LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

MAIŅAS PUNKTA NOTEIKŠANA LAIKRINDU ANALĪZĒ

DIPLOMDARBS

Autors: **Agris Vaselāns**

Stud. apl. av07050

Darba vadītājs: doc. Dr. math. Jānis Valeinis

RĪGA 2012

ANOTĀCIJA

Diplomdarba mērķis ir apskatīt maiņas punkta noteikšanas problemātiku laikrindu analīzē. Tā kā dažas maiņas punkta noteikšanas metodes atkarīgām datu struktūrām saistītas ar spektrālo analīzi, tad darba pirmā daļa veltīta stacionāru procesu spektrālajai analīzei. Pēc tam aplūkota maiņas punkta noteikšanas problemātika vispārējā formā. Diplomdarbā apskatītas vairākas metodes maiņas punkta noteikšanai lokācijas parametram: kumulatīvo summu metode, metode izmantojot pāreju uz spektrālo domēnu un Furjē koeficientu butstrapošanu. Ieviesta un pielietota jauna blokveida empīriskās ticamības funkcijas (EL) metode divu izlašu problēmām atkarīgu datu gadījumā, kas pielietota lokācijas parametra maiņas punkta noteikšanai. Metožu salīdzināšanai veiktas simulācijas un pielietoti iegūtie rezultāti praktiskām datu problēmām.

Atslēgas vārdi: Maiņas punkts, spektrs, empīriskās ticamības funkcija, datu blokošana, datu pārkārtošana, TFT butstraps, CUMSUM statistika.

ABSTRACT

The goal of this work is to examine change - point detection methods in time series analysis. In the first part of this work an insight into spectral analysis results is given, because some of the change - point detection methods use terms of the time series spectral analysis. The aspects of change - point detection are discussed, in particular, change point detection for the location parameter and the use of the CUMSUM statistic for independent data. More specifically, detection method with TFT bootstrap for dependent data is considered and new approach for the change - point detection problem is proposed. Its basic idea is to modify two - sample empirical likelihood method that is used to test equality of the two sample means. Methods are illustrated with real data analysis and simulation study.

Keywords: Change - point, spectrum, empirical likelihood, data blocking, data resampling, TFT bootstrap, CUMSUM.

Saturs

APZĪMĒJUMI 1

IEVADS 2

1. STACIONĀRU PROCESU SPEKTRĀLĀS ANALĪZES PAMATREZULTĀTI 4

1.1. Stacionāra stohastiska procesa spektrs un spektrālā blīvuma funkcija 4

1.2. $ARMA(p,q)$ spektri 6

1.3. Sezonālo $ARIMA(p,d,q) \times (P,D,Q)_s$ spektri 7

1.4. Periodogramma un tās gludināšana 9

1.5. Kodolu gludinātā periodogramma 11

1.6. Punktveida ticamības intervāli 16

1.7. Taperošana 17

1.8. Neparametriskās regresijas metodes periodogrammas gludināšana . 19

2. MAIŅAS PUNKTA ANALĪZE 24

2.1. Maiņas punkta analīze 24

2.2. Piemēri CUMSUM grafiki 27

2.3. TFT butstrapa metode 30

3. DIVU IZLAŠU LOKĀCIJAS PARAMETRU VIENĀDĪBAS TESTS AR EL, TĀ PIELIETOJUMS MAIŅAS PUNKTA ANALĪZĒ 33

3.1. Motivācija un īss vēsturisks ieskats 33

3.2. Problēmas nostādne 34

3.3. Datu bloku veidošana jeb blokošana 36

3.4. Metodes empīriskā analīze 39

3.5. pārklājuma precizitātes analīze 39

3.6. Maiņas punkta noteikšana ar divu izlašu testu un logiem 43

4. METOŽU PIELIETOJUMS 45

4.1. Reālu datu piemēri. 45

5. NOBEIGUMS 48

Izmantotā literatūra un avoti **49**

A PIELIKUMS. **51**

A1. EL pārklājuma precizitātes programpaketes R kods 52

A2. TFT metodes programpaketes R kods 59

A3. EL maiņas punkta programpaketes R kods 64

APZĪMĒJUMI

ACF - autokorelāciju funkcijas apzīmējums.

AIC - Aikake informācijas kritērijs

AMOC (angliski *At Most One Change*) - AMOC hipotēžu pārbaudes apzīmējums.

BIC - Beijesa informācijas kritērijs

$B(\cdot)$ - Brauna tilta procesa apzīmējums.

DFT - diskrētā Furjē transformācija

EL - empīriskā ticamības funkcija

iid - neatkarīgi un vienādi sadalīti (dati).

$MSE(\cdot)$ - vidējās kvadrātiskās kļūdas apzīmējums.

TFT - (angliski *Time frequency Toggle*) - TFT butstrapa metodes apzīmējums.

$\rightarrow^d$ konverģence pēc sadalījuma

IEVADS

Maiņas punkta noteikšana pēdējajā desmitgadē kļuvusi par arvien aktuālāku tematu matemātiskajā satistikā, it īpaši atkarīgu datu struktūru gadījumā (skatīt, piemēram, [1]. Šajā diplomdarbā apskatīta tieši maiņas punkta noteikšana laikrindu gadījumā, galvenokārt, apskatot tieši lokācijas parametra maiņas punkta noteikšanu.

Tā kā atsevišķas maiņas punkta analīzes metodes cieši saistītas ar stohastisku procesu spektrālo analīzi, tad diplomdarba sākuma daļa veltīta tieši stacionāru procesu spektrālās analīzes pamatjedzienu un problemātikas apskatam. Spektrālās analīzes pirmsākumi meklējami jau 19. gadsimta beigās, kad Šusters (Schuster) 1898. gadā ieviesa periodogrammas jēdzienu, kas aprakstīts viņa 1906. gada darbā izpētot periodiskumu datos par saules aktivitāti [2, 181. lpp.], tomēr aktualitāte saglabājas arī mūsdienās, pateicoties dažādas skaitļošanas un statistiskās pieejas stohastisku procesu analīzes metožu attīstībai. Spektrālā analīze galvenokārt balstās uz Furjē analīzi (diskrēto Furjē transformāciju realizāciju analīzē) un dod iespēju analizēt stohastiskus procesus nevis laika apgabalā (domēnā), bet gan frekvenču apgabalā. Tiek apskatīts procesa cikliskums un tādējādi iegūta jauna pieeja procesa analīzē izmantojot frekvenču jēdzienu ar mērķi atklāt procesā apslēptu ciklu. Spektrālā analīze ir svarīga fizikā trokšņu spektru analīzē, ekonometrijā ekonomisko procesu ciklu izpētē, inženierzinātnēs signālu pārraidīšanas analīzē, kā arī astronomijā un citās nozarēs.

Savukārt visās maiņas punkta analīžu metodēs galvenā problemātika saistīta ar spēju pārliecināties, vai modelis ir laikā nemainīgs, vai arī ir notikušas izmaiņas. Kad noskaidrots, ka modelī ir notikušas izmaiņas, varam apskatīt galvenos jautājumus, kurus cenšas risināt ar maiņas punkta analīzes palīdzību, piemēram, "Kad modelī notikušas izmaiņas?", "Vai izmaiņas ir vienas vienīgas?", "Cik maiņas punkti bijuši?"utt. Lai sniegtu atbildes uz šiem jautājumiem, lieti noder spektrālās analīzes pamati, kurus izmantojot, var veikt maiņas punkta analīzi ar TFT (angliski *Time Frequency Toggle*) butstrapa metodi. Šo metodi pavisam nesen piedāvāja publikācijā [3]. Šīs metodes idejas pamatā ir procesa spektra novērtējuma aplūkošana un Furjē koeficientu butstrapošna, kas dod labākus rezultātus nekā jau plašāk pazīstamie periodogrammas butstrapošnas metodes.

Darbā apskatītas maiņas punkta noteikšanas metodes lokācijas parametram un izmantots jau matemātiķa statistiķa programmas zinātniski - pētnieciskā praksē apskatītais divu izlašu tests izmantojot empīriskās ticamības funkciju (EL) atkarīgu datu gadījumā. Šī ideja attīstīta tālāk un ieviesta empīriskās ticamības funkcijas metode dažādām divu

izlašu problēmām atkarīgu datu gadījumā, kas pielietota maiņas punkta noteikšanai, veikts metožu salīdzinājums diplomdarba praktiskajā daļā. Tātad par šī diplomdarba mērķi tika izvirzīts:

- aplūkot stohastisku stacionāru procesu spektrālo analīzi;
- apskatīt maiņas punkta analīzes vispārējo problemātiku un CUMSUM statistiku;
- analizēt maiņas punkta analīzes metodes lokācijas parametram laikrindu (atkarīgu datu gadījumā) gadījumā;
- veikt metožu realizāciju datorprogrammu paketē R un veikt to savstarpējo salīdzinājumu gan izmantojot simulācijas, gan reālus datu piemērus.

Diplomdarba pirmajā nodaļā apskatīta stacionāru stohastisku procesu spektrālā analīze. Otrā nodaļa veltīta maiņas punkta analīzes vispārīgajai formai un kumulatīvo summu metodei neatkarīgu datu gadījumā, un TFT butstrapa metodei atkarīgiem datiem. Savukārt trešajā nodaļā apskatīta ieviestā blokveida empīriskās ticamības funkcijas metode atkarīgu datu gadījumā. Ceturtajā nodaļā veikts metožu pielietojums reālu datu gadījumā.

1. STACIONĀRU PROCESU SPEKTRĀLĀS ANALĪZES PAMATREZULTĀTI

1.1. Stacionāra stohastiska procesa spektrs un spektrālā blīvuma funkcija

Šajā diplomdarba nodaļā apskatīsim galvenās spektrālās analīzes definīcijas un pamatrezultātus, kas tālāk nepieciešami maiņas punkta analīzes metodēs. Nodaļā aplūkotas arī dažas spektrālās analīzes priekšrocības, kā arī trūkumi.

Lai ievestu spektra un spektrālās blīvuma funkcijas jēdzienu, apskatīsim dažus laikrindu analīzes jēdzienus un teorēmas.

