LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

BIKELA-ROZENBLATA TESTS

DIPLOMDARBS

Autors: **Audris Ločmelis**

Stud. apl. al06065

Darba vadītājs: docents Dr.math. Jānis Valeinis

RĪGA 2011

Anotācija

Bikels un Rozenblats 1973. gadā ieviesa testu par sadalījuma pārbaudi, kas ir balstīts uz integrēto kvadrātisko kļūdu starp blīvuma funkciju un kodolu blīvuma funkcijas novērtējumu. Statistikas robežsadalījums sakrīt neatkarīgiem novērojumiem un absolūti regulāriem procesiem, turklāt testu var izmantot vienkāršu un saliktu hipotēžu pārbaudei. Neskatoties uz tā labajām īpašībām, tests nav praksē pielietojams neatrisinātās joslas platuma izvēles dēļ. Līdzīgai statistikai 2008. gadā Gao un Gijbelsa atrisināja joslas platuma izvēles problēmu, atrodot izvedumus jaudas un nozīmīguma funkcijām. Darbā ar simulāciju palīdzību pētīta testa asimptotiskā uzvedība un jauda. Lai testu varētu padarīt pielietojamu, apskatīts, kā Gao un Gibelsas rezultāts pielietojams Bikela–Rozenblata statistikai.

Atslēgas vārdi: tests par sadalījuma pārbaudi; kodolu blīvuma funkcijas novērtējums; joslas platuma izvēle

Abstract

Bickel and Rosenblatt (1973) proposed a goodness-of-fit test based on the integrated squared error between the kernel density estimate and the parametric density. The test has the same limiting distribution for independent observations and absolutely regular processes. This holds true for both simple and composite hypothesis. Despite it's good properties this test has a big drawback – it can not be applied practically due to the problematic bandwidth parameter choice. Recently for similar test statistics Gao and Gijbels (2008) solved the problem of bandwidth choice by establishing closed–form expressions for both size and power functions. In this thesis the asymptotic behavior of the test is investigated by simulation study. Moreover it is shown how the result of Gao and Gijbels could be applied for the Bickel–Rosenblatt statistic.

Keywords: goodness-of-fit test; kernel density estimate; bandwidth selection

Saturs

Apzīmejumi

LU Latvijas Universitāte

$\mathbb{R}$ reālo skaitļu kopa

$\mathbb{Z}$ veselo skaitļu kopa

$\xrightarrow{d}$ konverģence pēc sadalījuma

$\mathbb{E}$ matemātiskā cerība

$\mathbb{D}$ dispersija

$I_A(x)$ indikatorfunkcija, vienāda ar 1, kad $x \in A$, un 0 pretējā gadījumā

$N(\mu, \sigma)$ normālais sadalījums ar vidējo vērtību μ un dispersiju σ^2

$v.s.\,[a, b]$ vienmērīgais sadalījums intervālā $[a, b]$

Ievads

Hipotēžu pārbaude par izlases sadalījumu ir viena no biežāk sastopamajām problēmām statistikā. Pazīstamākie no testiem par izlases sadalījuma atbilstību kādai parametriskai funkcijai ir 1933. gadā definētais Kolmogorova-Smirnova un 1900. gadā ieviestais Pīrsona χ^2 tests. Vēl viens klasisks tests ir 1937. gadā ieviestais Neimaņa tests, kurš īpaši aktuāls kļuvis, kopš Ledvina 1994. gadā ieviesa metodi nezināmā parametra k noteikšanai [1] un 1996. gadā kopā ar Inglotu (Inglot) parādīja, ka maziem izlašu apjomiem tas ir efektīvāks par citiem klasiskajiem testiem pārbaudei par sadalījuma likumu [2].

Salīdzinoši jaunāks ir Bikela (Bickel) un Rozenblata (Rosenblatt) 1973. gadā ieviestais tests [3]. Tā pamatā ir Rozenblata 1956. gadā atklātais neparametriskais kodolu blīvuma funkcijas novērtējums f_n [4], precīzāk, L_2–attālums (integrētā kvadrātiskā kļūda) starp f_n un hipotētisko blīvuma funkciju. Bikela–Rozenblata tests savā ziņā ir līdzīgs Kolmogorova–Smirnova testam, taču nepārbauda hipotēzi par sadalījuma, bet gan par blīvuma funkciju.

Klasiskais Bikela–Rozenblata tests paredzēts vienkāršu hipotēžu pārbaudei neatkarīgiem datiem, taču tas bez transformācijām izmantojams arī absolūti regulāriem procesiem, ko 2000. gadā parādījuši Noimans (Neumann) un Paparoditis (Paparoditis) [5].

Pavisam nesen, 2009. gadā, publicētajā Munka (Munk) u.c. rakstā [6] ir atklāta testa statistika, kas ļauj modificēt Neimaņa statistiku, lai to varētu pielietot atkarīgiem datiem. Bikela–Rozenblata statistika ir daudz salīdzināta ar Kolmogorova–Smirnova un Pīrsona testiem, taču ne ar Neimaņa statistiku, tādēļ viens no darba mērķiem ir ar simulāciju palīdzību salīdzināt šo testu jaudas atkarīgiem un neatkarīgiem datiem vienkāršām hipotēzēm, kas līdz šim vēl nav darīts. Tests nav sastopams nevienā no statistikas programmām, tādēļ pirmais uzdevums pirms simulāciju veikšanas bija uzrakstīt programmu testa realizācijai. Lai palielinātu ātrdarbību un precizitāti, programmā vienmērīgā kodola gadījums uzlabots ar analītiskiem pārveidojumiem.

Bikela–Rozenblata tests derīgs arī saliktu hipotēžu pārbaudei un daudzdimensionāliem gadījuma lielumiem neatkarīgiem un atkarīgiem datiem. Šis ir vienīgais zināmais tests ar tik labām īpašībām, tādēļ varētu kļūt par ideālu testu statistikas praksē. Taču neparametriskajā statistikā ļoti pazīstama ir optimālā joslas platuma h izvēles problēma. Izņēmums nav arī Bikela–Rozenblata tests, tādēļ otrs darba mērķis ir apskatīt šo problēmu un piedāvāt iespējamos risinājumus. Tas, cik būtiska loma ir pareiza joslas platuma izvēlei, tiks sākotnēji parādīts ar simulāciju palīdzību.

Tā kā, pētot Bikela–Rozenblata testu, atklājās, ka nav ieviesta procedūra optimālā joslas platuma izvēlei, darbā tiks meklētas iespējas atrast h optimālo novērtējumu. Gao (Gao) un Gijbelsa (Gijbels) 2008. gadā ievieš metodi, ar kuras palīdzību viņu apskatītajai statistikai tiek atrasts optimālais h. Pielietojot *Edgeworth* izvirzījumus tiek novērtētas nozīmīguma un jaudas funkcijas, kuras tālāk tiek lietotas optimizācijas uzdevuma sastādīšanai, kuru atrisinot iegūst optimālo joslas platumu. Apskatītā statistika arī izmanto kodolu gludināšanu, taču paredzēta citai, regresijas, problemātikai. Darbā apskatītas iespējas izmantot šo rezultātu Bikela–Rozenblata testam, kas ļautu padarīt testu izmantojamu.

Darbs sastāv no četrām nodaļām un pielikuma. Pirmajā nodaļā apskatītas divas klasiskās Bikela–Rozenblata statistikas un to asimptotiskais sadalījums. Otrajā nodaļā parādīti daži jaunākie rezultāti par Bikela–Rozenblata testu, kā arī definēta darbā apskatītā Bikela–Rozenblata statistika. Trešajā nodaļā ar simulāciju palīdzību pētīta statistikas asimptotiskā uzvedība un analizēta jauda, kā arī veikts salīdzinājums ar Neimaņa testu. Ceturtā nodaļa veltīta optimālā joslas platuma analīzei. Pielikumā atspoguļota daļa no simulāciju rezultātiem un pievienots programmu kods.

1. KLASISKAIS BIKELA–ROZENBLATA TESTS

Pieņemsim, ka $X_1, \dots, X_n$ ir neatkarīgi un vienādi sadalīti gadījuma lielumi ar nepārtrauktu blīvuma funkciju f. Hipotēzes

$$H_0 : f = f_0$$

$$H_1 : f \neq f_0$$

pārbaudei Bikels un Rozenblats [3] ievieš testa statistiku

$$T_n^{br} = nh_n \int [f_n(x) - \mathbb{E}(f_n(x))]^2 a(x) dx, \tag{1.0.1}$$

kur $a(x)$ ir brīvi izvēlēta integrējama, reālvērtīga funkcija. Blīvuma funkcija tiek novērtēta ar ierobežotas, integrējamas kodola funkcijas K palīdzību,

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right),$$

kas sīkāk tiks apskatīts apakšnodaļā 1.1.

1.1. Blīvuma funkcijas kodolu novērtējums

Pieņemsim, ka doti neatkarīgi vienādi sadalīti gadījuma lielumi $X_1, \dots, X_n$ ar sadalījuma funkciju $f(x) = F'(x)$ un blīvuma funkciju $f(x) = F'(x)$. Blīvuma funkcijas neparametriskās novērtēšanas mērķis ir novērtēt blīvuma funkciju f ar iespējami mazāk pieņēmumiem par f. Lai saprastu novērtējuma ideju, visvieglāk to uztvert kā histogrammas vispārinājumu.

Definīcija 1. Ja reālo skaitļu asi sadala intervālos B_j ar platumu $2h$,

$$B_j = [x_0 + 2(j-1)h, x_0 + 2jh), \quad j \in \mathbb{Z},$$

tad formāli histogramma tiek uzdota ar vienādojumu

$$\hat{f}_h(x) = \frac{1}{2nh} \sum_{i=1}^{n} \sum_{j} I_{B_j}(X_i) I_{B_j}(x).$$

Taču idejiski histogrammu var pārrakstīt ap x centrētiem intervāliem $[x - h, x + h)$,

$$\hat{f}_h(x) = \frac{1}{2nh} \#\{X_i \in [x - h, x + h)\}, \tag{1.1.1}$$

kur $\#A$ apzīmē kopas A apjomu. Jeb mazliet formālāk, izmantojot svaru funkciju

$$K^*(u) = \frac{1}{2} I_{[-1,1]}(u),$$

sauktu arī par vienmērīgo kodolu, jo tā sakrīt ar vienmērīgā sadalījuma intervālā $[-1, 1]$ blīvuma funkciju, un $u = (x - X_i)/h$, histogrammu (1.1.1) var pārrakstīt formā

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

Definīcija 2. Nenegatīvu funkciju $K(u)$ sauc par kodolu, ja

- $\int_{-\infty}^{\infty} K(u)dx = 1$,
- $\forall u \quad K(u) = K(-u)$.

Ja taisnstūrveida svaru funkcijas K^* vietā izvēlas vispārēju, tipiski gludāku, kodola funkciju $K(\cdot)$, iegūst kodolu blīvuma funkcijas definīciju.

Definīcija 3. Ja K ir kodols un F ir izlases sadalījuma funkcija, tad par kodolu blīvuma funkcijas novērtējumu sauc

$$\hat{f}_h(x) = \frac{1}{hn} \sum_{i=1}^{n} K \left(\frac{x - X_i}{h} \right) \tag{1.1.2}$$

$$= \frac{1}{h} \int K \left(\frac{x - t}{h} \right) dF(t). \tag{1.1.3}$$

Interesanti, ka kodolu blīvuma funkcijas novērtējuma idejas autors ir tieši Rozenblats, kas pirmo reizi apskatīta 1956. gadā [4].

Optimālais h. Teorētiski ir zināms [7], ka optimālā h izvēle blīvuma funkcijas novērtēšanai (1.1.3) ir

$$h_{opt} = \left(\frac{\|K\|_2^2}{\|f''\|_2^2 [\mu_2(K)]^2 n} \right)^{1/5} \sim n^{-1/5}, \tag{1.1.4}$$

kur $\|g\|_2^2 := \int g(s)ds$. Taču, lai arī var aprēķināt otro momentu $\mu_2(K)$, nav zināma īstā izlases blīvuma funkcija f, līdz ar to arī tās otrais atvasinājums f''.

Literatūrā aprakstītas dažādas metodes h novērtēšanai, no kurām pazīstamākās ir ievietošanas un krosvalidācijas metodes. Turpmāk apskatīsim mazāko kvadrātu krosvalidāciju.

Lai iegūtu h novērtējumu dotai izlasei, novērtē $\hat{f}_h$ ar risku jeb integrēto vidējo kvadrātisko kļūdu (L_2–attālumu) $R = \mathbb{E}(L)$, kur

$$L = \int (\hat{f}_h(x) - f(x))^2 dx \tag{1.1.5}$$

ir integrētās kvadrātiskās kļūdas zaudējumu funkcija. Novērtējums būs atkarīgs no h, risks tiks minimizēts izvēloties h ar viena izlaišanas (*leave–one–out*) krosvalidācijas metodi.

Zaudējumu funkciju, kas atkarīga no gludinošā parametra h var pārrakstīt

$$\begin{aligned} L(h) &= \int (\hat{f}_h(x) - f(x))^2 dx \\ &= \int \hat{f}_h^2(x)dx - 2\int \hat{f}_h(x)f(x)dx + \int f^2(x)dx. \end{aligned}$$

Tā kā pēdējais saskaitāmais nav atkarīgs no h, bet $\int \hat{f}_h^2(x)dx$ var aprēķināt no datiem, paliek tikai saskaitāmais, kas atkarīgs no h un satur nezināmo lielumu f. Savukārt $\int \hat{f}_h(x)f(x)dx$ ir $\hat{f}_h(X)$ matemātiskā cerība, ko var novērtēt

$$\widehat{\mathbb{E}[\hat{f}_h(X)]} = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(X_i),$$

kur $\hat{f}_{h,-i}$ ir blīvuma novērtējums iegūts pēc i–tā novērtējuma izņemšanas.

Definīcija 4. Krosvalidācijas riska novērtējums ir

$$\hat{J}(h) = \int \hat{f}_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(X_i), \tag{1.1.6}$$

kur $\hat{J}(h) = L(h) - \int f^2(x)dx$.

Gludinošo parametru jeb joslas platumu h iegūst, minimizējot novērtēto risku (1.1.6) pēc h.

