlenia.

Najprv si odvodíme, ako sa MSE rozkladá. Nech $\hat{\theta}$ je odhad nejakého parametra θ, (v našom prípade bude θ predstavovať $f(x)$ a $\hat{\theta}$ bude $\hat{f}_h(x)$) potom si definujeme strednú štvorcovú chybu MSE odhadu $\hat{\theta}$ ako

$$MSE(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta)^2.$$

Tento vzťah postupne upravíme využitím vlastností strednej hodnoty:

$$\mathrm{E}(\hat{\theta} - \theta)^2 = \mathrm{E}(\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2) = \mathrm{E}\hat{\theta}^2 - \mathrm{E}(2\hat{\theta}\theta) + \mathrm{E}\theta^2 =$$
$$\mathrm{E}\hat{\theta}^2 - (\mathrm{E}\hat{\theta})^2 + (\mathrm{E}\hat{\theta})^2 - 2\theta\mathrm{E}\hat{\theta} + \theta^2 = \mathrm{var}(\hat{\theta}) + (\mathrm{E}\hat{\theta} - \theta)^2 = \mathrm{var}(\hat{\theta}) + \mathrm{b}^2(\hat{\theta}),$$

kde $\mathrm{var}(\hat{\theta})$ je rozptyl a $\mathrm{b}(\hat{\theta})$ je vychýlenie. Na určenie strednej štvorcovej chyby pre jadrový odhad $\hat{f}_h$ v bode x nám teda stačí určiť rozptyl $\mathrm{var}(\hat{f}_h(x))$ a štvorec vychýlenia odhadu $\mathrm{b}^2(\hat{f}_h(x)) = (\mathrm{E}\hat{f}_h(x) - f(x))^2$. Nato si ďalej potrebujeme definovať konvolúciu dvoch funkcií:

Definícia 3.1. Nech f a g sú jednorozmerné spojité funkcie reálnej premennej, potom definujeme *konvolúciu* týchto funkcií ako

$$(f * g)(x) = \int f(x-y)g(y)dy \tag{3.1}$$

Postupne si vyjadríme výrazy, ktoré sa nám zídu na vyjadrenie rozptylu a vychýlenia, pri čom využijeme nezávislosť náhodných veličín $X_1, ..., X_n$.

Najprv teda pre strednú hodnotu odhadu máme:

$$\begin{aligned}\mathrm{E}(\hat{f}_h(x)) &= \mathrm{E}(\tfrac{\sum_{i=1}^n K_h(x-X_i)}{n}) = \tfrac{1}{n}\sum_{i=1}^n \mathrm{E}(K_h(x-X_i)) = \mathrm{E}(K_h(x-X_1)) = \\ &= \textstyle\int K_h(x-y)f(y)dy,\end{aligned} \tag{3.2}$$

čo použitím vzťahu (3.1) upravíme na $\mathrm{E}(\hat{f}_h(x)) = (K_h * f)(x)$. Pre vychýlenie $\mathrm{b}(\hat{f}_h(x))$ teda máme

$$\mathrm{b}(\hat{f}_h(x)) = \mathrm{E}(\hat{f}_h(x)) - f(x) = (K_h * f)(x) - f(x). \tag{3.3}$$

Ďalej si podobným spôsobom upravíme vyjadrenie rozptylu:

$$\mathrm{var}(\hat{f}_h(x)) = \mathrm{var}\Big(\frac{1}{n}\sum_{i=1}^n K_h(x-X_i)\Big) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{var}(K_h(x-X_i)) =$$

$$= \frac{1}{n}\mathrm{var}(K_h(x-X_1)) = \frac{1}{n}\big[\mathrm{E}(K_h(x-X_1))^2 - (\mathrm{E}(K_h(x-X_1)))^2\big].$$

Potrebujeme teda ešte vyjadriť $\mathrm{E}[K_h(x-X_1)]^2$, kde dostávame:

$$\mathrm{E}[K_h(x-X_1)]^2 = \int (K_h(x-y))^2 f(y)dy = (K_h^2 * f)(x).$$

Teda celkovo pre rozptyl odhadu máme:

$$var(\hat{f}_h(x)) = \frac{1}{n}\Big[(K_h^2 * f)(x) - ((K_h * f)(x))^2\Big]. \tag{3.4}$$

Strednú štvorcovú chybu odhadu $\hat{f}_h$ v bode $x \in R$ máme vyjadrenú takto:

$$\begin{aligned} MSE(\hat{f}_h(x)) = &\ \tfrac{1}{n} var(\hat{f}_h(x)) + (E\hat{f}_h(x) - f(x))^2 = \\ = &\ \tfrac{1}{n}((K_h^2 * f)(x) - ((K_h * f)(x))^2) + ((K_h * f)(x) - f(x))^2. \end{aligned} \tag{3.5}$$

Poznámka: Exsituje viacero možných spôsobov, ako určiť odchýlku funkcií v danom bode. Ďalším z nich je napríklad výpočet strednej absolútnej chyby MAE (mean absolute error), danej

$$MAE(\hat{f}_h(x)) = E|\hat{f}_h(x) - f(x)|,$$

kde $\hat{f}_h(x)$ je odhad pre $f(x)$. Výpočet MAE je ale často komplikovaný a vďaka menšej zložitosti a svojmu rozkladu na rozptyl a vychýlenie sa teda väčšinou uprednostňuje uvedená MSE.

