УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Аница Костић

Оцењивање густине расподеле

— мастер рад —

Београд, 2017.

Садржај

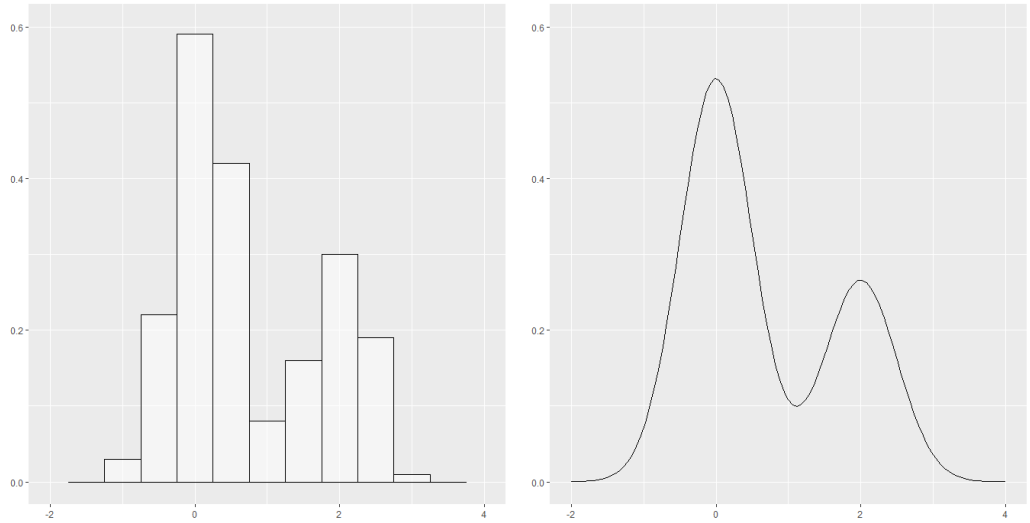
Глава 1

Увод

Основни проблем сатистике је како направити модел на основу датог узорка тако да се, имајући у виду тај модел, може наставити даља анализа и статистичко извођење закључака. Један од основних појмова статистике је густина расподеле. Проблем моделовања може се односити на одређивање густине расподеле из које долази узорак, што је и тема овог рада. Параметарски приступ овом проблему је претпостављање да је непозната густина из неке параметарске фамилије расподела. Дакле, облик густине је познат, а непознате параметре можемо оценити неком од познатих метода, на пример методом максималне веродостојности. Мана параметарског приступа је што се може десити да не можемо да нађемо одговарајућу фамилију расподела којој би узорак одговарао. Пирсон [10] је још 1895. године дефинисао хистограм, који је до данас остао један од најпопуларнијих начина графичког приказивања података. Његова значајност лежи у томе што нам он заправо говори о облику непознате густине расподеле обележја. Зато можемо рећи да је хистограм најстарији метод непараметарског оцењивања густине и његова корисност је неспорна. Укратко ћемо дефинисати хистограм и описати његове недостатке, ради мотивације увођења других оцена густине.

1.1 Хистограм и уопштења

За основни тип хистограма потребно је задати почетну тачку x_0 и ширину интервала h. Реалну осу делимо на интервале облика $[x_0 + mh, x_0 + (m+1)h)$, $m \in \mathbb{Z}$, а са n_m означимо број елемената узорка који припадају сваком од интервала. Хистограм је реална, степенаста функција, задата изразом



Слика 1.1: Хистограм података генерисаних из мешовите нормалне расподеле (лево) и густина те расподеле (десно).

$$\hat{f}(x) = \frac{1}{nh}\{\text{укупан број оних } X_i \text{ који су у истом интервалу као } x\}.$$

За произвољно $x \in \mathbb{R}$ можемо наћи интервал поделе коме припада и то је $[x_0 + m_x h, x_0 + (m_x + 1)h)$, где је $m_x = \lfloor \frac{x - x_0}{h} \rfloor$. Означимо са n_{m_x} број елемената узорка који припадају истом интервалу као x. Хистограм се једноставније може записати и као

$$\hat{f}(x) = \frac{n_{m_x}}{nh}.$$

До првог уопштења хистограма долазимо ако дозволимо различите ширине интервала, које можемо задати пре или након вађења узорка. Дакле, реална оса је подељена на интервале облика $[x_0 + mh_m, x_0 + (m + 1)h_m)$. Дужину интервала у ком се налази x означимо са h_{m_x}. Једначина хистограма са променљивом ширином интервала је

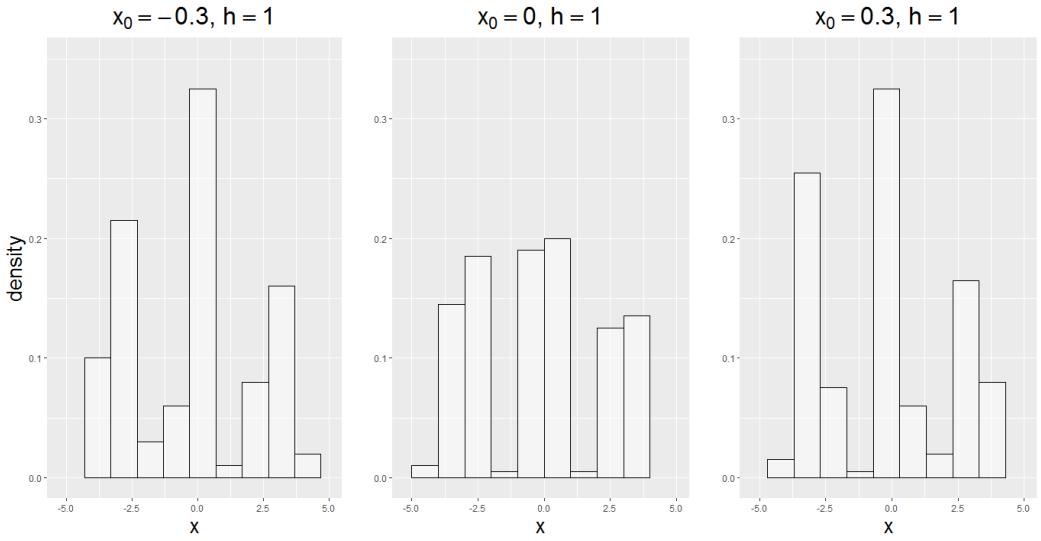
$$\hat{f}(x) = \frac{n_{m_x}}{nh_{m_x}}.$$

Хистограм има особину густине, да је његов интеграл по скупу $\mathbb{R}$ једнак 1. Доказаћемо ово тврђење у случају променљиве дужине интервала, а

одатле следи да исто важи и за фиксну дужину интервала.

$$\begin{aligned}\int_{\mathbb{R}} \hat{f}(x)\,dx &= \frac{1}{n}\sum_{m\in\mathbb{Z}} \frac{n_m}{h_m} h_m \\ &= \frac{1}{n} n \\ &= 1.\end{aligned}$$

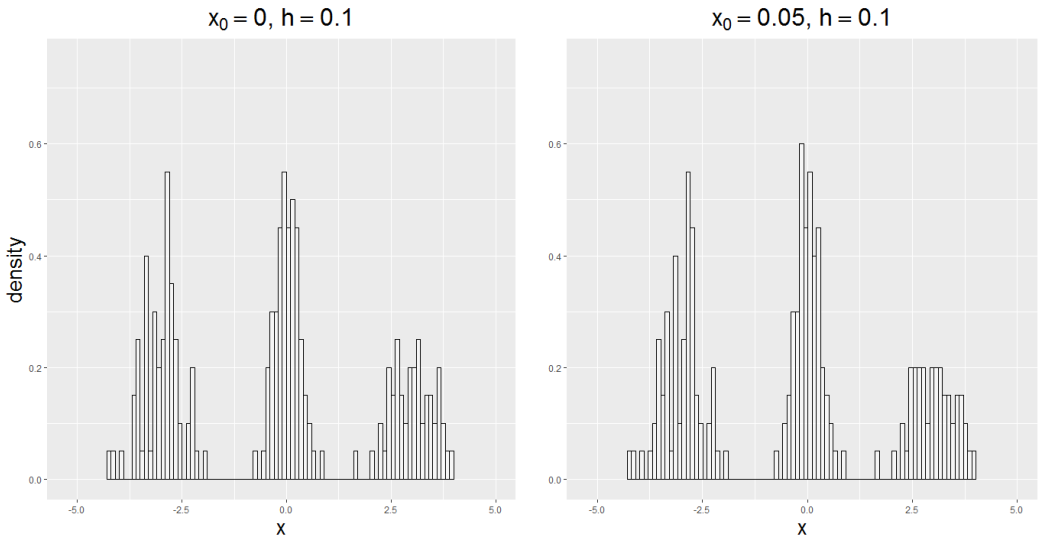
Хистограм се слично може дефинисати и у више димензија, као реална функција више променљивих и представља оцену непознате вишедимензионе густине, али се може користити за графичко приказивање узорка из највише дводимензионе расподеле. Иако је користан метод за први корак у анализи података, има неколико мана. Први недостатак хистограма је што, груписањем узорка по подеоним интервалима губимо информације које носи узорак. Још један недостатак је зависност од одабира x_0 и h. За фиксирано h, зависност од одабира x_0 се најјасније види ако постоје наизменичне области са већом и мањом фреквенцијом елемената узорка. Тада, у зависности од њиховог груписања по интервалима, за различите x_0 могуће је добити значајно различите хистограме. На слици 1.2 приказани су хистограми једног узорка из мешовите нормалне расподеле, за исто h и различите x_0.



Слика 1.2: Хистограми узорка из мешовите нормалне расподеле за различити избор почетне тачке x_0.

Зависност од одабира h је јасна. Што је h мање, то ће хистограм узимати више вредности и биће боље описано понашање на мањим интерва-

лима. Могли бисмо претпоставити да је најбоље одабрати што мање h, међутим тиме долазимо до другог проблема. За мале вредности h јавиће се интервали у којима нема тачака из узорка, па је ту оцена густине хистограмом једнака нули, и тиме се заправо удаљавамо од онога што хистограм треба да представља - оцену непрекидне функције густине. Ту се огледа и трећи недостатак хистограма, тај што није непрекидна функција, што може бити проблем када нам требају оцене извода функције густине. На слици 1.3 видимо да, када се смањи h, за различите x_0 хистограми изгледају слично, и како се јављају интервали у којима је оцена густине хистограмом 0. Параметар h зове се и параметар равнања (ПР). Осим наведених примера у којима се виде недостаци хистограма, његова неефикасност може се и математички извести. У овом раду се нећемо бавити асимптотским понашањем хистограма, само ћемо напоменути да, у неком смислу конвергенције, хистограм спорије тежи стварној густини у односу на неке друге могуће оцене густине. Више о особинама хистограма може се наћи у [6].



Слика 1.3: Хистограми узорка из мешовите нормалне расподеле за мале вредности h и различите почетне тачке x_0.

До следећег уопштења хистограма можемо доћи посматрањем једнакости

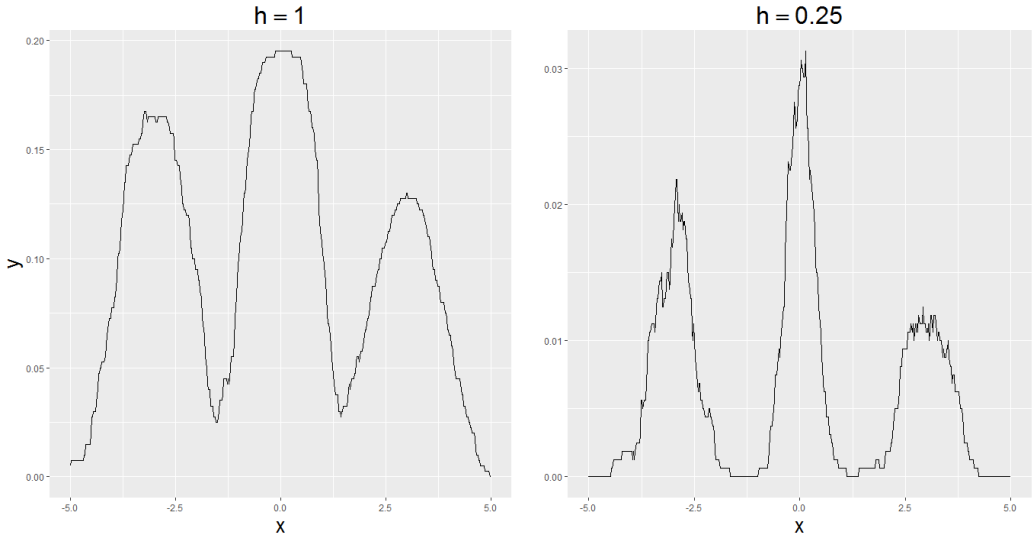
$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h},$$

где F функција расподеле која одговара густини f. Нека је h_n низ реал-

них бројева који тежи 0. Као оцену за $f(x)$ можемо узети

$$\hat{f}_n(x) = \frac{\hat{F}_n(x + h_n) - \hat{F}_n(x - h_n)}{2h_n}.$$

Ову оцену је дефинисао Розенблат (1956) [11], и то је један од првих радова на тему непараметарског оцењивања густине. Представићемо неке особине ове оцене. Она је, као и хистограм, степенаста функција, добија се тако што се изнад сваке опсервације поставља кутија ширине $2h_n$ (по h_n са обе стране) и висине $(2nh_n)^{-1}$, а вредност оцене у тачки x се добија сабирањем висина свих правоугаоника који садрже x. Њени прекиди су у тачкама $X_i \pm h_n$. Ово је уопштење хистограма којим се ослобађамо зависности од избора почетне тачке x_0. Веза између ове оцене и хистограма је та да је једнака вредност ове оцене и хистограма у истој тачки x, где је почетна тачка хистограма таква да је x средиште подеоног интервала ком припада, а ширина интервала је $2h_n$. На слици 1.4 приказане су две оцене густине добијене овом методом, за различите вредности параметра равнања h које има има исту улогу и назив и у случају ове оцене. За извођење особина оцене $\hat{f}_n$ биће нам потребне неке



Слика 1.4: Розенблатова оцена густине за различит избор h.

особине емпиријске функције расподеле. $\hat{F}_n$ је случајна величина и, за

свако x фиксирано, $n\hat{F}_n(x)$ има биномну $\mathcal{B}(n, F(x))$ расподелу.

$$\begin{aligned}
E(\hat{F}_n(x)) &= F(x);\\
E(\hat{F}_n(x)\hat{F}_n(y)) &= \frac{1}{n^2}E(\sum_{i=1}^{n} I\{X_i \le x\} \sum_{j=1}^{n} I\{X_j \le y\})\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E(I\{X_i \le x\}I\{X_j \le y\})\\
&= \frac{1}{n^2}(n(n-1)F(x)F(y) + nE(I\{X_i \le x\}I\{X_i \le y\}))\\
&= \frac{n-1}{n}F(x)F(y) + \frac{1}{n}F(x \wedge y);\\
cov(\hat{F}_n(x), \hat{F}_n(y)) &= \frac{1}{n}(F(x \wedge y) - F(x)F(y)),
\end{aligned}$$

где је $x \wedge y = \min\{x, y\}$. Помоћу ових једнакости можемо извести одговарајуће изразе за оцену густине $\hat{f}_n(x)$.

$$\begin{aligned}
E(\hat{f}_n(x)) &= [F(x + h_n) - F(x - h_n)]/2h_n;\\
D(\hat{f}_n(x)) &= \frac{1}{4h_n^2}D(\hat{F}_n(x + h_n) - \hat{F}_n(x - h_n))\\
&= \frac{1}{4h_n^2}[cov(\hat{F}_n(x + h_n), \hat{F}_n(x + h_n))\\
&\quad - 2cov(\hat{F}_n(x + h_n), \hat{F}_n(x - h_n)) + cov(\hat{F}_n(x - h_n), \hat{F}_n(x - h_n))]\\
&= \frac{1}{4nh_n^2}[F(x + h_n) - F(x - h_n) + (F(x + h_n) - F(x - h_n))^2].
\end{aligned}$$

Дефинишемо функцију

$$K(x) = \begin{cases} \frac{1}{2}, & -1 \le x < 1 \\ 0, & \text{иначе} \end{cases}.$$

Оцену густине можемо записати и преко функције K,

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

На основу израчунатог очекивања види се да је, за свако x, $\hat{f}_n(x)$ асимптотски непристрасна оцена за $f(x)$. Из асимптотске непристрасности и чињенице да дисперзија тежи нули, имамо да је $\hat{f}_n(x)$ и постојана оцена за $f(x)$.

Глава 2

Оцењивање густине језгром

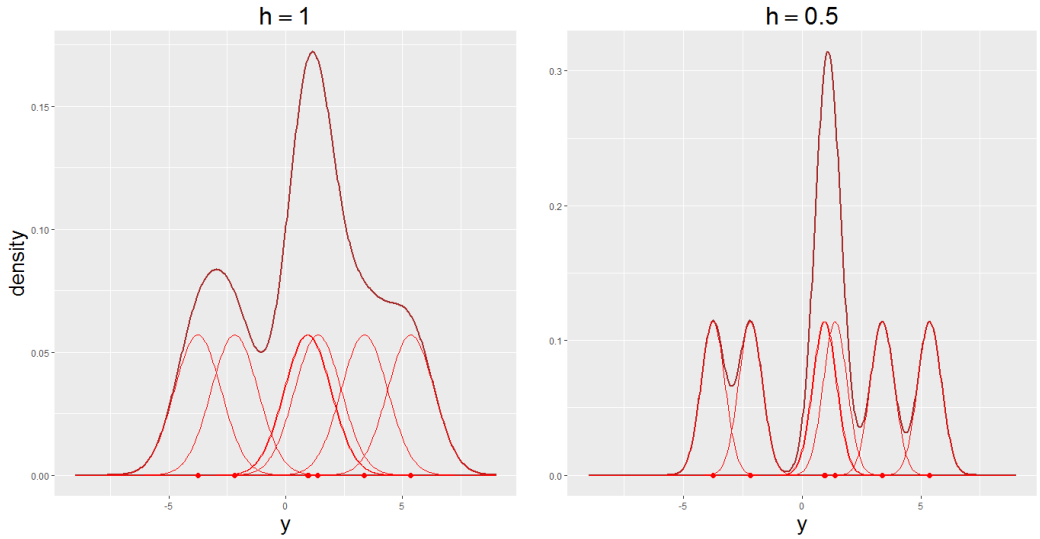
Приметимо да је K из претходног поглавља густина униформне расподеле на $[-1, 1]$. До следећег уопштења долазимо тако што дозволимо да K буде и нека друга густина расподеле. Густину K зовемо језгро а метод оцењивања зовемо оцењивање густине језгром. За језгро се углавном претпоставља да је ограничена, симетрична функција и да важи $|x|K(x) \to 0$, кад $x \to \infty$ и $\int x^2 K(x)\, dx < \infty$. За дати узорак $X_1, ..., X_n$ оцена густине језгром (ОГЈ) је

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{2.1}$$

Својства непрекидности или диференцијабилности наслеђују се од одговарајућих особина језгра. Уколико је, на пример, K густина стандардне нормалне расподеле, што зовемо и Гаусовим језгром, $\hat{f}_n(x)$ је непрекидна и бесконачно пута диференцијабилна функција. Због његове једноставности и ових особина, то је и најчешће коришћено језгро. Касније ћемо навести још нека језгра.