Turpmāk būs nepieciešams apskatīt joslas platumu, kas atkarīgs no izlases apjoma, h_n, tādēļ lietosim blīvuma funkcijas novērtējumu

$$f_n(x) := \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right). \tag{1.1.7}$$

1.2. Testa statistikas asimptotiskais sadalījums

Vispirms tiks definēti nosacījumi, kam jāizpildās, lai būtu spēkā teorēmas par testa statistiku asimptotisko sadalījumu.

Pieņēmums 1. Kodola funkcijai izpildās $\int K(x)dx = 1$ un K vai nu

a) konverģē uz nulli ārpus intervāla $[A, -A]$ un ir absolūti nepārtraukta un diferencējama $[A, -A]$, vai

b) ir absolūti nepārtraukta un diferencējama intervālā $(\infty, -\infty)$ un $\int |K'(t)|^k dt < \infty,\ k = 1, 2$.

Pieņēmums 2. Blīvuma funkcija f ir nepārtraukta, pozitīva un ierobežota.

Pieņēmums 3. Funkcija $f^{\frac{1}{2}}$ ir absolūti nepārtraukta un tās atvasinājuma $\frac{1}{2}f'/f^{\frac{1}{2}}$ absolūtā vērtība ir ierobežota. Vēl vairāk,

$$\int_{|z|\geq 3} |z|^{\frac{3}{2}}[\ln\ln|z|]^{\frac{1}{2}}[|w'(z)| + |w(z)|]dz \leq \infty.$$

Pieņēmums 4. Blīvuma funkcijai f eksistē otrais atvasinājums f'' un ir ierobežots. Turklāt kodola funkcija K ir simetriska ap nulli un $z^2w(z)$ ir integrējama.

Teorēma 1 ([3])**.** *Ja izpildās pieņēmumi 1-3 un $a(x)$ ir integrējama, gabaliem nepārtraukta un ierobežota. Tad pie patiesas hipotēzes H_0*

$$h_n^{-\frac{1}{2}}\left[T_n^{br} - \int f(x)a(x)dx \int K^2(z)dz\right]$$

ir asimptotiski normāli sadalīts ar vidējo vērtību 0 un dispersiju

$$2\int\left[\int K(x)K(x+y)dx\right]^2 dy \int a^2(x)f^2(x)dx, \tag{1.2.1}$$

kad $n \to \infty$.

Reizē ar statistiku T_n^{br} Bikels un Rozenblats ieviesa arī statistiku

$$\hat{T}_n^{br} = nh_n \int [f_n(x) - f_0(x)]^2 a(x)dx, \tag{1.2.2}$$

kura, izpildoties papildu pieņēmumam 4, asimptotiski uzvedas ļoti līdzīgi kā T_n^{br}. Formā (1.2.2) statistika ir pētīta arī nesenās publikācijās, taču biežāk sastopama ir $\hat{T}_n^{br}$ jaunākā iterācija T_n (2.2.2), kas tiks apskatīta nākamajā nodaļā.

Teorēma 2 ([3])**.** *Ja ir spēkā pieņēmumi 1-4 un $h_n = o(n^{-\frac{2}{9}})$, $n^{-\frac{1}{2}}(\ln n)^{\frac{1}{2}}(\ln\ln n)^{\frac{1}{4}} = o(h_n)$, kad $n \to \infty$, tad*

$$h_n^{-\frac{1}{2}}\left[\hat{T}_n^{br} - \int f(x)a(x)dx \int K^2(z)dz\right]$$

ir asimptotiski normāli sadalīts ar vidējo vērtību 0 un dispersiju (1.2.1)*, kad $n \to \infty$.*

2. BIKELA–ROZENBLATA TESTS ATKARĪGIEM DATIEM

Lai apskatītu Bikela–Rozenblata testu atkarīgiem datiem, pieņemsim, ka dots stacionārs process $\{X_i, -\infty < i < \infty\}$. Vispārīgā veidā atkarīgus procesus var definēt, izmantojot jauktos koeficientus atkarības struktūras klasifikācijai, ko 1956. gadā ieviesis tieši Rozenblats [8].

2.1. Jauktie procesi

Klasiski Bikela–Rozenblata statistika paredzēta neatkarīgiem vienādi sadalītiem datiem, taču praksē bieži nākas saskarties ar atkarīgiem datiem. Salīdzinoši nesen, 2000. gadā, Noimans un Paparoditis [5] pierādījuši, ka Bikela–Rozenblata statistikas robežsadalījums sakrīt neatkarīgiem datiem un absolūti regulāriem procesiem.

Absolūti regulāri procesi jeb β–jauktie procesi ir viens no stohastisko jaukto procesu veidiem. Daži pazīstamākie β–jaukto procesu piemēri ir Markova ķēdes, $ARMA$ un $GARCH$ procesi [9].

Latviski plašāku šo procesu apkopojumu izstrādājusi Rone [10] savā maģistra darbā, taču fundamentāli jaukto procesu teorija apskatījis, Bradlejs (Bradley) [11, 9], pēc kā arī tiks definēti interesējošie jaukto procesu koeficienti.

Definīcija 5. Par gadījuma procesu sauc gadījuma lielumu saimi $(X_t, t \in \mathbb{Z})$, kas uzdota varbūtību telpā $(\Omega, \mathcal{F}, P)$.

Definīcija 6. Pieņemsim, ka dota varbūtību telpa $(\Omega, \mathcal{F}, P)$, tad atkarības koeficientus starp divām σ–algebrām $\mathcal{A}$ un $\mathcal{B}$ definē kā

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup_{A \in \mathcal{A}, B \in \mathcal{B}} |P(A \cap B) - P(A)P(B)|,$$
$$\beta(\mathcal{A}, \mathcal{B}) = \sup_{A \in \mathcal{A}, B \in \mathcal{B}} \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |P(A_i \cap B_j) - P(A_i)P(B_j)|,$$

kur $\forall i \in \{1, I\}$, $A_i \in \mathcal{A}$ un $\forall j \in \{1, J\}$, $B_i \in \mathcal{B}$ ir nešķeļošas kopas, kas sadala Ω.

Pieņemsim, ka $(X_t, t \in \mathbb{Z})$ ir gadījuma lielumu virkne uzdota varbūtību telpā $(\Omega, \mathcal{F}, P)$. σ–algebra $\mathcal{F}_J^L = \sigma(X_k, J \leq k \leq L)$, $\forall J, L$ $(-\infty \leq J \leq L \leq \infty)$. Darbā apskatītajai problemātikai nepieciešamos jaukto procesu atkarības koeficientus priekš $(X_t, t \in \mathbb{Z})$ $\forall n$ definē kā

$$\alpha(n) := \sup_{J \in \mathbb{Z}} \alpha(\mathcal{F}_{-\infty}^{J}, \mathcal{F}_{J+n}^{\infty}), \tag{2.1.1}$$

$$\beta(n) := \sup_{J \in \mathbb{Z}} \beta(\mathcal{F}_{-\infty}^{J}, \mathcal{F}_{J+n}^{\infty}). \tag{2.1.2}$$

Definīcija 7. Procesu $(X_t, t \in \mathbb{Z})$ sauc par

α–jaukto procesu, ja $\alpha(n) \xrightarrow{n\to\infty} 0$,

β–jaukto procesu, ja $\beta(n) \xrightarrow{n\to\infty} 0$.

Apgalvojums 3 ([9])**.** *Pieņemsim, ka dota varbūtību telpa* $(\Omega, \mathcal{F}, P)$*. Tad* σ*–algebras* $\mathcal{A}, \mathcal{B} \subset \mathcal{F}$ *ir neatkarīgas, ja a)* $\alpha(\mathcal{A}, \mathcal{B}) = 0$ *vai b)* $\beta(\mathcal{A}, \mathcal{B}) = 0$.

Apgalvojums 4 ([9])**.** σ*–algebrām* $\mathcal{A}, \mathcal{B} \subset \mathcal{F}$ *dotā varbūtību telpā* $(\Omega, \mathcal{F}, P)$ *spēkā nevienādība*

$$2\alpha(\mathcal{A}, \mathcal{B}) \leq \beta(\mathcal{A}, \mathcal{B}).$$

Tas, ka α koeficients vienmēr ir mazāks par β koeficientu nosaka, ka α–jauktie procesi ir vispārīgāki par β–jauktajiem procesiem, t.i., visi β–jauktie procesi ir arī α–jauktie procesi:

$$\beta\text{–jauktais process} \Longrightarrow \alpha\text{–jauktais process}.$$

2.2. Vienkāršās hipotēzes pārbaude

Apskatām vienkāršu hipotēzi par atkarīgu vienādi sadalītu gadījuma lielumu $X_1, \ldots, X_n$ blīvuma funkciju

$$H_0 : f = f_0,$$

$$H_1 : f \neq f_0$$

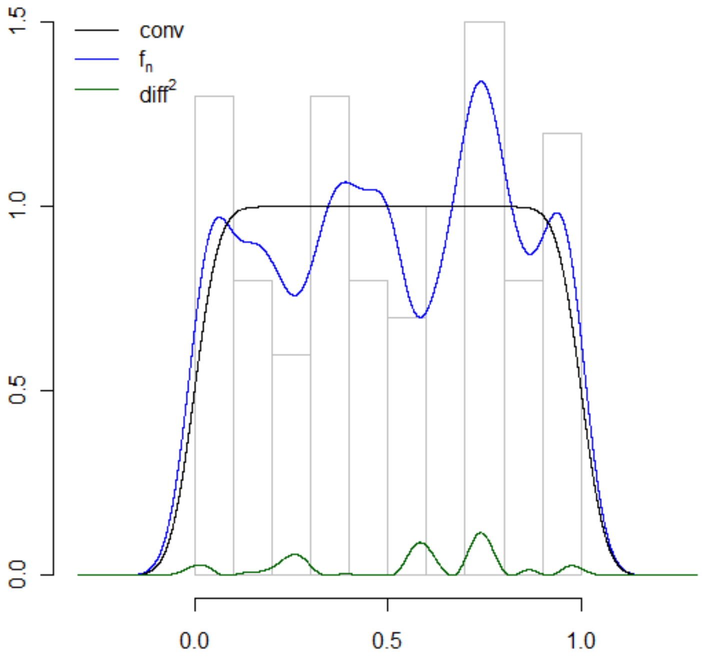
pie noteikta nozīmības līmeņa α un zināmas blīvuma funkcijas f_0.

Lai novērstu novirzes problēmas klasiskajā Bikela–Rozenblata statistikā $\hat{T}_n^{br}$ (1.2.2), hipotētiskās blīvuma funkcijas f_0 vietā Noimans un Paparoditis [5] iesaka lietot nogludinātu f_0 versiju, kas ir blīvuma funkcijas konvolūcija ar kodola funkciju K_{h_n}, t.i.,

$$(K_{h_n} * g)(\cdot) = \int h_n^{-1} K\left(\frac{\cdot - z}{h_n}\right) g(z)dz, \tag{2.2.1}$$

kur ar $*$ apzīmē konvolūcijas operatoru un ar $K_{h_n}(x)$ apzīmē $h_n^{-1}K\left(x/h_n\right)$.

Šādu nogludinātu funkciju līdzīgai statistikai, kas paredzēta priekš regresijas problemātikas, 1993.gadā lietojuši Hardle (Hardle) un Mamens (Mammen) [12] un Bikela–Rozenblata statistikai to izmantojis arī Fans (Fan) [13]. Kā piemērs blīvuma funkcijas gludināšanai 2.1.. attēlā parādīta vienmērīgā sadalījuma blīvuma funkcijas konvolūcija ar $N(0,1)$ kodolu.



2.1. att.: Simulēta vienmērīgi sadalīta izlase apjomā $n = 50$. Par gludinātāju izvēlēts normālais $N(0,1)$ kodols. Melnā krāsā attēlota konvolūcija $(K_{h_n} * f_0)(x)$, kur f_0 ir $v.s.[0,1]$ blīvuma funkcija. Zilā krāsā attēlots blīvuma funkcijas kodolu novērtējums $f_n(x)$, $h_n = 0.05$. Zaļā krāsā attēlota abu lielumu starpība, kāpināta kvadrātā.

Jau definēts blīvuma funkcijas f kodolu novērtējums f_n ar joslas platumu h_n atkarīgu no n (1.1.7). Līdz ar to statistikas $\hat{T}_n^{br}$ modifikāciju T_n var uzrakstīt formā

$$T_n = nh_n^{d/2} \int \left[f_n(x) - (K_{h_n} * f_0)(x)\right]^2 dx, \tag{2.2.2}$$

kur d apzīmē datu dimensiju. Turklāt par svaru funkciju Noimans un Paparoditis izvēlējušies $a(x) \equiv 1$. Šādu izvēli pamato arī diskusija, kas pārcelta uz datu simulēšanas aprakstu 3. nodaļā. Statistikas konstrukcijā izmantotie atsevišķie elementi parādīti 2.1. attēlā.

Pieņēmums 5. Jauktā procesa atkarības koeficients $\beta(k)$ ir eksponenciāli dilstošs, t.i.,

$$\beta(k) \leq C \exp(-Ck).$$

Par stacionārā procesa blīvuma funkciju pieņem f, un $f_{X_{i_1},\ldots,X_{i_m}}$ pieņem par $(X_{i_1}, \ldots, X_{i_m})$ daudzdimensionālo sadalījumu.

Pieņēmums 6. (i) f ir nepārtraukta,

(ii) $\sup_{X_{i_1},\ldots,X_{i_m}} \{f_{X_{i_1},\ldots,X_{i_m}}(x_1, \ldots, x_m)\} < \infty$ $\forall m$ un $i_1 < \ldots < i_m$.

Pieņēmums 7. Blīvuma funkcija K ir ierobežots un kompakts attēlojums (vienāda ar nulli ārpus kādas kompaktas kopas).

Pieņēmums 8. (i) $h_n = o([\ln(n)]^{-3})$,

(ii) $h_n^{-d} = o(n)$.

Asimptotisko uzvedību līdzīgai statistikai kā T_n jaukto procesu speciālgadījumam 1987. gadā jau pētījuši Takahata (Takahata) un Jošihara (Yoshihara) [14], taču Noimans un Paparoditis 2000. gadā to apskata vispārīgākam – β–jauktajam procesam, turklāt ar atvieglotiem pieņēmumiem, kas ļauj apskatīt arī salikto hipotēžu gadījumu.