3.2 Stredná integrovaná štvorcová chyba MISE

Pri jadrových odhadoch ale spravidla potrebujeme určiť odchýlku odhadovanej funkcie a odhadu na celom $\mathbb{R}$. Jedným zo spôsobov, ako túto odchýlku určiť je stredná hodnota štvorca L_2 vzdialeností medzi týmito funkciami, označovaná MISE (mean integrated squared error)

$$\begin{aligned} MISE(\hat{f}_h(.)) = \mathrm{E}[ISE(\hat{f}_h(.))] &= \mathrm{E}\int(\hat{f}_h(x) - f(x))^2 dx = \\ = \int \mathrm{E}(\hat{f}_h(x) - f(x))^2 dx &= \int MSE(\hat{f}_h(x)). \end{aligned}$$

Teda podľa (3.5) máme:

$$MISE(\hat{f}_h(.)) = \int \frac{1}{n}\Big\{\big[(K_h^2 * f)(x) - ((K_h * f)(x))^2\big] + \big[(K_h * f)(x) - f(x)\big]^2\Big\} dx \tag{3.6}$$

Úpravami sa to dá o trochu zjednodušiť, no toto zjednodušenie neprinesie významné vylepšenie v tom zmysle, že výsledný tvar bude stále príliš komplikovaný na to, aby bolo vidieť ako jadrové odhady závisia od voľby vyhladzovacieho okna. V nasledujúcej časti sa pokúsime MSE a MISE aproximovať. Tým získame jednoduchšie vyjadrenia pre tieto chyby, ktoré nám uľahčia interpretáciu vplyvu vyhladzovacieho okna na jadrový odhad.

3.3 Asymptotické chyby MSE a MISE

V tejto časti aproximujeme chyby MSE a MISE. Najprv si uvedieme Taylorovu vetu v tvare, v ktorom ju budeme neskôr používať, takýto tvar Taylorovej vety môžme nájsť napríklad v [1] kapitola 2, strana 19.

Veta 3.1 (Taylorova veta). *Nech f je reálna funkcia definovaná na $\mathbb{R}$ a nech $x \in \mathbb{R}$. Nech f má na intervale $(x-\delta, x+\delta)$ spojité derivácie p-tého rádu, $p \in \mathbb{N}$, $\delta > 0$. Potom pre postupnosť reálnych čísel α_n, konvergujúcu k nule platí:*

$$f(x+\alpha_n) = \sum_{j=0}^{p} \frac{\alpha_n^j}{j!} f^{(j)}(x) + o(\alpha_n^p) \tag{3.7}$$

Dôkaz. Dôkaz je uvedený v knihe [4], veta 153 a poznámka 3, pričom uvažujeme, že $y(x)$ je $o(x)$, $y = o(x)$ ak platí $\frac{|y|}{|x|} \to 0$ pre $x \to 0$.

❑

Vyššie odvodené chyby jadrového odhadu MSE (3.5) a MISE (3.6) závisia na vyhladzovacom parametri veľmi komplikovaným spôsobom. Preto sa budeme snažiť tento problém odstrániť použitím asymptotických chýb namiesto ich presných vyjadrení pomocou MSE a MISE. Na túto úlohu však budeme potrebovať dodatočné predpoklady:

P.1 hustota f má spojitú druhú deriváciu, ktorá je štvorcovo integrovateľná a existuje $M > 0$ tak, že f je monotónna na oboch intervaloch $(-\infty, -M)$ a (M, ∞);

P.2 vyhladzovací parameter h bude funkciou n, potom $h = h_n$ je nenáhodná postupnosť, $h_n > 0$ pre $\forall n \in \mathbb{N}$ a platí

$$\lim_{n\to\infty} h_n = 0, \quad \lim_{n\to\infty} h_n n = \infty;$$

P.3 jadro K je hustota, ktorá je obmedzená, má konečný štvrtý moment a je symetrická okolo počiatku (párna).

Zo vzťahu (3.2)

$$E(\hat{f}_h(x)) = \int K_h(x-y) f(y) dy = \int K(z) f(x-hz) dz, \tag{3.8}$$

kde sme použili substitúciu $z = \frac{x-y}{h}$, $dz = \frac{1}{h}$ a vzťah $K_h(u) = \frac{1}{h} K(\frac{u}{h})$.