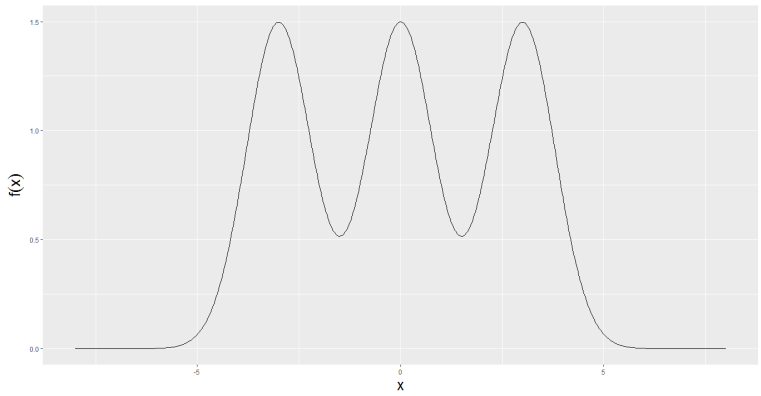
На слици 2.1 илустровано је оцењивање густине Гаусовим језгром за два различита избора параметра h. Црвеним тачкама на x-оси обележене су тачке из узорка, и изнад сваке опсервације центрирано је "брдашце", у ствари график функције $(nh)^{-1}K((x - X_i)/h)$. Оцену густине, такође дату на овој слици, добијамо сабирањем тих скалираних језгара. За скалирано језгро, ради једноставнијег записа, увешћемо ознаку $K_h(\cdot) = 1/hK(\cdot/h)$.

Иако има много могућности за избор K и h, испоставља се да оцена густине више зависи од избора параметра равнања, него од избора језгра. Вредност h одређује колико елемената узорка у близини неке тачке x ће утицати на оцену у тој тачки, пошто је то носач скалираног језгра. У

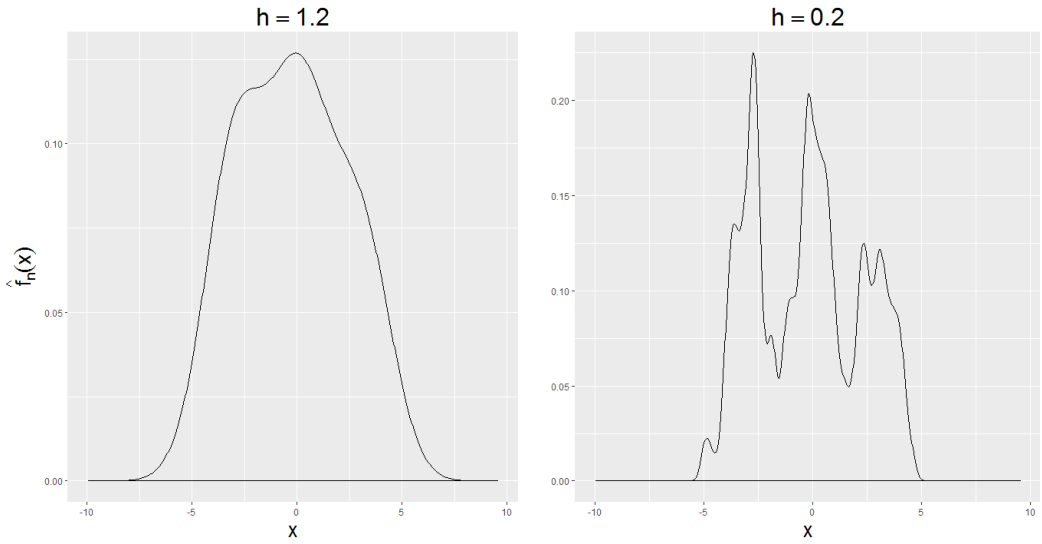


Слика 2.1: ОГЈ за узорак из мешовите нормалне расподеле, за различите параметре равнања.

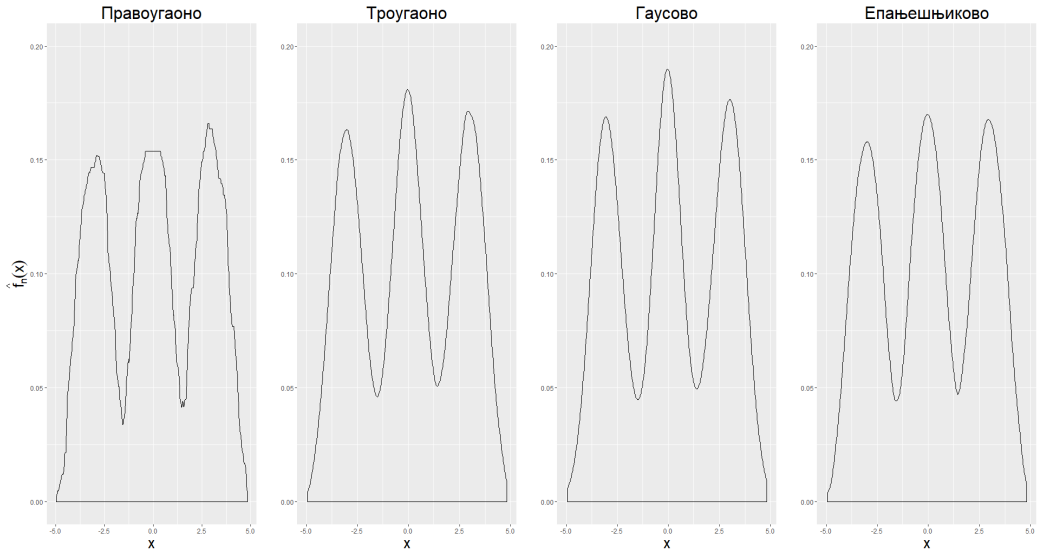
даљем тексту биће речено у ком смислу се тврди да језгро не утиче значајно на оцену. За узорак обима 200 из тримодалне расподеле са густином датом на слици 2.2, приказана је ОГЈ за различите h на слици 2.3. Ако одаберемо превише велико h могуће је да ћемо изгубити неке битне карактеристике узорка, што се види на левој слици, а са друге стране, ако одаберемо премало h придајемо велики значај нечему што је само последица варијабилности у узорку, што се види на десној слици. На слици 2.4 приказане су оцене густине језгром за једнаке вредности h и различита језгра.



Слика 2.2: Густина мешовите нормалне расподеле из које се генерише узорак.



Слика 2.3: ОГЈ узорка из мешовите нормалне расподеле, за различите вредности h.



Слика 2.4: ОГЈ узорка из мешовите нормалне расподеле, за различита језгра.

ОГЈ није непристрасна оцена за $f(x)$ за свако x. Може се доказати да таква оцена и не постоји.

Теорема 2.1. (Пракаса Рао (1983) [3]) *Нека је $X_1, ..., X_n$ прост случајан узорак из расподеле са густином f. Не постоји оцена за $f(x)$ која је*

непристрасна за свако x .

Доказ. Нека је $T_n(x, x_1, ..., x_n)$ ненегативна, Борелова функција. Претпоставимо да је статистика $T_n(x, X_1, ..., X_n)$ непристрасна оцена за $f(x)$ за свако x и сваку густину f. Вектор статистика поретка је комплетна довољна статистика у односу на фамилију свих апсолутно непрекидних расподела (в. [8]). Без умањења општости можемо претпоставити да је T_n симетрична функција узорка, тј. да је функција статистика поретка (у супротном је симетризујемо). За произвољне $a < b$, на основу Фубинијеве теореме, имамо

$$E\left[\int_a^b T_n(x)\,dx\right] = \int_a^b f(x)\,dx = F(b) - F(a),$$

где је F функција расподеле којој одговара густина f. Пошто је емпиријска функција расподеле $\hat{F}_n$ непристрасна оцена за F имамо и

$$E(\hat{F}_n(b) - \hat{F}_n(a)) = F(b) - F(a).$$

Како су и T_n и $\hat{F}_n(b) - \hat{F}_n(a)$ непристрасне оцене, функције вектора статистика поретка, из комплетности следи

$$\hat{F}_n(b) - \hat{F}_n(a) = \int_a^b T_n(x)\,dx \quad \text{с.с. .}$$

Горња једнакост је контрадикција, јер емпиријска функција расподеле није апсолутно непрекидна. □

За свако x фиксирано, ОГЈ $\hat{f}_n(x)$ је случајна величина. То је ненегативна, мерљива функција, дефинисана на $\mathbb{R} \times \Omega^n$. Као мера одступања оцене густине од стварне вредности у тачки x користи се средњеквадратна грешка

$$\begin{aligned} MSE_x(\hat{f}_n) &= E(\hat{f}_n(x) - f(x))^2 \\ &= D\hat{f}_n(x) + (E\hat{f}(x) - f(x)^2) \\ &= D\hat{f}_n(x) + bias_x(\hat{f}_n)^2. \end{aligned} \tag{2.2}$$

Као мера укупног одступања оцене од стварне густине на целом интервалу, може се користити интегрисана средњеквадратна грешка

$$\begin{aligned} MISE(\hat{f}_n) &= E\left(\int (\hat{f}_n(x) - f(x))^2\,dx\right) \\ &= \int MSE_x(\hat{f}_n)\,dx \\ &= \int D\hat{f}_n(x)\,dx + \int bias_x(\hat{f}_n)^2\,dx. \end{aligned} \tag{2.3}$$

Замена места очекивања и интеграла у другој једнакости је последица Фубинијеве теореме, која се може применити због ненегативности и мерљивости функције $(\hat{f}_n(x) - f(x))^2$. Изразићемо моменте $\hat{f}_n(x)$ од којих зависе мере одступања.

$$E(\hat{f}_n(x)) = \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)\, dy; \tag{2.4}$$

$$\begin{aligned} D(\hat{f}_n(x)) = &\int \frac{1}{nh^2} K\left(\frac{x-y}{h}\right)^2 f(y)\, dy \\ &- \left[\frac{1}{nh^2} K\left(\frac{x-y}{h}\right) f(y)\, dy\right]^2. \end{aligned} \tag{2.5}$$

Ови изрази се могу заменити у (2.2) и (2.3) али се добију компликовани изрази који се не могу у општем случају упростити. Приметимо да је функција $E(\hat{f}_n(x))$ конволуција две функције густине, $f(\cdot)$ и $K_h(\cdot)$. На овој чињеници се заснива један од алгоритама за оцену густине језгром који ћемо описати у наредном поглављу.

2.1 Алгоритам за ОГЈ

Најједноставнији алгоритам за рачунање ОГЈ је рачунање по једначини (2.1). Како смо ограничени дискретним подацима, рецимо да рачунамо ОГЈ у неких M тачака $y_1, ..., y_M$, за дати узорак $x_1, ..., x_n$ и параметар равнања h. Дакле, потребно је израчунати низ $\hat{f}_n(y_j)$.

$$\hat{f}_n(y_j) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - y_j}{h}\right). \tag{2.6}$$

Мана овог приступа је што је сложеност одговарајућег алгоритма $O(nM)$. У даљем тексту ћемо представити алгоритам сложености $O(M \log M)$ који је предложио Силверман (1986) [1] и који се користи и за рачунање ОГЈ функцијом `density` из стандардног пакета програмског језика `R`.

Посматрајмо једначину (2.4). Очекивање ОГЈ у тачки x је конволуција скалираног језгра и непознате функције густине. За сада занемаримо чињеницу да је функција густине непозната. У случају када је потребно израчунати конволуцију две познате функције најчешће се користи теорија Фуријеових трансформација. Прво ћемо навести потребну теорију у непрекидном случају, а потом у дискретном. Фуријеова трансформација функције $h : \mathbb{C} \to \mathbb{R}$ је функција $H : \mathbb{C} \to \mathbb{R}$, где је

$$H(f) = \int_{-\infty}^{\infty} h(t)e^{-2\pi ift}\,dt. \tag{2.7}$$

Функција h је потпуно одређена својом Фуријеовом трансформацијом и полазећи од функције H до ње можемо доћи такозваном инверзном Фуријеовом трансформацијом функције H,

$$h(t) = \int_{-\infty}^{\infty} H(f)e^{2\pi ift}\,df.$$

Теорема о конволуцији тврди да је Фуријеова трансформација конволуције две функције $g * h$ једнака производу конволуција тих функција, GH.

Једна од препрека које треба превазићи да би се дошло до алгоритма је та што смо ограничени дискретним подацима. Дакле, у случају када је потребно израчунати конволуцију две познате функције, која се не може лако израчунати аналитички, користимо дискретне Фуријеове трансформације. Уместо са функцијама са непрекидним аргументом, радимо са дискретним подацима. За сваку од функција можемо направити низ чији су елементи вредности те функције у тачкама поделе интервала на коме она није нула. Уколико носач функције није ограничен, онда се због њене интеграбилности може издвојити неки интервал такав да су вредности функције изван њега занемарљиве. Битно је да је дужина интервала поделе једнака за оба низа, а интервали на којима се рачунају вредности функција могу бити различити. Сада ћемо навести одговарајуће формуле у дискретном случају. Фуријеова трансформација низа тачака $h_0, ..., h_{N-1}$ је

$$H_n = \sum_{k=0}^{N-1} h_k e^{-2\pi ikn/N}.$$

Инверзном дискретном Фуријеовом трансформацијом, полазећи од низа H_n долазимо до низа h_k,

$$h_k = \frac{1}{N}\sum_{n=0}^{N-1} H_n e^{2\pi ikn/N}.$$

Пар низова h, H зовемо Фуријеов пар и означавамо $h \Longleftrightarrow H$. Навешћемо и дискретну верзију теореме о конволуцији која има неке додатне претпоставке. Нека је s периодичан низ са периодом N, потпуно одређен вредностима $s_0, s_1, ..., s_{N-1}$, и нека је низ r коначне дужине N, $r_{-N/2+1}, ..., r_{N/2}$.

Тада Фуријеов пар чине следећи низови

$$\sum_{k=-N/2+1}^{N/2} s_{j-k} r_k \Longleftrightarrow S_n R_n. \tag{2.8}$$

где је S_n дискретна Фуријеова трансформација низа s_k, $k = 0, ..., N-1$, а R_n дискретна Фуријеова трансформација низа r_k, $k = 0, ..., N-1$. Притом су вредности $r_{N/2+1}, ..., r_N$ једнаке $r_{-N/2+1}, ..., r_{-1}$, респективно.

Један начин за рачунање конволуције је директно по суми (2.8), и такав алгоритам је сложености $O(N^2)$. Други начин је уз помоћ популарног FFT (Fast Fourier Transform) алгоритма. Овим алгоритмом се рачунају дискретне Фуријеове трансформације низова и његова сложеност је $O(N \log N)$. Да бисмо израчунали конволуцију два низа, на основу поменуте теореме о конволуцији, довољно је FFT алгоритмом израчунати дискретне Фуријеове трансформације оба низа и онда их изможити члан по члан. Међутим, претпоставка теореме је да је један од низова периодичан, а као што ћемо касније видети, у нашем случају, а и често у пракси, тај услов није испуњен. Овај проблем се решава додавањем одређеног броја нула овим низовима пре примене FFT алгоритма. Функција језгра је позната, па на описани начин можемо формирати низ вредности ове функције, али као што је речено, стварна густина је непозната и то је други проблем који треба превазићи. Ипак, ми имамо познат узорак, и идеја је да се од њега направи прва оцена густине која се онда побољша конволуцијом. Алгоритам за оцену густине можемо поделити у 3 дела.

- Прва оцена густине (дискретизација узорка)
- Дискретно језгро
- Конволуција

Прва оцена густине

Нека је $[a, b]$ интервал ком припадају све тачке из узорка. Може се узети, на пример $a = \min X_i - 4h$, $b = \max X_i + 4h$. Бирамо $M = 2^r$, за неко $r \in \mathbb{N}$, број тачака у којима ће се рачунати оцена густине.

$$\delta = (b-a)/(M-1);$$
$$t_i = a + i\delta, \quad i = 0, 1, ..., M-1.$$

Узорак дискретизујемо тако што тежину сваке тачке из узорка разбијамо на границе интервала поделе ком припада. Ако X припада интервалу $[t_i, t_{i+1}]$, левој граници t_i доделимо тежину $(X - t_i)n^{-1}\delta^{-1}$, а десној $(t_{i+1} - X)n^{-1}\delta^{-1}$. Тежина сваке тачке из узорка је n^{-1}. Овим добијамо низ c_i, који зовемо дискретизован узорак, он има M елемената и његова сума је 1.

Дискретно језгро

Претпоставимо да је језгро симетрично. Можемо одабрати интервал $[-\tau, \tau]$ изван кога су његове вредности занемарљиво мале. У претходном кораку смо одабрали дужину интервала поделе δ. Због симетрије језгра довољно је израчунати вредности у позитивним тачкама,

$$k_i = K_h\left(\frac{(b-a)i}{M-1}\right), \quad i = 0, ..., L,$$

где је $L = \lfloor \tau h(M-1)/(b-a) \rfloor$.

Конволуција

Уопштено, када треба наћи конволуцију два низа коначне дужине, пошто претпоставке о периодичности нису испуњене, не можемо одмах применити теорему о конволуцији и једнакост (2.8). Без образложења описаћемо поменути поступак додавања нула ради превазилажења овог проблема. Нека је P степен двојке такав да је $P \geq M + L$. Дефинишемо низове

$$c = (c_1, ..., c_M, \mathbf{0}_{P-M});$$
$$k = (k_0, ..., k_{L-1}, \mathbf{0}_{P-2L-1}, k_L, ..., k_1).$$

Наћи низове C и K који су Фуријеове трансформације низова c и k, редом. Нека је $\tilde{F}$ производ низова C и K члан по члан. Коначно, низ $\tilde{f} = (\tilde{f}_1, ..., \tilde{f}_M)$ који чине првих M чланова инверзне Фуријеове трансформација низа $\tilde{F}$ је тражена оцена густине.