Teorēma 5 ([5]). *Ja ir spēkā nosacījumi 5–8, tad izpildoties H_0,*

$$T_n - \mu \xrightarrow{d} N(0, \sigma^2),$$

kur

$$\mu = h_n^{-d/2} \int K^2(u) du$$

un

$$\sigma^2 = 2 \int f_0^2(x) dx \times \int \left[\int K(u) K(u+v) du \right]^2 dv. \tag{2.2.3}$$

Tas nozīmē, ka Bikela–Rozenblata testam vienkāršai hipotēzei

$$H_0 \text{ tiek noraidīta, ja } T_n \geq \mu + z_\alpha \sigma,$$

kur z_α ir standart normālā sadalījuma $1 - \alpha$ kvantile, definēta kā

$$\Phi(z_\alpha) = 1 - \alpha, \quad \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx.$$

Šeit arī parādās jēga T_n rakstīt formā (2.2.2), jo σ^2 nav atkarīga no h_n, līdz ar to iespējams fiksēt kritisko vērtību $z_\alpha \sigma(K)$, kas ir ērti, analizējot jaudu ar simulāciju palīdzību.

2.3. Saliktās hipotēzes pārbaude

Viens no galvenajiem iemesliem, kas pamato Bikela–Rozenblata testa un līdz ar to arī darba tēmas aktualitāti ir lieliskā īpašība, ka ar to var pārbaudīt arī saliktas hipotēzes. Tas padara šo testu par universālāko no visiem sadalījuma pārbaudes testiem, aptverot gan neatkarīgus, gan atkarīgus daudzdimensionālus novērojumus, kā arī vienkāršas un saliktas hipotēzes.

$$H_0 : f \in \mathfrak{F},$$

kur $\mathfrak{F}$ ir parametrisku blīvuma funkciju klase. Praksē bieži lieto galīgi dimensionālas parametriskas hipotēzes, t.i., $\mathfrak{F} = \mathfrak{F}^{\gamma} = \{f_\gamma, \gamma \in \Gamma\}$, kur $\subseteq \mathbb{R}^p$.

Gadījumā, kad $f \in \mathfrak{F}^{\gamma}$, pieņemam, ka $\gamma_0 \in \Gamma$ ir tāds, ka $f_{\gamma_0} = f$. Tad ļoti līdzīgi kā vienkāršas hipotēzes gadījumā testa statistika balstīta uz L_2–attālumu (integrēto kvadrātisko kļūdu) starp novērtējumu $f_n(x)$ un nogludinātu parametrisko blīvumu $f_{\hat{\gamma}}$,

$$T_{n,\hat{\gamma}} = nh_n^{d/2} \int \left[f_n(x) - (K_{h_n} * f_{\hat{\gamma}})(x)\right]^2 dx,$$

izmantots tas pats gludinošais operators, kas iepriekš (2.2.1). Salīdzinot abas statistikas

$$T_{n,\hat{\gamma}} - T_n = 2nh_n^{d/2} \int [f_n(x) - (K_{h_n} * f_{\gamma_0})(x)][(K_{h_n} * f_{\gamma_0})(x) - (K_{h_n} * f_{\hat{\gamma}})(x)]dx$$
$$+ nh^{d/2} \int [K_{h_n} * (f_{\hat{\gamma}} - f_{\gamma_0}))(x)]^2 dx,$$

un formulējot papildus nosacījumus, var parādīt T_n un $T_{n,\hat{\gamma}}$ ekvivalenci. Šim mērķim jāpārraksta $f_\gamma(x)$ formā

$$f_\gamma(x) = f_{\gamma_0}(x) + (\gamma - \gamma_0) f'_{\gamma_0}(x) + R(\gamma, \gamma_0, x),$$

kur $\int R^2(\hat{\gamma}, \gamma_0, x)dx = o_P(n^{-1})$ un jādefinē daži papildu pieņēmumi:

Pieņēmums 9. (i) $\int [(K_{h_n} * f_{\hat{\gamma}})(x) - (K_{h_n} * f_{\gamma_0})(x)]^2 dx = o_P(n^{-1} h_n^{-d/2})$,

(ii) $\hat{\gamma} - \gamma_0 = o_P(n^{-1/2} h_n^{-d/2})$,

(iii) $\sup_x \{f'_{\gamma_0}(x)\} < \infty$,

(iv) $\int R^2(\hat{\gamma}, \gamma_0, x)dx = o_P(n^{-1})$.

Teorēma 6 ([5]). *Ja ir spēkā nosacījumi 5–9, tad izpildoties H_0,*

$$T_{n,\hat{\gamma}} - h_n^{-d/2} \int K^2(u)du \xrightarrow{d} N(0, \sigma^2),$$

un σ^2 definēts kā teorēmā 5.

Pierādījuma ideja. Teorēmu 5 un 6 pierādījumiem Noimans un Paparoditis izmanto gandrīz identisku ceļu kā Takahata un Jošihara [14]. Vispirms statistika tiek izteikta formā

$$T_n = \sum_{1 \le i < j \le n} H_n(Y_i, Y_j) + \frac{1}{2} \sum_{i=1}^{n} H_n(Y_i, Y_i),$$

kur

$$H_n(x,y) = \frac{2}{n h_n^{3d/2}} \int \left[K\left(\frac{u-x}{h_n}\right) - \mathbb{E}K\left(\frac{u-Y_1}{h_n}\right)\right] \left[K\left(\frac{u-y}{h_n}\right) - \mathbb{E}K\left(\frac{u-Y_1}{h_n}\right)\right] du.$$

Un starpība $T_n - h_n^{-d/2} \int K^2(u)du$ tiek aproksimēta ar

$$U_n = \sum_k S_k,$$

kur

$$S_k = \sum_{i=1}^{b_{k-1}} \sum_{j=a_k}^{b_k} H_n(Y_i, Y_j).$$

Šeit $Y_i, \ldots, Y_n$ tiek sadalīti pamīšus sakārtotos lielos un mazos blokos, kuru robežas nosaka bloku butstrapa metodes parametri a_k un b_k. Atsaucoties uz Dvoretzka (Dvorretzky) [15] rezultātiem par atkarīgu gadījuma lielumu summu asimptotiku,

$$U_n \xrightarrow{d} N(0, \sigma^2).$$

Savukārt starpība starp $T_n - h_n^{-d/2} \int K^2(u)du$ un U_n ir $o_P(1)$.

Detalizēti 6. teorēmu Noimans un Paparoditis pierādījuši tehniskajā ziņojumā [16], raksta agrīnākā nepublicētā versijā.

3. SIMULĀCIJU REZULTĀTI

Simulācijām tiks lietota Bikela–Rozenblata statistika T_n (2.2.2), kurai par svaru funkciju izvēlēta $a(x) \equiv 1$. Izrādās, ka šāda izvēle ir vislabākā iespējamā svaru funkcija intervālā $[0,1]$. Pamatojums atrodams divās secīgās publikācijās, no kurām pirmajā Gošs (Ghosh) un Huangs (Huang) [17] 1991. gadā parāda, ka Bikela–Rozenblata testam optimālais kodols ir

$$K^*(x) = \frac{1}{2} I_{[-1,1]}(x),$$

saukts arī par histogrammas vai vienmērīgo kodolu, jo apraksta vienmērīgā sadalījuma intervālā $[-1,1]$ blīvuma funkciju. Šis rezultāts ir ļoti pārsteidzošs, jo neparametriskajā statistikā pieņemts uzskatīt, ka kodola izvēle nav būtiska. Čebana [18] 2004. gadā ideju attīstījis tālāk un ieguvis, ka histogrammas kodolam optimālā svaru funkcijas izvēle ir $a(x) = 1/f_0$.

Simulētiem datiem $X_1, \ldots, X_n$ tiks pārbaudīta vienkāršā hipotēze

$$H_0 : f = f_0 \quad \text{pret} \quad H_1 : f \neq f_0,$$

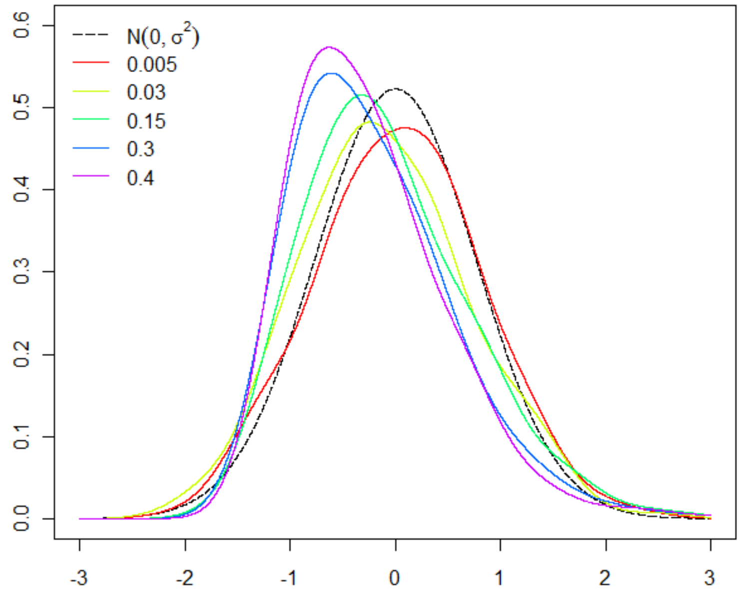
kur f_0 ir vienmērēgā sadalījuma intervālā $[0,1]$ blīvuma funkcija. Turklāt darba mērķiem būs nepieciešams izmantot tikai viendimensionālos sadalījumus, turpmāk $d = 1$.

Simulācijām par kodola funkciju izvēlēsimies optimālo, tas ir, vienmērīgo kodolu $K^*(x)$ un $a(x) \equiv 1$. Jāpiezīmē, ka visas simulācijas veiktas arī kodola funkcijai ar $N(0,1)$ sadalījumu, $K(x) = (2\pi)^{-1/2} e^{-x^2/2}$, taču galvenā uzmanība tiks pievērsta testam ar optimālo kodolu K^*. Iegūtie rezultāti izmantoti, lai iegūtu pilnīgāku skatu un salīdzinātu abas kodola izvēles, galvenās tabulas atrodamas A2. pielikumā.

3.1. Statistiku "slīdēšana"

No statistikas $T_n = n h_n^{1/2} \int [f_n(x) - (K_{h_n} * f_0)(x)]^2 \, dx$ redzams, ka vienīgais nezināmais ir gludinošais parametrs, tādēļ pirmais uzdevums ir saprast, cik liela nozīme ir h_n izvēlei.

Kā izrādās, h_n izvēlei ir būtiska nozīme. Pat pie liela izlases apjoma $n = 1000$, ja h_n izvēlēts nepareizi, statistiku blīvuma funkcija nekonverģē uz teorētisko. Lai to pārbaudītu



3.1. att.: Statistikas T_n, kurām atņemta vidējā vērtība μ, 1000 $v.s.[0,1]$ neatkarīgām izlasēm ar $n = 1000$ elementiem; apskatītas, mainoties 5 dažādiem h_0, $h_n = h_0 n^{-1/4}$; izmantots K^*.

tiek ģenerētas 1000 $v.s.[0,1]$ izlases apjomā $n = 1000$ un tām visām aprēķināta izteiksmes

$$T_n - \mu, \tag{3.1.1}$$

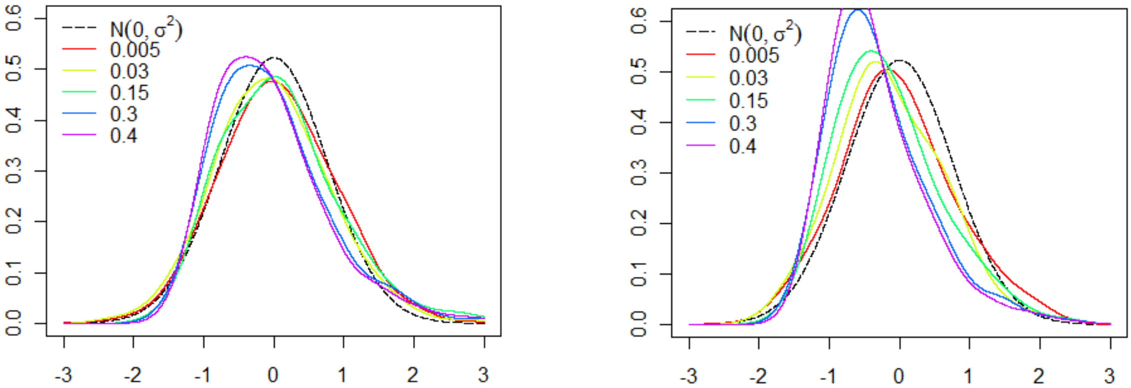
kur $\mu = h_n^{-1/2} \int K^2(u)du$ vērtība. Pie lieliem izlases apjomiem pēc teorēmas 5, tai būtu jātiecas uz $N(0, \sigma^2)$, taču to nevar viennozīmīgi apgalvot pie visām h_n vētībām. Grafiski simulētie dati parādīti 3.1. attēlā. Izteiksmes (3.1.1) 1000 realizāciju pie dažādiem joslas platumiem $h_n = h_0 n^{-1/4}$ neparametriskie blīvuma funkcijas novērtējumi parādīti ar krāsainajām līnijām.

Ja h_n izvēlēts pārāk liels, gan blīvuma funkcijas novērtējums f_n, gan konvolūcija $(K_{h_n} * f_0)(x)$ tiek pārgludināti. Rezultātā statistiku vērtības ir tuvas nullei un simulētie dati veido pīķi ap $(-\mu)$. Savukārt pie ļoti maziem h_n ir tendence veidoties “smagām astēm”, kas nav raksturīgi normālajam sadalījumam. Šo problēmu aprakstījis arī Fans [13].