Definīcija 1. [2, 20.lpp.] Par gadījuma procesa $\{x_t, t \in T\}$ kovariāciju funkciju sauc

$$\gamma_x(s,t) = E[x_s - \mu_s][x_t - \mu_t],$$

kur μ_s ir procesa vidējās vērtības funkcija $\mu_s = \int_{-\infty}^{\infty} x f_s(x)dx$ un $f_s(x) = \partial F_s(x)/\partial x$ ir procesa sadalījuma funkcijas atvasinājums jeb blīvuma funkcija. Ja nav šaubu, kuru procesu apskata, γ_x vietā lietosim γ.

Definīcija 2. [2, 24.lpp.] Gadījuma procesu $\{x_t, t \in T\}$ sauc par stacionāru vājā nozīmē, ja

- $Ex_t = c$, kur c ir konstante, un nav atkarīga no parametra t,
- kovariāciju funkcija $\gamma(s,t)$ ir atkarīga tikai no laika atstarpes $|s-t|$.

Turpmāk ar terminu *stacionārs* sapratīsim, ka process ir stacionārs vājā nozīmē. Tad aplūkojot satcionāra gadījuma procesa $\{x_t, t \in T\}$ kovariāciju funkciju un ņemot $s = t+h$, iegūsim, ka $\gamma(t+h,t) = \gamma(h,0)$, kas noved pie autokovariāciju funkcijas jēdziena.

Definīcija 3. [2, 25.lpp.] Par stacionāra procesa $\{x_t, t \in T\}$ autokovariāciju funkciju sauc

$$\gamma(h) = E[x_{t+h} - \mu][x_t - \mu],$$

kur h sauc par *lagu* jeb laika atstarpi.

Tālāk aplūkosim vienu no spektrālās analīzes pamatteorēmām, kas definē stacionāra procesa spektrālo blīvuma funkciju un tās saistību ar daudz plašāk pazīstamo autokovariāciju funkciju.

Teorēma 1. *[2, 182.-183.lpp] Ja $\gamma(h)$ ir stacionāra procesa $\{x_t\}$ autokovariāciju funkcija, turklāt*

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty,$$

tad

$$\gamma(h) = \int_{-1/2}^{1/2} e^{2\pi i\omega h} f(\omega) d\omega,\ h = 0, \pm 1, \pm 2, \ldots$$

un procesa $\{x_t\}$ spektrālā blīvuma funkcija ir

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i\omega h}, -1/2 \leqslant \omega \leqslant 1/2.$$

Pierādījums. Teorēmas pierādījums atrodams, piemēram, grāmatā [2, 537.-538.lpp]. □

Tātad spektrālā blīvuma funkcija zināmā mērā raksturo procesa dispersiju

Piezīme 2. *Samērā bieži spektrālo blīvuma funkciju definē nedaudz savādāk, izmantojot frekvences $\lambda = 2\pi\omega$, kur cikli tiek mērīti π radiānos. Šādi definē, piemēram, [4, 192-193.lpp], arī populārajā Hamiltona laikrindu analīzes grāmatā [5] ir šāda pieeja.*

Spektrālās blīvuma funkcijas īpašības. Spektrālā blīvuma funkcija ir analoģiska sadalījuma (varbūtību) blīvuma funkcijai:

- $f(\omega) \geqslant 0, \forall\omega$;
- $f(\omega) = f(-\omega)$;
- $f(\omega + 1) = f(\omega)$ (periods 1).

Precizēsim kā spektrālā blīvuma funkcija zināmā mērā raksturo procesa dispersiju, proti, izvēloties $h = 0$ autokovariāciju funkcijā, zināms, ka iegūstam procesa dispersiju, tad

$$\gamma(0) = D(x_t) = \int_{-1/2}^{1/2} f(\omega) d\omega.$$

Tātad integrējot spektrālo blīvumu funkciju pa visām frekvencēm, iegūsim procesa dispersiju un, ja izpildīti visi teorēmas nosacījumi, tad autokovariāciju un spektrālā blīvuma funkcijas satur vienu un to pašu informāciju. Tikai autokovariāciju funkcija informāciju sniedz, izmantojot *laga* jēdzienu, bet spektrālā blīvuma funkcija - frekvences (cikli novērotā procesa laika atstatumā (cikli/gadā; cikli/mēnesī utt.))

1.2. $ARMA(p,q)$ spektri

Šajā apakšnodaļā tiks pētīta plaši pazīstamo autoregresīvā slīdošā vidējā $ARMA(p,q)$ procesu spektri. Atgādināsim $ARMA(p,q)$ procesa definīciju, lai precizētu apzīmējumus un vēlāk dotu $ARMA(p,q)$ teorētiskā spektra formulu.

Definīcija 4. Process (laikrinda) $\{x_t; t = 0, \pm 1, \pm 2, \ldots\}$ ir $ARMA(p,q)$, ja tas ir stacionārs un

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q},$$

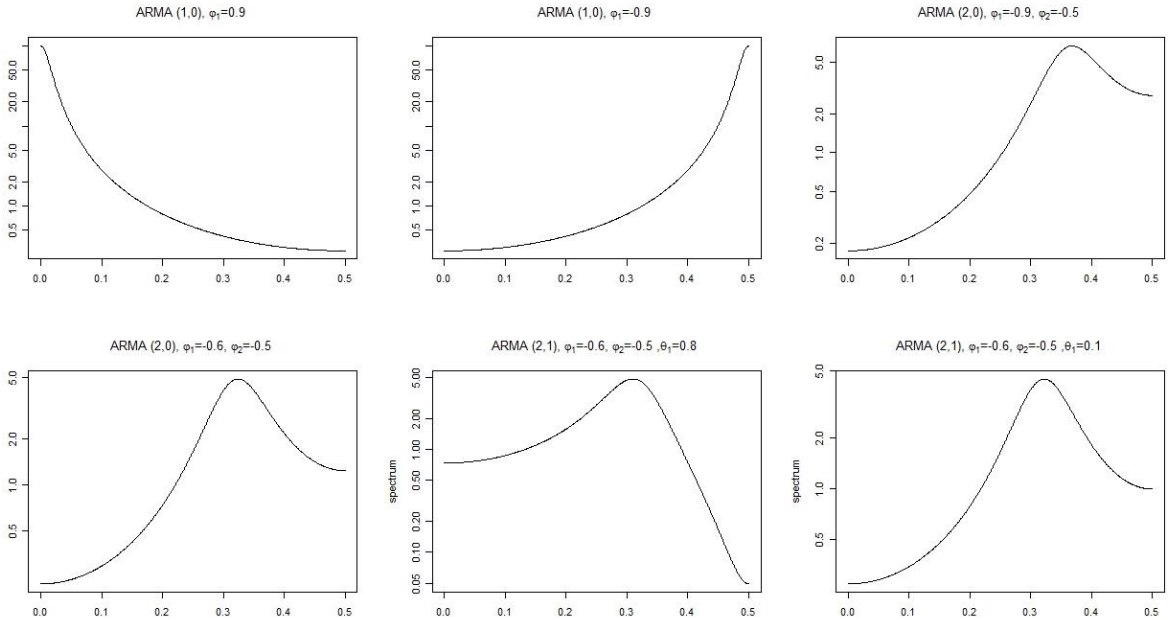
kur $\theta_q \neq 0$, $\phi_p \neq 0$ un $\sigma_\varepsilon^2 > 0$, un ε_t ir baltā trokšņa process ar $E\varepsilon_t = 0, E\varepsilon_t^2 = \sigma_\varepsilon^2$.

Apgalvojums 3. *[2, 229.lpp]ARMA(p,q) teorētiskais spektrs ir pierakstāms formā*

$$f(\omega) = \sigma_\varepsilon^2 \frac{|1 + \sum_{k=1}^{q} \theta_k e^{-2\pi i \omega k}|}{|1 - \sum_{m=1}^{p} \phi_m e^{-2\pi i \omega m}|},$$

kur $-1/2 \leq w \leq 1/2$.

Tātad jebkuram $ARMA(p,q)$ procesam var precīzi aprakstīt un uzzīmēt tā teorētisko spektru. Apskatīsim dažus piemērus.



1. att. Dažādi $ARMA(p,q)$ spektri.

Lai gan jebkuram procesam zināms tā teorētiskais spektrs, praktiski tos izmantot ir visai grūti, jo līdzīgiem spektriem var atbilst samērā dažādi procesi. Turklāt vēlāk veicot simulācijas redzēsim, ka arī katrā realizācijā procesu spektri ir nedaudz atšķirīgi un

tādēļ izmantot procesa spektru, lai noteiktu $ARMA(p,q)$ kārtas p un q un vēl jo vairāk parametru novērtējumus būs pietiekoši apgrūtinoši. Kā minēts arī literatūrā, spektrus drīzāk var izmantot procesa analīzes beigās, kad jāizšķiras starp dažiem vispiemērotākajiem modeļiem vai interesē pētāmā procesa spektrs. Grāmatā [6, 363-368.lpp] sākumā noteikta procesa kārta ar AIC kritēriju un tad analizēti gludinātie spektri.

1.3. Sezonālo $ARIMA(p,d,q)\times(P,D,Q)_s$ spektri

Atgādināsim sezonālo autoregresīvo integrēto slīdošā vidējā procesa (SARIMA) definīciju, lai precizētu apzīmējumus.

Definīcija 5. [2, 159.lpp.] Process (laikrinda) $\{x_t; t=0,\pm1,\pm2,\ldots\}$ ir $ARIMA(p,d,q)\times(P,D,Q)_s$, ja tas ir uzdodams formā

$$\Phi_P(B^s)\phi(B)\Delta_s^D\Delta^d x_t = \alpha + \Theta_Q(B^s)\theta(B)\varepsilon_t,$$

kur

$$\Phi_P(B) = 1 - \Phi_1 B - \Phi_2 B^2 - \ldots - \Phi_P B^P,$$

$$\Theta_Q(B) = 1 + \Theta_1 B + \Theta_2 B_2 + \ldots + \Theta_Q B_Q,$$

un

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_P B^{Ps},$$

$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \ldots + \Theta_Q B^{Qs},$$

un diferenču operatori

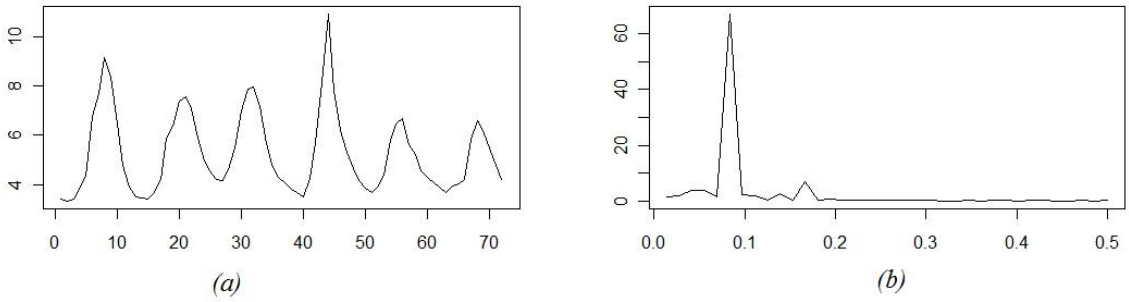
$$\Delta^d = (1-B)^d \ , \Delta_s^D = (1-B^s)^D,$$

kur $\Delta x_t = x_t - x_{t-1}$ un $Bx_t = x_{t-1}$, un ε_t ir baltā trokšņa process ar $E\varepsilon_t = 0, E\varepsilon_t^2 = \sigma_\varepsilon^2$.