2.2 Први асимптотски резултати

Претпоставимо да је језгро K густина расподеле симетрична око 0, за коју важи

$$\int tK(t)\,dt = 0, \qquad \int t^2K(t)\,dt = \mu_2 \neq 0.$$

Дефинишемо ред језгра као најмањи природан број k такав да су сви моменти језгра нижег реда једнаки нули, а k-ти моменат је не-нула и коначан. Дакле, наше језгро K је другог реда. Претпоставимо и да је непозната густина непрекидна, да има непрекидне изводе првог и другог реда и постоји извод трећег реда у свакој тачки x. Овај услов ће нам бити потребан да бисмо имали прва три члана Тејлоровог развоја. Претпоставимо и да је параметар равнања низ такав да $h_n \to 0$, кад $n \to \infty$. Пристрасност ОГЈ је

$$\begin{aligned} bias(x) &= E(\hat{f}_n(x)) - f(x) \\ &= \int \frac{1}{h_n} K\left(\frac{x-y}{h_n}\right) f(y)\, dy - f(x) \\ &= \int K(t)(f(x-h_n t) - f(x))\, dt. \end{aligned}$$

У последњој једнакости смо увели смену $y = x - h_n t$. Сада, примењујући Тејлорову формулу добијамо

$$\begin{aligned} bias(x) &= \int K(t)\left(-h_n t f'(x) + \frac{(h_n t)^2}{2} f''(x) + o(h_n^2)\right)\, dt \\ &= -h_n f'(x) \int tK(t)\, dt + \frac{h_n^2}{2} f''(x) \int t^2 K(t)\, dt + o(h_n^2) \\ &= \frac{h_n^2}{2} f''(x)\mu_2(k) + o(h_n^2). \end{aligned} \tag{2.9}$$

Под овим јаким претпоставкама имамо да је $\hat{f}_n(x)$ асимптотски непристрасна оцена за $f(x)$, за свако x, а због члана h_n^2 кажемо да је одступање од средње вредности другог реда.

$$\begin{aligned} D(\hat{f}_n(x)) &= \int \frac{1}{nh_n^2} K^2\left(\frac{x-y}{h_n}\right) f(y)\, dy - \frac{1}{n}(f(x) + bias(x))^2 \\ &= \int \frac{1}{nh_n^2} K^2(t)(f(x) - h_n t f'(x) + o(h_n))\, dt - \frac{1}{n}(f(x) + o(h_n))^2 \\ &= \int \frac{1}{nh_n^2} K^2(t)(f(x) - h_n t f'(x) + o(h_n))\, dt - O(n^{-1}) \\ &= \frac{1}{nh_n} f(x) \int K^2(t)\, dt + o((nh_n)^{-1}) \\ &\sim \frac{1}{nh_n} f(x) \int K^2(t)\, dt. \end{aligned}$$

Друга једнакост у претходном низу следи из смене $y = x - h_n t$ у интегралу, Тејлорове формуле, и асимптотског понашања $bias(x)$, а трећа и четврта применом особина асимптотских ознака. Ако уз већ претпостављене улове важи и $nh_n \to \infty$ онда одавде имамо постојаност оцене $\hat{f}_n$. Утицај ПР на оцену види се и у овим асимтотским изразима за очекивање и дисперзију. За мале вредности h_n оцењена функција је шиљата, али пристрасност је мала, са повећањем h_n изравнавамо оцену чиме смањујемо њену дисперзију по цени повећања пристрасности. Добијене изразе можемо убацити у израз за $MISE(\hat{f}_n)$ и тако добијамо њено асимптотско понашање.

$$\begin{aligned} MISE(\hat{f}_n) &= \int D(\hat{f}_n(x))\,dx + \int bias(x)^2\,dx \\ &\sim \frac{1}{nh_n} R(K) + \frac{h_n^4 \mu_2}{4} R(f'') \\ &= AMISE(\hat{f}_n), \end{aligned} \tag{2.10}$$

где је $R(g) = \int g^2(x)\,dx$. Раније смо објаснили утицај h_n на ОГЈ, али нисмо још увек дали рецепт по коме се h_n рачуна. $MISE(\hat{f}_n)$ можемо посматрати као функцију од h_n, и наћи тачку у којој се достиже минимум. Тиме добијамо h_{AMISE}, оптималну вредност параметра равнања у смислу минимизирања интегралне средњеквадратне грешке. Једна од предности $MISE$ као мере блискости $\hat{f}_n$ и f је та што се из њеног асимптотског израза оптимално h_n може експлицитно изразити,

$$h_{AMISE} = \left[\frac{R(K)}{\mu_2^2 R(f'') n} \right]^{1/5}, \tag{2.11}$$

где је μ_2 други моменат језгра. Када заменимо ову вредност у израз за $MISE$ добијамо

$$\inf_{h>0} MISE(\hat{f}_n) \sim \frac{5}{4} C(K) R(f'')^{1/5} n^{-4/5},$$

где је $C(K)$ константа која зависи само од језгра,

$$C(K) = k_2^{2/5} R(K)^{4/5}.$$

Приметимо да h_{AMISE} зависи од непознате густине, тако да га, када имамо само познат узорак, не можемо израчунати. У следећем поглављу бавићемо се овим проблемом. Да бисмо добили што мању грешку такође хоћемо да минимизирамо и $C(K)$. Можемо претпоставити да је μ_2 једнак јединици, иначе језгро можемо скалирати. Сада се проблем

минимизирања $C(K)$ своди на проблем тражења минимума $\int K(t)^2\,dt$ при условима $\int K(t)\,dt = 1$ и $\int t^2K(t)\,dt = 1$. Као решење добија се Епањешњиково језгро

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

Ово језгро је добило име по В.А. Епањешњикову који је први проучавао његова оптимална својства. Можемо дефинисати ефикасност других симетричних густина као језгара у односу на Епањешњиково,

$$\begin{aligned} \text{eff}(K) &= \{C(K_e)/C(K)\}^{5/4} \\ &= \frac{3}{5\sqrt{5}}\left\{\int t^2K(t)\,dt\right\}^{-1/2}\left\{K^2(t)\,dt\right\}^{-1}. \end{aligned}$$

Испоставља се да већина језгара која се користе, међу којима су она наведена у табели 2.1, имају ефикасност преко 0.9. Дакле, посматрајући у односу на интегралну средњеквадратну грешку, избор језгра не утиче пуно на оцену. Чак, упркос својој оптималности, чешће од Епањешњиковог користи се Гаусово језгро. Вероватно главни разлог за то је што већ први извод Епањешњиковог језгра није непрекидан, а изводи реда већег од 3 су идентички једнаки нули. С друге стране, Гаусово језгро је бесконачно пута диференцијабилно, што је добра особина када, на пример, оцењујемо изводе густине, о чему ће касније бити речи.

Језгро	$K(t)$
Троугаоно	$\begin{cases} 1-\lvert t\rvert & \lvert t\rvert < 1 \\ 0 & \text{иначе} \end{cases}$
Правоугаоно	$\begin{cases} 1/2 & \lvert t\rvert < 1 \\ 0 & \text{иначе} \end{cases}$
Биквадратно	$\begin{cases} 15/16(1-t^2)^2 & \lvert t\rvert < 1 \\ 0 & \text{иначе} \end{cases}$
Гаусово	$\frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \quad t \in \mathbb{R}$

Табела 2.1: Уобичајена језгра

Глава 3

Избор параметра равнања и језгра

3.1 Избор параметра равнања

Постоји више метода за избор параметра равнања. Уколико имамо неко првобитно знање о густини, можемо да подесимо параметар тако да график оцене густине највише одговара том нашем првобитном знању. Ово је субјективни метод избора ПР и у неким случајевима је и то довољно. Као што је већ речено, проблем са избором h_n је што ако одаберемо мало h_n оцена густине ће својим обликом превелики значај дати неким структурама које су само последица варијабилности у узорку, док ако одаберемо превише велико h битне структуре ће се изгубити јер ћемо превише изравнати. У наредном тексту описана су четири метода за избор параметра равнања. Већина описаних метода, сви осим ММВ оцене, се заснива на једначини (2.11). Она је значајна јер, за почетак, из те једначине видимо да за параметар равнања треба да важи $h_n = O(n^{-1/5})$. Такође, на основу ње можемо да израчунамо оптималну вредност ПР за неку унапред познату густину. То може бити корисно када испитујемо да ли је неки предложени метод рачунања ПР добар - поређењем оцењене и асимптотски оптималне вредности. У даљем тексту ћемо уместо досадашње ознаке h_n писати h ради једноставнијег записа, али имаћемо у виду да је h функција од n.

3.1.1 Позивање на нормалну расподелу

Овај метод је брз и једноставан, али не гарантује никакву оптималност добијеног резултата. Базира се на једначини (2.11). Њом смо добили оптималну вредност за h у смислу минимизирања $AMISE$, али је

проблем што она зависи од непознате густине f. Идеја овог метода је да у изразу $\int f''(x)^2\,dx$ уместо непознате, користимо други извод нормалне густине са одговарајућим параметрима. Како вредност овог интеграла у случају нормалне расподеле неће зависити од средње вредности, довољно је на основу узорка оценити параметар σ^2 нормалне расподеле.

$$\begin{aligned}\int f''(x)^2\,dx &= \sigma^{-5}\int \phi''(x)^2\,dx \\ &= \frac{3}{8}\pi^{-1/2}\sigma^{-5} \sim 0.212\sigma^{-5}. \end{aligned} \tag{3.1}$$

У случају Гаусовог језгра, када вредност из (3.1) убацимо у израз за h_{AMISE} (2.11), добијемо

$$\begin{aligned} h_{AMISE} &= (4\pi)^{-1/10}\left(\frac{3}{8}\pi^{-1/2}\right)^{-1/5}\sigma n^{-1/5} \\ &= \left(\frac{4}{3}\right)^{1/5}\sigma n^{-1/5} \\ &= 1.06\sigma n^{-1/5}. \end{aligned}$$

Одговарајућа оцена параметра равнања је

$$\hat{h}_{NS} = 1.06\hat{\sigma}n^{-1/5},$$

где је $\hat{\sigma}$ оцена стандардне девијације расподеле узорка. То може бити узорачка стандардна девијација или нека робусна оцена, на пример стандардизовано интерквартилно растојање $\hat{\sigma}_{IQR} = (\text{узорачко интерквартилно растојање})/(\Phi^{-1}(3/4) - \Phi^{-1}(1/4))$, где је Φ^{-1} инверзна функција $\mathcal{N}(0,1)$ расподеле. Ако је узорак заиста из нормалне или неке друге унимодалне расподеле, добијена вредност је добра оцена за h. Међутим, у случају када је стварна густина мултимодална, вредност $\int f''(x)^2\,dx$ ће бити мања него она добијена поменутом апроксимацијом. Дакле, одабрано h ће бити веће него што би требало. Једна од предности је што се избором већег h неће преувеличати неке карактеристике које су последица варијабилности узорка. Следећа оцена се у литератури налази под различитим називима, као Силверманов или као Скотов параметар равнања.

$$\hat{h}_{Scott} = 1.06An^{-1/5},$$

где је $A = \min(\overline{s}_n, \hat{\sigma}_{IQR})$. Испитује се и осетљивост ове оцене у зависности од различитих вредности параметара асиметрије и спљоштености непознате густине, при различитим оценама за σ. Као најбољу оцену, којом

смањујемо шансе превеликог равнања, Силверман [1] предлаже

$$\hat{h}_{Silv} = 0.9An^{-1/5}.$$

Овај метод оцењивања је брз и једноставан, али није асимптотски оправдан - не може се доказати да иједна од ове две оцене са повећањем узорка тежи ка стварној оптималној вредности h_{MISE}.

3.1.2 ММВ са унакрсном провером

Уопштено, идеја метода максималне веродостојности (ММВ) је да изаберемо модел за који је вероватноћа добијеног узорка највећа. Код параметарских модела знамо заједничку густину узорка, па и функцију веродостојности до на параметар, међутим у овом случају не знамо. Ако је Y случајна величина која има расподелу одређену непознатом густином f, тада је $f(Y)$ функција веродостојности тог узорка обима 1. Са друге стране, $\hat{f}_n$ је оцена густине и то је, за фиксиране $X_1, ..., X_n$ и језгро K, параметарски модел за f, са параметром h. У случају да поред узорка $X_1, ..., X_n$ имамо и независну од узорка случајну величину Y, оптималну вредност h тражимо као тачку у којој се достиже максимум $\ln \hat{f}_n(Y)$. Пошто имамо само $X_1, ..., X_n$ идеја овог метода је да на основу $n-1$ елемената у узорку оценимо густину са $\hat{f}_{-i}$, и да h добијемо као тачку минимума $\ln \hat{f}_{-i}(X_i)$, где је X_i елемент узорка изостављен из формирања густине. Свеједно је који елемент узорка ћемо изоставити, тако да, коначно, оцену h_{MLCV} добијамо као тачку минимума функције

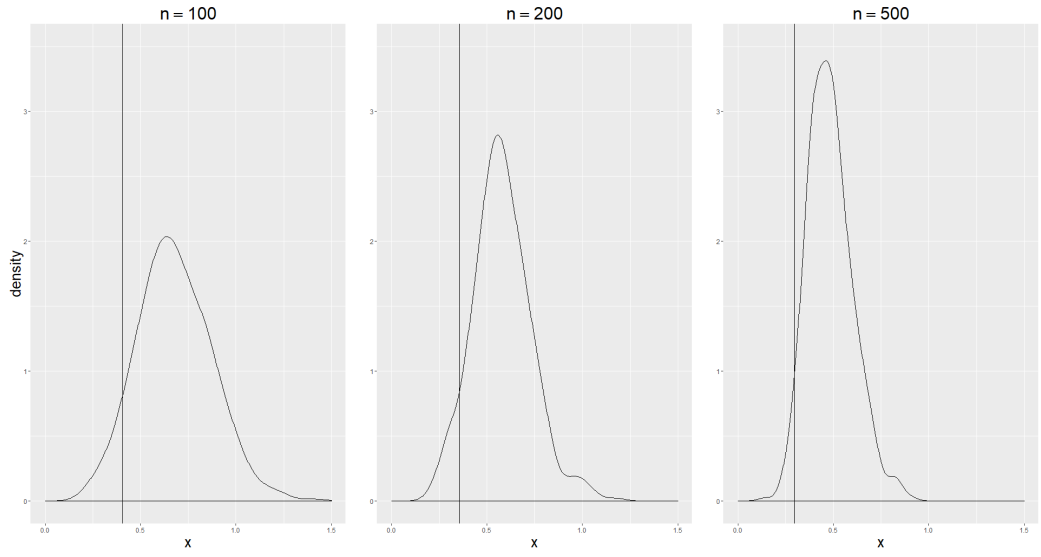
$$CV(h) = n^{-1} \sum_{i=1}^{n} \ln(\hat{f}_{-i}(X_i)).$$

Термин унакрсна провера се односи на идеју о коришћењу једног дела узорка ради добијања информација о другом делу узорка. У нашем случају $\hat{f}_{-i}(X_i)$ даје информацију о вредности густине у изостављеној тачки узорка. Може се показати да је за h добијено овом методом оцена густине блиска стварној густини у смислу минимизирања Кулбак-Лајблеровог растојања дефинисаног са

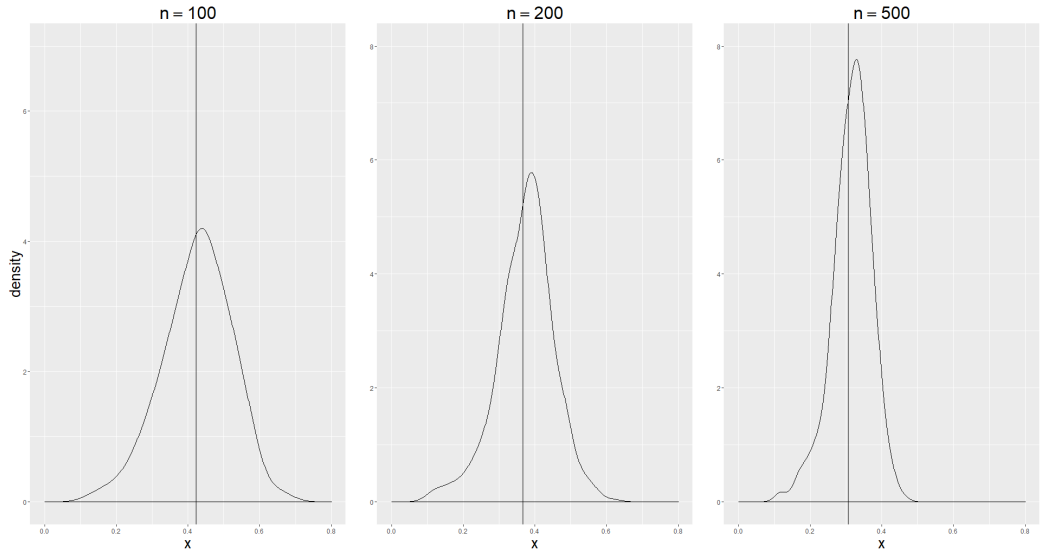
$$I(f, \hat{f}_n) = \int f(x) \log f(x)/\hat{f}_n(x)\, dx;$$

$$\begin{aligned} E\{CV(h)\} &= E \ln(\hat{f}_{-i}(X_i)) = E \int f(x) \ln \hat{f}_{-i}(X_i)\, dx \\ &\approx E \int f(x) \ln \hat{f}(x)\, dx = -EI(f, \hat{f}) + \int f(x) \ln f(x)\, dx. \end{aligned}$$

Претходно примењена апроксимација није строго математички оправдана, али на овај начин видимо како је $-CV(h)$ до на константу непристрасна оцена очекиваног Кулбак-Лајблеровог растојања. Иако интуитиван, овај начин оцењивања параметра равнања има многе недостатке. Уколико је носач за f неограничен, а за K ограничен, $I(f, \hat{f}_n)$ ће бити $-\infty$ за свако h. Такође, показује се да је ова оцена осетљива на аутлајере као и да у случају када репови расподеле теже нули експоненцијално или спорије, овај метод води до оцене густине која није постојана, пошто h добијено максимизирањем $CV(h)$ неће тежити нули кад $n \to \infty$. Дакле, лоше понашање се односи на већину густина, осим нормалне и оних са ограниченим носачем. Показаћемо на примеру Лапласове расподеле лоше понашање ове оцене. За 500 узорака обима 100, 200 и 500 из Лапласове(0,1) расподеле израчунати су овом методом параметри равнања, а онда је за та три низа од 500 елемената оцењена густина. Као параметар равнања користимо h_{Scott}. На слици 3.1 приказане су оцене густине. Вертикалним линијама означене су асимптотски оптималне вредности h_{AMISE} које се разликију у зависности од обима узорка. Видимо како се са порастом обима узорка смањује дисперзија, а средња вредност је ближа h_{AMISE}, ипак већина параметара равнања је и за велико n доста већа од h_{AMISE}. Много боље понашање може се видети у случају $\mathcal{N}(0,1)$ расподеле. Исти поступак одрађен је и у случају ове расподеле и резултати су приказани на слици 3.2.