Z predpokladu P.1 máme, že funkcia $f(x-hz)$ má spojitú druhú deriváciu. Ďalej z P.2 predpokladu vyplýva, že pre každé $z \in \mathbb{R}$ je

$$\lim_{n\to\infty} zh = z \lim_{n\to\infty} h = 0.$$

Máme teda splnené predpoklady Taylorovej vety (3.1) a funkciu $f(x-hz)$ môžme rozvinúť do Taylorovho radu:

$$f(x-zh) = f(x) + f'(x)(-hz) + \frac{1}{2} f''(x)(hz)^2 + o(h^2). \tag{3.9}$$

Takže po dosadení do výrazu pre strednú hodnotu máme:

$$\mathrm{E}(\hat{f}_h(x)) = \int K(z)[f(x) - hzf'(x) + \frac{1}{2} f''(x)(hz)^2 + o(h^2)]dz. \tag{3.10}$$

Ďalej teda

$$\mathrm{E}(\hat{f}_h(x)) = \int K(z)f(x)dz - \int K(z)hzf'(x)dz + \int K(z)\frac{1}{2}f''(x)(hz)^2dz + o(h^2)$$

$$= f(x)\int K(z)dz - hf'(x)\int zK(z)dz + \frac{1}{2}h^2f''(x)\int z^2K(z)dz.$$

Pretože $K(z)$ je hustota, platí $\int K(z)dz = 1$. Ďalej z predpokladu P.3 vieme, že K má konečný štvrtý moment, teda platí $\int zK(z)dz < \infty$ a $\int z^2K(z)dz < \infty$. Naviac, K má byť párna, obmedzená, teda potom $\int zK(z)dz = 0$, pretože pre funkciu $g(z) = zK(z)$ platí: $g(-z) = (-z)K(-z) = -zK(z)$, lebo $K(z)$ je párna, t.j. $g(-z) = -g(z)$, z čoho máme, že g je nepárna.

Celkovo, pre asymptotický odhad strednej hodnoty jadrového odhadu $\hat{f}_h$ máme:

$$\mathrm{E}(\hat{f}_h(x)) = f(x) + \frac{1}{2}h^2f''(x)\int z^2K(z)dz + o(h^2). \tag{3.11}$$

Vychýlenie teda vyzerá takto:

$$\mathrm{b}(\hat{f}_h(x)) = \mathrm{E}(\hat{f}_h(x) - f(x)) = \frac{1}{2}h^2f''(x)\int z^2K(z)dz + o(h^2). \tag{3.12}$$

Pre rozptyl máme z (3.4):

$$\mathrm{var}(\hat{f}_h(x)) = \frac{1}{n}(K_h^2 * f)(x) - ((K_h * f)(x))^2 =$$

$$= \frac{1}{n}\Big[\int K_h^2(x-y)f(y)dy - (\int K_h(x-y)f(y)dy)^2\Big].$$

Opäť použijeme substitúciu ako v (3.8) a dostaneme:

$$\mathrm{var}(\hat{f}_h(x)) = \frac{1}{nh}\int K(z)^2f(x-hz)dz - \frac{1}{n}(\mathrm{E}(\hat{f}_h(x))^2 =$$

$$= \frac{1}{nh}\int K(z)^2[f(x) - hzf'(x) + o(h^2)]dz - \frac{1}{n}[f(x) + o(h^2)]^2 =$$

$$= \frac{1}{nh}\int K(z)^2(f(x) + o(1))dz - \frac{1}{n}(f(x) + o(1))^2$$

$$= \frac{1}{nh}f(x)\int K(z)^2dz + o\{(nh)^{-1}\}. \tag{3.13}$$

Zo vzťahov (3.12) a (3.13) konečne dostávame nasledujúce vyjadrenia pre rozptyl a vychýlenie:

$$\mathrm{b}(\hat{f}_h(x)) = \tfrac{1}{2}h^2f''(x)\textstyle\int z^2K(z)dz + o(h^2)$$
$$\mathrm{var}(\hat{f}_h(x)) = \tfrac{1}{nh}f(x)\textstyle\int K(z)^2dz + o((nh)^{-1}).$$

Predpokladajme, že sa snažíme zvoliť parameter h tak, aby sme minimalizovali strednú integrovanú chybu odhadu MISE, ktorú tvorí integrál zo súčtu štvorca vychýlenia $[\mathrm{b}(\hat{f}_h)]^2$ a rozptylu $\mathrm{var}(\hat{f}_h)$. Z ich vyjadrení ale vidíme, že rozptyl sa zmenšuje pri zväčšovaní parametra h, kdežto vychýlenie sa zmenšuje ak sa zmenšuje aj h. To spôsobuje, že, ako sme videli na obrázku (3), keď sme h zmenšili, zväčšil sa rozptyl oproti pôvodnému stavu no vychýlenie sa zmenšilo, naopak na druhom obrázku sme zvýšili hodnotu h, krivku sme „vyhladili“ - rozptyl sa zmenšil ale vychýlenie sa zvýšilo. Znamená to teda, že voľba parametra h predstavuje kompromis medzi náhodnou a systematickou chybou.