Sezonālo ARIMA procesu spektru piemēri doti grāmatā [6, 339-340.lpp]. Sezonalitāte grafiski izpaužas kā *pīķa* parauga atkārtojumi ar mazāku svārstību amplitūdu ik pēc $1/k$ frekvencēm, kur k apzīmē sezonalitātes kārtu. Šajā grāmatā labi ilustrēti gadījumi $ARIMA(1,0,0)\times(1,0,0)_{12}$ un $ARIMA(0,0,1)\times(0,0,1)_{12}$.

Savukārt šeit apskatīsim reālu piemēru par citronu cenām (dati pieejami `http://faculty.arts.ubc.ca/ediewert/concepts3.pdf` [atsauce 23.01.2011.]). (Var atzīmēt, ka šim piemēram samērā labi atbilst modelis $ARIMA(2,0,0)\times(2,0,0)_{12}$.) Kā redzams, tad periodogrammā pie vērtības $\omega \approx 0.82$ ir liels *pīķis*, kas tieši norāda uz datu periodiskumu,

proti, 12 mēnešiem (jo $1/12 \approx 0.83$), savukārt nākošais leciens ir šī lielā *pīķa atbalsošanās.*



2. att. Citronu cenu laikrindas attēls (a) un periodogrammai (b).

1.4. Periodogramma un tās gludināšana

Par spektra vienkāršāko novērtējumu kalpo periodogramma. Šajā sadaļā apskatīsim periodogrammu un tās gludināšanu.

Definīcija 6. [2, 187.lpp] Ja doti dati $x_1, x_2, \ldots, x_n$, tad par diskrēto Furjē transformāciju (DFT) sauc

$$\begin{aligned} d(\omega_j) &= n^{-1/2} \sum_{t=1}^{n} x_t e^{-2\pi i \omega_j t} && (1.1) \\ &= a(j) + ib(j) && (1.2) \\ &= n^{-1/2} \sum_{t=1}^{n} V(t)(cos(2\pi\omega_j t) + isin(2\pi\omega_j t))), && (1.3) \end{aligned}$$

kur $j = 0, 1, \ldots, n-1$ un frekvences $\omega_j = j/n$ tiek sauktas par Furjē jeb fundamentālajām frekvencēm.

Definīcija 7. [2, 188.lpp] Ja doti dati $x_1, x_2, \ldots, x_n$, tad par periodogrammu sauc

$$I(\omega_j) = |d(\omega_j)|^2, \tag{1.4}$$

kur $j = 0, 1, \ldots, n-1$.

Šeit jāpiezīmē, ka $I(0) = n\overline{x}^2$, kur $\overline{x}$ ir izlases vidējā vērtība. Bet, ja $j \neq 0$, tad $I(\omega_j) = \sum_{h=-(n-1)}^{n-1} \hat{\gamma}(h) e^{2\pi i \omega_j h}$, kur $\hat{\gamma}(h)$ ir izlases autokovariāciju funkcija

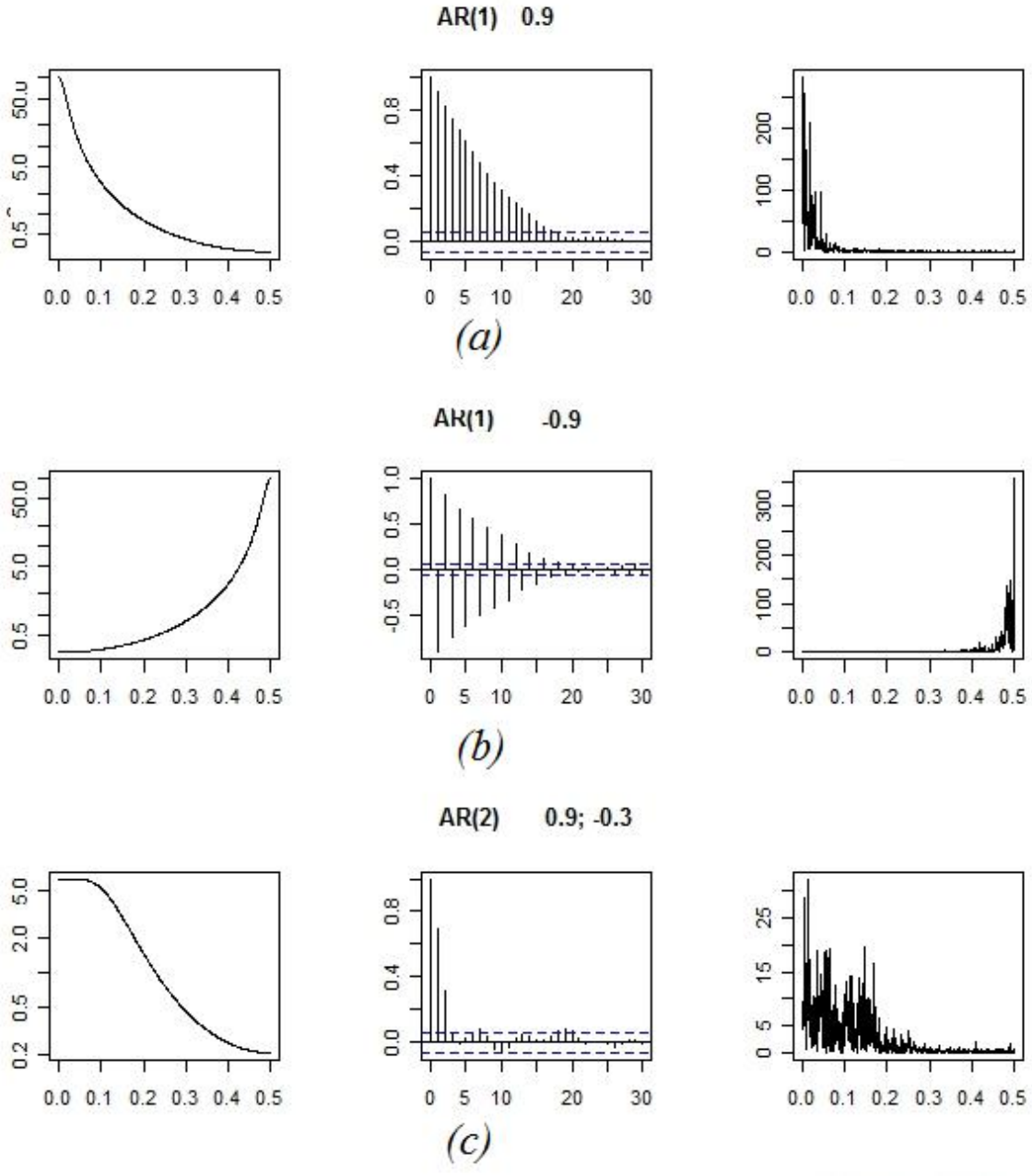
$$\hat{\gamma}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \overline{x})(x_t - \overline{x}).$$

Piezīme 4. *Līdzīgi kā definējot spektrālo blīvuma funkciju, definīcijā var izmantot frekvences π radiānos $\lambda = 2\pi\omega$, tā arī, definējot periodogrammu, pastāv šīs pašas atšķirības. Tātad nepieciešams izvēlēties vienu no pieejām un saskaņot šīs definīcijas. (Piemēram, grāmatā [4, 185.lpp] dota gadījuma procesa periodogramma izmantojot tieši frekvences π radiānos.)*

Var teikt, ka periodogramma sadala procesa dispersiju pa frekvencēm, bet periodogrammai kā spektrālās blīvuma funkcijas novērtējumam ir 2 būtiski trūkumi. Pirmais ir tas, ka aplūkoto frekvenču skaits ir mainīgs un ierobežots, jo tas ir atkarīgs no izlases novērojumu skaita. Un otrais trūkums ir tas, ka periodogramma nekļūst gludāka (netiecas uz teorētisko spektru), ja palielina novērojumu skaitu, t.i., periodogramma nav būtisks spektrālās blīvuma funkcijas novērtējums [4, 175.lpp], [6, 342.lpp]. Tādēļ svarīgi veikt

periodogrammas gludināšanu ar dažādām metodēm, kas arī sagādā vislielākās grūtības spektrālajā analīzē. Daži gludināšanas veidi tiks apskatīti vēlāk.

Vienkāršākajos gadījumos var aplūkot procesa autokorelāciju funkciju un spektru, tādejādi mēģinot apskatīt, vai pastāv kāda saistība, jo, kā jau iepriekš minēts, pēc būtības autokovariāciju funkcija un spektrālā blīvuma funkcija satur vienu un to pašu informāciju. Līdz ar to arī autokorelāciju funkcijai varētu aplūkot saistību ar spektru.



3. att.: Dažādi $AR(p)$ procesu teorētiskais spektrs, autokovariāciju funkcija un periodogramma.

Bet, kā redzams (skat. att.3), praktiski nekādas saistības nav, jo autorelāciju funkcija izmanto *lagus*, bet periodogramma (kā spektrālās blīvuma funkcijas novērtējums) infor-

māciju izsaka izmantojot frekvences.