Слика 3.1: ОГЈ параметара h_{MLCV} за узорке обима 100 (лево), 200 (средина) и 500 (десно) из Лапласове(0,1) расподеле.



Слика 3.2: ОГЈ параметара h_{MLCV} за узорке обима 100 (лево), 200 (средина) и 500 (десно) из $\mathcal{N}(0,1)$ расподеле.

3.1.3 Метод најмањих квадрата са унакрсном провером

Овај метод се такође заснива на интегрисаној средњеквадратној грешци као мери одступања оцене од стварне густине,

$$\begin{aligned} MISE(\hat{f}_n) &= E\int (\hat{f}_n(x) - f(x))^2\,dx \\ &= E\int \hat{f}_n(x)^2\,dx - 2E\int \hat{f}_n(x)f(x)\,dx + E\int f(x)^2\,dx. \end{aligned}$$

Идеја је наћи h за које је ова вредност минимизирана. Како ова вредност зависи и од стварне, непознате густине, то неће бити могуће, али постоји начин да се модификацијом ове идеје дође до асимптотске оптималности. За почетак, пошто последњи сабирак не зависи од оцене, оптимална вредност h у смислу минимизирања ове грешке је оно h за које је минимално

$$M(\hat{f}_n) = E\int \hat{f}_n(x)^2\,dx - 2E\int \hat{f}_n(x)f(x)\,dx. \tag{3.2}$$

Израз (3.2) ћемо оценити на основу узорка и онда тражити минимум по h те оцене. Први сабирак зависи само од узорка, а други сабирак зависи

и од непознате густине. Дефинишемо оцену језгром $\hat{f}_{-i}(x)$, конструисану од $n-1$ елемената узорка, без елемента X_i

$$\hat{f}_{-i}(x) = \frac{1}{n-1} \sum_{j \neq i} K_h(x - X_j).$$

Једна непристрасна оцена за $M(\hat{f}_n)$ је

$$LSCV(h) = \int \hat{f}_n(x)^2 \, dx - \frac{2}{n} \sum_i \hat{f}_{-i}(X_i). \tag{3.3}$$

Доказаћемо непристрасност ове оцене.

$$\begin{aligned} E(LSCV(h)) &= E\left(\int \hat{f}_n(x)^2 \, dx - \frac{2}{n} \sum_i \hat{f}_{-i}(X_i) \right) \\ &= E \int \hat{f}_n(x)^2 \, dx - 2E\hat{f}_{-n}(X_n) \\ &= E \int \hat{f}_n(x)^2 - 2 \int \hat{f}_n(x) f(x) \, dx \\ &= M(\hat{f}_n). \end{aligned} \tag{3.4}$$

Трећа једнакост у претходном низу важи јер је

$$\begin{aligned} E\hat{f}_{-n}(X_n) &= E\left(\frac{1}{n-1} \sum_{i \neq n} K_h(X_n - X_i) \right) \\ &= \int \cdots \int \frac{1}{n-1} \sum_{i \neq i} K_h(x_n - x_i) f(x_1) \cdots f(x_n) \, dx_1 \ldots \, dx_n \\ &= \int E\hat{f}_{-n}(x) f(x) \, dx \\ &= \int E\hat{f}_n(x) f(x) \, dx \\ &= E \int \hat{f}_n(x) f(x) \, dx. \end{aligned}$$

Дакле, минимизирањем функције $LSCV$ добијамо оцену параметра равнања h_{LSCV}. Иако не можемо да нађемо експлицитан израз за тачке локалног минимума ове функције, можемо још мало поједноставити из-

раз,

$$\begin{aligned}\int \hat{f}_n(x)^2\,dx &= \int \frac{1}{n}\sum_i K_h(x-X_i)\,dx \\ &= \frac{1}{n^2}\sum_i\sum_j \int K_h(x-X_i)K_h(x-X_j)\,dx \\ &= \frac{1}{n^2h^2}\sum_i\sum_j \int K(u-h^{-1}X_i)K(u-h^{-1}X_j)\,du \\ &= \frac{1}{n^2h}\sum_i\sum_j \int K(t)K(h^{-1}(X_j-X_i)-t)\,dt \\ &= \frac{1}{n^2h}\sum_i\sum_j K^{(2)}(h^{-1}(X_j-X_i)). \end{aligned} \tag{3.5}$$

Дакле, у општем случају треба рачунати конволуцију, и за то се може користити већ поменута теорија Фуријеових трансформација. У специјалном случају, када користимо Гаусово језгро, рачун се поједностављује, пошто је конволуција тог језгра са самим собом густина нормалне $\mathcal{N}(0,2)$ расподеле. Сада ћемо да поједноставимо и други сабирак у (3.3),

$$\begin{aligned}\frac{1}{n}\sum_i \hat{f}_{-i}(X_i) &= \frac{1}{n}\sum_i \frac{1}{n-1}\sum_{j\neq i} K_h(x-X_i) \\ &= \frac{1}{n(n-1)}\sum_i\sum_j h_n^{-1}K(h^{-1}(X_i-X_j)) \\ &\quad - \frac{1}{n(n-1)}n^{-1}K_h(0). \end{aligned} \tag{3.6}$$

Коначно разлика (3.5) и (3.6) је $LSCV(h)$ за Гаусово језгро. Ради једноставности фактор $(n-1)^{-1}$ заменићемо са n^{-1}. Тиме добијамо

$$\begin{aligned}LSCV(h) &\approx n^{-2}h^{-1}\sum_i\sum_j K^{(2)}(h^{-1}(X_j-X_i)) \\ &\quad - 2n^{-2}h^{-1}\sum_i\sum_j K(h^{-1}(X_j-X_i)) - 2n^{-1}h^{-1}K(0) \\ &= n^{-2}h^{-1}\sum_i\sum_j K^*(h^{-1}(X_j-X_i)) - 2n^{-1}h^{-1}K(0),\end{aligned}$$

где је $K^*(t) = K^{(2)}(t) - 2K(t)$.

Због непристрасности коју смо и доказали, (3.4), овај метод се зове и

непристрасна унакрсна провера. Може се десити да функција $LSCV(h)$ има више од једног локалног минимума. У том случају се за h предлаже не тачка глобалног минимума већ највећа тачка локалног минимума (в. [2]). Стоун (1984) [14] је доказао да је овај избор параметра равнања асимптотски оптималан, у смислу да важи

$$\frac{h_{LSCV}}{h_{AMISE}} \xrightarrow{\text{c.c.}} 1, \quad n \to \infty.$$

Међутим доказано је и да h_{LSCV} има велику асимптотску дисперзију, па се из тог разлога сматра да овај метод ипак нема довољно добра ни теоријска (в. [15]) ни практична својства (в. [19]), и предложено је више побољшања.

Једна модификација овог метода је пристрасна унакрсна провера која оцењује h тако што минимизује функцију $AMISE(\hat{f}_n)$ (2.10) где се непознато $R(f'')$ оцењује се са

$$\begin{aligned}\widetilde{R(f'')} &= R(\hat{f}_n'') - (nh^5)^{-1}R(K'') \\ &= n^{-2}\sum\sum_{i\neq j} K_h'' * K_h''(X_i - X_j).\end{aligned}$$

Друга једнакост се добија након мало дужег извођења сличног оном у (3.5). Функција $BCV(h)$ којој треба наћи минимум, добија се заменом $R(f'')$ са $\widetilde{R(f'')}$ у изразу за $AMISE(\hat{f}_n)$. Минимум те функције је оцена методом пристрасне унакрсне провере и означавамо је са h_{BCV}. Доказано је да је асимптотска дисперзија ове оцене мања од h_{LSCV}, али да има позитивну пристрасност, по чему је ова оцена и добила име.

3.1.4 Метод уврштене оцене

Овај метод развили су Хол и Марон (1987) [17] и Шедер и Џонс (1991) [18] и сличан је методу пристрасне унакрсне провере. Идеја је да се у изразу за h_{AMISE} функционал непознате густине оцени на мало другачији начин, итеративно. Прво ћемо да опишемо оцењивање функционала густине облика

$$R(f^{(s)}) = \int f^{(s)}(x)^2\,dx.$$

Јасно је како нам ово може бити од користи, пошто нам баш овај функционал за $s = 2$ прави проблем. Парцијалном интеграцијом овај израз се своди на

$$R(f^{(s)}) = (-1)^s \int f^{(2s)}(x)f(x)\,dx,$$

па се претходни проблем своди на оцењивање

$$\psi_r = \int f^{(r)}(x) f(x)\, dx,$$

за r парно. Како је претходни израз једнак $Ef^{(r)}(X)$ намеће се следећа оцена

$$\hat{\psi}_r(g) = n^{-1} \sum_{i=1}^{n} \hat{f}_n^{(r)}(X_i, g) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} L_g^{(r)}(X_i - X_j).$$

где су g и L редом параметар равнања и језгро који се користе у оцени и који могу бити различити од h и K. Од значаја ће нам бити асимтптотско понашање средњеквадратне грешке оцене $\hat{\psi}_r(g)$. Може се доказати да под неким условима (в. [2]) важи да је за познато L, оптимална вредност g у смислу минимизирања ове грешке

$$g_{AMSE} = \left[\frac{k! L^{(r)}(0)}{-\mu_k(L) \psi_{r+k} n} \right]^{1/(r+k+1)}, \tag{3.7}$$

где је k ред језгра L. Вратимо се на проблем оцењивања параметра равнања. После увођења нових ознака имамо

$$h_{AMISE} = \left[\frac{R(K)}{k_2^2 \psi_4 n} \right]^{1/5}. \tag{3.8}$$

Функционалу непознате густине придружујемо његову оцену $\hat{\psi}_4(g)$. Овим добијамо оцену за параметар раванања

$$\hat{h}_{DPI} = \left[\frac{R(K)}{k_2^2 \hat{\psi}_4(g) n} \right]^{1/5}. \tag{3.9}$$

Проблем је што она зависи од избора параметра g. Оптимална оцена за тај параметар била би g_{AMSE} међутим то поново зависи од функционала непознате густине. Настављајући даље овај проблем не нестаје, само са проблема оцењивања ψ_r прелазимо на проблем оцењивања ψ_{r+2}. Једно решење овог проблема је да за неко $2l + 4$, ψ_{2l+4} оценимо неком једноставном методом, на пример позивањем на нормалну расподелу, онда ту оцену искористимо за налажење оптималне вредности параметра g оцене $\hat{\psi}_{2l+2}$ и тако даље док не дођемо до оцене за ψ_4, коју онда убацујемо у израз за $\hat{h}_{DPI}$. Ово значи да се l пута рачунају помоћни парамети равнања по формули g_{AMSE}. Питање је колико корака уназад је потребно

ићи, тј. коју вредност изабрати за l. Препоручује се (в. [2]) да l буде најмање 2 и то је и уобичајена вредност за l.
Сада ћемо на једном примеру упоредити описане методе. Како два метода описана у одељку 3.1 не гарантују никакву оптималност, њих ћемо искључити из разматрања. Нека је f густина мешовите нормалне расподеле

$$f(x) = \frac{3}{4}\phi(x) + \frac{1}{4}\phi_{1/3}\left(x - \frac{2}{3}\right), \tag{3.10}$$

где је ϕ густина $\mathcal{N}(0,1)$ расподеле а $\phi_{1/3}$ густина $\mathcal{N}(0,(1/3)^2)$ расподеле. Да бисмо оценили колико је неки метод рачунања ПР добар, требало би упоредити добијене вредности са стварном оптималном вредношћу за ову расподелу. Лако можемо израчунати вредност h_{AMISE}, међутим и ту је примењена апроксимација, то није стварна тачка минимума функције $MISE(h)$. Срећом, у случају мешовите нормалне расподеле и Гаусовог језгра, због појаве конволуција нормалних густина, може се извести општи израз за $MISE(h)$ (в. [2]). За узорак обима 100 из расподеле са густином f то је $h_{MISE} = 0.318$. За 500 узорака обима 100 из ове расподеле израчунаћемо различитим методима параметре равнања. На слици 3.3 приказане су у различитим бојама оцене густина тих параметара равнања. Вертикалном линијом обележена је оптимална вредност h_{MISE}. Најгоре је понашање метода пристрасне унакрсне провере (BCV), а на овом примеру се не види ни значајно већа дисперзија метода непристрасне унакрсне провере (UCV) у односу на BCV. Пристрасност, по којој је BCV и добио име се јасно може видети са слике. Уопште, на узорцима мањег обима методом BCV се увек превише изравна, док UCV метод има велику дисперзију (в. [16]). Добро понашање метода MB (MLCV) је овде очекивано јер се ради о расподели за чије репове важи $f(\pm x) = o(e^{-x})$, $x \to \infty$, иначе са овим методом треба бити опрезан. Као најбољи на слици 3.3, истиче се метод увртене оцене (SJ). У свом раду Парк и Марон (1990) [19] баве се упоређивањем метода за избор ПР. Они тврде да је, под неким јачим условима постављеним над непознатом густином, SJ метод заиста бољи од осталих наведених. Јачи услови су потребни због оцењивања функционала густине.

3.2 Избор језгра

Иако смо у Глави 2 доказали да је језгро K мање утицајна компонента ОГЈ у односу на параметар равнања, уколико је носач непознате густине ограничен са бар једне стране, јавља се проблем при оцењивању густине на граници. Вероватно најчешћи случај је кад је познато да узо-

рак може узимати само позитивне вредности. Тачке из узорка у близини границе ће утицати на то да се одабраним симетричним језгром додели тежина тачкама изван носача. Једно могће решење, које је предложио Шустер (1985) [21] је рефлексија података. Марон, Руперт и Ванд (1991) [23] су предложили трансформисање узорка тако да он буде из неке расподеле са неограниченим носачем, и онда се до оцене долази обичном ОГЈ трансформисаног узорка и инверзном трансформацијом. Марон и Руперт (1994) [22] су ове две идеје објединили, трансформацијом узорка пре рефлектовања ради бољих особина оцене. Други приступ проблему има Чен (1999) [24] и (2000) [25], који предлаже коришћење асиметричних језгара. Уколико је носач густине ограничен, језгро је густина бета расподеле, а ако је ограничен само са једне стране, онда је језгро густина гама расподеле са одговарајућим параметрима. У наредном тексту ћемо укратко описати неке од поменутих метода.

3.2.1 Рефлексија података. Понашање на граници.

Код стандардне оцене густине језгром, у зависности од избора језгра разликоваће се носач оцене. На пример, ако је носач језгра $[-1,1]$ онда је носач оцене густине $[\min_i X_i - h, \max_i X_i + h]$ што може бити изван граница за густину f. Уколико користимо Гаусово језгро оцена ће сигурно узети не нула вредности и изван носача. Претпоставимо да је познато да је носач непознате густине облика $[c, +\infty)$. Једно могуће решење за овај проблем је симетрично пресликавање $\hat{f}_n(x)$ у односу на праву $x = c$, са интервала $(-\infty, c)$ на интервал $[c, \infty)$ и додавање оцени густине у тачкама интервала $(c, +\infty)$. Једначина описане оцене је

$$\begin{aligned} f_n^*(x) &= f_n(x) + f_n(2c - x) \\ &= \frac{1}{n}\sum_{i=1}^{n}[K_h(x - X_i) + K_h(2c - x - X_i)], \end{aligned}$$

за $x \in [c, +\infty)$ и 0 иначе.

Овај метод илустрован је на слици 3.4, где су приказане ОГЈ узорка обима 100 из експоненцијалне $\mathcal{E}(1)$ расподеле са и без рефлектовања узорка. На графику лево види се како оцена густине узима и негативне вредности, док смо на десној слици овај проблем једноставно решили рефлектовањем узорка. У случају да је носач густине облика $[d, +\infty)$ на претходно описани начин се оцењује густина за узорак $Y_1 = -X_1, ..., Y_n = -X_n$, означимо ту оцену са g_n^*. Тражена оцена непознате густине тада је

$$f_n^*(x) = g_n^*(-x),$$

за $x \in (-\infty, d]$ и 0 иначе.

Када је носач густине интервал облика $[c, d]$, метод рефлектовања података сугерише следећу оцену

$$\begin{aligned} f_n^*(x) &= f_n(x) + f_n(2c - x) + f_n(2d - x) \\ &= \frac{1}{n} \sum_{i=1}^{n} [K_h(x - X_i) + K_h(2c - x - X_i) + K_h(2d - x - X_i)], \end{aligned}$$

$x \in [c, d]$ и 0 иначе. У овом случају узима се да је носач језгра $[-1, 1]$, док је у другим случајевима свеједно. Претпоставимо да је носач густине $[0, 1]$. Нека је h фиксирано и $x \in [0, h]$. Тада се x може записати као $x = Ch$, за неко $C \in [0, 1]$. Да бисмо образложили лоше понашање на границама ОГЈ, показаћемо да је очекивана вредност $\hat{f}_n(0)$ једнака $1/2f(0)$, а слично се може извести и за десну границу. Приметимо да, уколико је $f(0) \neq 0$, због претпоставке о носачу, f није непрекидна функција, па неће бити испуњени услови потребни за асимптотску непристрасност коју смо извели у одељку 2.2.

$$\begin{aligned} E\hat{f}_n(x) &= \frac{1}{h} \int_0^1 K\left(\frac{x - y}{h}\right) f(y)\, dy = \frac{1}{h} \int_0^{h(1+C)} K\left(C - \frac{y}{h}\right) f(y)\, dy \\ &= \int_{-1}^{C} K(u) f(x - uh)\, du \\ &\quad + \int_{-1}^{C} K(u)[f(x) - f'(x)uh + \frac{1}{2} f''(x)(uh)^2]\, du + o(h^2) \\ &= f(x)\mu_0(C) - f'(x)h\mu_1(C) + \frac{1}{2} f''(x)h^2\mu_2(C) + o(h^2), \end{aligned}$$

где је $\mu_k(C) = \int_{-1}^{C} u^k K(u)\, du$. За $C \geq 1$ добија се израз за очекивање као у (2.9). За $C = 0$, пошто је $\mu_0(C) = 1/2$ добијамо да је $E\hat{f}_n(0) = 1/2f(0)$. Како је за $C \in [0, 1)$, $\mu_0(C) < 1$, члан који не зависи од h није једнак $f(x)$, па кажемо да је пристрасност *нултог реда*. Методом рефлексије побољшава се понашање оцене на граници. Сличним извођењем добија се

$$\begin{aligned} E\hat{f}_n(x) = f(x) &+ \frac{h^2}{2}\mu_2(1)f''(x) - 2h[C\mu_0(-C) + \mu_1(-C)]f'(x) \\ &+ 2h^2[C^2\mu_0(-C) + C\mu_1(-C)]f''(x) + o(h^2), \end{aligned}$$

за $x = Ch$, $C \in [0, 1]$. За $C \geq 1$ као и у претходном случају, своди се на (2.9), а за $C < 1$ смањили смо одступање од средње вредности јер је оно сада реда $O(h)$ - *првог реда*. У случају када је $f'(0) = 0$, одступање је

другог реда. Ову чињеницу Марон и Руперт (1994) [22] су искористили да направе алгоритам за оцену густине којим се побољшава понашање на крајевима интервала. Укратко ћемо описати кораке у алгоритму.