Definīcija 8. Stacionāru procesu $\{X_t\}_{t \in Z}$, kas definēts kā

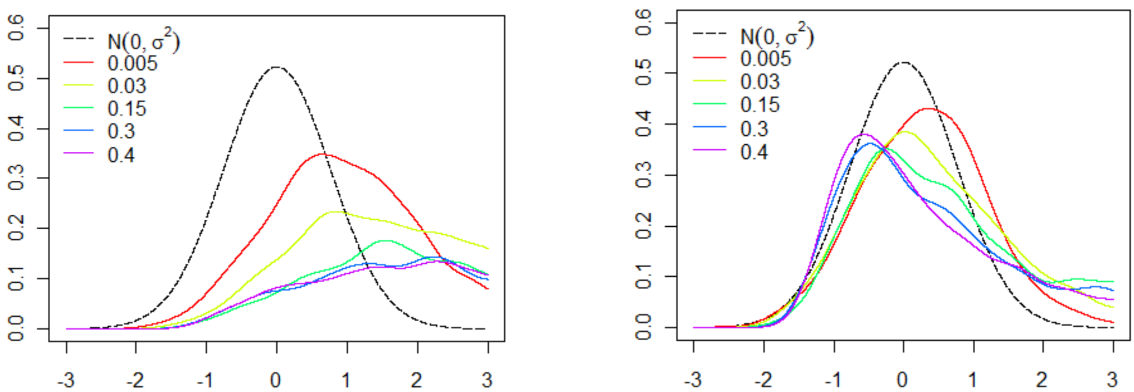
$$X_t := \varphi X_{t-1} + Z_t, \tag{3.1.2}$$

kur $\{Z_t\}_{t \in Z}$ ir vāji stacionārs atjaunojumu process ar vidējo vērtību 0 un autokovariāciju $\mathbb{E}(Z_t Z_{t+h}) = \sigma_Z^2 < \infty$, ja $h = 0$ un 0 pretējā gadījumā, $|\varphi| \leq 1$, sauc par autoregresīvo procesu $AR(1)$.



(a) $AR(1)$, $\varphi = 0.3$ (b) $AR(1)$, $\varphi = -0.3$

3.2. att.: Statistikas T_n, kurām atņemta vidējā vērtība μ, 1000 $v.s.[0,1]$ atkarīgām izlasēm ar $n = 1000$ elementiem; apskatītas, mainoties 5 dažādiem h_0, $h_n = h_0 n^{-1/4}$; izmantots K^*.



(a) $AR(1)$, $\varphi = 0.9$ (b) $AR(1)$, $\varphi = -0.9$

3.3. att.: Statistikas T_n, kurām atņemta vidējā vērtība μ, 1000 $v.s.[0,1]$ atkarīgām izlasēm ar $n = 1000$ elementiem; apskatītas, mainoties 5 dažādiem h_0, $h_n = h_0 n^{-1/4}$; izmantots K^*.

Atkarīgiem datiem (atkarīgu datu ģenerēšana aprakstīta 18. lpp.) ar $\varphi = \pm 0.3$ situācija ir ļoti līdzīga kā neatkarīgo datu gadījumā (3.2. attēls). Negatīvai atkarībai grafiks ir gandrīz identisks neatkarīgajam gadījumam, bet pozitīvai grafiki mazliet atšķiras. Sākotnēji var šķist, ka ar atkarības struktūru $AR(1)$, $\varphi = 0.3$ normālais sadalījums tiek aproksimēts vislabāk, bet turpmāk būs redzams (3.1., A1., 3.2. tabulās), ka neatkarīgiem un atkarīgiem ($AR(1)$ ar $\varphi = -0.3$) datiem statistikas uzvedība ir līdzīgāka, turklāt $1-\alpha$ kvantiles ar teorētiskā sadalījuma kvantilēm sakrīt labāk.

Atkarīgiem datiem ar $\varphi = \pm 0.9$ statistikas sadalījums pie izlases apjoma $n = 1000$ nav tuvs teorētiskajam, kas nozīmē, ka nav pareizi izvēlēts h_n. Tā kā tik slikti rezultāti iegūti tikai ar atkarības struktūru, kur $\varphi = \pm 0.9$, ir pamats domāt, ka labā h_n novērtējumā vajadzētu iekļaut atkarības struktūras novērtējumu. Šo un arī eksplozīvu procesu ($|\varphi| = 1.1$) gadījumu, plašāk pētījuši Lī (Lee) un Na (Na) [19].

3.2. Jaudas analīze

Jaudas analīzei izmantotas gludas alternatīvas, kuras savā 1995. gada publikācijā apskatījuši Ledvina (Ledwina) un Kalenbergs (Kallenberg) [20],

$$g_1(x) = 1 + \rho\cos(j\pi x) \tag{3.2.1}$$

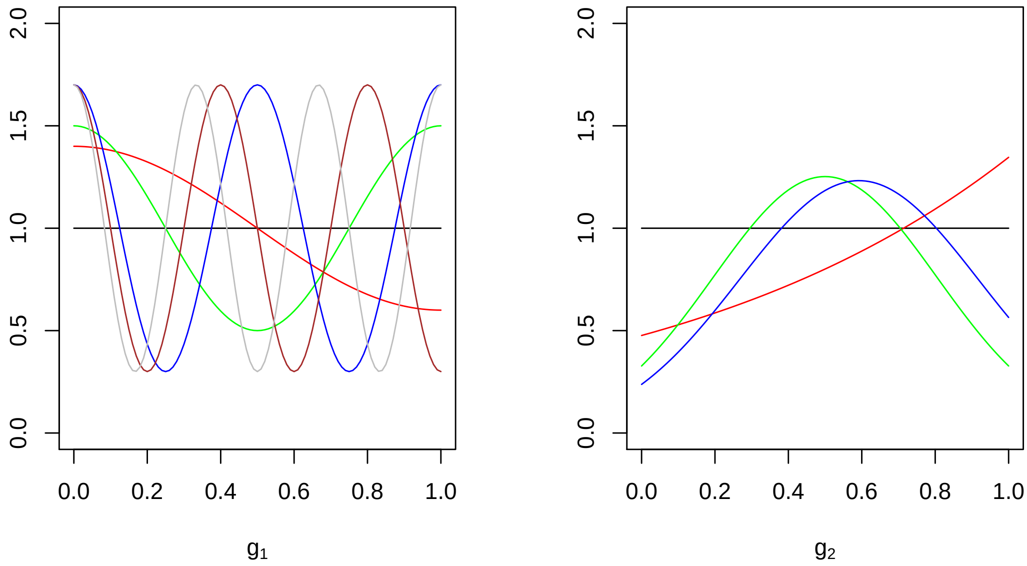
un

$$g_2(x) = \exp\left(\sum_{j=1}^{k} \theta_j \phi_j(x) - \psi_k(\theta)\right), \tag{3.2.2}$$

kur $\{\phi_j\}$ ir ortonormēti Ležandra polinomi intervālā $[0,1]$ formā (3.2.7), $\theta \in \mathbb{R}^k$, $\psi_k(\theta) = \ln \int_0^1 \exp(\theta \circ \phi(x))dx$ un $\circ$ apzīmē skalāro reizinājumu. Grafiski alternatīvas atspoguļotas 3.4. attēlā.

Datu ģenerēšana. Ņemot vērā, ka tests bez transformācijām derīgs arī atkarīgiem datiem, hipotēžu pārbaude veikta gan neatkarīgām izlasēm, gan atkarīgiem datiem, kas ģenerēti izmantojot autoregresīvo procesu $AR(1)$ (3.1.2). Datu ģenerēšana aprakstīta Munka u.c. publikācijā [6].

Vispirms simulē $X_1, \ldots, X_n$ no stacionāra $AR(1)$ procesa $\{X_t\}_{t \in Z}$. Tālāk ģenerē datus no (3.1.2) ar pieaugumu sadalījumu $Z_t \sim N(0, 1-\varphi^2)$, līdz ar to procesam $\{X_t\}_{t \in Z}$ būs $N(0,1)$ sadalījums. Lai iegūtu atkarīgus datus ar vienmērīgo sadalījumu intervālā $[0,1]$, tiek veikta datu transformācija ar $N(0,1)$ sadalījuma funkciju Φ. Iegūst izlasi



3.4. att.: Apskatītās alternatīvas g_1 (pa kreisi) un g_2 (pa labi) ar parametriem kā (3.2.4) un (3.2.5), un $v.s.[0,1]$ blīvuma funkcija.

$X_1', \ldots, X_n' = \Phi(X_1), \ldots, \Phi(X_n)$, kurai ir $v.s.[0,1]$ sadalījums. Darbā analizēti četri atkarības gadījumi ar $\varphi = \pm 0.3$ un $\varphi = \pm 0.9$.

Jaudas analīzei ar alternatīvām g_1 (3.2.1) un g_2 (3.2.2) arī nepieciešams lietot atkarīgus un neatkarīgus datus. To iegūšanai lieto inverso transformāciju iepriekš aprakstītajai metodei $Y = F_X(X)$, $Y \sim v.s.[0,1]$, tas ir,

$$F_X^{-1}(Y) = X \sim F_X, \tag{3.2.3}$$

kur F_X ir gadījuma lieluma X kumulatīvā sadalījuma funkcija. Alternatīvu ģenerēšanai pielieto kumulatīvās sadalījuma funkcijas F_{g_1} un F_{g_2}, Y vietā ievietojot iepriekš aprakstīto izlasi $Y = (X_1', \ldots, X_n')$, lai iegūtu atkarīgus datus, un $Y = (Y_1, \ldots Y_n)$, $Y \sim v.s.[0,1]$, lai iegūtu neatkarīgus alternatīvu datus apjomā n.

Joslas platuma h_n izvēle. Neparametriskajā statistikā ir zināms, ka labs joslas platuma h novērtējums ir $h = cn^{-1/5}$ (1.1.4), taču tas ir pretrunā ar ierobežojumiem uz h_n izvēli, tādēļ neapmierina vājās konverģences nosacījumus, uz ko norādījis arī Čebana (Chebana) [18]. Līdz ar to Bikela–Rozenblata testam neder tāds pats joslas platums kā labākajam blīvuma funkcijas novērtējumam. To apstiprināja arī empīriskā analīze, veicot simulācijas. Sākotnējā loģiskā doma bija izvēlēt h ar krosvalidācijas metodi (1.1.6), bet tā deva ļoti sliktus rezultātus. Šī testa vajadzībām blīvuma funkcijas novērtējums bija par daudz nogludināts, tas ir, tika izvēlēts pārāk liels joslas platums. Arī literatūrā, piemēram, Čebana [18], apstiprināts, ka šis h novērtējums ir labs blīvuma funkcijas novērtēšanai, bet ne hipotēžu pārbaudei.

Literatūrā nav aprakstīta procedūra, kā izvēlēties optimālu h_n, kas vienlaicīgi ierobežotu pirmā veida kļudu un minimizētu otrā veida kļūdu. Dažādos rakstos par Bikela–Rozenblata testu ir sastopami dažādi veidi, kā tiek izvēlēts h_n, piemēram, Noimans un Paparoditis bez diskusijas simulācijām izvēlas konstantu $h = 0.03$, atsevišķi konstanta h ietekmi ar lielām, mazām un vidējām vērtībām $h = \{2.0, 1.0, 0.2, 0.05\}$ apskatījis Tenreiro (Tenreiro) [21]. Taču vislabāk aprakstīta un piedevām ar vislielākajām manevra iespējām ir joslas platuma izvēle $h_n = h_0 n^{-\delta}$, kur $\delta = 1/4$, kā to darījuši Fans [13] un Čebana [18]. Šāda izvēle tiks lietota arī darbā. Ar $\delta = 1/4$ un konstantām h_0 vērtībām

$$h_0 = \{0.005; 0.01; 0.02; 0.03; 0.05; 0.1; 0.15; 0.2; 0.25; 0.3; 0.35; 0.4\}.$$

Simulētās empīriskās kritiskās vērtības. Lai varētu pielietot testa statistiku, jāatrod empīriskās kritiskās vērtības. To izdara, aprēķinot $1 - \alpha$ kvantili no statistikām, kas izskaitļotas 10000 izlasēm ar hipotētisko sadalījumu katrai n un h_n kombinācijai. Teorētiskās vērtības pielietot būtu nepamatoti, jo pie maza izlases apjoma n statistiku vērtības vēl nekonverģē uz asimptotisko sadalījumu $N(0, \sigma^2)$.

3.1. tabula: T_n procentuālais patiesās hipotēzes H_0 noraidīšanas apjoms pie nozīmības līmeņa $\alpha = 0.05$ neatkarīgām izlasēm apjomā n; simulētas 10000 $v.s.[0,1]$ izlases; $h_n = h_0 n^{-1/4}$; izmantots $v.s.[-1,1]$ kodols.

	h_0											
n	0.005	0.01	0.02	0.03	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4
20	6.99	6.55	6.17	5.78	**5.09**	4.17	3.58	3.03	2.69	2.30	1.85	1.53
50	6.90	6.61	6.03	5.59	**4.86**	4.34	3.93	3.49	2.98	2.65	2.39	2.25
100	6.94	6.45	5.56	5.34	**5.21**	4.41	3.93	3.65	3.40	3.15	2.84	2.53
500	6.06	6.21	6.35	5.60	5.67	**4.99**	4.51	4.45	4.21	3.88	3.70	3.45
1000	6.58	6.69	6.50	5.22	**4.99**	5.28	5.58	4.68	4.41	4.21	4.09	3.87

Kā tas parādīts 3.1. nodaļā, liels izlases apjoms vēl negarantē sakritību ar normālo sadalījumu, jābūt arī pareizi izvēlētam joslas platumam. Kritērijs labai h_n izvēlei lielām izlasēm ir empīriskās un teorētiskās kritiskās vērtības sakrišana. Citiem vārdiem sakot, kad pirmā veida kļūda ir $\alpha = P(T_n \geq \mu + z_\alpha \sigma | H_0)$. Pretējā gadījumā iespējams izvēlēties kritisko vērtību, kas labi pielāgota izlasei, bet šādam testam nebūtu jēgas.

Neatkarīgu datu gadījumam 3.1. tabulā apkopota patiesas hipotēzes noraidīšana procentos, izmantojot teorētisko kritisko vērtību. Ja skaitlis tuvojas $\alpha \times 100\%$, tad empīriskā kritiskā vērtība ir tuva teorētiskajai. Darbā izvēlētais nozīmības līmenis ir $\alpha = 0.05$, līdz ar to h_0 izvēlei tabulā jāmeklē vērtība 5%. No tabulas var redzēt, ka pamatoti būtu izvēlēt $h_n = 0.05n^{-1/4}$, līdz ar to arī testa jaudas analīzē būtu jāņem vērā vērtības, kas iegūtas pie fiksēta $h_0 = 0.05$.