Kā jau iepriekš noskaidrojām, visbūtiskāk ir noskaidrot ar kādām metodēm veikt periodogrammas gludināšanu. Pastāv divas būtiski atšķirīgas metodes. Visbiežāk tiek apskatīta tā sauktā " kodolu gludināšana" , kur, līdzīgi kā neparametriskajā kodolu gludināšanā blīvuma funkcijai, tiek izvēlēts kāds noteikts joslas platums un veikta periodogramas "vidējošana"attiecīgajā frekvencē dažādi izsvarojot vai neizsvarojot apkārtējo frekvenču ietekmi. Tas tiek darīts, izvēloties apkārtējo frekvenču skaitu m, kas arī sagādā vislielākās grūtības šajā metodē, jo nav viennozīmīgas metodes kā noteikt parametru m. Otra fundamentālā pieeja ir periodogrammas gludināšana izmantojot neparametrisko regresiju. Tad, savukārt, parasti svarīgi pareizi novērtēt joslas platuma parametru h.

1.5. Kodolu gludinātā periodogramma

Aprakstīsim visbiežāk sastopamo periodogrammas gludināšanas veidu - kodolu gludināšanu. Šeit tiek izvēlēta frekvenču josla $\mathcal{B}$ no $L << n$ sekojošām fundamentālām frekvencēm, kas centrētas ap $\omega_j = j/n$, kas tuva interesējošai frekvencei ω:

$$\mathcal{B} = \{\omega : \omega_j - m/n \leqslant \omega \leqslant \omega_j + m/n\},$$

kur $L = 2m + 1$. Lielumu $\mathcal{B}_\omega = L/n$ sauc par joslas platumu. [2, 197. lpp]

Definīcija 8. [2, 197. lpp] Ja $\mathcal{B}$ ir frekvenču josla no $L << n$ sekojošām fundamentālām frekvencēm, kas centrētas ap $\omega_j = j/n$, tad par vidējoto periodogrammu sauc

$$\overline{f}(\omega) = \frac{1}{L} \sum_{k=-m}^{m} I(\omega_j + k/n).$$

Lai iegūtu precīzāku spektra novērtējumu, periodogrammu var gludināt izmantojot dažādus svarus, tādēļ definēsim svērto periodogrammu.

Definīcija 9. [2, 203. lpp] Par svērto periodogrammu sauc

$$\hat{f}(\omega) = \sum_{k=-m}^{m} h_k I(\omega_j + k/n), \tag{1.5}$$

kur svari h_k apmierina nosacījumus $\forall\, k\, :\, h_{-k} = h_k$ un $\sum_{k=-m}^{m} h_k = 1$.

Svaru h_k izvēlei tiek ieteikti Daniela, Fejer, Dirihlē un modificētais Daniela kodols. Šajā gadījumā parametrs L jāaizvieto ar $L_h = (\sum_{k=-m}^{m} h_k^2)^{-1}$, tad joslas platums $\mathcal{B}$ svērtajai periodogrammai būs $\mathcal{B} = L_h/n$.

Parametra m izvēle

Tātad svarīga ir parametra L vai m (angliski tiek arī saukts par *window closing*, jo, samazinot m, samazinās joslas platums) izvēle. Neliels ieskats par parametru m tiek dots grāmatā [6, 355. lpp]. Dženkins (*Jenkins*) un Vatts (*Watts*) (1968) esot ieteikuši mēģināt ar trim dažādām m vērtībām un tad izvēlēties atbilstošāko. Mazāka vērtība dos priekšstatu par to, kur atrodas lielākie pīķi, bet periodogrammā būs daudz pīķu, un tā būs samērā negluda. Liela parametra vērtība var veidot līkni, kas būs *pārgludināta*. Savukārt Čatfīlds (*Chatfield*) (2004) iesakot ņemt kompromisu, izvēlotie vuenu no $m = \sqrt{n}$, $m = 2\sqrt{n}$ un $m = 0.5\sqrt{n}$, kas dos priekšstatu par spektru. Un piedevām tiek minēts spilgts citāts, kas raksturo šo problemātiku: " Experience is the real teacher and cannot be got from a book."(no angļu valodas " *Pieredze ir patiesais skolotājs, un tā nav iegūstama grāmatās* ")

Tātad varam apkopot ieteikumus:

- izvēlēties $m = \sqrt{n}$;
- vēl var aplūkot $m = 2\sqrt{n}$, $m = 0.5\sqrt{n}$.

Gludināšanas kodoli

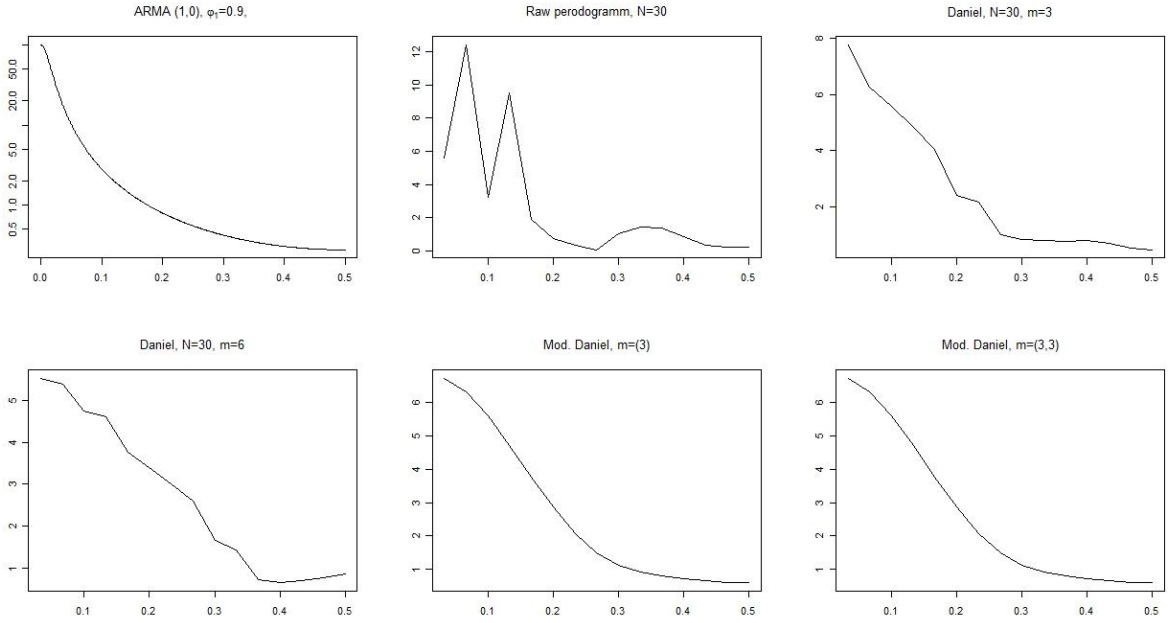
Kā jau ierasts statistikā kodolu gludināšanā paša kodola izvēle nav tik būtiska kā joslas platuma (šajā gadījumā parametra m) izvēle. Literatūrā visbiežāk tiek minēti Daniela un modificētais daniela kodols. Daniela kodols ir visvienkāršākais, tas visām frekvencēm no joslas piekārto vienādus svarus, proti, tiek piekārtoti svari $1/L$, kur $L = 2m+1$. Savukārt modificētais Daniela kodols pirmajam un pēdējajam elementam piešķir pusi no pārējiem svariem, t.i., pirmais un pēdējais svara koeficienti ir uz pusi mazāki nekā pārējie, kas ir vienādi savā starpā. Protams arī modificētā Daniela kodola svaru summa ir 1.

Vēl bieži kodolu modifikācija notiek vienkārši atkārtojot kodolu gludināšanu periodogrammai. Matemātiski tas nozīmē, ka tiek ņemta kodolu konvolūcija. Grāmatā [6, 354. lpp] attēlots kā izmainās svara koeficienti atkārtojot izsvarošanu ar modificēto Daniela kodolu pie $m = 3$. Atkārtojot kodolu otru reizi svari veido trijstūra formas grafiku, bet jau pie trešās atkārtošanas, tā forma jau atgādina normālo sadalījumu.

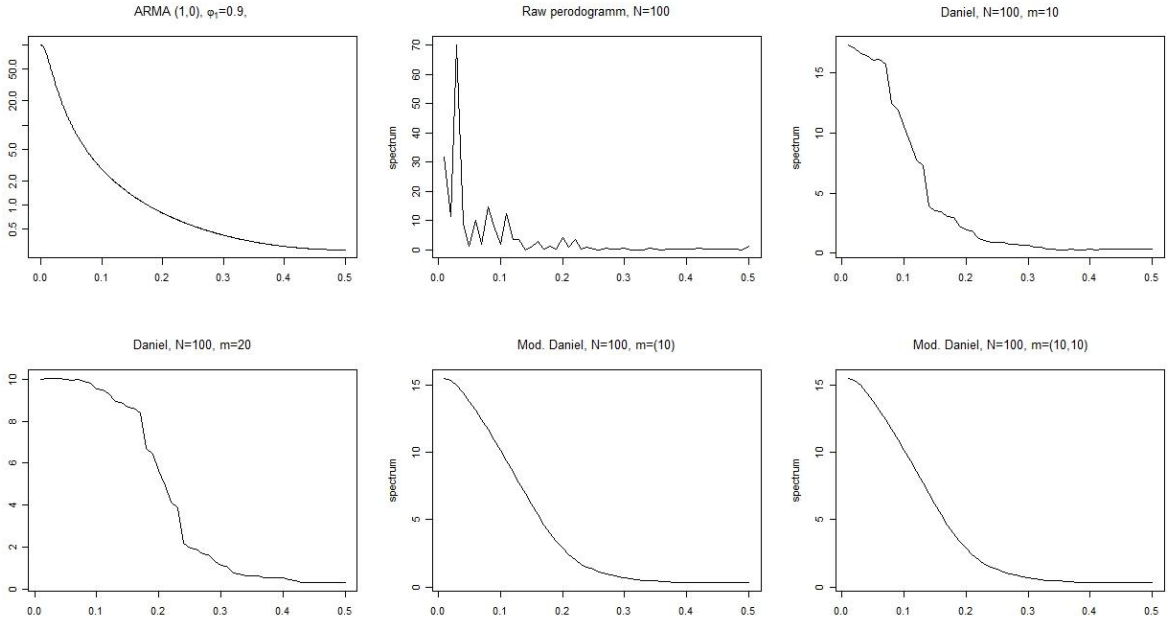
Vēl literatūrā atrodami arī Fejer un Dirihlē kodoli (skat., piemēram, [2, 207.-215. lpp]

Kodolu gludināto periodogrammas (gludinātie spektra novērtējumi)

Tagad nedaudz ilustrēsim kodolu gludināšanas metodi ar dažiem piemēriem. Sākumā apskatīsim $ARMA(p, q)$ procesu simulācijas, līdzīgi kā apskatījām to teorētiskos spektrus jau iepriekš.