- Трансфомисање узорка, $Y_i = g(X_i)$. Функција g је одабрана из одређене параметарске фамилије тако да је извод густине Y_i приближно једнак 0 на граници њеног носача.
- Оценити густину на основу узорка Y_i методом рефлексије, $f_Y(x)$
- Користећи везу између f_Y и f_X, $\hat{f}_X(x) = \hat{f}_Y(g(x))g'(x)$, трансформацијом $\hat{f}_X(x) = \hat{f}_Y(g(x))g'(x)$ долазимо до тражене оцене.

3.2.2 Асиметрична језгра

Други могући приступ овом проблему је коришћење других језгара. Симетрична језгра су одговарајућа ако је носач густине $\mathbb{R}$, али нису ако је ограничен слева или здесна. Укратко ћемо описати асиметрична језгра које је предложио Чен. Претпоставимо опет да је носач густине $[0, 1]$ и да f има непрекидан други извод. Обележимо са $K_{p,q}$ густину бета $\beta(p, q)$ расподеле. Осим што носач језгра гарантује да ће и носач оцене густине бити $[0, 1]$, идеја је да се не користи исто језгро у свакој тачки, већ да фукција језгра зависи од тачке у којој оцењујемо густину. Конкретно, као језгро за оцену у тачки $x \in [0, 1]$ узима се функција $K_{x/h+1,(1-x)/h+1}$ где је h параметар равнања за који важи $h \to 0$ кад $n \to \infty$. Оцена густине тада је

$$\hat{f}_B(x) = n^{-1} \sum_{i=1}^{n} K_{x/h+1,(1-x)/h+1}(X_i). \tag{3.11}$$

Дакле ова оцена је као уобичајена ОГЈ само фиксирано језгро мења бета језгрима. Нећемо изводити али ћемо навести асимптотска својства оцене овом методом. Као и обично испитују се пристраност и дисперзија од локалних, а $MISE$ од глобалних својстава.

1. Пристрасност је првог реда, за свако $x \in [0, 1]$, дакле ова оцена је асимптотски непристрасна.
2. Иако су оцене бета језгром непристрасне, њихова дисперзија на граници је реда $n^{-1}h^{-1/2}$ а у унутрашњости интервала је $n^{-1}h^{-1}$.
3. Оптимална вредност h се експлицитно може извести минимизирањем $MISE$ и добија се да треба да важи $h = O(n^{-2/5})$, док је, подсетимо се, код других оцена језгром $h = O(n^{-1/5})$.

4. Заменом оптималног h у $MISE$ добија се $MISE = O(n^{-4/5})$, што је најбоља могућа брзина конвергенције $MISE$ ка нули и за симетрична језгра.

Као побољшање ове оцене, ради додатног смањивања пристрасности на границама, предлаже се још једна оцена која тачкама близу граница додељује бета језгро али са мало другачијим параметрима (в. [24]). На слици 3.5 ради упоређивања приказане су оцена густине бета језгрима, обична ОГЈ и стварна густина. Јасно је како оцена бета језгрима има много боље понашање на границама од обичне ОГЈ.

Остало нам је још да уведемо оцену густине гама језгрима. Претпоставимо да је узорак из расподеле чији је носач $[0, \infty)$. Претпоставимо и да непозната густина f има непрекидни други извод и да су интеграли $\int f'^2(x)\,dx$ и $\int (xf''(x))^2\,dx$ коначни. Бета језгра у овом случају мењамо гама језгрима, да би носач језгара као и у прошлом случају био исти као носач густине. Обележимо са $K_{p,q}$ густину гама $\gamma(p, q)$ расподеле где је q параметар размере. Овом оценом се као језгро за оцену у тачки $x \in [0, \infty)$ узима функција $K_{x/h+1,h}$ где је h параметар равнања за који важи $h \to 0$ и $nh \to \infty$ кад $n \to \infty$. Оцена густине је

$$\hat{f}_G(x) = n^{-1} \sum_{i=1}^{n} K_{x/h+1,h}(X_i). \tag{3.12}$$

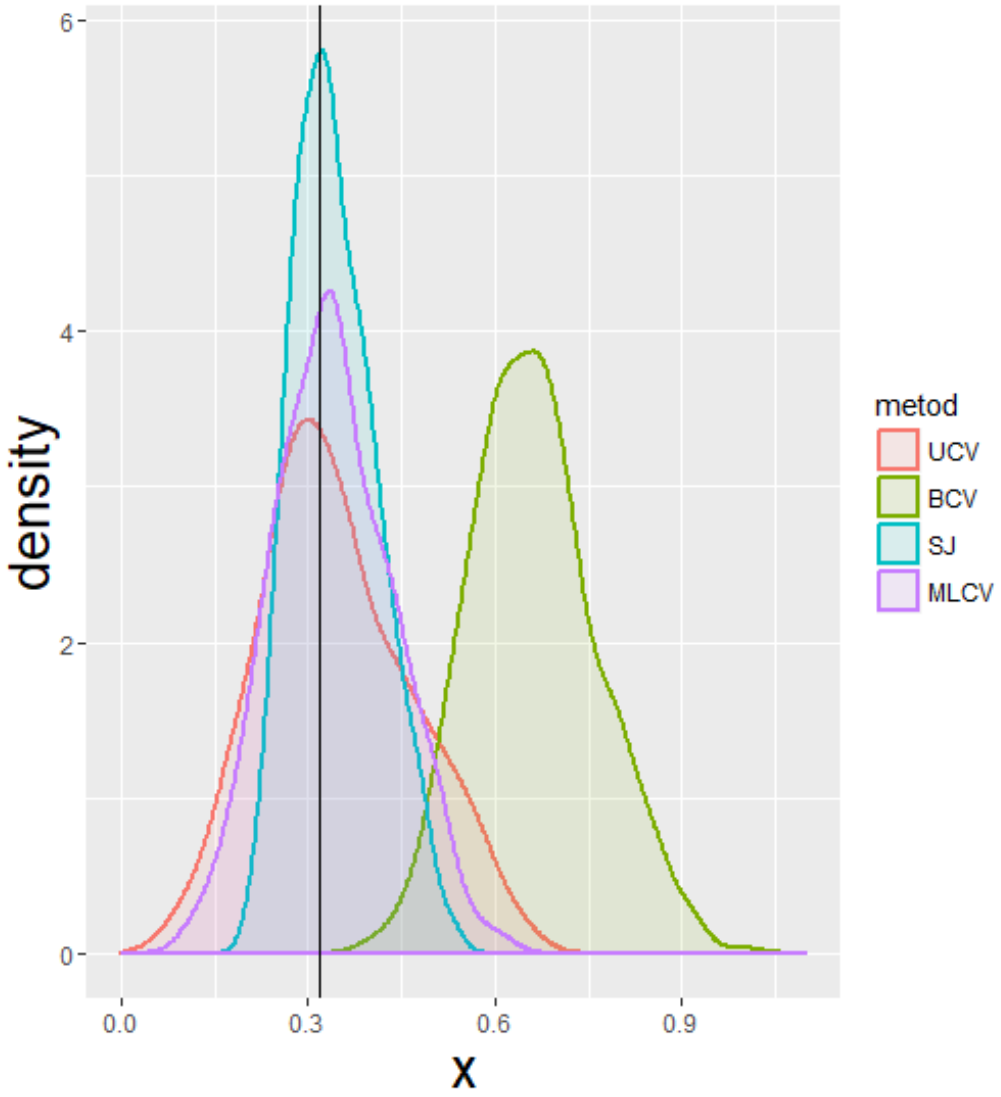
Испоставља се да оцене бета и гама језгрима имају слична асимптотска својства. Све ставке о оцени густине бета језгрима важе у истом облику и за оцену гама језгрима. Такође, слично као и код бета језгара, Чен предлаже побољшање ове оцене коришћењем мало другачијих гама језгара на граници носача. Користе се гама језгра са параметрима $K_{\rho_h(x),h}$, где је

$$\rho_h(x) = \begin{cases} x/h & x \geq 2h \\ \frac{1}{4}(x/h)^2 + 1 & x \in [0, 2h) \end{cases}. \tag{3.13}$$

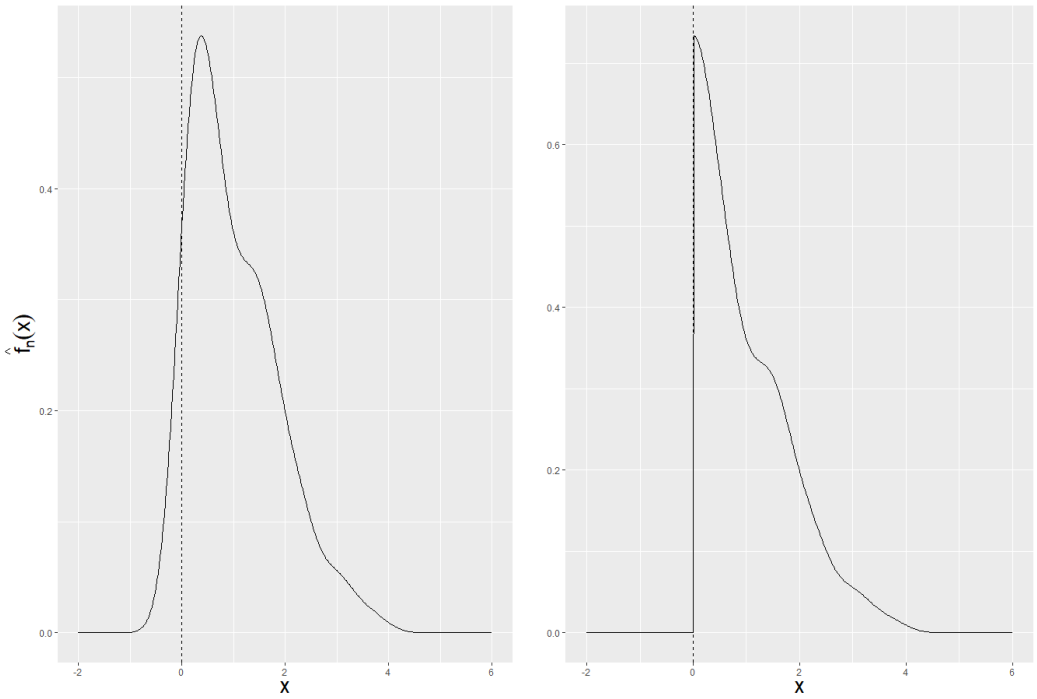
Одговарајућа оцена је

$$\hat{f}_G^{\star}(x) = n^{-1} \sum_{i=1}^{n} K_{\rho_h(x),h}(X_i). \tag{3.14}$$

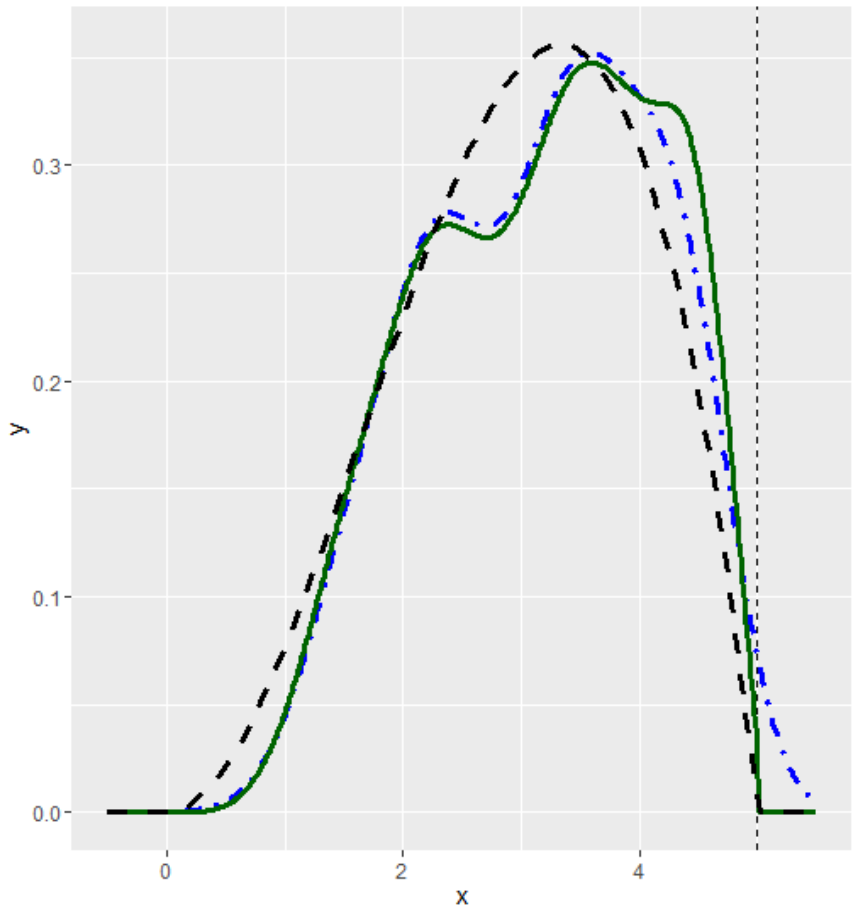
На слици 3.6 ради упоређивања приказане су оцена густине бета језгром $\hat{f}_G$, обична ОГЈ и стварна густина. Јасно је како оцена гама језгрима има много боље понашање на граници од обичне ОГЈ.



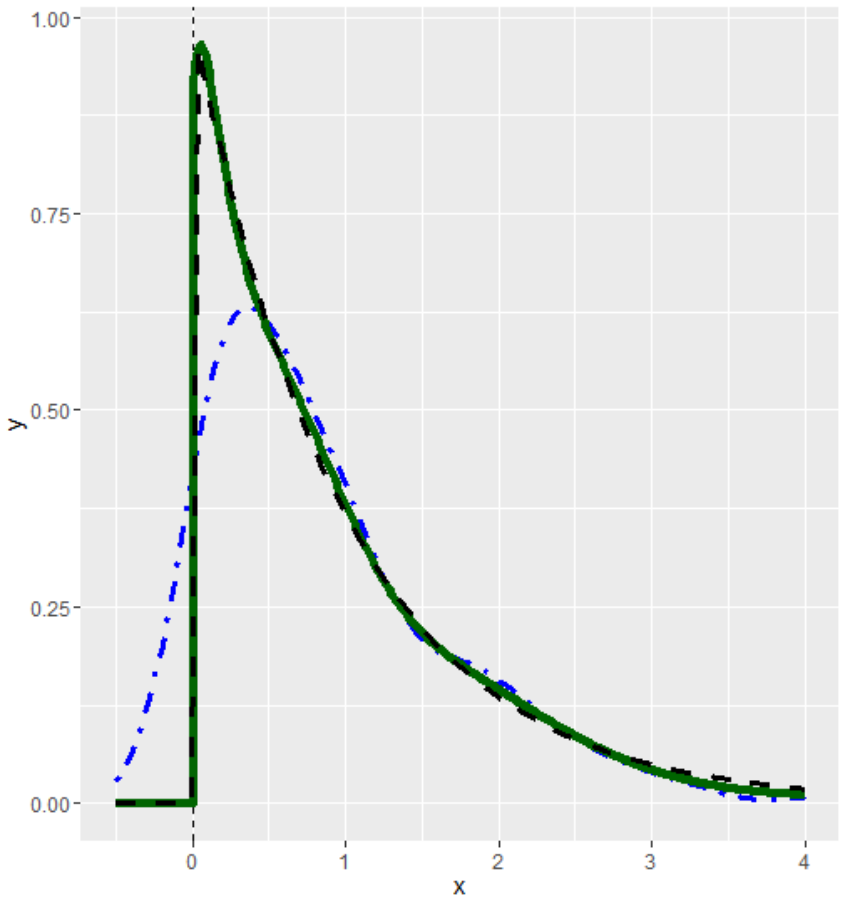
Слика 3.3: Оцењене густине параметара равнања добијених различитим методама, за узорке обима 100 из мешовите нормалне расподеле са густином f_1.



Слика 3.4: оцењена густина без рефлектовања (лево) и са рефлектовањем (десно) података



Слика 3.5: Оцена густине бета језгрима, за узорак обима 100 из скалиране бета расподеле, чији је носач $[0, 5]$. Оцена густине бета језгром (пуна линија), обична ОГЈ (тачка-црта-тачка) и стварна густина (испрекидана линија).



Слика 3.6: Оцена густине гама језгрима, за узорак обима 100 из $\varepsilon(1)$ расподеле. Оцена густине бета језгром (пуна линија), обична ОГЈ (тачка-црта-тачка) и стварна густина (испрекидана линија).

Глава 4

Асимптотско понашање ОГЈ

У одељку 1.4 извели смо асимптотско понашање очекивања и дисперзије ОГЈ. То нам је било потребно за увођење оптималног језгра и описивање више метода избора параметра равнања. У овом одељку навешћемо неке нове асимптотске резултате, а неке већ поменуте ћемо поново доказати али под блажим условима постављеним над густином f. Сва наведена тврђења могу се наћи у [3]. Следећа лема из математичке анализе ће нам омогућити поменуто ослабљивање услова и користи се у доста теорема о асимптотици ОГЈ.

Лема 4.0.1. *Нека је функција $K : (\mathbb{R}, \mathcal{B}) \mapsto (\mathbb{R}, \mathcal{B})$ мерљива функција за коју важи*

$$|K(z)| \leq M, \quad z \in \mathbb{R}; \tag{4.1}$$

$$\int_{\mathbb{R}} |K(z)|\, dz < \infty; \tag{4.2}$$

$$|zK(z)| \to 0, \quad z \to \infty; \tag{4.3}$$

Нека је функција $g : (\mathbb{R}, \mathcal{B}) \mapsto (\mathbb{R}, \mathcal{B})$ мерљива функција за коју важи

$$\int_{\mathbb{R}} |g(z)|\, dz < \infty. \tag{4.4}$$

Дефинишемо и низ функција g_n

$$g_n = \frac{1}{h_n} \int_{\mathbb{R}} K\left(\frac{z}{h_n}\right) g(x - z)\, dz, \tag{4.5}$$

где је $h_n > 0$ и $h_n \to 0$ кад $n \to \infty$. Тада, ако је g непрекидна функција важи

$$\lim_{n\to\infty} g_n(x) = g(x) \int_{\mathbb{R}} K(z)\, dz. \tag{4.6}$$

Ако је додатно g равномерно непрекидна функција, онда је конвергенција у (4.6) равномерна.