Pre MSE teda máme vzťah:

$$MSE(\hat{f}_h(x)) = \frac{1}{nh} f(x) \int K(z)^2 dz + o\{(nh)^{-1}\} + \Big[\frac{1}{2} h^2 f''(x) \int z^2 K(z) dz + o(h^2)\Big]^2$$

Keďže $MISE(\hat{f}_h(.)) = \int MSE(\hat{f}_h(.))$ a v predpoklade P.1 požadujeme, že f je hustota (t.j. $\int f(x)dx = 1$), dostávame:

$$MISE(\hat{f}_h(.)) = \frac{1}{nh} \int K(z)^2 dz + \frac{1}{4} h^4 \big[\int z^2 K(z) dz \big]^2 \Big(\int f''(x) dx \Big)^2 + o\{(nh)^{-1} + h^4\}$$

Za asymptotickú aproximáciu chyby MISE teda budeme považovať výraz

$$AMISE(\hat{f}_h(.)) = \frac{1}{nh} \int K(z)^2 dz + \frac{1}{4} h^4 \big[\int z^2 K(z) dz \big]^2 \Big(\int f''(x) dx \Big)^2 \quad (3.14)$$

AMISE závisí od parametra h jednoduchším spôsobom ako MISE v (3.6). Pretože chceme zistiť ako zvoliť h čo najlepšie, to znamená tak, aby sme minimalizovali AMISE, výraz pre AMISE zderivujeme podľa h a položíme ho rovno nule. Pre prehľadnosť si najprv označím členy, ktoré nezávisia na h postupne

$$A = \frac{1}{n} \int K(z)^2 dz, B = \big[\int z^2 K(z) dz \big]^2, C = [\int f''(x) dx]^2. \quad (3.15)$$

Potom $[AMISE(h)]' = -\frac{1}{h^2} A + h^3 BC = 0$. Odtiaľ $h_{min} = \sqrt[5]{\frac{A}{BC}}$.

Ešte sa presvedčíme, že h_{min} je skutočne minimum. Vieme, že členy A, B, C sú všetky kladné. Ďalej predpokladáme, že $h > 0$. Potom druhá derivácia $AMISE$ je $[AMISE(h)]'' = \frac{2}{h^3} A + 3h^2 BC > 0$ pre $\forall h > 0$, teda druhá derivácia je kladná pre všetky h, daná funkcia je vzhľadom k h konvexná a teda v bode h_{min} je skutočne minimum.

Teda pri zvolení $h = h_{min}$ nám vyjde minimálna odchýlka AMISE. Z vyjadrenia h_{min} vidíme, že h_{min} závisí na C, teda na neznámej hustote $f(x)$. V nasledujúcej kapitole budeme hľadať spôsob, ako túto závislosť na neznámej hustote obísť a zvoliť vyhladzovací parameter čo najlepšie.

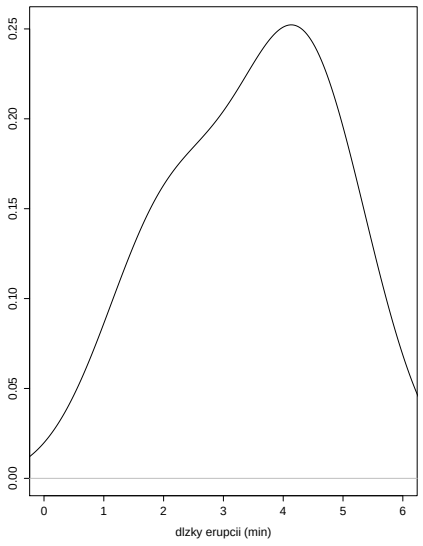
Kapitola 4

Metódy výberu vyhladzovacieho parametra

Voľba správneho vyhladzovacieho okna v jadrových odhadoch hustoty je nesmierne dôležitá. Nesprávne zvolené h môže viesť k mylnej interpretácií javu, k zlému záveru a podobne. V predchádzajúcej kapitole 3 sme ale videli, že vyjadrenie optimálnej hodnoty h vzhľadom k minimálnej AMISE závisí na neznámej hustote $f(x)$. Existujú viaceré prístupy, ktoré tento problém riešia. Všeobecne by sa dali roztriedť do dvoch skupín, v článku [3] ich nazývajú „rules of thumb“ a „cross-validation“ metódy. Obe tieto skupiny metód majú svoje uplatnenie v praxi, ktoré záleží od povahy riešeného problému.

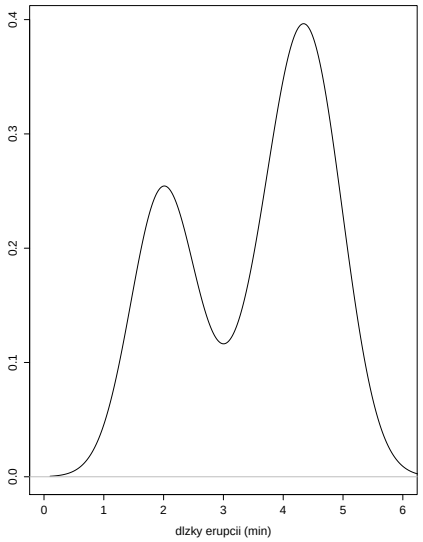
Význam správnej voľby vyhladzovacieho okna môžeme ilustrovať na nasledujúcich obrázkoch. Použili dáta $faithful$, implementované v programe R. Sú to dĺžky erupcií gejzíru $Old\ Faithful$, ktorý sa nachádza v Yellowstoneskom národnom parku v USA. Dĺžky erupcií sú v rozpätí 1 až 5 minút. Bude nás zaujímať odhad hustoty trvania erupcií. Na obrázku 4.1 sme zvolili vyhladzovací parameter „veľký“, $h = 1$.