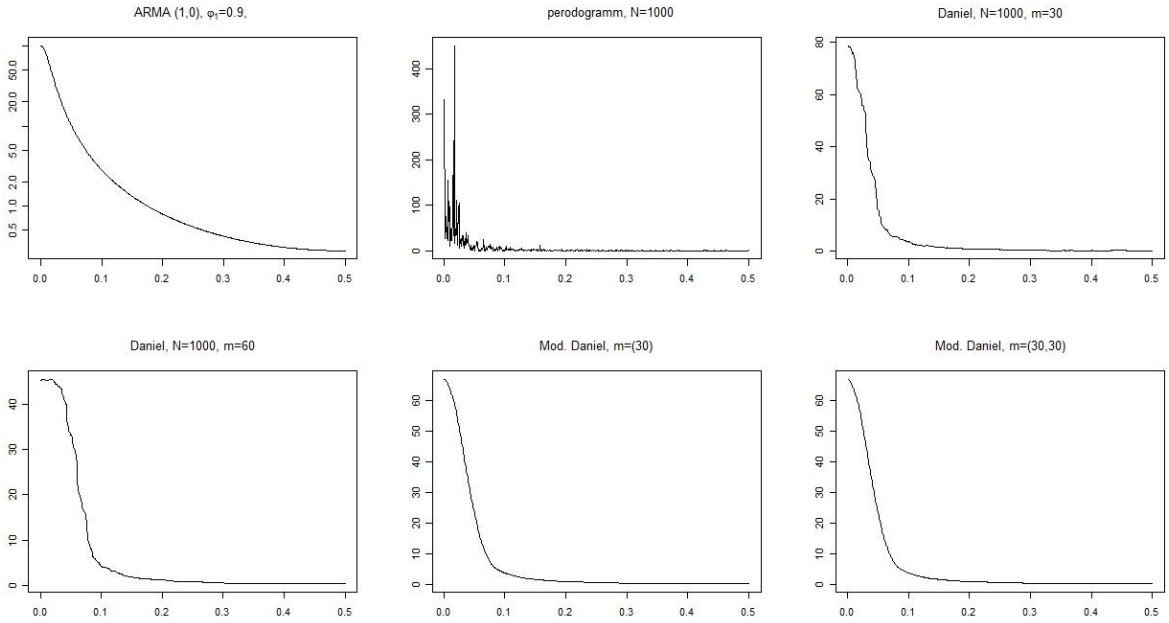


4. att.: $AR(1)$ teorētiskais spektrs un simulāciju gludināšana pie izlases apjoma $N = 30$.

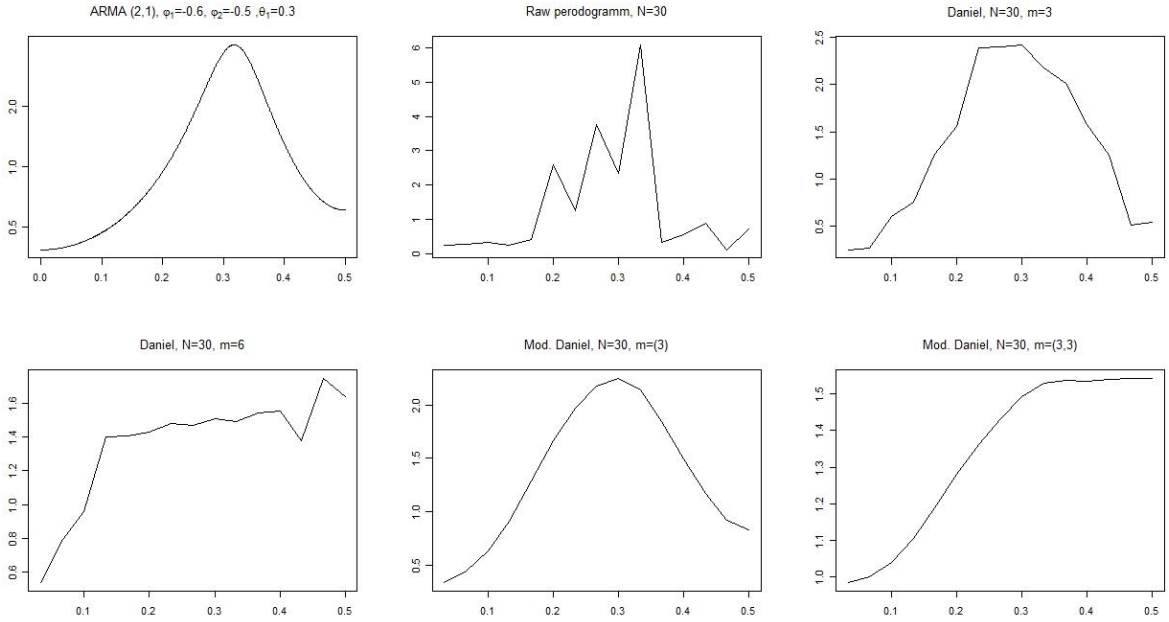


5. att.: $AR(1)$ teorētiskais spektrs un simulāciju kodolu gludināšana pie izlases apjoma $N = 100$.

Kā redzams, negludinātā periodogramma, protams, ir samērā slikts spektra novērtējums,

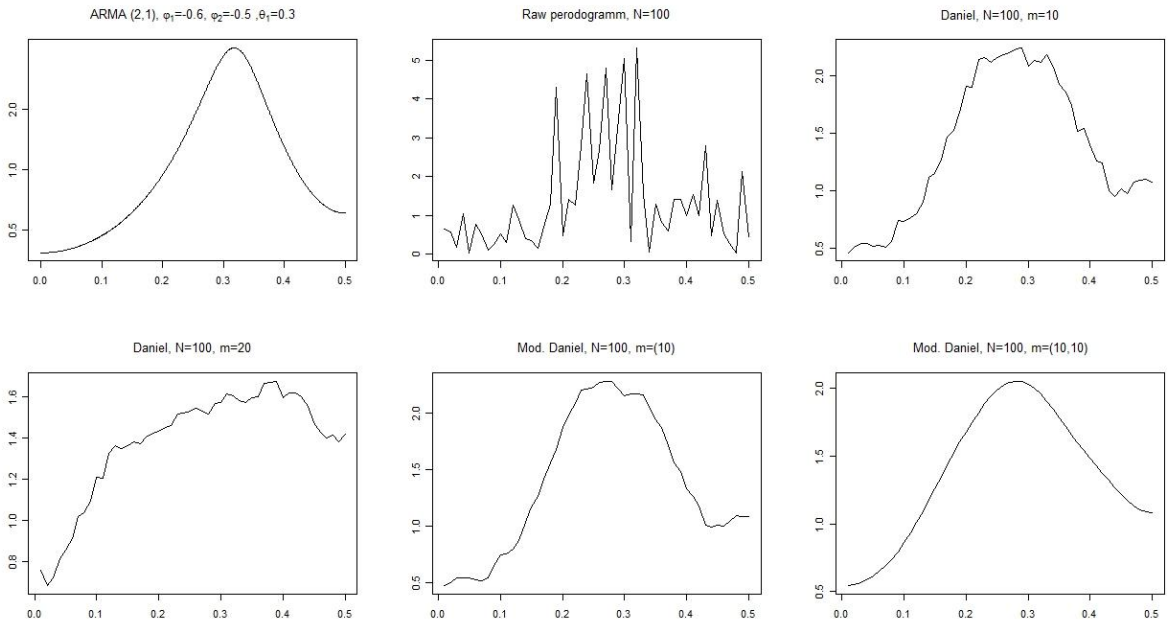


6. att.: $AR(1)$ teorētiskais spektrs un simulāciju gludināšana pie izlases apjoma $N = 1000$.

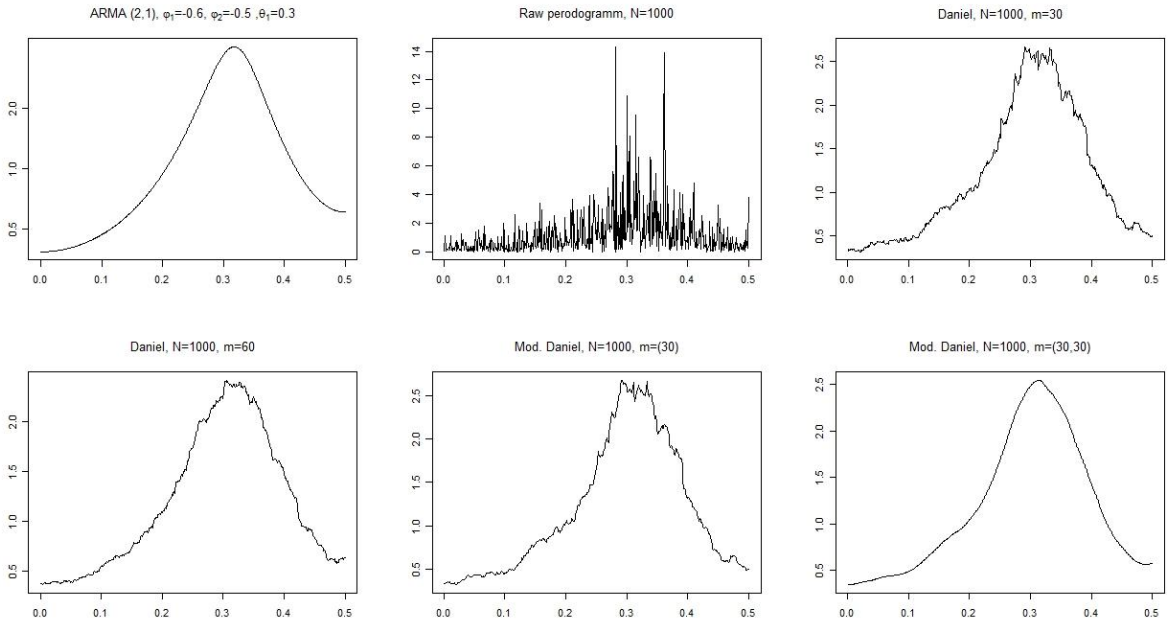


7. att.: $ARMA(2,1)$ teorētiskais spektrs un simulāciju gludināšana pie izlases apjoma $N = 30$.

bet jau vienkāršākā gludināšana ar Daniela kodolu un ieteikto $m \approx \sqrt{N}$ dod priekšstatu par spektra formu. Izmantojot modificēto Daniela kodolu vienkāršākajā $AR(1)$ procesa gadījumā jau iegūst samērā labu spektra novērtējumu. Savukārt jau nedaudz komplicētāks piemērs ar $ARMA(2,1)$ ir jau interesantāks. Pie izlases apjoma $N = 30$ Daniela kodols ar $m \approx \sqrt{N}$ dod kādu priekšstatu par spektra formu, bet otrais ieteiktais variants ar $m \approx 2\sqrt{N}$ jau ir *pārgludināts*. Tāpat līdzīgi modificētā Daniela kodola gadījumā pie



8. att.: $ARMA(2,1)$ teorētiskais spektrs un simulāciju gludināšana pie izlases apjoma $N = 100$.