Доказ.

$$\begin{aligned}\left|g_n(x)-g(x))\int_{\mathbb{R}}K(z)\,dz\right| \leq & \left|\int_{|z|<\delta}\frac{1}{h_n}K\left(\frac{z}{h_n}\right)g(x-z)dz\right.\\ & \left.-g(x)\int_{|z|<\delta h_n^{-1}}K(z)\,dz\right|\\ & +\left|\int_{|z|>\delta}\frac{1}{h_n}K\left(\frac{z}{h_n}\right)g(x-z)dz\right|\\ & +\left|g(x)\int_{|z|>\delta h_n^{-1}}K(z)\,dz\right|.\end{aligned}$$

Ова неједнакост важи за свако $\delta>0$. Увођењем смене $u=z/h_n$ у првом интегралу првог сабирка, добијамо

$$\begin{aligned}\mathrm{I} &= \left|\int_{|z|<\delta}\frac{1}{h_n}K\left(\frac{z}{h_n}\right)g(x-z)dz-g(x)\int_{|z|<\delta h_n^{-1}}K(z)\,du\right|\\ &\leq \int_{|u|<\delta h_n^{-1}}|K(u)||g(x-uh_n)-g(x)|\,du\\ &\leq \sup_{|u|<\delta}|g(x-u)-g(x)|\int_{|u|<\delta h_n^{-1}}|K(z)|\,du.\end{aligned}$$

Увођењем исте смене у другом сабирку, добијамо

$$\begin{aligned}\mathrm{II} &= \left|\int_{|u|>\delta h_n^{-1}}\frac{1}{h_n}K(u)g(x-uh_n)\,du\right|\\ &\leq \sup_{|u|>\delta h_n^{-1}}\left(\frac{1}{h_n}|K(u)|\right)\int_{|u|>\delta h_n^{-1}}|g(x-uh_n)|\,du\\ &\leq \sup_{|u|>\delta h_n^{-1}}\left(|uK(u)|\right)\delta^{-1}\int_{\mathbb{R}}|g(u)|\,du.\end{aligned}$$

Спајањем ових резултата добијамо

$$\begin{aligned}\left|g_n(x)-g(x)\int_{\mathbb{R}}K(z)\,dz\right| \leq & \sup_{|z|<\delta}|g(x-z)-g(x)|\int_{|z|<\delta h_n^{-1}}|K(z)|\,dz\\ & +\sup_{|z|>\delta h_n^{-1}}\left(|zK(z)|\right)\delta^{-1}\int_{\mathbb{R}}|g(z)|\,dz\\ & +|g(x)|\int_{|z|>\delta h_n^{-1}}|K(z)|\,dz. \qquad (4.7)\end{aligned}$$

Завршићемо доказ одмах у случају равномерне непрекидности функције g, а у случају само обичне непрекидности слично следи. Ако је g равномерно непрекидна функција онда из (4.4) следи да је и ограничена, па за свако $\delta > 0$ из ограничености g и (4.1) следи да последњи сабирак у (4.7) тежи равномерно по x ка 0. Слично, из услова (4.3) и (4.4) следи равномерна конвергенција по x другог сабирка ка 0. Пошто неједнакост (4.7) важи за свако δ, када $\delta \to 0$ из равномерне непрекидности функције g, и услова (4.2), имамо да и први сабирак тежи ка 0, равномерно по x. Овим завршавамо доказ јер смо доказали да је конвергенција у (4.6) равномерна. □

Применом ове леме доказаћемо следећу теорему.

Теорема 4.1. *Нека је f непрекидна густина расподеле. Тада је ОГЈ постојана оцена за $f(x)$, за свако x, ако је језгро K симетрична густина и ако важи $nh_n \to \infty$ и $h_n \to 0$.*

Доказ. Прво ћемо показати асимптотску непристрасност. Сви услови за K из Леме 4.0.1 су испуњени и примењујемо је на језгро K и густину f, која такође задовољава постављени услов о непрекидности.

$$E\hat{f}_n(x) = \int \frac{1}{h_n} K\left(\frac{x-y}{h_n}\right) f(y)\, dy,$$

па директном применом Леме 4.0.1 добијамо да кад $n \to \infty$ очекивање тежи ка

$$f(x) \int K(y)\, dy = f(x).$$

Дакле, довољан услов да ОГЈ буде асимптотски непристрасна је да је густина f непрекидна и да $h_n \to 0$. Још је потребно показати да $D(\hat{f}_n) \to 0, \quad n \to \infty$.

$$\begin{aligned} D(\hat{f}_n) &= \frac{1}{n} D\left(\frac{1}{h_n} K\left(\frac{x-X_1}{h_n}\right)\right) \\ &\leq E\left[\frac{1}{h_n} K\left(\frac{x-X_1}{h_n}\right)\right]^2 \\ &= \frac{1}{nh_n} \int \frac{1}{h_n} K^2\left(\frac{x-y}{h_n}\right) f(y)\, dy, \end{aligned}$$

па поново, применом Леме 4.0.1 на функције K^2 и f добијамо

$$\int \frac{1}{h_n} K^2\left(\frac{x-y}{h_n}\right) f(y)\, dy \to f(x) \int K^2(y)\, dy,$$

када $n \to \infty$. Сви услови за примену леме су задовољени, а специјално из услова (4.2) следи ограниченост последњег интеграла. Према томе

$$D(\hat{f}_n) \to 0, \quad \text{кад } nh_n \to \infty.$$

□

У одељку 1.4 извели смо асимптотско понашање $MISE$, одакле је јасно да $MISE$, као функција од h_n, тежи 0 кад $n \to \infty$. Исти резултат, као што је већ био случај у овом одељку, можемо извести и под слабијим условима.

Теорема 4.2. *Нека за густину f важи да $f, f^{(1)} \in L^2(\mathbb{R})$ и нека њена карактеристична функција ϕ припада $L^1(\mathbb{R})$ и важи $\int \lambda^2 |\phi(\lambda)|^2 \, d\lambda < \infty$. Нека је језгро K ограничено, има коначан други моменат и $h_n \to 0$ кад $n \to \infty$. Тада важи*

$$\int E[\hat{f}_n(x) - f(x)]^2 \, dx \to 0, \quad \textit{кад } n \to \infty. \tag{4.8}$$

Доказ. Користећи неједнакост Коши-Шварца добија се

$$\begin{aligned}
\int E[\hat{f}_n(x) - f(x)]^2 \, dx &= \int \left\{ \int K(t)[f(x - h_n t) - f(x)] \, dt \right\}^2 dx \\
&\leq \int \left\{ \int [f(x - h_n t) - f(x)]^2 K(t) \, dt \right\} dx \\
&= \int K(t) \left\{ \int [f(x - h_n t) - f(x)]^2 \, dx \right\} dt. \qquad (4.9)
\end{aligned}$$

Пошто је, по претпоставци, карактеристична функција ϕ густине f апсолутно интеграбилна, према Теореми 10.6. из [9] веза између ове две функције је

$$f(x) = \frac{1}{2\pi} \int e^{-ix\lambda} \phi(\lambda) \, d\lambda.$$

Дакле, f је Фуријеова трансформација функције ϕ [1]. Према Парсевалової једнакости која по овој дефиницији Фуријеове трансформације гласи $\int |\phi(\lambda)|^2 \, d\lambda = \frac{1}{2\pi} \int f^2(x) \, dx$ имамо да важи

$$\int [f(x - h_n t) - f(x)]^2 \, dx = \int |\phi(\lambda)|^2 \left| e^{ih_n t\lambda} - 1 \right|^2 d\lambda. \tag{4.10}$$

[1] Ово је још један могући начин дефинисања Фуријеове трансформације, а разликује се од оног коришћеног у одељку 2.1

Из (4.10) и (4.9) имамо

$$\int E[\hat{f}_n(x) - f(x)]^2\,dx \le \int \left[\int |\phi(\lambda)|^2 \left|e^{ih_n t\lambda} - 1\right|^2 d\lambda\right] K(t)\,dt$$
$$\le \left[\int t^2 K(t)\,dt\right]\left[\int |\phi(\lambda)|^2 \lambda^2\,d\lambda\right] h_n^2.$$

У последњој неједнакости примењено је $|e^{iu} - 1| \le |u|$. Пошто $h_n \to 0$ кад $n \to \infty$ имамо

$$\int E[\hat{f}_n(x) - f(x)]^2\,dx \to 0. \tag{4.11}$$

Лако се доказује да важи

$$\int D\hat{f}_n(x)\,dx \le \frac{1}{nh_n}\int K^2(x)\,dx, \tag{4.12}$$

одакле следи

$$\int D\hat{f}_n(x)\,dx \to 0, \quad \text{кад } nh_n \to \infty, \tag{4.13}$$

па из (4.11) и (4.13) следи тражена конвергенција ка 0. □

До сада смо као меру одступања ОГЈ од стварне густине испитивали непристрасност и постојаност тачка по тачка, а као једину меру одступања оцене на целом интервалу увели смо $MISE$, очекивање квадрата растојања $\hat{f}_n$ од f у L^2 мери. L^2 теорија је у најразвијенија и најчешће помињана у литератури. Уместо ове може се користити и L^1 мера, и тако добијамо средњу интегрисану апсолутну грешку,

$$MIAE(\hat{f}_n) = E\int \left|\hat{f}_n(x) - f(x)\right| dx.$$

За разлику од L^2, испитивање L^1 мере одступања је значајно компликованије. Предности ове глобалне мере у односу на L^2 и многи резултати L^1 теорије могу се наћи у [7]. Осим ове, могу се испитивати још неке мере одступања, на пример, јака постојаност, равномерна асимптотска непристрасност, равномерна постојаност (слаба и јака). Сада ћемо доказати и равномерну јаку постојаност ОГЈ.

Теорема 4.3. *Нека је језгро K ограничене варијације и ред $\sum_{i=1}^{\infty} exp(-\gamma n h_n^2)$ конвергира за свако $\gamma > 0$. Тада*

$$V_n = \sup_x |f_n(x) - f(x)| \to 0,$$

скоро сигурно када $n \to \infty$, ако је густина f равномерно непрекидна.

Доказ. Дефинишемо

$$\begin{aligned}
\overline{V}_n &= \sup_x \left| \hat{f}_n(x) - E[\hat{f}_n(x)] \right| \\
&= \sup_x \left| \frac{1}{h_n} \int K\left(\frac{x-y}{h_n}\right) d\hat{F}_n(y) - \frac{1}{h_n} \int K\left(\frac{x-y}{h_n}\right) dF(y) \right| \\
&= \left| K\left(\frac{x-y}{h_n}\right) (\hat{F}_n(y) - F(y)) \Big|_{y=-\infty}^{y=+\infty} - \int (\hat{F}_n(y) - F(y))\, dK\left(\frac{x-y}{h_n}\right) \right| \\
&\leq \sup_x \left| \hat{F}_n(x) - F(x) \right| \frac{1}{h_n} V(K), \qquad (4.14)
\end{aligned}$$

где је $V(K)$ тотална варијација функције K. Означимо

$$D_n = \sup_x \left| \hat{F}_n(x) - F(x) \right|.$$

Доказано је у [13] да постоје константе C и $\alpha \in (0, 2]$ такве да важи

$$P\{D_n > \lambda n^{-1/2}\} \leq C \exp(-\alpha\lambda^2), \qquad (4.15)$$

за свако $\lambda > 0$ и сваку непрекидну функцију расподеле F. Из (4.14) и (4.15) имамо

$$\begin{aligned}
P\left(\sup_x \left| \hat{f}_n(x) - E\hat{f}_n(x) \right| > \varepsilon \right) &\leq P\left(\sup_x \left| \hat{F}_n(x) - F(x) \right| > \varepsilon h_n (V(K))^{-1} \right) \\
&\leq C_1 \exp\left(-\beta n h_n^2\right),
\end{aligned}$$

где је $\beta = \alpha\varepsilon^2 (VK)^{-2}$. Из претпоставке да ред $\sum_{n=1}^{\infty} \exp(-\gamma n h_n^2)$ конвергира за свако $\gamma > 0$, имамо да је $\sum_{i=1}^{\infty} P\left(\sup_x \left| \hat{f}_n(x) - E\hat{f}_n(x) \right| > \varepsilon\right) < \infty$ па из Теореме 12.2 о довољном услову за скоро сигурну конвергенцију из [9] следи

$$\overline{V}_n \to 0 \quad \text{с.с.} \quad n \to \infty. \qquad (4.16)$$

Ако је f равномерно непрекидна, пошто је и апсолутно интеграбилна, онда је и ограничена, $\sup_x f(x) = M < \infty$. Из већ доказане асимптотске непристрасности имамо $E\hat{f}_n(x) \to f(x)$, кад $n \to \infty$, а ако је f равномерно непрекидна онда је та конвергенција и равномерна. То значи да

$$\sup_x \left| E\hat{f}_n(x) - f(x) \right| \to 0 \quad \text{кад } n \to \infty. \qquad (4.17)$$

Из неједнакости троугла и (4.16) и (4.17) имамо коначно да

$$V_n \to 0 \quad \text{с.с кад} \quad n \to \infty.$$

□

Напоменимо да важи и обрат ове теореме (в. [3]).

Сада ћемо одредити услове под којима можемо на ОГЈ применити централну граничну теорему. Оцену можемо написати и у следећем облику

$$f_n(x) = n^{-1} \sum_{k=1}^{n} V_{nk}, \qquad V_{nk} = \frac{1}{h_n} K\left(\frac{x - X}{h_n}\right),$$

где су V_{nk} независне случајне величине са расподелом као

$$V_n = \frac{1}{h_n} K\left(\frac{x - X_i}{h_n}\right).$$

Да би важила ЦГТ довољно је да важи Љапуновљева теорема, што значи да је потребно да $E|V_n|^{2+\delta} < \infty$, за неко $\delta > 0$, и да важи

$$\frac{E|V_n - EV_n|^{2+\delta}}{n^{\delta/2}\sigma^{2+\delta}(V_n)} \to 0, \quad \text{кад } n \to \infty. \tag{4.18}$$

где је $\sigma^r(X) = \int |X - EX|^r \, dF(x)$ за случајну величину X са функцијом расподеле F, и $r \in (0, \infty)$. Следеће наводимо асимптотско понашање $\sigma^2(V_n)$ и $E|V_n|^{2+\delta}$ које се се изводи применом Леме 4.0.1.

$$\sigma^2(V_n) \sim \frac{1}{h_n} f(x) \int_{\mathbb{R}} K(y)^2 \, dy; \tag{4.19}$$

$$E|V_n|^{2+\delta} = \int_{\mathbb{R}} \left| \frac{1}{h_n} K\left(\frac{x - y}{h_n}\right) \right| f(y) \, dy \sim \frac{1}{h_n^{1+\delta}} f(x) \int_{\mathbb{R}} |K(y)|^{2+\delta} \, dy. \tag{4.20}$$

Израз 4.18 можемо записати као

$$\frac{h^{1+\delta} E|V_n - EV_n|^{2+\delta}}{(nh)^{\delta/2} h^{1+\delta/2} \sigma^{2+\delta}(V_n)}. \tag{4.21}$$

На основу неједнакости Минковског имамо

$$\begin{aligned}
(E|V_n - EV_n|^p)^{1/p} &\le (E|V_n|^p)^{1/p} + (E|-EV_n|^p)^{1/p} \\
&\le (E|V_n|^p)^{1/p} + (E(E|V_n|)^p)^{1/p} \\
&= (E|V_n|^p)^{1/p} + E|V_n| \\
&\le (E|V_n|^p)^{1/p} + (E|V_n|^p)^{1/p} \\
&\le 2(E|V_n|^p)^{1/p},
\end{aligned}$$

где претпоследња неједнакост следи из примењене Хелдерове неједнакости. Одавде, за $p = 2 + \delta$ имамо

$$E|V_n - EV_n|^{2+\delta} \le 2^{2+\delta} E|V_n|^{2+\delta}. \tag{4.22}$$

Када заменимо (4.19) у (4.18) и искористимо процену (4.22), уз услове $nh_n \to \infty$ и $\int_{\mathbb{R}} |K(y)|^{2+\delta}\, dy < \infty$ добијамо да је Љапуновљев услов испуњен.

Глава 5

Примена ОГЈ на тестирање симетрије

Тестирање симетрије расподеле је популаран проблем непараметарске статистике. Први тестови симетрије настали су још почетком 20. века и засновани су на разним својствима симетрије. Коришћени су, на пример, узорачки моменти и распоред статистика поретка. Увођење оцене густине језгром дало је нове могућности за приступ овом проблему и у овом поглављу ћемо описати неке предложене тестове симетрије.

Тестови које ћемо овде разматрати баве се упоређивањем теоријских и узорачких коефицијената преклапања. За почетак ћемо дефинисати могуће коефицијенте преклапања између две функције густине расподеле. Нека су f и g две густине. Предложени коефицијенти преклапања (мере сродности) су:

$$\text{Матусита [26]:} \quad \rho = \int \sqrt{f(x)g(x)}\,dx; \tag{5.1}$$

$$\text{Ахмад, Ван Бел [27]:} \quad \lambda = \frac{2\int f(x)g(x)\,dx}{\int f^2(x)\,dx + \int g^2(x)\,dx}; \tag{5.2}$$

$$\text{Вајцман [28]:} \quad \theta = \int \min\{f(x), g(x)\}\,dx. \tag{5.3}$$

Сваки од наведених коефицијената, за произвољне функције густине, узима вредности у интервалу $(0, 1]$, где се вредност 1 достиже ако и само ако је $f(x) = g(x)$ с.с.. Још једна мера блискости, са мало другачијим особинама од поменутих, је квадрат L^2 растојања између непознатих густина

$$I = \int (f(x) - g(x))^2\,dx. \tag{5.4}$$

Важи да је $I \geq 0$ и $I = 0$ акко $f(x) = g(x)$ с.с.. За тестирање једнакости две непознате густине формирамо нулту и алтернативну хипотезу

$$\begin{aligned} H_0 &: f(x) = g(x) \quad \text{с.с.} \\ H_1 &: f(x) \neq g(x) \quad \text{за } x \text{ из скупа мере веће од } 0 \end{aligned}$$

У потрази за тест статистиком има смисла посматрати неку од дефинисаних мера сродности. Уместо непознатих густина, можемо користити оцене густина језгром, и тако заменом $f(x)$ са $\hat{f}(x)$, а $g(x)$ са $\hat{g}(x)$ добијамо оцене ових мера. Уколико је нулта хипотеза тачна, вредност сваког од понуђених коефицијената је 1, осим за меру I где је 0, међутим не знамо шта се дешава са узорачким панданима, нити коју ће они расподелу имати. Могуће тест статистике и њихове расподеле у случају тестирања једнакости густина два независна узорка, за меру λ (5.2), испитивали су Фан и Генчај (1993) [30], а за меру θ (5.3) испитивали су Андерсон, Линтон и Ванг (2009) [33]. Осим за тестирање једнакости две густине, ове мере можемо искористити и за тестирање симетрије расподеле једног узорка. Формирамо нулту и алтернативну хипотезу

$$\begin{aligned} H_0' &: f(x) = f(-x) \quad \text{с.с.} \\ H_1' &: f(x) \neq f(-x) \quad \text{за } x \text{ из скупа мере веће од } 0 \end{aligned}$$

Идеја је иста као у претходном случају, само сада мењамо $\hat{g}(x)$ са $\hat{f}(-x)$ и тако добијамо могуће тест статистике за тестирање симетрије густине. У поменутим радовима [30] и [32], изведене су асимптотске расподеле ових тест статистика и у случају тестирања симетрије. Слично, по две тест статистике за тестирање симетрије на основу мере сродности (5.4) предложили су Ахмад и Ли (1997) [32], и Фернандес, Мендес и Скаје (2015) [35]. У наредна три одељка навешћемо тест статистике, њихове асимптотске расподеле и оценити мере и моћи предложених тестова Монте Карло методом.