Obrázok 4.1: Jadrový odhad pre dáta $faithful$ s vyhladzovacím parametrom $h = 1$. Ako jadro bola použitá hustota $N(0, 1)$ rozdelenia.



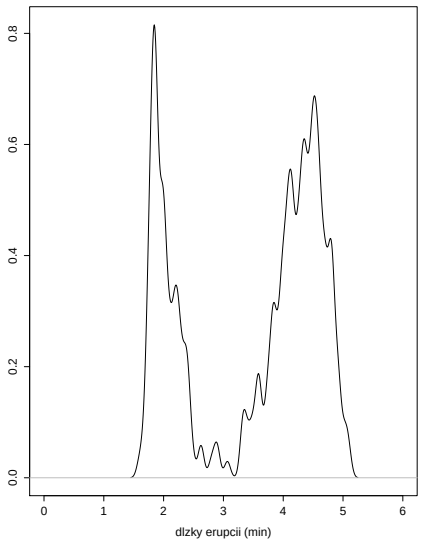
Na obrázku (4.2) sme ho zmenšili na polovicu, t.j. $h = 0.5$. Vidíme, že oproti predošlému obrázku tento naznačuje v použitých dátach bimodálne rozdelenie.

Obrázok 4.2: Jadrový odhad pre dáta $faithful$ s vyhladzovacím parametrom $h = 0.5$. Ako jadro bola použitá hustota $N(0, 1)$ rozdelenia.



Keď budeme parameter h ďalej zmenšovať, odhad hustoty sa bude viac „rozčleňovať“. Na treťom obrázku 4.3 je vyhladzovací parameter $h = 0.05$.

Obrázok 4.3: Jadrový odhad pre dáta $faithful$ s vyhladzovacím parametrom $h = 0.05$. Ako jadro bola použitá hustota $N(0, 1)$ rozdelenia.



Vidíme, že jadrové odhady pre tieto dáta sa líšia v závislosti na vyhladzovacom parametri. Ktorý teda predstavuje ten „najsprávnejší“ z nich?

4.1 „Rules of thumb“

Jedna skupina metód, ktoré sa zaoberajú výberom vyhladzovacieho parametra pre jadrové odhady sú „rules of thumb“, niekedy označované ako „quick and simple“. „Táto prvá trieda obsahuje relatívne jednoduché, ľahko vypočítateľné vzorce, ktoré sa zameriavajú na nájdenie vyhladzovacieho parametra, ktorý je „rozumný“ pre široké spektrum situácií, ale bez žiadnych matematických garancií, že bude blízko optimálnej hodnote.“[1], kapitola 3, strana 59. Tieto metódy spočívajú v tom, že neznáma časť C (výraz (3.15)) vo vyjadrení $AMISE$ sa nahradí jej parametrickým odhad, ktorý sa urobí na základe dát. Sú užitočné, ak potrebujeme nejakú počiatočnú hodnotu pre vyhladzovací parameter, ktorú potom budeme prispôsobovať podľa potrieb.

Potrebujeme priradiť hodnotu výrazu C vo vyjadrení h_{min}, pretože ten je závislý od neznámej hustoty $f(x)$. Jedným z možných spôsobov ako to urobiť je predstaviť si, že funkcia $f(x)$ je hustotou z rozdelenia $N(0, \sigma^2)$, teda

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}.$$

Pre hustotu normálneho rozdelenia $N(\mu, \sigma^2)$ platí $f(x) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$, kde $\phi(t)$ je hustota normovaného normálneho rozdelenia $N(0, 1)$. Potom teda

$$C = \int f''(x)^2 dx = \int \Big[\frac{1}{\sigma^3}\phi''\big(\frac{x}{\sigma}\big)dx\Big]^2 = \frac{1}{\sigma^5}\int \phi''(t)^2 dt = \frac{1}{\sigma^5}\frac{3}{8}\frac{1}{\sqrt{\pi}},$$

kde

$$\phi''(t) = \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\right]'' = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}(t^2 - 1).$$

Pre h_{min} máme:

$$h_{min} = \sqrt[5]{\frac{A}{BC}} = \sqrt[5]{\frac{\frac{1}{n}\int K(z)^2 dz}{\big[\int z^2 K(z)dz\big]^2 \Big[\int f''(x)dx\Big]^2}}.$$