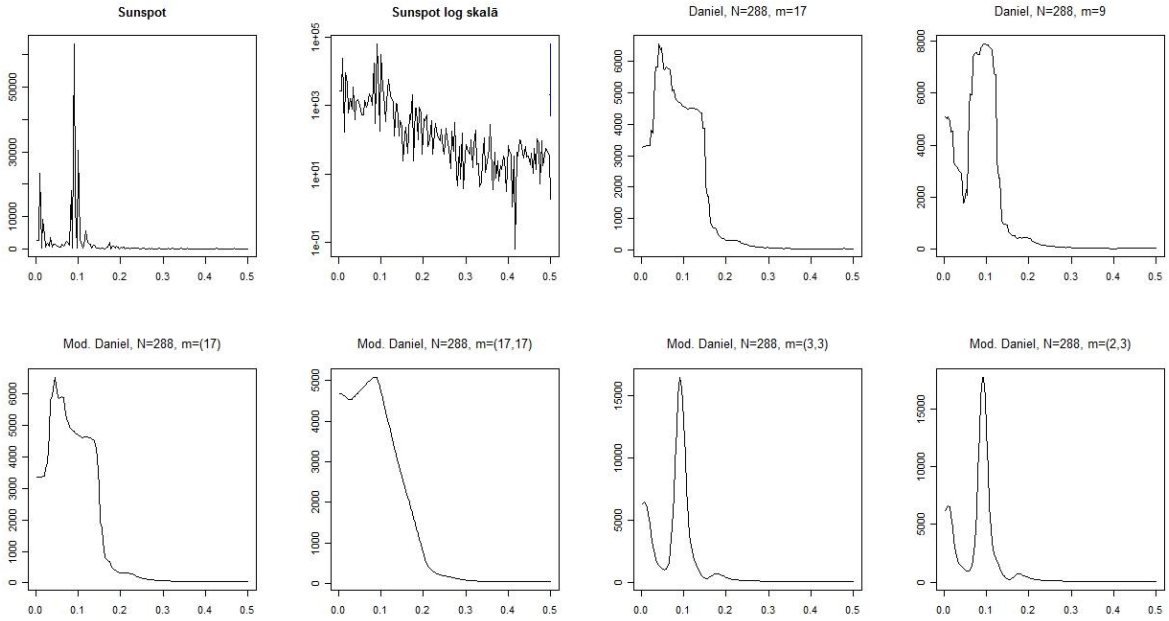


9. att.: $ARMA(2,1)$ teorētiskais spektrs un simulāciju gludināšana pie izlases apjoma $N = 1000$.

$m \approx \sqrt{m}$ iegūst jau samērā ļoti labu rezultātu, bet pie $m \approx 2\sqrt{N}$ atkal periodogramma ir *pārgludināta*. Savukārt $ARMA(2,1)$ simulācijai pie izlases apjoma $N = 100$ tieši atkārtoti pielietots modificētais Daniela kodols dod vislabāko rezultātu, līdzīgi arī tas ir, ja izlases apjomu palielina līdz $N = 1000$. Interesanti, ka simulētajam procesam pie $N = 1000$ negludinātā periodogramma vispār nedod priekšstatu par procesa spektru. Patiesībā tas

jau bija sagaidāms, jo, kā jau iepriekš atzīmējām, periodogramma nav būtisks spektrālās blīvuma funkcijas novērtējums.

Tagad apskatīsim nedaudz interesantāku piemēru par saules aktivitātes datiem, kas bieži tiek izmantoti laikrindu analīzē (datus var iegūt, piemēram, `http://www.napscience.com/astro/sunspots/sunspotdata.htm`, ņemot datus par 1700.-1987. gadu).



10. att.: Sunspot periodogramma lineārajā un logaritmiskajā skalā, periodogrammas lineārajā skalā gludināšana.

Kā redzams, saules aktivitātes datiem piemērotāka ir lineārā skala (skat. *Logaritmiskā vai lineārā skala periodogrammas attēlošanā*). Izmantojot ieteikumu, ka $m \approx \sqrt{N}$ vai $m \approx 0.5\sqrt{N}$, un tad veicot periodogrammas gludināšanu, var ievērot, ka parametrs m ir izvēlēts pārāk liels. Šis piemērs labi ilustrē jau iepriekš minēto, ka parametra m noteikšana ir samērā nenoteikta un ka tā jābalsta arī uz iepriekšējo pētnieka pieredzi un *apriorajām* zināšanām par procesu. Var redzēt, ka, pielietojot modificēto Daniela kodolu ar $m = 3$ atkārtoti (divas reizes), iegūtais tuvinājums ir samērā labi pielāgots datiem, arī gludinot ar modificēto Daniela kodolu ar $m = 2$ un atkārtoti ar $m = 3$, iegūtais rezultāts nav slikts.

1.6. Punktveida ticamības intervāli

Zinot periodogrammas īpašības, var konstruēt periodogrammas punktveida ticamības intervālus. Ja izpildās iepriekšminētie nosacījumi par svariem h_k, turklāt $m \to \infty$, kad $n \to \infty$ un $m/n \to 0$, tad var pierādīt (sīkāk skat.[2, 197.-198.lpp, 543.lpp, teorēma C.4]

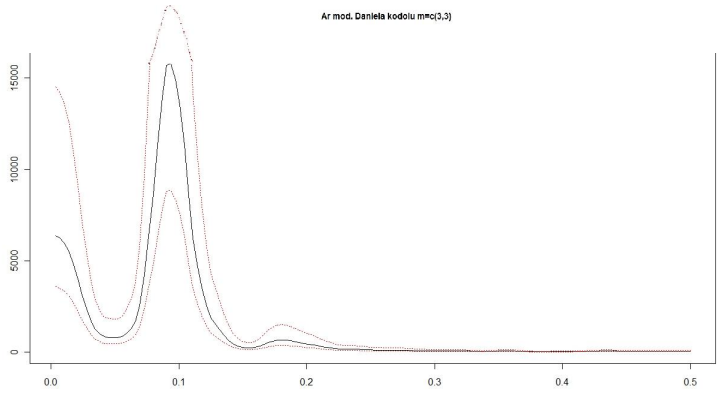
vai [6, 356. lpp.], ka

$$\frac{2L_h\hat{f}(\omega)}{f(\omega)} \to^d \chi^2_{2L}, \quad , L = 2m+1$$

Tātad ticamības intervāls

$$\frac{2L\hat{f}(\omega)}{\chi^2_{2L}(1-\alpha/2)} \leq f(\omega) \leq \frac{2L\hat{f}(\omega)}{\chi^2_{2L}(\alpha/2)}.$$

Ja vidējotās periodogrammas vietā apskata gludināšanu ar dažādiem svariem, tad L vietā lieto $L_h = (\sum_{k=-m}^{m} h_k^2)^{-}1$, kur h_k ir izvēlētie svari. Šo ilustrēsim tikai ar vienu piemēru. Atkal izmantosim jau apskatītos datus par saules aktivitāti.



11. att.: Sunspot periodogrammas lineārajā skalā divkārša gludināšana ar modificēto Daniela kodolu ar $m = 3$.

Šeit attēlota jau gludinātā periodogramma ar atbilstošiem punktveida ticamības intervāliem. Kā var redzēt galvenās neskaidrības ir frekvencēs, kur periodogramma pieņem lielas vērtības, proti, tur šie intervāli ir samērā plaši.

1.7. Taperošana

Tā kā spektra novērtējumā periodogrammas definēšanai izmatojām diskrēto Furjē transformāciju, tad spektra novērtējumā diezgan būtiskas novirzes dod novērotā procesa sākuma un beigu vērtības. Lai mazinātu šo datu ietekmi, lieto tā saukto "taperošanu "l . Aizvieto novērotos datus x_t ar $y_t = h_t x_t$, $t = 1, 2, \ldots, n$ un tad pielieto DFT:

$$d_y(\omega_j) = n^{-1/2} \sum_{t=1}^{n} h_t x_t e^{-2\pi i \omega_j t}.$$

Tad definē šiem datiem periodogrammu $I_y(\omega_j) = |d(\omega_j)|^2$. Būtībā tiek samazināta novērotā procesa sākuma un beigu vērtību ietekme. Tas tiek darīts, jo nav zināms vai pro-

cess novērots tieši atbilstoši tā ciklu sākumam un beigām, proti, vai novērojumu sākums sakrīt arī ar procesa ciklu sākumu un novērojumi pārtraukti arī noslēdzoties ciklam. Tā kā reāli veicot novērojumus, visticamāk, par procesu tas nav zināms, var gadīties, ka dati fiksēti, kad process atrodas jau kādā cikla fāzē. Tāpēc piemēro sākumu un beigu datu, visbiežāk tieši 10%, datu ietekmes mazināšanu. Piemēram, 10% no datiem sākumā un 10% datiem beigās tiek piekārtoti svari h_t tā, ka, jo tuvāk datu sākumam/beigām ir novērojums, jo mazāks svara koeficients tam tiek piešķirts. Kā piemērs bieži tiek minēts ([2, 207.lpp] un [6, 360.lpp]) kosinusa zvans (angliski *cosine bell*) vai nošķeltais kosinusa zvans (angliski *split cosine bell*). Kosinusa zvana svari tiek uzdoti pēc likuma

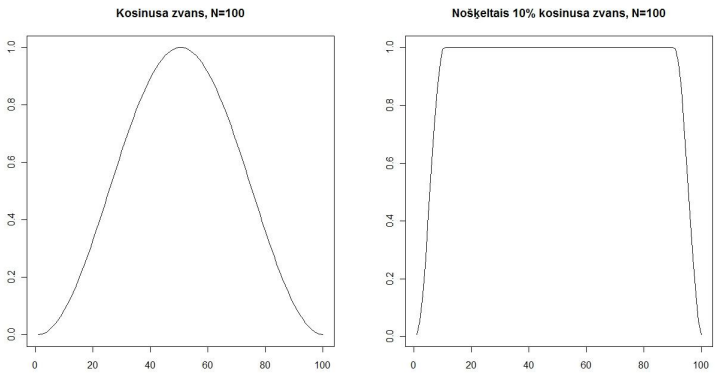
$$h_t = \frac{1}{2}\left(1 - \cos\left[\frac{2\pi(t-0.5)}{n}\right]\right),$$

bet nošķeltais kosinusa zvans pēc formulas

$$h_t = \begin{cases} \frac{1}{2}\left(1 - \cos\left[\frac{\pi(t-0.5)}{m}\right]\right), ja\, 1 \leq t \leq m \\ 1,\, ja\, m+1 \leq t \leq n-m \\ \frac{1}{2}\left(1 - \cos\left[\frac{\pi(n-t+0.5)}{m}\right]\right), ja\, n-m+1 \leq t \leq n, \end{cases}$$

kurš tiek saukts par $100p\%$ kosinusa zvana taperošanu, kur $p = 2m/n$. Praksē visbiežāk izmanto 10% vai 20% datu taperošanu.