3.2. tabula: T_n procentuālais patiesās hipotēzes H_0 noraidīšanas apjoms pie nozīmības līmeņa $\alpha = 0.05$ atkarīgām $\big(AR(1),\ \varphi = 0.3\big)$ izlasēm apjomā n; simulētas 10000 $v.s.[0,1]$ izlases; $h_n = h_0 n^{-1/4}$; izmantots $v.s.[-1,1]$ kodols.

	h_0											
n	0.005	0.01	0.02	0.03	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4
20	7.51	6.97	7.29	7.36	7.21	7.13	6.83	6.75	6.50	6.05	5.75	5.36
50	7.15	7.23	7.35	7.31	7.46	7.98	7.79	7.64	7.50	7.11	6.80	6.32
100	6.98	6.73	7.08	7.34	7.32	7.80	7.71	7.64	7.59	7.33	7.06	6.94
500	7.68	7.65	8.02	7.06	7.80	8.01	7.75	7.95	8.09	8.00	8.08	8.07
1000	7.48	8.12	8.48	6.85	6.71	8.27	9.63	8.54	8.87	8.85	8.77	8.66

Ne tik viennozīmīga situācija kā neatkarīgo datu gadījumā ir atkarīgiem datiem, kuru ģenerēšanai izmantots $AR(1)$, ar $\varphi = -0.3$, process. No A1. tabulas var redzēt, ka par labu h_n visdrīzāk jāuzskata tādu, kur $h_0 = 0.03$, kaut arī vērtības ir tuvas 5%, kas liecina par pamatotu testa lietošanu vāji atkarīgiem datiem. Tomēr šeit iezīmējas problēma testa praktiskā pielietošanā. Bikela–Rozenblata testam literatūrā nav atrodama procedūra, ar kuras palīdzību izvēlēties optimālu h_n. Īpaši svarīgi tas ir maziem izlases apjomiem, kur statistikas īstais sadalījums var atšķirties no teorētiskā.

Gadījums ar atkarīgiem datiem, kad $\varphi = 0.3$, dažādiem n apkopots 3.2. tabulā. Šīs vērtības ir samērā līdzīgas kā neatkarīgo datu gadījumā, taču neviena nav ļoti tuva 5% robežai. Gluži atšķirīgu ainu rāda $\varphi = 0.9$ un $\varphi = -0.9$ gadījumi 3.3. un A2. tabulās, kur skaidri redzams, ka empīriskās kritiskās vērtības nemaz nav tuvas teorētiskajām. To uzskatāmi parāda arī 3.3. attēls, kur redzams, ka statistiku vērtības izlasēm apjomā $n =$

3.3. tabula: T_n procentuālais patiesās hipotēzes H_0 noraidīšanas apjoms pie nozīmības līmeņa $\alpha = 0.05$ atkarīgām $(\varphi = 0.9)$ izlasēm apjomā n.

	h_0											
n	0.005	0.01	0.02	0.03	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4
20	32.57	40.44	49.02	54.49	61.33	68.45	70.69	70.59	70.31	69.21	67.58	65.88
50	37.47	46.68	56.37	62.36	68.14	74.47	75.82	76.36	75.81	75.00	74.01	72.62
100	40.76	50.36	60.14	66.03	71.49	76.90	78.41	78.92	78.17	77.61	76.72	75.86
500	42.55	53.74	64.22	67.86	73.94	78.53	80.32	80.99	81.11	81.03	80.57	80.03
1000	41.00	52.78	63.50	66.02	71.38	77.63	80.92	80.52	81.01	80.97	80.77	80.59

1000 nekonverģē uz teorētisko sadalījumu.

3.4. tabula: Empīriskā jauda statistikai T_n (%) alternatīvām $g_1(\rho, j)$ un $g_2(\theta)$, kas iegūta no 1000 neatkarīgu izlašu apjomā $n = 50$ statistikām ar vienmērīgo kodolu $[-1, 1]$.

		h_0											
ρ	j	0.005	0.01	0.02	0.03	**0.05**	0.1	0.15	0.2	0.25	0.3	0.35	0.4
0.4	1	9.2	10.3	13.9	15.6	**20.5**	28.1	33.2	36.1	38.7	41.7	43.5	45.4
0.5	2	11.6	12.3	17.6	21.4	**26.8**	37.9	43.9	48.8	51.8	54.0	55.3	55.5
0.7	4	17.0	24.6	34.8	43.2	**55.5**	70.0	73.3	73.6	69.6	64.9	55.2	40.4
0.7	5	17.6	23.9	34.9	43.1	**55.0**	66.7	69.2	66.4	55.7	38.9	23.2	13.3
0.7	6	16.9	22.6	32.8	40.9	**52.7**	62.4	60.3	50.9	33.5	18.6	10.2	7.2
θ													
(0, 3)		25.9	23.5	24.4	26.0	**28.4**	32.8	36.4	40.1	42.5	44.6	46.6	48.2
(0,-0. 4)		10.4	12.4	17.5	21.9	**26.8**	34.0	39.1	41.9	42.9	45.3	46.0	45.8
(0.25,-0. 35)		10.0	12.0	17.6	21.0	**27.2**	33.7	38.2	42.5	45.2	48.5	49.6	50.5

Testa jauda. Lai noskaidrotu, cik labi Bikela–Rozenblata statistika der hipotēzes par blīvuma funkciju pārbaudei, tiks lietotas iepriekš aprakstītās alternatīvas g_1 (3.2.1) un g_2 (3.2.2), kur parametri izvēlēti kā [20, 6], lai būtu iespējams salīdzināt iegūtos rezultātus,

$$g_1(\rho, j), \text{ kur } (\rho, j) = \{(0.4; 1), (0.5; 2), (0.7; 4), (0.7; 5), (0.7; 6)\}, \tag{3.2.4}$$

$$g_2(\theta), \text{ kur } \theta = \{(0; 3), (0; -0.4), (0.25; -0.35)\}. \tag{3.2.5}$$

Jaudas analizētas neatkarīgiem un atkarīgiem ($\varphi = \pm 0.3$, $\varphi = \pm 0.9$) datiem. Ar jaudu saprot daudzumu, cik bieži tiek noraidīta nepatiesā hipotēze H_1 katrai no alternatīvām

$g_1(\rho, j)$ un $g_2(\theta)$,

$$\beta_n^{br}(h) = P(T_n(h) > l_\alpha | H_1), \tag{3.2.6}$$

kur l_α ir empīriskā kritiskā vērtība jeb $1-\alpha$ kvantile simulētās izlases sadalījumam. Darbā tiek apskatīta jauda procentos, tas ir, $\beta_n(h) \cdot 100$ (%). Piemēram, 3.4. tabulā pie $h_0 = 0.05$ un alternatīvas $g_1(0.7, 5)$, 550 no 1000 neatkarīgu gadījuma lielumu izlasēm ($n = 50$) tiek noraidīta hipotēze H_0 par vienmērīgo sadalījumu.

Bikela–Rozenblata statistika literatūrā ir daudz salīdzināta ar klasiskajiem Pīrsona un Kolmogorova–Smirnova testiem. Jau savā klasiskajā publikācijā [3] Bikels un Rozenblats teorētiski parāda, ka viņu tests ir stingri labāks par Pīrsona χ^2 testu, bet ne par Kolmogorova–Smirnova testu noteiktai alternatīvu klasei (Pitmana alternatīvas). Pēc diviem gadiem Rozenblats [22] parāda alternatīvu klasi, kurai Bikela–Rozenblata tests ir labāks par testiem, kas balstīti uz empīrisko sadalījuma funkciju (piemēram, Kolmogorova–Smirnova tests). Salīdzinājumu ar Kolmogorova–Smirnova testu sniedzis arī Fans [13], kā arī Bahmans (Bachmann) un Dete (Dette) [23], kuri parādījuši, ka alternatīvām ar χ^2 sadalījumu Bikela–Rozenblata tests strādā labāk.

3.5. tabula: Empīriskā jauda statistikai T_n (%) alternatīvām $g_1(\rho, j)$ un $g_2(\theta)$, kas iegūta no 1000 atkarīgu izlašu ($AR(1)$, $\varphi = 0.3$) apjomā $n = 50$ statistikām ar vienmērīgo kodolu $[-1, 1]$.

		h_0											
ρ	j	0.005	0.01	0.02	0.03	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4
0.4	1	8.1	9.3	12.4	13.8	16.7	20.9	24.4	26.5	27.9	29.7	30.3	30.8
0.5	2	10.8	11.9	15.9	19.5	25.3	31.8	36.2	38.6	39.7	40.5	39.9	38.9
0.7	4	18.1	22.8	31.7	39.5	48.4	61.1	66.3	64.0	59.0	50.9	37.9	25.7
0.7	5	16.7	21.7	32.4	38.8	48.3	58.8	60.3	52.9	41.1	26.5	15.9	10.4
0.7	6	16.6	22.1	30.2	39.1	48.2	56.7	52.4	40.7	24.2	11.7	7.8	6.4
θ													
(0, 3)		27.6	23.3	22.5	24.0	26.1	28.4	32.5	33.3	33.6	35.1	36.5	37.1
(0,-0. 4)		10.9	10.9	14.0	16.2	19.4	24.9	31.3	32.9	33.7	34.1	33.8	32.3
(0.25,-0. 35)		9.5	10.6	14.3	16.5	20.0	26.9	32.9	34.9	37.1	38.2	38.5	38.0

Neimaņa tests Darbā Bikela–Rozenblata tests tiek salīdzināts ar citu testu, kas arī derīgs atkarīgiem datiem. Ledvina 1994. gadā [1] ieviesa procedūru, kā no izlases datiem izvēlēties ortonormētas sistēmas kārtu k klasiskajam Neimaņa testam. Līdz ar to arī bija iespējams parādīt, ka Neimaņa testam maziem izlašu apjomiem ir labākas īpašības kā citiem klasiskajiem testiem [2].

Neimaņa tests hipotēzes pārbaudei $H_0 : F = v.s.[0,1]$, $H_1 : F \neq v.s.[0,1]$ par gadījuma lielumu $X_1, \ldots, X_n$ sadalījumu ir formā

$$R_k = \sum_{j=1}^{k} \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \phi_j(X_i) \right]^2, \quad k = 1, 2, \ldots,$$

kur $\phi_j,\ j = 1, \ldots, k$ ir ir ortonormēti polinomi intervālā $L_2[0,1]$ un $\phi_0(x) = 1$. Testam tiek izvēlēti ortonormēti Ležandra polinomi intervālā $[0,1]$. Tie ir rekursīvi definējami polinomi, kas priekš $j = 1, 2, 3, 4$ ir formā

$$\begin{aligned}
\phi_1(x) &= \sqrt{12}(x - 1/2), \\
\phi_2(x) &= \sqrt{5}(6(x - 1/2)^2 - 1/2), \\
\phi_3(x) &= \sqrt{7}(20(x - 1/2)^3 - 3(x - 1/2)), \\
\phi_4(x) &= 210(x - 1/2)^4 - 45(x - 1/2)^2 + 9/8.
\end{aligned} \tag{3.2.7}$$

Nesenā publikācijā 2009. gadā Munks u.c. [6] ievieš Neimaņa testa statistiku atkarīgiem datiem. Gadījuma procesam $(X_t)_{t\in\mathbb{Z}}$, kas ir α–jauktais process (definīcija 7) tiek definēta jauna statistika

$$N_k = \frac{1}{12\sigma^2} R_k = \frac{1}{12\sigma^2} \sum_{j=1}^{k} \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \phi_j(X_i) \right]^2,$$

kur

$$\sigma^2 = \sum_{t=-\infty}^{+\infty} Cov(X_0, X_t).$$

Reizinātājs $(12\sigma^2)^{-1}$ koriģē statistiku R_k, lai, izpildoties H_0, statistikas N_n robežsadalījums būtu tāds pats kā R_k neatkarīgajā gadījumā.

Lai izvēlētos nezināmo parametru k, Ledvina [1] izmanto Švarca selekcijas likumu, tā padarot Neimaņa testu par ļoti labu alternatīvu citiem testiem. Nesen publicētajos rezultātos, analizējot Neimaņa testu atkarīgiem datiem, Munks u.c. [6] lieto divas šī likuma modifikācijas modifikācijas

$$S_{mod} = \min\{k : 1 \leq k \leq d(n);\ R_k - k \ln n \geq R_j - j \ln n;\ j = 1, \ldots, d(n)\}$$

un

$$S_{mod2} = \min\{k : 1 \leq k \leq d(n);\ N_k - k \ln n \geq N_j - j \ln n;\ j = 1, \ldots, d(n)\}.$$

Jauda Neimaņa testam neatkarīgiem un atkarīgiem ($AR(1),\ \varphi = 0.3$) datiem apkopota 3.6. tabulā. Izvēlētas alternatīvas $g_1(\rho, j)$, kur $(\rho, j) = \{(0.4;1), (0.5;2), (0.7;4)\}$, un $g_2(\theta)$, kur $\theta_1 = 0.3,\ \theta_2 = (0; 0.4),\ \theta_3 = (0.25; -0.35)$.

3.6. tabula: [24] Empīriskā jauda (%) alternatīvām $g_1(\rho, j)$ un $g_2(\theta)$ ar izvēlētiem parametriem, kas iegūta no (a) 1000 neatkarīgu izlašu statistikām R_k ar k izvēles likumu $Smod$; (b) 1000 atkarīgu izlašu ($AR(1),\ \varphi = 0.3$) statistikām N_k ar k izvēles likumu $Smod2$; $d(n) = 10$; $n = 50$.

	$g_1(\rho, j)$			$g_2(\theta)$		
Statistika	(0.4;1)	(0.5; 2)	(0.7;4)	θ_1	θ_2	θ_3
R_{Smod}	34.23	58.16	52.73	38.02	58.16	57.62
N_{Smod2}	31.82	30.35	14.94	33.95	56.18	23.95

Neatkarīgiem datiem apjomā $n = 50$ Neimaņa tests (R_{Smod}) uzrāda labākus rezultātus par Bikela–Rozenblata testu ar $h_0 = 0.05$ pie alternatīvām $g_1((0.4;1))$, $g_1((0.5;2))$ un visām $g_2(\theta)$, taču ir pārāks pie alternatīvas $g_1((0.7;4))$. Tomēr var redzēt, ka izvēloties atšķirīgus h_0, Bikela–Rozenblata testa jauda ir ļoti tuva Neimaņa testa jaudai.