Ako jadro zvolíme hustotu normovaného normálneho rozdelenia, teda

$$K(z) = \tfrac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

Pre takto zvolené jadro vypočítame A a B, teda máme:

$$nA = \int K(z)^2 dz = \int \big[\tfrac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\big]^2 dz = \int \frac{1}{2\pi} e^{-z^2} = \frac{1}{2\sqrt{\pi}} \int \frac{1}{\sqrt{2\pi\frac{1}{2}}} e^{-\frac{z^2}{2\frac{1}{2}}} dz$$

a $\frac{1}{\sqrt{2\pi\frac{1}{2}}} e^{-\frac{z^2}{2\frac{1}{2}}}$ je hustota rozdelenia $N(0, \frac{1}{2})$, teda celkovo $A = \frac{1}{n}\frac{1}{2\sqrt{\pi}}$.
Ďalej, pre B máme:

$$B = \Big[\int z^2 K(z)dz\Big]^2 = \Big[\frac{1}{\sqrt{2\pi}}\int z^2 e^{-\frac{z^2}{2}} dz\Big]^2 = \Big[\frac{2\sqrt{2}}{\sqrt{2\pi}}\int_0^\infty \sqrt{t} e^{-t} dt\Big]^2 = \Big[\frac{2}{\sqrt{\pi}}\Gamma\big[\frac{3}{2}\big]\Big]^2$$

Celkovo je teda $B = 1$.

Takže máme:

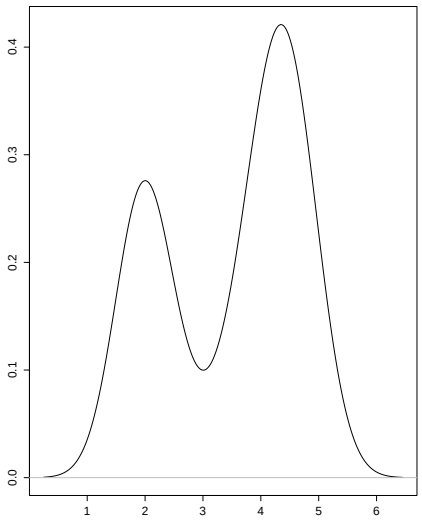
$$A = \frac{1}{n}\frac{1}{2\sqrt{\pi}},\ B = 1 \text{ a } C = \frac{1}{\sigma^5}\frac{3}{8}\frac{1}{\sqrt{\pi}}$$

a pre h_{min} dostávame:

$$h_{min} \doteq \sigma n^{-1/5} 1.06$$

Čiže pre voľbu vyhladzovacieho parametra potrebujeme odhadnúť parameter σ z dát, používa sa výberový rozptyl S_n^2. Spočítame si teda optimálnu hodnotu vyhladzovacieho parametra pre túto metódu. Výberový rozptyl získame pomocou softvéru$R$ a následne spočítame hodnotu $h_{min} \doteq 0.45$. Na obrázku 4.4 skonštruujeme jadrový odhad pre dáta $faithful$ použitím tohto získaného vyhladzovacieho parametra.

Obrázok 4.4: Jadrový odhad s vyhladzovacím parametrom $h \doteq 0.45$, za jadro bola zvolená hustota $N(0, 1)$.



Existujú rôzne modifikácie, ako pomocou tejto metódy získať optimálnu hodnotu vyhladzovacieho okna. Jednou z nich je napríklad namiesto odhadu rozptylu vo vyjadrení h_{min} použiť medzikvartilové rozpätie (IQR). Vzorec pre h_{min} potom vyzerá $h_{min} = 0.79(IQR)n^{-\frac{1}{5}}$. Toto použitie medzikvartilového rozpätia namiesto rozptylu je výhodnejšie v tom, že nezohľadňuje odľahlé pozorovania. Použitie takto získanej hodnoty vyhladzovacieho parametra môže teda priniesť lepšie výsledky pre zošikmené dáta a dáta s „dlhými chvostami“. Avšak, podľa knihy [2], kapitola 3.4.2 to citlivosť na bimodalitu v pozorovaniach nezlepší.

4.2 „Krosvalidačné“ metódy

„Cross-validation“ niekedy prekladaná ako „krížová validácia“, alebo „krosvalidácia“ odkazuje na metódu štatistickej analýzy, ktorá využíva časť napozorovaných dát k získaniu informácií o zvyšnej časti. V našom prípade to označuje ďalší spôsob, ako určiť optimálnu hodnotu vyhladzovacieho parametra. Konkrétne sa budeme zaoberať metódou „least-squares cross-validation“ ($LSCV$). $LSCV$ je narozdiel od „rules of thumb“ metód plne automatický spôsob získania hodnoty paramtera h. „Krosvalidácia“ tu spočíva v tom, že si definujeme odhad hustoty f ako odhad $\hat{f}_i$ vytvorený pomocou dát $X_1, ..., X_n$ s tým, že X_i-te pozorovanie sa vynechá, t.j.

$$\hat{f}_{-i}(x) = \frac{1}{(n-1)h} \sum_{i \neq j} K\Big(\frac{x - X_j}{h}\Big) \tag{4.1}$$