5.1 Мера сродности λ

Меру λ можемо записати и у следећем облику

$$\lambda = \frac{2\delta}{\Delta(f) + \Delta(g)},$$

где је $\delta = \int f(x)g(x)\,dx$, $\Delta(f) = \int f^2(x)\,dx$, $\Delta(g) = \int g^2(x)\,dx$. Као оцену за λ, Ахмад (1980) [29] предлаже

$$\hat{\lambda} = \frac{2\hat{\delta}}{\hat{\Delta}(f) + \hat{\Delta}(g)},$$

где је $\hat{\delta} = \frac{1}{2}\left[\int \hat{f}(x)\, d\hat{G}_n(x) + \int \hat{g}(x)\, d\hat{F}_n(x)\right]$, $\hat{\Delta}(f) = \int \hat{f}^2(x)\, dx$, $\hat{\Delta}(g) = \int \hat{g}^2(x)\, dx$. Уместо $\hat{\Delta}$ као оцену за Δ можемо користити и

$$\tilde{\Delta} = \int \hat{f}(x)\, d\hat{F}_n(x) = (n^2 h)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} K\left((X_i - X_j)/h\right).$$

Доказана је постојаност у средње квадратном смислу оцене $\hat{\Delta}$ (в. [20]), а доказано је и да под истим условима она важи и за оцену $\tilde{\Delta}$ (в. [29]).

Осим тестирања једнакости непознатих густина два узорка, могућа су и тестирања хипотезе о симетрији густине расподеле на једном узорку. Заменом $g(x)$ са $f(-x)$ и $\hat{g}(x)$ са $\hat{f}(-x)$, добијамо израз за λ^* и оцену $\hat{\lambda}^*$

$$\lambda^* = \frac{\delta^*}{\Delta(f)};$$
$$\hat{\lambda}^* = \frac{\hat{\delta}^*}{\hat{\Delta}(f)},$$

где је $\delta^* = \int f(x) f(-x)\, dx$, а $\hat{\delta}^* = (n^2 h)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} K((X_i + X_j)/h)$. Ахмад (1980) [29] је доказао да под неким условима $\hat{\lambda}$ и $\hat{\lambda}^*$ имају нормалну расподелу, међутим Фан и Генчај (1993) [30] су ово оповргнули доказом да је $\sqrt{n}(\hat{\lambda} - 1) = o_P(1)$, и да исто важи и за $\hat{\lambda}^*$. У ово се можемо уверити симулацијама. У табели 5.1 приказан је распон узорка статистике $\sqrt{n}(\hat{\lambda} - 1)$ израчунате за узорке из нормалне расподеле, за 100 понављања теста и различите обиме узорка. Види се да је са порастом обима узорка распон све мањи. Слични резултати добију се и за друге симетричне расподеле.

n	min	max	max − min
1000	-0.953	0.248	1.201
2000	-0.373	0.147	0.520
3000	-0.380	0.063	0.443
4000	-0.331	0.042	0.373

Табела 5.1: Распон узорка статистике $\sqrt{n}(\hat{\lambda} - 1)$ израчунате за узорке из нормалне расподеле, за 100 понављања теста.

Дакле, ова статистика се не може користити за конструисање асимптотских тестова симетрије. Такође, они тврде да не постоји низ a_n којим би

се помножило $\hat{\lambda} - 1$ тако да гранична расподела те статистике буде нека од познатих расподела. Они предлажу следећу модификацију оцене $\hat{\lambda}^*$,

$$\hat{\lambda}^*_\gamma = \frac{\hat{\delta}^*_\gamma}{\hat{\Delta}(f)},$$

где је $\hat{\delta}^*_\gamma = n_\gamma^{-1} \sum_{i=1}^n C_i(\gamma) \hat{f}(-X_i)$ и

$$C_i(\gamma) = \begin{cases} 1+\gamma & \text{за непарно } i \\ 1-\gamma & \text{за парно } i \end{cases}; \qquad n_\gamma = \begin{cases} n & \text{за парно } n \\ n+\gamma & \text{за непарно } n \end{cases},$$

и $\gamma \in (0, 1]$. За $\gamma = 0$ ова оцена се своди на Ахмадову оцену $\hat{\lambda}^*$. Навешћемо, без доказа, теорему о асимптотској расподели ове оцене. Дефинишемо услове:

(K) K је симетрична и ограничена густина расподеле и важи $|x|K(x) \to 0$ кад $|x| \to \infty$, $\int xK(x)\,dx = 0$, $\int x^2 K(x)\,dx < \infty$.

(F) Густина f је два пута диференцијабилна са ограниченим другим изводом и $\int f^4(x)\,dx < \infty$.

(H) $h \to 0$, $nh^2 \to \infty$, $nh^4 \to 0$ кад $n \to \infty$.

Теорема 5.1. (Фан, Генчај (1983) [30]) *Ако су испуњени услови* (K), (F), (H), *при* H_0', *важи*

$$\sqrt{n}(\hat{\lambda}^*_\gamma - 1) \xrightarrow{D} \mathcal{N}(0, \sigma^{*2}_{\gamma,0}),$$

где је

$$\sigma^{*2}_{\gamma,0} = \frac{\gamma^2 \sum_{i=1}^n \left[\hat{f}(X_i) - \hat{\delta}^*_\gamma\right]^2}{n\hat{\Delta}^2(f)}.$$

Нека је $\hat{T}_{\gamma,*} := \sqrt{n}(\hat{\lambda}^*_\gamma - 1)/\sigma^{*2}_{\gamma,0}$. На основу теореме, $\hat{T}_{\gamma,*}$ је наша тест статистика која има асимптотски стандардну нормалну расподелу. Критичне су нам мале вредности $\hat{\lambda}^*_\gamma$, па при нивоу значајности α, нулту хипотезу одбацујемо ако је $\hat{T}_{\gamma,*}(\boldsymbol{x}) < -z_\alpha$, где је z_α горњи α-ти квантил $\mathcal{N}(0,1)$ расподеле.

Тест статистика зависи од параметра γ ком треба доделити неку вредност у симулацијама. Фан и Генчај (1995) [31] су симулацијама нашли да за обиме узорака $n = 50$ и $n = 100$ тестови дају најбоље резултате када је $0.55 < \gamma < 0.70$, а конкретно се одлучују за $\gamma = 0.65$ што ћемо и ми урадити.

Приметимо да не можемо користити неке од познатих начина за избор параметра равнања, јер је тада $h = O(n^{-1/5})$ па неће бити испуњен трећи у низу услова Н. Фан и Генчај (1995) [31] предлажу параметар равнања облика

$$h = \eta\sigma n^{-1/3},$$

где је η константа и σ стандардна девијација узорка. Они такође предлажу и вредност $\eta = 1.7$, јер се симулацијама показало да ова вредност даје добре резултате.

Расподела	5%	10%	m	$\overline{s}_n$
		$n = 50$		
$\mathcal{N}(0,1)$	0.106	0.132	0.345	2.384
t_5	0.049	0.074	0.825	2.104
Лапласова	0.029	0.051	1.163	1.804
		$n = 100$		
$\mathcal{N}(0,1)$	0.094	0.140	0.340	1.925
t_5	0.05	0.070	0.837	2.001
Лапласова	0.014	0.031	1.302	1.640
		$n = 200$		
$\mathcal{N}(0,1)$	0.068	0.108	0.242	1.587
t_5	0.021	0.040	0.820	1.441
Лапласова	0.008	0.015	1.303	1.373
		$n = 500$		
$\mathcal{N}(0,1)$	0.062	0.103	0.138	1.259
t_5	0.03	0.046	0.670	1.289
Лапласова	0.003	0.011	1.257	1.131

Табела 5.2: Оцењена средња вредност и стандардна девијација статистике $\hat{T}_{\gamma,*}$ и мере предложеног теста симетрије са 1000 понављања.

У табели 5.2 приказани су резултати тестирања симетричних расподела, $\mathcal{N}(0,1), t_5$ и Лапласове$(0,1)$. Рађено је са узорцима обима 50, 100, 200 и 500, и са 1000 понављања. Осим приказаних резултата, добија се да $\hat{\lambda}^*_\gamma$ има расподелу померену улево, али се коефицијент асиметрије приближава нули са повећањем узорка. Видимо да се у зависности од расподеле разликују оцене грешке прве врсте, највеће су за $\mathcal{N}(0,1)$, али са порастом обима узорка се и оне приближавају теоријским вредностима. Како је овај тест направљен за тестирање симетрије око 0, алтернативне хипотезе смо тестирали тако што смо центрирали одговарајуће асиме-

Расподела	5%	10%	m	$\overline{s}_n$
		$n = 50$		
$\chi^2(2)$	0.765	0.808	-3.127	4.503
Логнормална	0.757	0.786	-3.246	9.766
$\mathcal{E}(1)$	0.777	0.819	-3.234	4.726
		$n = 100$		
$\chi^2(2)$	0.973	0.982	-5.843	4.153
Логнормална	0.938	0.946	-6.233	9.061
$\mathcal{E}(1)$	0.965	0.977	-5.830	4.026

Табела 5.3: Оцењена средња вредност и стандардна девијација статистике $\hat{T}_{\gamma,*}$ и моћи предложеног теста симетрије са 1000 понављања.

тричне расподеле тако да им средња вредност буде 0. Те расподеле су $\chi^2(2) - 2$, $\mathcal{E}(1) - 1$ и Логнормална$(0, 1) - e^{1/2}$. У табели 5.3 дати су резултати тестирања. За узорке обима 200 моћи тестова су већ јако близу 1.

5.2 Мера сродности I

Ахмад и Ли (1997) [32] посматрају функционал

$$\begin{aligned} I &= \frac{1}{2}\int [f(x) - f(-x)]^2\, dx \\ &= \int [f(x) - f(-x)]\, dF(x). \end{aligned}$$

Оцена овог функционала методом замене је

$$\begin{aligned} \hat{I}_n &= \int [\hat{f}(x) - \hat{f}(-x)]\, dF_n(x) \\ &= (n^2 h)^{-1} \sum_{i=1}^{n} \left[K(0) - K\left(\frac{2X_i}{h}\right)\right] \\ &\quad + (n^2 h)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{n} \left[K\left(\frac{X_i - X_j}{h}\right) - K\left(\frac{X_i + X_j}{h}\right)\right] \\ &= \hat{I}_{1n} + \hat{I}_{2n}. \end{aligned}$$

Навешћемо, без доказа, теорему о асимптотској расподели ове оцене. Дефинишемо услове:

(K) K је симетрична и ограничена густина расподеле и важи $|x|K(x) \to 0$, кад $|x| \to \infty$, $\int xK(x)\,dx = 0$, $\int x^2K(x)\,dx < \infty$.

(F) Густина f је ограничена и непрекидна на $\mathbb{R}$.

(H) $h \to 0$, $nh \to \infty$, кад $n \to \infty$.

Теорема 5.2. Ахмад, Ли (1997) [32] *Ако су испуњени услови* (K), (F), (H), *при* H_0' *важи*

$$nh^{1/2}(\hat{I}_n - (nh)^{-1}K(0))/(2\hat{\sigma}) \xrightarrow{D} \mathcal{N}(0,1), \tag{5.5}$$

где је $\hat{\sigma}^2 = R(K)\sum_{i=1}^{n}\hat{f}(X_i)$.

	J_{2n}				J_n			
Расподела	5%	10%	m	$\overline{s}_n$	5%	10%	m	$\overline{s}_n$
				$n = 50$				
$\mathcal{N}(0,1)$	0.035	0.56	-0.018	0.515	0.025	0.04	-0.263	0.464
t_5	0.025	0.043	-0.042	0.433	0.016	0.031	-0.298	0.450
Лапласова	0.021	0.037	-0.032	0.337	0.02	0.038	-0.289	0.401
				$n = 100$				
$\mathcal{N}(0,1)$	0.035	0.61	-0.016	0.516	0.024	0.041	-0.208	0.625
t_5	0.036	0.059	-0.030	0.472	0.026	0.04	-0.293	0.469
Лапласова	0.029	0.048	-0.005	0.413	0.032	0.044	-0.272	0.534

Табела 5.4: Оцењена средња вредност и стандардна девијација статистика J_n и J_{2n} и мере предложених тестова симетрије са 1000 понављања.

Осим ове тест статистике, као корак у извођењу њене асимптотске расподеле, доказује се да под истим условима $nh^{1/2}\hat{I}_{2n}/2\hat{\sigma}$ има стандардну нормалну расподелу, тако да и ово може бити тест статистика. Уведимо ознаке $J_n := nh^{1/2}(\hat{I}_n - (nh)^{-1}K(0))/(2\hat{\sigma})$ и $J_{2n} := nh^{1/2}\hat{I}_{2n}/2\hat{\sigma}$. За параметар равнања одабрали смо $\hat{h}_{Scott}$ описан у одељку 3.1.1. Критичне су велике вредности тест статистика J_n и J_{2n} тако да нулту хипотезу одбацујемо ако је $J_n > z_\alpha$, односно $J_{2n} > z_\alpha$. Рађено је са узорцима обима 50 и 100, и 1000 понављања. У табели 5.4 дати су резултати при тестирању расподела из нулте хипотезе. Видимо да J_n има већу негативну пристрасност него J_{2n}, па је препоручено користити J_{2n} као тест

	J_{2n}				J_n			
Расподела	5%	10%	m	$\overline{s}_n$	5%	10%	m	$\overline{s}_n$
			$n = 50$					
χ^2_2	0.880	0.916	4.705	7.274	0.856	0.901	4.531	7.432
Логнормална	0.988	0.993	7.549	7.776	0.984	0.991	7.452	7.966
$\mathcal{E}(1)$	0.893	0.918	4.804	6.640	0.856	0.897	4.565	7.327

Табела 5.5: Оцењена средња вредност и стандардна девијација статистика J_n и J_{2n} и моћи предложених тестова симетрије са 1000 понављања.

статистику. Осим тога, оцењене мере су ближе онима за нормалну расподелу у случају J_{2n} статистике. У табели 5.5 дате су оцењене моћи теста за расподеле као у претходном одељку. Навели смо само резултате за узорак обима 50, пошто су су за узорак обима 100 моћи већ јако близу 1. У односу на претходни, овај тест има мере ближе теоријским, као и веће моћи за узорке мањег обима.

5.3 Мера сродности θ

Као оцену ове мере средности узимамо

$$\hat{\theta} = \int \min\{\hat{f}(x), \hat{g}(x)\}\, dx. \tag{5.6}$$

Андерсон, Линтон и Ванг [33] су извели две теореме. Навешћемо прву теорему у случају када је испуњена хипотеза H_0. Уведимо ознаке:

$$a_n = h^{-1/2}||K||_2 E(\min\{Z_1, Z_2\}) \int f^{1/2}(x)\, dx;$$

$$\sigma_0^2 = ||K||_2^2 \int_{-1}^{1} cov\Big(\min\{Z_1, Z_2\}, \min\Big\{\rho(t)Z_1 + \sqrt{1-\rho(t)^2}Z_3, \rho(t)Z_2 + \sqrt{1-\rho(t)^2}Z_4\Big\}\Big)\, dt,$$

где су Z_1, Z_2, Z_3, Z_4 независне случајне величине са $\mathcal{N}(0,1)$ расподелом. Услове под којима следећа теорема важи нећемо наводити због њихове гломазности.

Теорема 5.3. (Андерсон, Линтон, Ванг (2009) [33]) *Нека су испуњени услови A1-A5 (в. [33]). Тада при H_0' важи*

$$\sqrt{n}(\hat{\theta} - \theta) - a_n \xrightarrow{D} \mathcal{N}(0, \sigma_0^2).$$

Означимо $U := \sqrt{n}(\hat{\theta} - \theta) - a_n$. Ову тест статистику можемо применити и на тестирање симетрије. Претпоставимо да је позната могућа тачка симетрије. Без умањења општости претпоставимо да је то 0. Узорак ћемо поделити на два дела, позитивне и негативне елементе. Нека су $X_1, ..., X_{n_1}$ позитивни а $Y_1, ..., Y_{n_2}$ апсолутне вредности негативних елемената узорка. Посебно ћемо оценити језгром густине за ова два узорка, а онда ћемо симетрију расподеле тестирати применом Теореме 5.3. Нека је $\hat{f}_+$ оцењена густина позитивних а $\hat{f}_-$ оцењена густина апсолутних вредности негативних елемената узорка. Један од услова теореме је да је носач језгра $[-1/2, 1/2]$, па ћемо за језгро узети густину униформне $U[-0.5, 0.5]$ расподеле, а параметар равнања бирамо позивањем на нормалну расподелу, и он за ово језгро износи $h_n = 1.84\hat{\sigma}n^{-1/5}$. За ова два узорка имаћемо различите параметре равнања. Претходна теорема је изведена под претпоставком да су два узорка једнаког обима, међутим и у случају када они нису истог обима слично важи, са променама у a_n и σ_0^2. Нека је $n_2/n_1 \to \tau \in (0, \infty)$. Уместо пристрасности a_n, пристрасност је $a_n\frac{1+1/\tau}{2}$, а дисперзија је уместо σ_0^2, $\sigma_0^2(1 + 1/\tau)/2$. Ови резултати се могу наћи у [33], као и то да је за случај униформног језгра $\sigma_0^2 = 0.6135$. Једини проблем који је остао је оцењивање интеграла $\int f_+^{1/2}(x)\,dx$ који се налази у изразу за a_n, где је f_+ стварна вредност густине позитивних елемената $X_1, ..., X_{n_1}$. Ову вредност оценићемо са $\int \hat{f}_+^{1/2}(x)\,dx$. Овај, као и интеграл за рачунање $\hat{\theta}$ рачунамо методом трапеза.