Atkarīgiem datiem ar $\varphi = 0.3$ jaudas pārākums Neimaņa testam vairs nav tik izteikts kā neatkarīgiem datiem. Izlasēm apjomā $n = 50$ Bikela–Rozenblata tests pie atsevišķiem h_0 uzrāda labāku rezultātu kā N_{Smod2} alternatīvām g_1 un $g_2(\theta_3)$. Taču jāņem vērā rezultāti no 3.2. tabulas, kas parāda, ka pirmā veida kļūda pie ši φ visiem h_0 ir lielāka par 5%, tādēļ ir grūti adekvāti spriest par testa jaudu. Savukārt datiem ar atkarības koeficientu $\varphi = \pm 0.9$ 3.3. un A2. tabulas uzrāda tik sliktus rezultātus, ka nebūtu atbilstoši runāt par jaudu, kamēr nav pārliecības, kā izvēlēties kritiskās vērtības. Šo problēmu uzskatāmi atspoguļo 3.3. attēls, kur redzams, ka T_n sadalījums nesakrīt ar neatkarīgo datu gadījumu.

Šie vērojumi un grūti interpretējamie rezultāti liek secināt, ka Bikela–Rozenblata tests bez procedūras optimālā h_n izvēlei ir praksē neizmantojams, taču ar potenciālu uzrādīt labas īpašības, tādēļ būtu nepieciešams jauns rezultāts, metode h_n izvēlei.

4. JOSLAS PLATUMA IZVĒLE KODOLA TESTIEM

Iepriekš apspriestas Bikela–Rozenblata testa labās īpašības, kas padara šo testu unikālu. Tas bez modifikācijām derīgs gan atkarīgiem, gan neatkarīgiem datiem. Vēl vairāk, tas lietojams d–dimensionāliem datiem vienkāršām un saliktām hipotēzēm. Tomēr bez procedūras, kā izvēlēties piemērotu joslas platumu h_n, šis tests praksē nav pielietojams.

4.1. Optimālā h atrašanas metode

Nesenā publikācijā 2008. gadā Gao (Gao) un Gijbelsa (Gijbels) [25] izstrādāja procedūru, kā izvēlēties optimālo h testiem, kas salīdzina parametrisko regresijas modeli ar neparametrisko. Arī viņu apskatītajā testā lietota integrētā kvadrātiskā kļūda (L_2–attālums), tāpat kā Bikela–Rozenblata testā. Tas padara šos rezultātus vēl interesantākus. Turpmāk tiks paskaidrota nesen ieviestās procedūras galvenā ideja.

Dota statistika vispārīgā formā

$$\hat{T}_n(h) = \sum_{i=1}^{n}\sum_{j=1}^{n} e_i \omega_n(X_i, X_j) e_j,$$

kurai, līdzīgi kā aprakstīts 3.2. nodaļā, $l_\alpha = l_\alpha(h)$ ir simulētās statistikas sadalījuma $1-\alpha$ kvantile. Tiek lietota arī kvantiles butstrapa aproksimācija $l_\alpha^* = l_\alpha^*(h)$. Katram h definē nozīmīguma (*size*) un jaudas funkcijas

$$\alpha_n(h) := P(\hat{T}_n(h) > l_\alpha | H_0)$$

un

$$\beta_n(h) := P(\hat{T}_n(h) > l_\alpha | H_1).$$

Nākamais solis ir atrast *Edgeworth* izvirzījumus funkcijām $\alpha_n(h)$ un $\beta_n(h)$ un izvest l_α koriģēto vērtību. Šo izvirzījumu būtība ir aproksimēt sadalījuma funkciju, šajā gadījumā statistikas $\hat{T}_n$ sadalījuma funkciju $P(\hat{T}_n \leq x)$, tā iegūstot uzlabotu statistikas sadalījuma funkciju, kas ir īpaši svarīgi maziem izlašu apjomiem. Vispārīgai statistikai S_n ar standartnormālo sadalījumu uzlabotā sadalījuma funkcija izsakāma formā

$$P(S_n \leq x) = \Phi(x) + n^{-1/2} p_1(x)\phi(x) + n^{-1} p_2(x)\phi(x) + \ldots + n^{-j/2} p_j(x)\phi(x) + \ldots,$$

kur p_j ir polinomi ar kārtu ne lielāku par $3j-1$, bet Φ un ϕ ir $N(0,1)$ sadalījuma un blīvuma funkcijas.

Pasaulē šī ir aktuāla izpētes tēma un aizvien vairāk statistikām tiek atrasti uzlabotie sadalījumi. Latviski *Edgeworth* izvirzījumi ir plašāk analizēti Vucānes maģistra darbā [26] un šeit netiks apskatīti smalkāk.

Pēdējais solis ir atrast optimālo h. Tas tiek darīts, ierobežojot nozīmīgumu $\alpha_n(h)$ un maksimizējot jaudu $\beta_n(h)$. Optimālo joslas platumu atrod, atrisinot optimizācijas problēmu

$$h_{ew} = \arg \max_{h \in H_n(\alpha)} \beta_n(h),$$

kur $H_n(\alpha) = \{h : \alpha - c_{\min} < \alpha_n(h) < \alpha + c_{\min}\}$ mazam $c_{\min} > 0$. Publikācijā [25] $c_{\min}$ tiek izvēlēts kā $c_{\min} = \alpha/10$, lai arī sākotnējā izvēles rekomendācija ir $0 < c_{\min} < \alpha$.

Kad ir izvests h, atliek testa statistikai aprēķināt optimālā h vērtību un lietot asimptotiski uzlaboto l_α hipotēžu pārbaudei.

Aprakstīto procedūru joslas platuma atrašanai bija iespējams realizēt, jo varēja pielietot 2007. gada Gotzes (Gotze), Tihomirova (Tikhomirov) un Jurčenko (Yurchenko) rezultātu [27]. Neatkarīgiem vienādi sadalītiem gadījuma lielumiem $X_1, X_2, \ldots$, kuriem $\mathbb{E}X_1 = 0$ un $\mathbb{E}X_1^2 = \mu_2$ tika apskatīta statistika formā

$$Q_n = S_n + R_n,$$

kur

$$S_n = \sum_{j=1}^{N} a_{jj}(X_j^2 - \mu_2) \quad \text{un} \quad R_n = \sum_{1 \le j \ne k \le N} a_{jk} X_j X_k,$$

kur $\mathbf{A} := \mathbf{A}^{(n)} := (a_{jk})_{j,k=1}^{N} := (a_{jk}^{(n)})_{j,k=1}^{N}$, $\|\mathbf{A}\|^2 = \sum_{j,k=1}^{N} a_{j,k}^2 < \infty$. Šai statistikai, kas ir U–statistika (4.2.1), tika atrasts *Edgeworth* izvirzījums. Tā kā Gao un Gijbelsas apskatīto statistiku bija iespējams uzrakstīt formā Q_n, tad sekojot [27] varēja atrast *Edgeworth* izvirzījumus funkcijām $\alpha_n(h)$ un $\beta_n(h)$. Lai varētu runāt par šīs procedūras adaptēšanu Bikela–Rozenblata statistikai, tiks ieviestas U–statistikas un deģenerētas U–statistikas.

4.2. U–statistikas

Pieņemsim, ka $\mathcal{P}$ ir nepārtrauktu sadalījuma funkciju saime un ka $\theta(F)$ apzīmē reālvērtīgu funkciju, kur $F \in \mathcal{P}$ un $\mathcal{P}$ ir varbūtību mērs.

Definīcija 9. Funkciju $\theta(F)$ sauc par novērtējamu parametru, ja $\exists m \in \mathbb{Z}$ tāds, ka eksistē reālvērtīga integrējama funkcija $h(x_1, \ldots, x_m)$, kurai

$$\mathbb{E}_F(h(X_1, \ldots, X_m)) = \theta(F), \quad \forall F \in \mathcal{P}$$

un $X_1, \ldots, X_m$ ir neatkarīgi vienādi sadalīti gadījuma lielumi ar sadalījuma funkciju F. Mazāko $m \in \mathbb{Z}$ sauc par $\theta(F)$ pakāpi.

Definīcija 10. Pieņemsim, ka doti neatkarīgi gadījuma lielumi $X = (X_1, \ldots, X_n)$, $X \sim F$, $n \geq m$ un dota simetriska, mērojama, reālvērtīga funkcija $h(X_1, \ldots, X_m)$, tad par U–statistiku ar kodolu h sauc

$$U_n := U_n(h) := \frac{1}{C_n^m} \sum_{\mathbf{C}_{m,n}} h(X_{i_1}, \ldots, X_{i_m}), \tag{4.2.1}$$

kur summēšana notiek pa m veselo skaitļu $i_1, i_2, \ldots, i_m$, kas izvēlēti no $1, 2, \ldots, n$, kombināciju kopu $\mathbf{C}_{m,n}$.

Apgalvojums 7 ([28]). *Ja $\theta(F)$ ir novērtējams parametrs, tad U_n ir nenovirzīts parametra $\theta(F)$ novērtējums.*

Dotai simetriskai kodola funkcijai $h(x_1, \ldots, x_m)$ un novērtējamam parametram $\theta = \theta(F)$ pieņemsim, ka $\mathcal{P}$ ir sadalījuma funkciju saime, kam izpildās $\mathbb{D}(h(X_1, \ldots, X_m)) < \infty$. Definē ar h saistītu funkciju virkni,

$$h_c(x_1, \ldots, x_c) := \mathbb{E}h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m), \quad c = 1, 2, \ldots, m,$$

kur $X_{c+1}, \ldots, X_m$ ir neatkarīgi gadījuma lielumi ar sadalījuma funkciju F. Tad $h_0 = \theta$ un $h_m(x_1, \ldots, x_m) = h(x_1, \ldots, x_m)$.

Visu šo funkciju matemātiskā cerība ir θ,

$$\mathbb{E}h_c(X_1, \ldots, X_c) = \mathbb{E}h(X_1, \ldots, X_c, X_{c+1}, \ldots, X_m) = \theta,$$

taču funkcijas h_c nesauc par kodoliem, jo tās var būt atkarīgas no F.

U–statistikas (4.2.1) dispersija ir atkarīga no h_c dispersijām, katram $c = 1, 2, \ldots, m$,

$$\sigma_c^2 = \mathbb{D}(h_c(X_1, \ldots, X_c)) \tag{4.2.2}$$

un $\sigma_0^2 = 0$, $\sigma_m^2 = \mathbb{D}(h(X_1, \ldots, X_m))$.

Teorēma 8 ([28]). *Ja $\sigma_m^2 < \infty$, tad $\sqrt{n}(U_n - \theta) \xrightarrow{d} N(0, m^2\sigma_1^2)$.*

Definīcija 11. U–statistiku sauc par deģenerētu ar pakāpi k, ja $\sigma_1^2 = \cdots = \sigma_k^2 = 0$ un $\sigma_{k+1}^2 > 0$.

Deģenerētība bieži ir spēkā, kad ar U–statistikām tiek testētas hipotēzes un, izpildoties nulles hipotēzei, statistikas dispersija ir nulle.

Piemērs 1. Ja $\mathcal{P}$ ir visu reālvērtīgo sadalījumu ar galīgu k–to momentu kopa, tad dotiem gadījuma lielumiem $X_1, \ldots, X_n$ moments $\mu_k = \int x^k dF(x)$ ir pirmās pakāpes novērtējams parametrs ar U–statistiku $\frac{1}{n}\sum_1^n X_i^k$, kur kodols $h(x) = x$. Jeb empīriskie momenti ir U–statistikas.

Piemērs 2. Par kodolu pieņem $h(x_1, x_2) = x_1 x_2$, tad $h_1(x_1) = \mathbb{E}(x_1 X_2) = x_1 E(X_2) = x_1 \mu$ un $\sigma_1^2 = \mathbb{D}(h_1(X_1)) = \mu_2 \sigma^2$, kur $\sigma^2 = \mathbb{D}(X_1)$. No teorēmas 8 seko, ka

$$\sqrt{n}(U_n - \mu^2) \xrightarrow{d} N(0, 4\mu^2\sigma^2).$$

Taču pieņemot, ka izpildās nulles hipotēze par gadījuma lielumu nepārtrauktu sadalījumu F, kas ir simetrisks ap nulli, $\mu = \mathbb{E}(X_1) = 0$. Attiecīgi robežsadalījuma dispersija ir nulle. Taču pieņemot, ka $\sigma^2 > 0$, iegūst $\sigma_2^2 = \mathbb{D}(X_1 X_2) = \sigma^4 > 0$, līdz ar to deģenerētība ir ar pakāpi $k = 1$.

Lai atrastu statistikas U_n asimptotisko sadalījumu izlasei $X_1, X_2, \ldots$ ar vidējo vērtību 0 un dispersiju σ^2, statistiku pārraksta formā

$$\begin{aligned} U_n &= \frac{1}{n(n-1)} \sum\sum_{i \neq j} X_i X_j = \frac{1}{n(n-1)} \left[\left(\sum_{i=1}^n X_i \right)^2 - \sum_{i=1}^n X_i^2 \right] \\ &= \frac{1}{n-1} \left[\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \right)^2 - \frac{1}{n} \sum_{i=1}^n X_i^2 \right] \end{aligned}$$

No centrālās robežteorēmas izriet, ka $\frac{1}{\sqrt{n}}\sum_1^n X_i \xrightarrow{d} N(0, \sigma^2)$, un no lielo skaitļu likuma $\frac{1}{n}\sum_1^n X_i^2 \xrightarrow{d} \sigma^2$. Pielietojot Slutska teorēmu Fergusons (Ferguson) [28] iegūst

$$nU_n \xrightarrow{d} (Z^2 - 1)\sigma^2, \quad \text{kur } Z \in N(0, 1).$$

4.3. Joslas platums Bikela–Rozenblata testam

Sākotnējā doma, iedvesmojoties no Gao un Gijbelsas darba [25], bija arī Bikela–Rozenblata statistikai pielietot Gotzes, Tihomirova un Jurčenko rezultātu, bet tiešā veidā neizdevās to pārrakstīt formā Q_n, kas ir U–statistika. Kā vēlāk izrādījās tas nemaz nav iespējams. To izskaidro Bahmana un Detes [23] publikācija, kur parādīts, ka T_n ir deģenerēta U–statistika (Definīcija 11).