V druhej kapitole, v (3.2) sme videli, že stredná integrovaná štvorcová chyba jadrového odhadu $\hat{f}_h$ je rovná $MISE(\hat{f}_h(.)) = \mathrm{E} \int (\hat{f}_h(x) - f(x))^2 dx$.
Máme teda

$$MISE(\hat{f}_h(.)) = \mathrm{E} \int (\hat{f}_h(x) - f(x))^2 dx = \mathrm{E}\Big(\int [\hat{f}_h(x)^2 - 2\hat{f}_h(x) f(x) + f(x)^2] dx \Big)$$

odtiaľ

$$MISE(\hat{f}_h(.)) = \mathrm{E} \int [\hat{f}_h(x)^2 dx - 2\mathrm{E} \int \hat{f}_h(x) f(x) dx + \int f(x)^2 dx. \tag{4.2}$$

Posledný člen na vyhladzovacom parametri nezávisí. Hľadanie vhodného parametra h vzhľadom k minimalizácií $MISE$ preto prejde k minimalizácií výrazu $\mathrm{E}R(\hat{f}_h) = \mathrm{E}\big(\int [\hat{f}_h(x)^2 dx - 2 \int \hat{f}_h(x) f(x)] dx \big)$.

Pretože $R(\hat{f}_h)$ závisí na neznámej hustote f, odvodíme si nestranný odhad $R(\hat{f}_h)$ pomocou „krosvalidácie“. Použijeme na to odhad hustoty skonštruovaný pomocou (4.1).

Označme si

$$LSCV(h) = \int \hat{f}_h(x)^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i). \tag{4.3}$$

Najprv si vyjadríme čomu je rovná stredná hodnota z druhej časti vo výraze (4.3):

$$\mathrm{E}\Big(\frac{1}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i)\Big) = \mathrm{E}\Bigg(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{(n-1)h} \sum_{j \neq i} K\Big(\frac{X_i - X_j}{h}\Big)\Bigg) =$$

$$= \frac{1}{n(n-1)h} \sum_{i=1}^{n} \sum_{j \neq i} \mathrm{E}\Bigg(K\Big(\frac{X_i - X_j}{h}\Big)\Bigg) = \mathrm{E}\Bigg(K\Big(\frac{Y - X}{h}\Big)\Bigg) =$$

$$= \int \int K\Big(\frac{y - x}{h}\Big) f(x) f(y) dx dy = \mathrm{E} \int \hat{f}_h(x) f(x) dx$$

Ukážeme teda, že $LSCV$ je nestranným odhadom pre $R(\hat{f}_h) = MISE - \int f(x)^2 dx$, máme

$$\mathrm{E}(LSCV(h)) = \mathrm{E}\Bigg(\int \hat{f}_h(x)^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i) \Bigg) =$$

$$= \mathrm{E} \int \hat{f}_h(x)^2 dx - 2\mathrm{E}\Big(\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i)\Big)$$

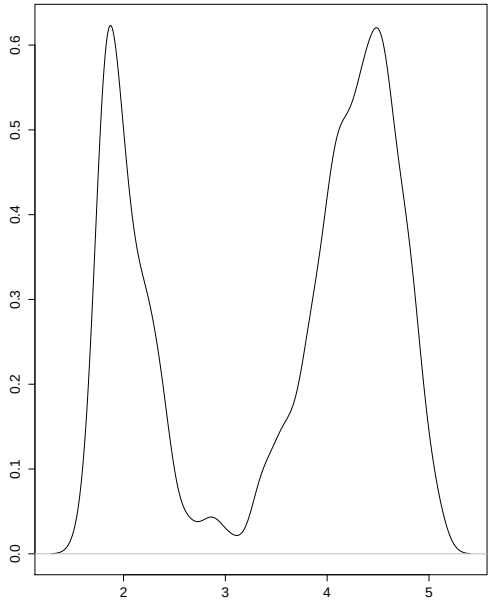
V (4.2) sme si vyjadrili, čomu je rovná stredná hodnota výrazu $\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i)$, po dosadení do rovnice pre E($LSCV$) teda dostaneme:

$$E(LSCV(h)) = E \int \hat{f}_h(x)^2 dx - 2E \int \hat{f}_h(x) f(x) dx,$$

čo je podľa (4.2) $MISE - \int f(x)^2 dx$. Dostali sme teda, že $LSCV(h)$ je nestranným odhadom pre $R(\hat{f}_h) = MISE - \int (f(x))^2 dx$. Takže hľadanie optimálnej hodnoty vyhladzovacieho parametra h pomocou minimalizácie E$[LSCV(h)]$ by malo korešpondovať s minimalizovaním E$[R(\hat{f}_h)]$ a pretože $\int f(x)^2$ je konštanta, v konečnom dôsledku by to teda malo korešpondovať s minimalizovaním MISE.

Na obrázku 4.5 zostrojíme jadrový odhad pre $faithful$ pomocou tejto metódy. Použili sme program R a funkciu *density*, kde sme vzhladzovací parameter nastavili pomocou funkcie *bw.ucv*, čo je v R implementovaný výpočet vyhladzovacieho parametra minimalizáciou $LSCV$. V našom prípade vyšla táto hodnota $h \doteq 0.102$.