Piezīme 5. *Datorprogrammu paketē R automātiskā periodogrammas attēlošanas komanda* `spec.pgram()` *veic* 10% *datu taperošanu pēc noklusējuma, kaut gan paziņojumā virs periodogrammas tiek norādīts, ka tā ir negludinātā periodogramma (angliski Raw periodogramm)!*



12. att.: Kosinusa zvana svaru funkcija un 10% nošķeltā kosinusa zvana svaru funkcija datiem ar apjomu $N = 100$.

1.8. Neparametriskās regresijas metodes periodogrammas gludināšana

Fana un Kreutzbergeres publikācijā [7] tiek minēts, ka eksitē vairāki veidi, ko pielieto periodogrammas gludināšanā, turklāt turpat minēts, ka līdz šim tā arī nav noteikts, kura pieeja ir labākā, jo tās ir grūti salīdzināt dažādo gludināšanas procedūru dēļ. Kā senākā pieeja, kas arī tika apskatīta iepriekš, tiek minēta kodolu gludināšana. Kā otrā pieeja tiek minēta logaritmiskās periodogrammas gludināšana ar mazāko kvadrātu metodi. Kā piemērs minēta Vābas darbs (Wahba (1980)), kur tiek izmantoti splaini. Trešā pieeja balstoties uz Vaitla (Whittle (1962)) ticamības funkcijas aproksimāciju periodogrammai. Kā arī minēts, ka lielā daļā literatūras spektrālā analīze tiek izmantota tikai pašā modeļa izvēles beigās, kad no atlikušajiem jāizvēlas atbilstošākais. Tātad sākumā tiek meklēts atbilstošs $ARMA(p,q)$ process, visbiežāk izmantojot Beijesa informācijas kritēriju (BIC) un informāciju, kas zināma par procesu, un tikai tad aplūko procesa gludināto periodogrammu un teorētisko spektru, kādu dod izvēlētie modeļi. Pēc tam izvēlas atbilstošāko. Bet šādas, parametriskās, pieejas galvenais trūkums ir tas, ka ne visi procesi pakļaujas labai $ARMA(p,q)$ aproksimācijai.

Viens no veidiem kā gludināt periodogrammu ir pielietot neparametriskās lokālās regresijas novērtējumu.Kā parasti, svarīga ir gludinošā parametra izvēle. Datorprogrammu paketē R pieejama iebūvētā funkcija `locfit`, ko apskatīsim nedaudz vēlāk.

Lokālā regresija [8, 94.-96. lpp.]

Pieņemsim, ka $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ ir neatkarīgi pa pāriem iegūti novērojumi. Sakarību starp rezultējošo jeb atbildes mainīgo Y un X var izteikt ar vienādojumiem

$$Y_i = m(X_i) + \epsilon_i,\ \mathbb{E}(\epsilon_i) = 0,\ i = 1,2,\ldots,n,$$

kur m ir regresijas funkcija. $m(x)$ var interpretēt kā Y vidējo vērtību, ja dota vērtība $X = x$, tas ir,

$$m(x) = \mathbb{E}(Y|X = x).$$

Aplūkosim nosacītās vidējās vērtības funkcijas $m(\cdot)$ izvirzījumu Teilora rindā patvaļīgam t kādā punkta x apkārtnē

$$m(t) \approx m(x) + m'(x)(t-x) + \ldots + m^{(p)}(x)(t-x)^p \frac{1}{p!}.$$

Šis var kalpot par iemeslu, lai aplūkotu lokālo lineāro regresiju (meklēt polinomiālu aproksimāciju punkta x apkārtnē). Ideja par lokālo lineāro regresiju punkta x apkārtnē tiek realizēta izmantojot tā sauktos *kodola svarus* minimizācijas problēmā

$$\min_{\beta} \sum_{i=1}^{n} \{Y_i - \beta_0 - \beta_1(X_i - x) - \ldots - \beta_p(X_i - x)^p\}^2 K_h(x - X_i),$$

kur β ir koeficientu vektors $(\beta_0, \beta_1, \ldots, \beta_p)^T$. Tātad tiek izmantota svērtā mazāko kvadrātu metode ar svariem $K_h(x - X_i)$. Izmantosim apzīmējumus

$$X = \begin{pmatrix} 1 & (X_1 - x) & (X_1 - x)^2 & \ldots & (X_1 - x)^p \\ 1 & (X_2 - x) & (X_2 - x)^2 & \ldots & (X_2 - x)^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n - x) & (X_n - x)^2 & \ldots & (X_n - x)^p \end{pmatrix}; \; Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix};$$

$$W = \begin{pmatrix} K_h(x - X_1) & 0 & 0 & \ldots & 0 \\ 0 & K_h(x - X_2) & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & K_h(x - X_n) \end{pmatrix}.$$

Tad, izmantojot tradicionālo svērto vidējo kvadrātu metodes formulu, iegūsim parametru vektora β novērtējumu

$$\widehat{\beta} = (X^T W X)^{-1} X^T W Y.$$

Jāuzsver, ka atšķirībā no novērtējuma, kas iegūts ar parametriskās regresijas palīdzību, šis novērtējums ir atkarīgs no mainīgā x. Tātad šī patiešām ir lokālā regresija punktā x. Apzīmēsim vektora $\widehat{\beta}$ komponentes ar $\widehat{\beta}_0(x), \widehat{\beta}_1(x), \ldots \widehat{\beta}_p(x)$. Tātad lokālais regresijas funkcijas m novērtējums ir

$$\widehat{m_{p,h}}(x) = \widehat{\beta}_0(x).$$

Tātad $m(x) \approx \beta_0(x)$.Vēl tikai atzīmēsim, ka šis novērtējums ir atkarīgs no parametra h. Tagad aplūkosim lokālo polinomiālo regresiju dažādiem p

- ja $p = 0$, tad iegūstam *lokālo konstantes* jeb Nadaraja-Watsona novērtējumu

$$\widehat{m_{0,h}}(x) = \widehat{m_h}(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i) Y_i}{\sum_{i=1}^{n} K_h(x - X_i)};$$

- pie $p = 1$ iegūstam lineāro lokālās regresijas novērtējumu. Sākumā apzīmēsim

$$S_{h,j}(x) = \sum_{i=1}^{n} K_h(x - X_i)(X_i - x)^j,$$

$$T_{h,j}(x) = \sum_{i=1}^{n} K_h(x - X_i)(X_i - x)^j Y_i,$$

tad iegūsim, ka

$$\widehat{\beta}(x) = \begin{pmatrix} S_{h,0}(x) & S_{h,1}(x) \\ S_{h,1}(x) & S_{h,2}(x) \end{pmatrix} \begin{pmatrix} T_{h,0}(x) \\ T_{h,1}(x) \end{pmatrix}.$$

Līdz ar to esam ieguvuši lokālo lineāro regresijas novērtējumu

$$\widehat{m_{1,h}}(x) = \widehat{\beta}_0(x) = \frac{T_{h,0}(x)S_{h,2}(x) - T_{h,1}(x)S_{h,1}(x)}{S_{h,0}(x)S_{h,2}(x) - S_{h,1}^2(x)}.$$

Piezīme 6. *Literatūrā [9] parasti iesaka lietot tieši lokālo lineāro regresiju, kurai nav robežu novirzes.*

Lokālās regresijas ticamības joslas

Lai aplūkotu lokālās regresijas ticamības joslas, ievērosim, ka lokālās regresijas novērtējums ir mazāko kvadrātu metodes atrisinājums un ir lineārs novērtējums. Tad definēsim svaru diagrammas vektoru.

Apgalvojums 7. *[9, 27. lpp.] Katram x eksistē svaru diagrammas vektors $l(x) = (l_1(x), \ldots, l_n(x))$, ka*

$$\widehat{m_{p,h}}(x) = \sum_{i=1}^{n} l_i(x) Y_i.$$

Apgalvojums 8. *[9, 34. lpp.] Svaru diagrammas vektora koeficientus var aprēķināt pēc formulas*

$$l(x)^T = e_1^T (X^T W X)^{-1} X^T W,$$

kur ar e_1 apzīmēts vienības vektors $e_1 = (1, 0, \ldots, 0)^T$.

Tagad samērā viegli aprēķināt matemātisko cerību un dispersiju.

Apgalvojums 9. *[9, 28. lpp.] Lokālās regresijas novērtējuma $\widehat{m_{p,h}}(x)$ matemātiskā cerība*

un dispersija ir izsakāma ar svaru diagrammas vektoru šādā formā

$$E(\widehat{m_{p,h}}(x)) = \sum_{i=1}^{n} l_i(x)\widehat{m_{p,h}}(x_i)$$

$$D(\widehat{m_{p,h}}(x)) = \sigma^2 \sum_{i=1}^{n} l_i(x)^2 =: \sigma^2 ||l(x)||^2.$$

Šeit pieņemts, ka Y_i ir neatkarīgi un ar konstantu dispersiju $\sigma^2 < \infty$.