Neatkarīgiem, vienādi sadalītiem gadījuma lielumiem $Z_1, Z_2, \dots$ ar blīvuma funkciju f, vidējo vērtību 0 un dispersiju $\sigma^2 > 0$ viņi parāda, ka statistikai T_n (2.2.2) spēkā

$$\frac{T_n}{n\sqrt{h}} - \frac{1}{nh}\int K^2(x)dx - \int [K_h * (f - f_0)]^2(x)dx = U_n^* + \frac{2}{n}\sum_{i=1}^{n} Y_i + O_P\left(\frac{1}{n}\right), \quad (4.3.1)$$

kur, nodefinējot lielumus $e_h := K_h * f$ un $g_h := K_h * (f - f_0)$, U_n^* ir deģenerēta U–statistika,

$$\begin{aligned} U_n^* &= \frac{2}{n^2}\sum_{i<j} H_n(Z_i, Z_j) \\ &= \frac{2}{n^2}\sum_{i<j}\int \left[K_h(x - Z_i) - e_h(x)\right]\left[K_h(x - Z_j) - e_h(x)\right]dx \end{aligned} \quad (4.3.2)$$

un

$$Y_i = (K_h * g_h)(Z_i) - \mathbb{E}[K_h * g_h(Z_i)].$$

Izpildoties nulles hipotēzei $H_0 : f = f_0$, gan g_h, gan trešais saskaitāmais izteiksmes (4.3.1) kreisajā pusē kļūst vienādi ar nulli, līdz ar to T_n var izteikt ar deģenerēto U–statistiku U_N^*, pārrakstot (4.3.1) formā

$$\frac{T_n}{n\sqrt{h}} - \frac{1}{nh}\int K^2(x)dx = U_n^* + O_P\left(\frac{1}{n}\right).$$

Bahmans un Dete [23], izmantojot šo izvedumu, ar vājākiem nosacījumiem arī pierāda klasisko Bikela–Rozenblata rezultātu (Teorēma 1) U–statistikai

$$n\sqrt{h}U_n \xrightarrow{d} N(0, \sigma^2),$$

kur σ^2 definēts kā (2.2.3).

Kaut arī eksistē *Edgeworth* izvirzījumi U–statistikām, literatūrā nav atrodami izvirzījumi deģenerētām U–statistikām formā (4.3.2). Tomēr ir rezultāti, kur deģenerētām, asimptotiski normāli sadalītām U–statistikām hipotēžu testēšanai ir atrasti *Edgeworth* izvirzījumi. Samērā nesen, 2003. gadā, publicētajā Fana (Fan) un Lintona (Linton) [29] darbā, tiek apskatīta deģenerēta U–statistika testam par regresijas funkciju. Šai statistikai viņi iegūst jaunus teorētiskus rezultātus – izved *Edgeworth* izvirzījumus un iegūst koriģētas kritiskās vērtības.

Līdzīgā veidā būtu jārīkojas arī ar Bikela–Rozenblata statistiku. Sekojot Fana un Lintona rezultātam jāatrod deģenerētās U–statistikas (4.3.2) nozīmīguma $\alpha_n(h)$ un jaudas

$\beta_n(h)$ funkciju *Edgeworth* izvirzījums. Tālāk, pielietojot Gao un Gijbelsas [25] piedāvāto metodi, jāatrod optimālais h_n un uzlabotās kritiskās vērtības. Šāds jauns rezultāts ļautu padarīt Bikela–Rozenblata testu pielietojamu praksē.

Secinājumi

Ar darbā sniegtajiem piemēriem parādīts, ka Bikela–Rozenblata testam joslas platuma izvēle ir kritiska, taču pagaidām literatūrā nav piedāvāta procedūra optimālā h_n atrašanai. Tas padara testu grūti salīdzināmu ar citiem testiem un arī praksē nepielietojamu. Neskatoties uz teorētiskiem rezultātiem, kas ļauj testu izmantot arī atkarīgiem datiem, simulācijas parāda, ka kritiskās vērtības un sadalījums statistikām no $AR(1)$, kur $\varphi = \pm 0.9$ iegūtiem datiem stipri atšķiras no neatkarīgo un $\varphi = \pm 0.3$ datu gadījumiem. Pat pie izlašu apjoma $n = 1000$ statistiku sadalījums neatbilst teorētiskajam, kas liek secināt, ka joslas platums izvēlēts nepareizi. Darbā apskatītā metode joslas platuma izvēlei bija labākā no literatūrā piedāvātajām, kas patiesībā ir apzināti izvēlēta, lai būtu līdzīga klasiskajai izvēlei neparametriskajā blīvuma funkcijas novērtēšanā. Lielā atšķirība starp koeficienta φ izvēlēm parāda, ka labai h_n izvēlei būtu jāņem vērā arī datu atkarības struktūra.

Kaut gan ir pierādīts statistikas T_n teorētiskais sadalījums, arī lieliem izlašu apjomiem simulāciju rezultāti parāda būtisko h_n izvēles lomu. Pārlieku lielām h_n vērtībām tiek pārgludināts blīvuma funkcijas novērtējums un nogludinātais hipotētiskais sadalījums, līdz ar to statistikas vērtības tiecas uz nulli un to sadalījums veido pīķi ap nulli. Nav arī skaidras atbildes, vai izvēle $h_n = h_0 n^{-1/4}$ nodrošina īsto konverģences ātrumu, tas liedz atrast optimālo h_0 visiem izlašu apjomiem.

Hipotēžu pārbaudē teorētiski pamatoti būtu lietot teorētiskās kritiskās vērtības, taču šādai izvēlei ir nozīmīgs pretarguments. Ja tiktu izmantotas teorētiskās kritiskās vērtības abiem, $\varphi = \pm 0.9$, atkarīgo datu gadījumiem, testa jauda būtu ļoti augsta, bet šādam testam nebūtu nekādas jēgas. Par pieņemamu metodi, lai h_n izvēli uzskatītu par labu noder empīriskās un teorētiskās kritiskās vērtības sakritība. Tas nozīmē, ka pie attiecīgās h_n izvēles statistikas sadalījums, izpildoties nulles hipotēzei, pietiekami labi aproksimē teorētisko sadalījumu.

Jaudas analīze parādīja, ka Bikela–Rozenblata tests ir salīdzināms ar Neimaņa testu gan atkarīgiem, gan neatkarīgiem datiem. Ja tiktu atrisināta joslas platuma izvēles problēma, šis tests varētu izrādīties nopietns konkurents Neimaņa testam, kas mazām izlasēm tiek uzskatīts par labāko testu.

Nesenie Gao un Gijbelsas rezultāti paver iespēju atrisināt problēmu par joslas platuma izvēli Bikela–Rozenblata testam. Darbā apskatīti teorētiskie rezultāti, kas iegūti līdzīgiem testiem regresijas problemātikai. Tie nevar tikt tiešā veidā piemēroti interesējošajai statistikai, taču dod pilnīgu skatu par nepieciešamajiem soļiem, lai atrastu optimālo joslas platumu. Svarīgi, ka ir skaidra metode, kā šo problēmu var atrisināt. Ir tikai jāaizpilda trūkstošie posmi, jāatrod *Edgeworth* izvirzījumi deģenerētām U–statistikām ar zināmu kodola funkciju. Kad izteikta jaudas un nozīmīguma funkcija, jāatrisina optimizācijas problēma, lai iegūtu optimālo joslas platumu. Tālāk atliek tikai atrast koriģētās kritiskās vērtības, lai varētu pielietot testu.

Šādi jauni rezultāti beidzot padarītu Bikela–Rozenblata testu pielietojamu un ļautu izmantot daudzās priekšrocības, kas tam piemīt pār citiem testiem.

Izmantotā literatūra un avoti

[1] T. Ledwina. Data-driven version of Neyman's smooth test of fit. *Journal of the American Statistical Association*, 89(427):1000–1005, 1994.

[2] T. Inglot and T. Ledwina. Asymptotic optimality of data-driven Neyman's tests for uniformity. *The Annals of Statistics*, 24(5):1982–2019, 1996.

[3] P. J. Bickel and M. Rosenblatt. On some global measures ot the deviations of density function estimates. *The Annals of Statistics*, 1(6):1071–1095, 1973.

[4] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27:832–837, 1956.

[5] M. H. Neumann and E. Paparoditis. On bootstrapping L_2-type statistics in density testing. *Statistics & Probability Letters*, 50(2):137–147, 2000.

[6] A. Munk, J. P. Stockis, J. Valeinis, and G. Giese. Neyman smooth goodness-of-fit tests for the marginal distribution of dependent data. *Annals of the Institute of Statistical Mathematics*, 2009. Published online.

[7] W. Hardle, M. Muller, S. Sperlich, and A. Werwatz. Nonparametric and semiparametric models: An introduction. http://fedc.wiwi.hu-berlin.de/xplore/ebooks/html/spm/, 2004.

[8] M. Rosenblatt. A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences USA*, 42:43–47, 1956.

[9] R. C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

[10] J. Rone. *Jaukto procesu analīze un to pielietojums statistikā: Maģistra darbs.* LU FMF Matemātikas nodaļa, Rīga, 2009.

[11] R. C. Bradley. *Introduction to strong mixing conditions. Vol. 1.* Kendrick Press., Heber City, UT, 2007.

[12] W. Hardle and E. Mammen. Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, 21(4):1926–1947, 1993.

[13] Y. Fan. Testing the Goodness of Fit of a Parametric Density Function by Kernel Method Yanqin Fan. *Econometric Theory*, 10(2):316–356, 1994.

[14] H. Takahata and K. Yoshihara. Central limit theorems for integrated square error of nonparametric density estimators based on absolutely regular random sequences. *Yokohama Mathematical Journal*, 35(1-2):95–111, 1987.

[15] A. Dvoretzky. Asymptotic normality for sums of dependent random variables. Proc. 6th Berkeley Sympos. math. Statist. Probab., Univ. Calif. 1970, 2, 513-535, 1972.

[16] M. Neumann and E. Paparoditis. A nonparametric test for the stationary density. Technical report, 1998.

[17] B.K. Ghosh and W.M. Huang. The power and optimal kernel of the Bickel-Rosenblatt test for goodness of fit. *The Annals of Statistics*, 19(2):999–1009, 1991.

[18] F. Chebana. On the optimization of the weighted bickel–rosenblatt test. *Statistics & Probability Letters*, 68(4):333–345, 2004.

[19] S. Lee and S. Na. On the bickel-rosenblatt test for first-order autoregressive models. *Statistics & Probability Letters*, 56(1):23–35, 2002.

[20] W. C. M. Kallenberg and T. Ledwina. Consistency and monte carlo simulation of a data driven version of smooth goodness-of-fit tests. *The Annals of Statistics*, 23(5):1594–1608, 1995.

[21] C. Tenreiro. On the role played by the fixed bandwidth in the bickel-rosenblatt goodness-of-fit test. *Statistics and Operations Research Transactions*, 29(2):201–216, 2005.

[22] M. Rosenblatt. A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *The Annals of Statistics*, 3:1–14, 1975.

[23] D. Bachmann and H. Dette. A note on the Bickel - Rosenblatt test in autoregressive time series. *Statistics & Probability Letters*, 74(3):221–234, 2005.

[24] J. Valeinis. Goodness-of-fit tests for weakly dependent data. Abstracts of the 10th International Vilnius Conference on Probability and Mathematical Statistics, 2010.

[25] J. Gao and I. Gijbels. Bandwidth selection in nonparametric kernel testing. *Journal of the American Statistical Association*, 103(484):1584–1594, 2008.

[26] S. Vucāne. *Bārtleta korekcija empīriskās ticamības metodei: Maģistra darbs.* LU FMF Matemātikas nodaļa, Rīga, 2009.

[27] F. Gotze, A. Tikhomirov, and V. Yurchenko. Asymptotic expansion in the central limit theorem for quadratic forms. *Zapiski Nauchnykh Seminarov POMI*, 341:81–114, 2007.

[28] T. S. Ferguson. Statistics 200c: Large sample theory. http://www.math.ucla.edu/ tom/Stat200C/Ustat.pdf, 2005.

[29] Y. Fan and O. Linton. Some higher-order theory for a consistent non-parametric model specification test. *Journal of Statistical Planning and Inference*, 109(1-2):125–154, 2003.