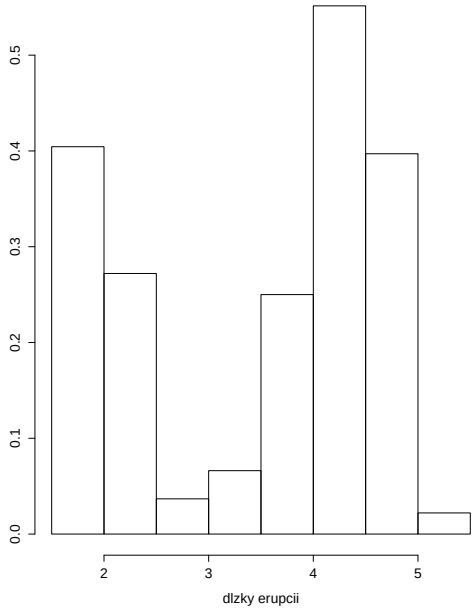
Obrázok 4.5: Jadrový odhad pre dáta faithful s $h \doteq 0.102$



Keď si porovnáme výsledné odhady získané metódou „rules of thumb“ a „krosvalidáciou“ na obrázkoch 4.4 a 4.5, je vidieť, že na obrázku 4.5 nie sú v okolí bodu 3 takmer žiadne pozorovania narozdiel od výsledku, ktorý ponúka obrázok 4.4. Po vyfiltrovaní dát $faithful$ tak, aby sme získali len tie, ktoré sú v blízkosti bodu 3, zvolili sme interval $(2.9, 3.1)$ sme zistili, že počet tých pozorovaní, ktoré do tohto intervalu spadajú je 1. Teda odhad s vyhladzovacím parametrom vypočítaným pomocou „rules of thumb“ metódy je „nadhladený“.

Pre porovnanie si ešte uvedieme, ako by vyzeral histogram pre tieto dáta. Histogram na obrázku 4.6 je pre $a = 0$ a $h = 0.5$.

Obrázok 4.6: Histogram pre dáta $faithful$, $a = 0$, $h = 0.5$.



Z tohto obrázku by sme teda mohli usúdiť, že tvarom mu je „podobnejší“ jadrový odhad z vyhladzovacím parametrom získaným metódou minimalizácie $LSCV$.

V knihe [2], kapitola 3.4.3, strana 51 sa uvádza, že v zmysle minimalizácie integrovanej štvorcovej chyby $MISE$ metóda minimalizácie $LSCV$ dosahuje asymptoticky najlepšiu možnú voľbu vyhladzovacieho parametra. To znamená, že ak máme teoretickú hodnotu optimálneho parametra, označime ju h_{opt} a odhad tejto hodnoty získaný pomocou minimalizovania $LSCV$, h_{lscv}, potom

$$\frac{MISE(\hat{f}_{h_{lscv}})}{MISE(\hat{f}_{h_{opt}})} \xrightarrow{n\to\infty} 1$$

Exsitujú ďalšie iné metódy, ktorými sa hľadá optimálna hodnota vyhladzovacieho parametra h pre jadrové odhady hustoty. My sme si ako kritérium správnosti zadali „vzdialenosť“ dvoch funkcií, odhadu a odhadovanej hustoty. Niekedy ale potrebujeme čo najlepší odhad v inom zmysle, napríklad aby mal odhad rovnaký počet módov a rôzne iné.

Kapitola 5

Záver

V práci sme sa zaoberali neparametrickými metódami odhadu hustoty, najmä jadrovými odhadmi. Na základe odporučenej literatúry sme zhrnuli základné poznatky a stručne popísali problematiku jadrových odhadov a voľby vyhladzovacieho parametra.Voľbu vyhladzovacieho parametra sme založili na požiadavke, aby sa minimalizovala odchýlka vzdialeností odhadu od skutočnej hustoty. Na základe tohto kritéria správnosti odhadu sme popísali metódy, ako vyhladzovací parameter zvoliť v závislosti na tom, ktorá z týchto metód sa nám na riešenie problému viac hodí. Doplnili sme chýbajúce kroky v niektorých odvodeniach a daný problém sme sa pokúsili ilustrovať na dátach, použitím softwaru R.

Problematika jadrových odhadov hustoty je ale veľmi široká. Existuje viacero kritérií na „správnosť“ odhadu a taktiež mnoho spôsobov ako voliť parametre odhadu, totiž vyhladzovacie okno h a jadro K. Práca by sa teda dala rozvíjať mnohými smermi, dalo by sa na ňu nadviazať napríklad popísaním niektorých ďalších metód voľby parametra h alebo doplnením teórie k jadrovým odhadom pre viacrozmerné dáta.

Literatúra

[1] M.P. Wand and M.C. Jones: *Kernel smoothing*, Chapman & Hall, 1995

[2] B.W. Silverman: *Density estimation for statistics and data analysis*, Chapman & Hall, 1986

[3] S.J. Sheater: *Density estimation, 2004: Statistical science vol.19*, No. 4 (Nov., 2004), str. 588-597

[4] V. Jarník: *Diferenciální počet I*, Academia, 1974