Критичне су нам мале вредности тест статистике, тако да нулту хипотезу одбацујемо ако је $U < -z_\alpha$. За узорке обима 50, 100, 200 и 500 и 1000 понављања оценићемо грешке прве врсте и моћи теста. У табели 5.6 дати су резултати симулација при тестирању узорака из нулте хипотезе. Видимо да су оцењене грешке прве врсте много мање од оних за нормалну расподелу и да средња вредност, иако се смањује повећањем обима узорка, није блиска 0. Осим резултата наведених у табели 5.6, кофицијенти асиметрије при нултој хипотези су негативни. У табели 5.7 дати су резултати симулација при тестирању узорака из алтернативних хипотеза, за исте несиметричне расподеле као раније. Моћи теста су јако мале за узорак обима 50, али са порастом узорка расту и достижу јединицу. Лоше понашање теста на мањем обиму узорка може бити последица тога што један битан услов ове теореме није испуњен, а то је независност два узорка на која се она примењује.

Расподела	5%	10%	m	$\overline{s}_n$
		$n = 50$		
$\mathcal{N}(0,1)$	0.001	0.003	0.166	0.131
t_5	0	0.001	0.194	0.099
Лапласова	0	0.002	1.193	0.107
		$n = 100$		
$\mathcal{N}(0,1)$	0	0.002	0.151	0.135
t_5	0.001	0.001	0.186	0.110
Лапласова	0	0.001	0.209	0.12
		$n = 200$		
$\mathcal{N}(0,1)$	0.001	0.005	0.156	0.141
t_5	0	0.002	0.170	0.138
Лапласова	0	0.004	0.173	0.162
		$n = 500$		
$\mathcal{N}(0,1)$	0.002	0.004	0.117	0.198
t_5	0.001	0.002	0.177	0.192
Лапласова	0	0.004	0.162	0.184
		$n = 1000$		
$\mathcal{N}(0,1)$	0.002	0.005	0.099	0.247
t_5	0	0.005	0.129	0.315
Лапласова	0.001	0.005	0.133	0.283

Табела 5.6: Оцењена средња вредност и стандардна девијација статистике U и мере предложеног теста симетрије са 1000 понављања.

5.4 Асиметрична језгра и мера сродности I

Можемо приметити да је још једна мана претходног тестирања то што за оцењивање густине два подузорка користимо симетрична језгра - њима смо ограничени због услова Теореме 5.3. Узорци $X_1, ..., X_{n_1}$ и $Y_1, ..., Y_{n_2}$ могу узимати само позитивне вредности, па би то требало да важи и за њихове оцењене густине. За крај, описаћемо још један приступ тестирању симетрије уз помоћ ОГЈ којим се овај проблем превазилази. Фернандес, Мендес и Скаје (2015) [35], су извели граничне расподеле две тест статистике за тестирање симетрије које се такође заснивају на мери одступања I. Као и раније, идеја је да се узорак подели, да се оцене густине посебно за позитивне и апсолутне вредности негативних чланова, и да се оне упореде. Разлика је у томе што се при оцењивању густине не користе симетрична, већ асиметрична језгра, на пример

Расподела	5%	10%	m	$\overline{s}_n$
		$n = 50$		
$\chi^2(2)$	0.076	0.23	-0.956	0.216
Логнормална	0.045	0.15	-0.804	0.205
$\mathcal{E}(1)$	0.07	0.208	-0.939	0.199
		$n = 100$		
$\chi^2(2)$	0.618	0.847	-1.810	0.258
Логнормална	0.429	0.712	-1.550	0.272
$\varepsilon(1)$	0.594	0.840	-1.792	0.253
		$n = 200$		
$\chi^2(2)$	0.997	0.999	-3.073	0.299
Логнормална	0.96	0.993	-2.658	0.345
$\mathcal{E}(1)$	0.999	1	-3.086	0.270
		$n = 500$		
$\chi^2(2)$	1	1	-5.840	0.408
Логнормална	1	1	-5.062	0.536
$\mathcal{E}(1)$	1	1	-5.056	0.376

Табела 5.7: Оцењена средња вредност и стандардна девијација статистике U и моћи предложеног теста симетрије са 1000 понављања.

фамилије гама језгара описане у одељку 3.2.2, а мера I се оцењује методом замене слично као у [32]. Њихову идеју надоградили су Хирукава и Сакудо (2016) [36]. У овом раду они су дефинисали фамилију уопштених гама језгара и за оцену језгрима из ове фамилије, они изводе тест статистику и предлажу алгоритам за одређивање параметра равнања. Пратићемо њихове резултате и описати процедуру тестирања симетрије за један специјални случај фамилије уопштених гама језгара, који и они посматрају, већ описану $K_{\rho_h(x),h}$.

Нека су $X_1, ..., X_{n_1}$ позитивни чланови узорка и $\hat{f}_G(x)$ њихова оцењена густина, а $Y_1, ..., Y_{n_2}$ апсолутне вредности негативних чланова узорка и $\hat{g}_G(x)$ њихова оцењена густина. Уведимо ознаку $K_u(v) = K_{\rho_h(u),h}(v)$,

$\hat{f}_G(x)$. Узорачки оцењена мера I је

$$I_{n_1,n_2} = \sum_{j=1}^{n_1}\sum_{i=1,i\neq j}^{n_1}\frac{1}{n_1^2}K_{X_j}(X_i) + \sum_{j=1}^{n_2}\sum_{i=1,i\neq j}^{n_2}\frac{1}{n_2^2}K_{Y_j}(Y_i) - \sum_{j=1}^{n_2}\sum_{i=1,i\neq j}^{n_1}\frac{1}{n_1 n_2}K_{Y_j}(X_i) - \sum_{j=1}^{n_1}\sum_{i=1,i\neq j}^{n_2}\frac{1}{n_1 n_2}K_{X_j}(Y_i). \tag{5.7}$$

Услове под којима следећа теорема важи нећемо наводити због њихове гломазности.

Теорема 5.4. (Хирукаща, Сакудо (2016) [36]) *Нека су испуњени услови 1-5 (в. [36]). Тада при H_0' важи*

$$n_1 h^{1/4} I_{n_1,n_2}/\sqrt{\hat{\sigma}^2} \xrightarrow{D} \mathcal{N}(0,1),$$

где је

$$\hat{\sigma}^2 = \frac{1}{\sqrt{\pi}n}\left\{\frac{1}{n_1}\sum_{i=1}^{n_1}X_i^{-1/2}\hat{f}_G^{\star}(X_i) + \frac{1}{n_2}\sum_{i=1}^{n_1}X_i^{-1/2}\hat{g}_G^{\star}(X_i) + \frac{n_1}{n_2^2}\sum_{i=1}^{n_2}Y_i^{-1/2}\hat{f}_G^{\star}(Y_i) + \frac{n_1^2}{n_2^3}\sum_{i=1}^{n_2}Y_i^{-1/2}\hat{g}_G^{\star}(Y_i)\right\}. \tag{5.8}$$

Уведимо ознаку $T_{n_1,n_2} := n_1 h^{1/4} I_{n_1,n_2}/\sqrt{\hat{\sigma}^2}$. Описаћемо и алгоритам за избор параметра равнања.

- Одабрати неко $\delta \in (0,1)$ и израчунати $M = \min\{\lfloor n_1^{\delta}\rfloor, \lfloor n_2^{\delta}\rfloor\}$.
- Сортирати узорке X_i и Y_i и направити M подузорака обима $(k_1,k_2) = (\lfloor n_1/M\rfloor, \lfloor n_2/M\rfloor)$, где је m-ти узорак дефинисан са $\left\{\left\{X_{m+(i-1)M}\right\}_{i=1}^{k_1}, \left\{Y_{m+(i-1)M}\right\}_{i=1}^{k_2}\right\}$, $m = 1,...,M$.
- Одабрати две константе $0 < \underline{H} < \overline{H} < 1$. Дефинишемо интервал $H_{k_1} = [\underline{H}, \overline{H}]$.
- Наћи $\hat{b}_{k_1} = \inf\left\{\arg\max_{b_{k_1}\in H_{k_1}} \hat{\pi}_M(b_{k_1})\right\}$ где је

$$\hat{\pi}_M(b_{k_1}) = \frac{1}{M}\sum_{m=1}^{M} I\{T_{k_1,k_2}(m) > 1.645\},$$

а $T_{k_1,k_2}(m)$ је тест статистика као она из теореме 5.4, само над m-тим подузорком.

- Оцењени ПР је $\hat{b}_{n_1} = \hat{B} n_1^{-q}$ где је $\hat{B} = b_{k_1}\hat{k}_1^q$.

За вредности непознатих параметара узимамо $\delta = 0.3$, $q = 4/9$, $H_{k_1} = [0.01, 0.64]$ које су предложене у [36]. Критичне су велике вредности тест статистике, па нулту хипотезу одбацујемо ако је $T_{n_1,n_2} > z_{\alpha/2}$. Рађено је са узорцима обима 50, 100 и 200, и са 1000 понављања. У табели 5.8 приказани су резултати симулација при тестирању узорака из нулте хипотезе. Оцењене мере теста, као и моменти тест статистика иду у прилог томе да је гранична расподела заиста $\mathcal{N}(0,1)$. У табели 5.9 приказани су резултати симулација при тестирању узорака из алтернативне хипотезе. Као и код претходног теста, и овде се претпоставља тестирање симетрије око нуле. Центриране асиметричне расподеле за које тестирамо су $\chi^2(3) - 3$, Логнормална$(0,1) - e^{1/2}$ и $\mathcal{E} - 1$. Моћ теста је мала за узорке обима 50, али расте са повећањем узорка и приближава се јединици. У односу на претходни тест заснован на мери θ, који такође дели узорак на два дела, овај има мере ближе теоријским, као и веће моћи за узорке мањег обима.

Расподела	5%	10%	m	$\overline{s}_n$
		$n = 50$		
$\mathcal{N}(0,1)$	0.033	0.069	0.005	0.779
t_{10}	0.032	0.063	-0.025	0.782
Лапласова	0.032	0.077	0.056	0.786
		$n = 100$		
$\mathcal{N}(0,1)$	0.04	0.074	-0.018	0.849
t_{10}	0.052	0.098	0.019	0.887
Лапласова	0.043	0.075	-0.027	0.840
		$n = 200$		
$\mathcal{N}(0,1)$	0.05	0.082	-0.012	0.877
t_{10}	0.057	0.095	0.035	0.903
Лапласова	0.04	0.074	-0.016	0.835

Табела 5.8: Оцењена средња вредност и стандардна девијација T_{n_1,n_2} и грешка прве врсте предложеног теста симетрије са 1000 понављања.

Расподела	5%	10%	m	$\overline{s}_n$
	$n = 50$			
$\chi^2(3)$	0.123	0.204	0.657	0.867
Логнормална	0.306	0.405	1.177	1.071
$\varepsilon(1)$	0.349	0.461	1.351	1.117
	$n = 100$			
$\chi^2(3)$	0.377	0.492	1.441	1.172
Логнормална	0.740	0.830	2.731	1.466
$\varepsilon(1)$	0.767	0.845	2.819	1.532
	$n = 200$			
$\chi^2(3)$	0.795	0.863	2.924	1.517
Логнормална	0.986	0.994	5.865	2.065
$\varepsilon(1)$	0.983	0.994	5.647	2.078

Табела 5.9: Оцењена средња вредност и стандардна девијација T_{n_1,n_2} и моћи предложеног теста симетрије са 1000 понављања.

Глава 6

Закључак

Овај рад бави се оцењивањем густине језгром, непараметарским методом оцењивања непознате густине неке случајне величине. Књиге на тему оцењивања густине језгром имају различити приступ овом проблему, више практични или више теоријски. У овом раду, предност је дата практичним проблемима који се јављају при коришћењу ОГЈ и методима за њихово решавање. У прва два поглавља дефинисана је ОГЈ и изведене су њене основне особине. Неки од метода за избор параметара од којих зависи ова оцена описани су у трећем поглављу. Четврто поглавље је у потпуности теоријско и бави се асимптотским понашањем ОГЈ. У петом поглављу разматрана је примена ОГЈ на тестирање симетрије расподеле. Избор параметра равнања посебно утиче на ОГЈ и зато би му требало посветити већу пажњу и код тестирања хипотеза. На пример, тест статистика из одељка 5.1 умногоме зависи од одабира параметра равнања, и добре оцењене вредности грешака прве врсте су последица добро одабране константе η. Тест описан у одељку 5.3 има лоше резултате на малом обиму узорка, и расподела тест статистике нема особине нормалне расподеле, ипак на то не утиче избор параметра равнања. Овај тест је првобитно конструисан за тестирање једнакости густина расподела два независна узорка. Идеја може бити да се поменута теорема прилагоди случају тестирања симетрије. Неопходно би било ослабити услов независности, без обзира на то да ли се узорак дели пре оцењивања густине или не. Побољшање би могло доћи и коришћењем асиметричних језгара, као у одељку 5.4, која су увек боља опција када се тестира симетрија дељењем узорка.

Литература

[1] B. W. Silverman, Density estimation for Statistics and Data Analysis, Chapmann & Hall/CRC, (1986)

[2] M. P. Wand, M. C. Jones, Kernel Smoothing, Chapman and Hall/CRC, (1994)

[3] B. L. S. Prakasa Rao, Nonparametric Functional Estimation, Academic Press, (1983)

[4] A. W. Bowman, A. Azzalini, Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations, Clarendon Press, (1997)

[5] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, Numerical Recipes in C: The Art of Scientific Computing, Cambridge University Press, (1992)

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

[7] L. Devroye, L. Gyfri, Nonparametric Density Estimation: The L_1 View, Wiley, New York, (1985)

[8] E. L. Lehmann, J. P. Romano, Testing statistical hypotheses, Springer-Verlag New York, 115-119, (2005)

[9] П. Младеновић, Вероватноћа и статистика, Математички факултет у Београду, (2002)

[10] K. Pearson, Contributions to the Mathematical Theory of Evolution II. Skew Variation in Homogeneous Material, Philosophical Transactions, 186. 343-414, (1895)

[11] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, Ann. Math. Statist. 27, 832-837, (1956)

[12] E. Parzen, On the estimation of a probability density function and the mode, Ann. Math. Statist. 33, 1065-1076, (1962)

[13] A. Dvoretzky, J. Kiefer, J. Wolfowitz, Asymptotic minimax character of the sample distribution function and of classical multinomial estimator, Ann. Math. Stat. 27, 642-669, (1956)

[14] C. J. Stone, An asymptotically optimal window selection rule for kernel density estimates, Ann. Statist., 1285-1297, (1984)

[15] P. Hall, J. S. Marron, Extent to Which Least-Squares Cross-Validation Minimises Integrated Square Error in Nonparametric Density Estimation, Probability Theory and Related Fields, 74, 567- 581, (1987)

[16] D. W. Scott, G. R. Terrell, Biased and unbiased cross validation in density estimation, J. Amer. Stat. Assoc., 82, 1131-1146

[17] P. Hall, J. S. Marron, Estimatio of integrated squared density derivatives. Statist. Probab. Lett. 6, 109-115, (1987)

[18] S. J. Sheather, M. C. Jones, A reliable data-based bandwidth selection method for kernel density estimatio. J. Roy. Statist. Soc. Ser. B 53, 683-690, (1991)

[19] B. U. Park, J. S. Marron, Comparison of data-driven bandwidth selectors, J. Amer. Stat. Assoc. 85, 66-72, (1990)

[20] G. K. Bhattacharayya, G. Roussas, Estimation of certain functional of probability density function, Skand. Aktuarietidskr., 52, 203-206, (1969)

[21] E. F. Schuster, Incorporating support constraints into nonparametric estimators of densities, Communications in Statistics - Theory and Methods, 14:5, 1123-1136, (1985)

[22] J. S. Marron, D. Rupert, Transformations to reduce boundary bias in kernel density estimation, Journal od the Royal Statistical Society. Series B (Methodological), 653-671, (1994)

[23] M. P. Wand, J. S. Marron, D. Rupert, Transformation in density estimation, Journal of the American Statistical Asociation Vol. 86, No. 414, Theory and Methods, 343-353, (1991)

[24] Song Xi Chen, Beta kernel estimators for density functions, Computational Statistics and Data Analysis 31, 131-145, (1999)

[25] Song Xi Chen, Probability density estimation using gamma kernels Ann, Inst. Statist. Math. Vol. 52, No. 3, 471-480, (2000)

[26] K. Matusita, Decision rules based on the distance for the problems of fit, two samples, and estimation, Ann. Math. Statist., 26, 631-640, (1955)

[27] I. A. Ahmad, G. Van Belle, Measuring affinity of distributions. Reliability and Biometry, Statistical Analysis of Life Testing, (eds. Proschan and R. J. Serfling), SIAM, Philadelphia, 651-668, (1974)

[28] M. Weitzman, Measures of Overlap of Income Distributions of White and Negro Families in the U.S.Technical Paper 22 Bureau of the Census, (1970)

[29] I. A. Ahmad, Nonparametric Estimation of Affinity Measure between Two Absolutely Continuous with Hypotheses Testing Applications, Ann. Inst. Statist. Math. 32, Part A, 223-240, (1980)

[30] Y. Fan, R. Gencay, Hypotheses testing based on modified nonparametric estimation of an affinity measure between two distributions, Nonparametric Statistics, Vol. 2, 389-403, (1993)

[31] Y. Fan, R. Gencay, A consistent nonparametric test of symmetry in linear regression models, Journal of the American Statistical Asociation Vol. 90, No. 430, Theory and Methods, (1995)

[32] I. A. Ahmad, Qi Li, Testing symmetry of an unknown density function by kernel method, Journal of Nonparametric Statistics, 7:3, 279-293, (1997)

[33] G. Anderson, O. Linton, and Y. Whang, Nonparametric estimation of a polarization measure, Discussion Paper, University of Toronto, The London School of Economics and Seoul National University,(2009)

[34] H. M. Samawi, A. Helu, R Vogel, A nonparametric test of symmetry based on the overlapping coefficient, J. Appl. Stat. Vol. 38(5), 885-898, (2010)

[35] M. Fernandes, E.F. Mendes, O. Scaillet, Testing for symmetry and conditional symmetry using asymmetric kernels, Annals of the Institute of Statistical Mathematics, 67, 649-671, (2015)

[36] M. Hirukawa, M. Sakudo, Testing symmetry of unknown densities via smoothing with the generalized gamma kernels, Econometrics Vol. 4(2), (2016)

Биографија

Аница Костић рођена је 6.1.1993. у Београду. Математичку гимназију и Средњу музичку школу ,,Др Војислав Вучковић" завршила је 2011. године. Математички факултет Универзитета у Београду, смер Статистика, актуарска и финансијска математика, уписала је 2011. године. Основне академске студије је завршила 2015. године са просечном оценом 9.95. Од 2015. године ради као сарадник у настави на катедри за Вероватноћу и статистику. Држала је вежбе на седам курсева у надлежности ове катедре.