A Tabulas

A1. Kodols $v.s.[-1, 1]$

A1. tabula: T_n procentuālais patiesās hipotēzes H_0 noraidīšanas apjoms pie nozīmības līmeņa $\alpha = 0.05$ atkarīgām $\big(AR(1),\ \varphi = -0.3\big)$ izlasēm apjomā n; simulētas 10000 $v.s.[0, 1]$ izlases; $h_n = h_0 n^{-1/4}$; izmantots $v.s.[-1, 1]$ kodols.

	h_0											
n	0.005	0.01	0.02	0.03	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4
20	6.53	6.10	5.42	**4.68**	4.26	2.97	2.39	1.86	1.48	1.08	0.90	0.66
50	6.26	5.97	**5.31**	**5.31**	4.59	3.59	2.91	2.48	2.07	1.63	1.49	1.29
100	6.02	5.77	**4.98**	4.82	4.44	3.69	3.40	3.01	2.56	2.33	1.96	1.66
500	5.91	5.94	6.00	5.21	**5.20**	4.34	3.91	3.53	3.13	2.85	2.54	2.29
1000	5.95	6.05	5.99	4.49	4.29	4.52	**4.88**	3.66	3.42	3.36	3.11	2.78

A2. tabula: T_n procentuālais patiesās hipotēzes H_0 noraidīšanas apjoms pie nozīmības līmeņa $\alpha = 0.05$ atkarīgām $\big(AR(1),\ \varphi = -0.9\big)$ izlasēm apjomā n; simulētas 10000 $v.s.[0, 1]$ izlases; $h_n = h_0 n^{-1/4}$; izmantots $v.s.[-1, 1]$ kodols.

	h_0											
n	0.005	0.01	0.02	0.03	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4
20	14.09	15.88	18.11	19.39	21.52	23.74	24.12	23.03	21.87	20.40	19.05	17.15
50	15.91	18.77	21.61	23.77	25.90	28.22	28.21	27.59	26.87	25.88	24.54	23.19
100	16.42	19.11	22.42	24.64	27.03	29.94	30.13	29.93	28.95	28.11	27.00	25.81
500	14.39	18.69	22.89	23.51	26.62	29.68	30.61	30.47	30.20	29.89	29.31	28.76
1000	14.68	18.91	22.86	22.64	25.40	30.01	32.57	31.48	31.47	31.43	31.08	30.75

A3. tabula: Empīriskā jauda statistikai T_n $(\%)$ alternatīvām $g_1(\rho, j)$ un $g_2(\theta)$, kas iegūta no 1000 atkarīgu izlašu $\big(AR(1),\ \varphi = -0.3\big)$ apjomā $n = 50$ statistikām ar vienmērīgo kodolu $[-1, 1]$.

		h_0											
ρ	j	0.005	0.01	0.02	0.03	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4
0.4	1	6.1	9.2	10.8	12.8	15.2	22.5	28.6	34.4	37.3	41.5	44.9	48.1
0.5	2	8.9	12.8	16.6	19.2	26.3	38.9	46.0	50.7	54.6	57.5	58.7	60.7
0.7	4	15.3	21.9	33.8	40.5	53.0	70.0	75.7	77.9	75.4	70.4	62.2	48.7
0.7	5	16.3	22.6	33.6	39.7	53.6	67.9	70.8	68.1	58.3	42.2	26.0	15.8
0.7	6	16.1	22.0	32.0	38.9	52.2	64.2	64.8	55.3	39.0	22.2	12.5	9.6
θ													
(0, 3)		25.1	25.3	25.2	24.5	27.6	34.6	40.1	45.3	49.8	53.7	57.4	60.9
(0,-0. 4)		7.8	11.1	15.4	17.1	23.2	33.4	38.4	42.5	44.8	47.0	47.7	48.3
(0.25,-0. 35)		9.3	11.7	15.5	17.4	23.7	33.8	40.2	45.7	48.5	51.6	53.3	55.5

A2. Kodols $N(0, 1)$

A4. tabula: T_n procentuālais patiesās hipotēzes H_0 noraidīšanas apjoms pie nozīmības līmeņa $\alpha = 0.05$ neatkarīgām izlasēm apjomā n; simulētas 10000 $v.s.[0, 1]$ izlases; $h_n = h_0 n^{-1/4}$; izmantots $N(0, 1)$ kodols.

	h_0											
n	0.005	0.01	0.02	0.03	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4
20	5.92	5.36	**4.99**	4.50	3.76	2.95	2.23	1.72	1.26	0.83	0.55	0.35
50	5.68	**5.33**	4.60	4.14	3.83	3.10	2.49	2.04	1.75	1.44	1.23	0.91
100	5.58	**5.02**	4.45	4.30	3.93	3.26	2.75	2.55	2.13	1.83	1.46	1.16

A5. tabula: Empīriskā jauda statistikai T_n $(\%)$ alternatīvām $g_1(\rho, j)$ un $g_2(\theta)$, kas iegūta no 1000 neatkarīgu izlašu apjomā $n = 50$ statistikām ar normālo kodolu $N(0, 1)$.

		h_0											
ρ	j	0.005	0.01	0.02	0.03	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4
0.4	1	9.2	10.3	13.9	15.6	20.5	28.1	33.2	36.1	38.7	41.7	43.5	45.4
0.5	2	11.6	12.3	17.6	21.4	26.8	37.9	43.9	48.8	51.8	54.0	55.3	55.5
0.7	4	17.0	24.6	34.8	43.2	55.5	70.0	73.3	73.6	69.6	64.9	55.2	40.4
0.7	5	17.6	23.9	34.9	43.1	55.0	66.7	69.2	66.4	55.7	38.9	23.2	13.3
0.7	6	16.9	22.6	32.8	40.9	52.7	62.4	60.3	50.9	33.5	18.6	10.2	7.2
θ													
(0, 3)		25.9	23.5	24.4	26.0	28.4	32.8	36.4	40.1	42.5	44.6	46.6	48.2
(0,-0. 4)		10.4	12.4	17.5	21.9	26.8	34.0	39.1	41.9	42.9	45.3	46.0	45.8
(0.25,-0. 35)		10.0	12.0	17.6	21.0	27.2	33.7	38.2	42.5	45.2	48.5	49.6	50.5

A6. tabula: Empīriskā jauda statistikai T_n $(\%)$ alternatīvām $g_1(\rho, j)$ un $g_2(\theta)$, kas iegūta no 1000 atkarīgu izlašu $(AR(1),\ \varphi = -0.3)$ apjomā $n = 50$ statistikām ar normālo kodolu $N(0, 1)$.

		h_0											
ρ	j	0.005	0.01	0.02	0.03	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4
0.4	1	14	20.7	27.9	33.6	44.5	58.4	66.1	70.4	72.7	74.8	76.8	78
0.5	2	19.7	29	43.9	52.2	64.5	77.2	82.7	86.9	88	87.6	87.1	85.8
0.7	4	42.8	62.5	81.9	89.1	95.4	98.4	98.5	97.6	95.1	85.7	68.2	41.2
0.7	5	42.5	64	82.9	89	94.2	97	95.5	89.4	71	38.3	16.9	10.5
0.7	6	41.2	63	83.4	90.8	95.3	96.6	91.1	71.6	31.5	13.4	7.2	5.2
φ													
(0, 3)		46.3	47	53	54.1	61.4	68.5	75.7	78.6	81.2	82.4	83.2	83.5
(0,-0. 4)		19.2	29.4	41.2	48.7	60.3	72.1	76.9	80	80.4	80.2	79.7	78.4
(0.25,-0. 35)		17.9	29.8	41.8	49.6	61.8	74.9	80.6	83	84.4	85.2	86.3	85.5

B Izveidoto programmu kods

```
theta<-0.3
g<-function(i)
{
x<-arima.sim(list(order=c(1,0,0),
 ar=theta),n= i, sd=sqrt(1-theta^2))
xx<-pnorm(x)
xx[1:i]
}

f0<-function(u) dunif(u)

K<-function(x)
 {
 x[abs(x)<=1]<-1
 x[abs(x)>1]<-0
 1/2*x
 }
K2<-function(x) K(x)^2

###conv
conv12<-function(x)
  {
    if ((x <= - h) | (x >= 1 + h))
      {
         0
      }
    else
      {
         if (x < h)
           {
             (x + h)/2/h
```

```
          }
         else
          {
            if (x > 1 - h)
              {
                (1 - x + h)/2/h
              }
            else
              {
                 1
               }
           }
       }
  }
conv<-Vectorize(conv12)

convprim<-function(x)
 {
    if ((x <= - h) | (x >= 1 + h))
      {
        if (x >= 1 + h)
          1
        else
          0
      }
    else
      {
         if (x < h)
          {
           x/2 + x^2/4/h + h/4
          }
         else
          {
            if (x > 1 - h)
              {
```

```
                x/2 + x/2/h - x^2/4/h
                + 1/2 - h/4 - 1/4/h
               }
             else
               {
                  x
               }
           }
       }
 }
primC<-Vectorize(convprim)

#BR-statistika
stat<-function(izl)
 {
 n<-length(izl)
 izm <- c( c(izl-h, izl+h), -h, h, 1-h, 1+h )
 augd <- c( c(rep(1, length(izl)),
 rep(-1, length(izl))), rep(0,4) )/n/h/2
 ooo  <- order(izm)
 izm <- izm[ooo]
 augd <- augd[ooo]
 caug <- cumsum(augd)
 nn <- length(izm)

 sask1<-sum((caug[1:(nn - 1)])^2 * diff(izm))
 sask2<-sum( caug[1:(length(izm) - 1)]
 * diff(primC(izm)) )
 sask3<-integrate(function(x) conv(x)^2,
 -Inf,Inf,subdivisions=50)$value

 Tn<-n*h^(1/2)*(sask1-2*sask2+sask3)
 Tn
 }
stat_v<-Vectorize(stat)
```

```
####Kritiskās vērtības###

alpha<-0.05
ni<-c(100,50,20)
h0<-c(0.005,0.01,0.02,0.03,0.05
 ,0.1,0.15,0.2,0.25,0.3,0.35,0.4)
delta<-1/4
rep<-1000

n1<-1
n2<-length(ni)
h1<-1
h2<-length(h0)

for (i in n1:n2)
 {
 n<-ni[i]
 Kr.v.e<-c()
# Kr.v.t<-c()
 for (j in h1:h2)
  {
  h<-h0[j]*n^(-delta)
  set.seed(123)
  rep1<-10000
  rez<-replicate(rep1,stat(g(n)))
  cat(rez,file=paste("T_rep"..."))
  rez<-scan(file=paste("T_rep"..."))
#  mu<-h^(-1/2)*integrate(K2,-1,1)$value
  Kr.v.e[j]<-quantile(rez,1-alpha)
print(paste(i,j,'Kr.v.e.',date()))
#  Kr.v.t[j]<-qnorm(1-alpha,mu,sigma)
  }
 cat(Kr.v.e,file=paste("emp_Kr.v..."))
# cat(Kr.v.t,file=paste("teor_Kr.v..."))
```

```
 print(paste(i,j,date()))
 }



###bisektrišu metode
bis<-function(x0,y0,fun,prec)
 {
 a<-x0;b<-y0
 while (b-a>prec)
  {
  if (fun(a+(b-a)/2)>0)
   b<-a+(b-a)/2 else a<-a+(b-a)/2
  }
 a+(b-a)/2
}

###datu ģenērēšana
data.gen<-function(n,alt)
 {
 DD<-g(n)
 Data<-c()
 for(i in 1:(n))
  {
  aimfun<-function(x)
   integrate(alt,0,x)[[1]]-DD[i]
  Data[i]<-bis(0,1,aimfun,0.001)
  }
 Data
 }

###Cos alternatīvas
alt.1<-function(x) 1+0.4*cos(1*pi*x)
alt.2<-function(x) 1+0.5*cos(2*pi*x)
alt.3<-function(x) 1+0.7*cos(4*pi*x)
```

```
alt.4<-function(x) 1+0.7*cos(5*pi*x)
alt.5<-function(x) 1+0.7*cos(6*pi*x)

###Eksp.alternatīvas
###Ležanra polinomi
P1<-function(x) (x-0.5)*sqrt(12)
P2<-function(x) sqrt(5)*(6*(x-0.5)^2-0.5)

theta11<-0.3;
f.const<-function(x) exp(theta11*P1(x))
const<-log(integrate(f.const,0,1)[[1]])
alt.exp1<-function(x) exp(theta11*P1(x)- const)
...

#####statistikas#####
for (i in n1:n2)
 {
 n<-ni[i]

# set.seed(123)
# dg<-data.gen(n*rep,alt.1)
# cat(dg,file=paste("dati_n..."))
 dg<-scan(file=paste("dati_..."))
 dati.alt1<-c()
 for (r in 1:rep) dati.alt1[[r]]<-dg[ ((r-1)*n+1) : (r*n) ]
...

# set.seed(123)
# dg<-data.gen(n*rep,alt.exp1)
# cat(dg,file=paste("dati_n"...))
 dg<-scan(file=paste("dati_n...))
 dati.exp1<-c()
 for (r in 1:rep) dati.exp1[[r]]<-dg[ ((r-1)*n+1) : (r*n) ]
...]
```

```
 ###statistikas alternatīvām
 for (j in h1:h2)
  {
  h<-h0[j]*n^(-delta)

  rez.alt1<-stat_v(dati.alt1)
  cat(rez.alt1,file=paste("jrez_rep..."))
  print(paste(i,j,'alt1',date()))

  rez.alt2<-stat_v(dati.alt2)
  cat(rez.alt2,file=paste("jrez_rep..."))
  print(paste(i,j,'alt2',date()))
  ...

  rez.exp1<-stat_v(dati.exp1)
  cat(rez.exp1,file=paste("jrez_rep..."))
  print(paste(i,j,'alt.exp1',date()))
  ...
  }

 }


###JAUDA###

rep<-1000
rep1<-10000

for (i in 1:length(ni))
 {
 emp_Kr.v<-scan(file=paste("emp_Kr.v..."))
 ee<-c()

###jauda.cos
 for (k in 1:5)
```

```
  {
  emp_jauda<-c()
  for (j in 1:length(h0))
   {
   rez.br<-scan(file=paste("jrez_rep..."))
   emp_jauda[j]<-length(rez.br[rez.br>=emp_Kr.v[j]])
   }

  cat(emp_jauda,file=paste("jauda(emp_Kr.v)..."))
  ee[[k]]<-emp_jauda
  }
 ee[[6]]<-replicate(length(h0),0)

 for (k in 1:3)
  {
  emp_jauda<-c()

  for (j in 1:length(h0))
   {
   rez.br<-scan(file=paste("jrez_rep..."))
   emp_jauda[j]<-length(rez.br[rez.br>=emp_Kr.v[j]])
   }

  cat(emp_jauda,file=paste("jauda(emp_Kr.v)_rep..."))
#  cat(teor_jauda,file=paste("jauda_rep..."))
  ee[[k+6]]<-emp_jauda
#  tt[[k+6]]<-teor_jauda
  }

 mtx.e<-c()
# mtx.t<-c()
 for (l in 1:length(h0))
  {
  xx.e<-c()
#  xx.t<-c()
```

```
  for (m in 1:9)
   {
   xx.e[m]<-ee[[m]][l]
#    xx.t[m]<-tt[[m]][l]
   }
  mtx.e[[l]]<-xx.e
#  mtx.t[[l]]<-xx.t
 }

 write.table(mtx.e,file=paste("jauda(emp_Kr.v)..."),
  row.names = F,col.names = F)


 }
```

Diplomdarbs "Bikela-Rozenblata tests" izstrādāts LU Fizikas un matemātikas fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti un iesniegtā darba elektroniskā kopija atbilst izdrukai.

Autors: Audris Ločmelis

____________________ ________________

(paraksts) (datums)

Rekomendēju darbu aizstāvēšanai.

Vadītājs: docents Dr.math. Jānis Valeinis

____________________ ________________

(paraksts) (datums)

Recenzente: docente Dr.math. Nedežda Siņenko

Darbs iesniegts Matemātikas nodaļā ________________

(datums)

____

(darbu pieņēma)

Diplomdarbs aizstāvēts valsts pārbaudījuma komisijas sēdē

________________ prot. Nr. ________, vērtējums________

(datums)

Komisijas sekretāre: asoc. prof. Dr.math. Inese Bula ________________

(paraksts)