Apgalvojums 10. *Lokālās regresijas novērtējuma dispersija aprēķināma pēc formulas*

$$D(\widehat{m_{p,h}}(x)) = \sigma^2 ||l(x)||^2 = \sigma^2 e_1^T (X^T W X)^{-1} (X^T W^2 X)(X^T W X)^{-1} e_1.$$

Tagad varam konstruēt ticamības joslas. Apzīmēsim ar $\overline{m}(x) = E(\widehat{m_{p,h}}(x))$.

Apgalvojums 11. *[10, 91.-92. lpp.] $\overline{m}(x)$ ticamības joslas ir formā*

$$I(x) = [\widehat{m_{p,h}}(x) - c\widehat{\sigma}||l(x)||, \widehat{m_{p,h}}(x) + c\widehat{\sigma}||l(x)||], x \in [a, b],$$

kur c ir konstante, kas novērtēta no vienādojuma

$$2(1 - \Phi(c)) + \frac{k_0}{\pi} e^{-\frac{c^2}{2}} = \alpha,$$

kur savukārt α ir uzdotais drošības līmenis un

$$k_0 = \int_a^b ||T'(x)|| dx,$$

kur $T'(x) = (T_1'(x), \ldots, T_n'(x))$, $T_i'(x) = \partial T_i(x)/\partial x$ un $T_i(x) = l_i(x)/||l(x)||$.

Teorēma 12. *[10, 85.-86. lpp.] Ja $\widehat{m_{p,h}}(x)$ ir lineāra regresija, tad*

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} (Y_i - \widehat{m_{p,h}}(x_i))^2}{n - 2\nu + \tilde{\nu}},$$

kur

$$\nu = tr(L), \quad \tilde{\nu} = tr(L^T L) = \sum_{i=1}^{n} ||l(x_i)||^2,$$

kur matricas L elementi ir $L_{i,j} = l_j(x_i)$. Turklāt, ja $m(x)$ ir pietiekami gluda, t.i., $\nu = o(n)$ un $\tilde{\nu} = o(n)$, tad $\widehat{\sigma}^2$ ir būtisks novērtējums.

Piezīme 13. *Ja atlikumu dispersija ir heteroskadiska, tad ticamības joslu konstruēšanā nepieciešams izmantot dispersijas novērtējumu $s(x)$, ko šeit neaplūkosim, bet var apskatīt, piemēram, [10, 92.lpp.].*

Piezīme 14. *Datorprogammu paketē R lokālo regresiju var konstruēt ar paketes* `locfit(·)` *palīdzību. Piemēram,* `locfit(y~x)` *nozīmē, ka tiek veikta y lokālā regresija pēc x. Pēc noklusēšanas tiek izmantota kvadrātiskā regresija, bet, lai veiktu lokālo lineāro regresiju, jānorāda* `locfit(y~x, deg=1)` .

Piezīme 15. *Datorprogammu paketē R lokālās regresijas ticamības joslas var konstruēt arī ar paketes* `locfit(·)` *palīdzību, t.i., jānorāda* `fit<-(locfit(y~x))` *un tad, attēlojot grafiku ar komandas* `plot` *palīdzību, norāda* `plot(fit, band= "global ")`*, kur* `global` *nozīmē, ka atlikumi ir ar konstantu dispersiju. Ja tiek norādīts* `band= "local "`*, tad tiek izdarīts pieņēmums, ka atlikumu dispersija ir mainīga, t.i., $\sigma^2 = \sigma^2(x)$.*

Piezīme 16. *[9, 20.-21. lpp.] Komanda* `locfit(·)` *joslas platuma $h(x)$ izvēlei izmanto* `tuvākā kaimiņa` *metodi ar parametru α. Tad kā gludinošais parametrs tiek norādīts tieši $\alpha \in [0, 1]$. Tas notiek pēc šādas metodes:*

1. *Tiek aprēķināti attālumi $d(x, x_i) = |x - x_i|$ starp interesējošo punktu x un datu punktiem x_i.*

2. *Joslas platums $h(x)$ tiek izvēlēts kā k-tais mazākais no attālumiem, kur $k = \lfloor n\alpha \rfloor$.*

2. MAIŅAS PUNKTA ANALĪZE

2.1. Maiņas punkta analīze

Šī diplomdarba mērķis ir apskatīt maiņas punkta noteikšanu laikrindām, bet vispirms nepieciešams apskatīt maiņas punkta analīzes vispārīgo uzstādījumu.

Visās maiņas punkta analīžu metodēs sākumā nepieciešams pārliecināties, vai modelis ir laikā nemainīgas, vai arī ir notikušas izmaiņas. Kad noskaidrots, ka modelī ir notikušas izmaiņas, varam apskatīt galvenos jautājumus, kurus cenšas risināt ar maiņas punkta analīzes palīdzību, un tie ir [11, 1.lpp.]:

- Kad modelī notikušas izmaiņas?
- Vai izmaiņas ir vienas vienīgas?
- Cik maiņas punkti bijuši? utt.

Sākumā varam aplūkot vienkāršāko maiņas punkta definīciju, kuru izmanto kvalitātes kontroles metodēs.

Definīcija 10. [11, 1.lpp.] Laika momentu, kad statistiskajā modelī notikušas izmaiņas sauc par maiņas punktu.

Tālāk apskatīsim vispārīgo problēmas nostādni [11, 11.lpp.] vidējās vērtības (lokācijas parametra) hipotēžu pārbaudes gadījumā, kad sākuma vidējā vērtība un dispersija nav zināmas, tad nulles hipotēze un alternatīva pierakstāma kā

$$H_0 : \exists m \in \{1, 2, \ldots, M\}, \tag{2.1}$$

ka

$$X_i = a + e_i,\ i = 1, 2, \ldots, m, \tag{2.2}$$

$$X_i = a + \delta + e_i,\ i = m + 1, m + 2, \ldots, M, \tag{2.3}$$

$$H_1 :\ X_i = a + e_i,\ i = 1, 2, \ldots, M,$$

turklāt e_i apzīmē atlikuma locekļus, kas ir *iid* (vienādi un neatkarīgi sadalīti) un M ir datu apjoms.. Papildus, lai konstruētu ticamības intervālus, izdarīsim papildus pieņēmumu, ka $\{e_i\} \sim N(0, \sigma^2)$

Piezīme 17. *Tādu m, kam izpildās (2.1), (2.2) un (2.3), tad sauc par procesa* $X_1, X_2, \ldots, X_{m-1}, X_m,$ *maiņas punktu.*

CUMSUM algoritms

Viens no vienkāršākajiem algoritmiem, lai noskaidrotu, vai maiņas punkts pastāv, un lai noskaidrotu tā atrašanās vietu, ir tā sauktais kumulatīvo summu jeb *CUMSUM* algoritms [12, 5.-10. lpp]:

1. Aprēķina vidējo vērtību $\overline{X} = 1/M \sum_{i+1}^{M} X_i$.

2. Definē sākotnējo jeb nulles kumulatīvo summu $S_0 = 0$.

3. Aprēķina atlikušās kumulatīvās summas

$$S_j = S_{j-1} + X_j - \overline{X},\ j = 1, 2, \ldots, M$$

   pārbaudei var izmantot, ka S_M vienmēr ir 0.

4. Aprēķina S_{dif}, S_{max}, S_{min}

$$S_{dif} = S_{max} - S_{min},$$

$$S_{max} = \max_{j=1,\ldots,M} S_j,$$

$$S_{min} = \min_{j=1,\ldots,M} S_j.$$

   Tālāk tiek realizēta algoritma daļa, kas ietver butstrapa jēdzienu, un šajā algoritmā tiek izmantots neparametriskais butstraps.

5. No sākotnējās izlases $X_1, X_2, \ldots, X_M$ iegūst butstrapa izlasi $X_1^*, X_2^*, \ldots, X_M^*$.

6. Definē sākotnējo butstrapa izlases kumulatīvo summu S_0^*.

7. Aprēķina atlikušās butstrapa izlases kumulatīvās summas

$$S_j^* = S_{j-1}^* + X_j^* - \overline{X},\ j = 1, 2, \ldots, M.$$

8. Analoģiski aprēķina S_{dif}^*, S_{max}^*, S_{min}^*.

9. Apskata, $k = \sum_{i=1}^{N} I_{\{S_{dif} > S_{dif}^*\}}$, kur N butstrapa reižu skaits, bet indikatorfunkcija

$$I_{\{S_{dif} > S_{dif}^*\}} = \begin{cases} 1, & S_{dif} > S_{dif}^* \\ 0, & S_{dif} \leq S_{dif}^* \end{cases}.$$

10. Iegūst ticamības līmeni $\widehat{1-\alpha} = 100\frac{k}{N}$.

Parasti tiek pieņemts, ka maiņas punkts ir nozīmīgs, ja $\widehat{1-\alpha} = 99\%$ vai $\widehat{1-\alpha} = 95\%$. Piemēru apskatīsim nedaudz vēlāk. Kad noskaidrots, vai procesam ir maiņas punkts, atliek precizēt maiņas punkta m atrašanās vietu. Lai to izdarītu, pastāv divas iespējas:

- pirmā metode nosaka, ka maiņas punkts m nosakāms no

$$|S_m| = \max_{j=0,\dots,M} |S_j|;$$

- otrajā metodē maiņas punktu m nosaka ar vidējās kvadrātiskās kļūdas funkcijas $MSE(m)$ palīdzību

$$m = \min_{j=0,\dots,M} MSE(j),$$

kur

$$MSE(m) = \sum_{i=1}^{m}(X_i - \overline{X_1})^2 + \sum_{i=m+1}^{M}(X_i - \overline{X_2})^2,$$

$$\overline{X_1} = 1/m \sum i = 1^m X_i; \quad \overline{X_2} = 1/(M-m) \sum_{i=m+1}^{M} X_i.$$

Piezīme 18. *Šo algoritmu viegli realizēt, izmantojot speciālo datorprogrammu paketi Change Point Analyzer, kas ir maksas programmatūra, bet izmēģinājuma versijā tiek piedāvāts 30 dienu bezmaksas lietošana.*

Ticamības intervāli

Ja papildus tiek izdarīts pieņēmums, ka dati ir ar normālo sadalījumu, tad var izmantot tā saukto 3 *sigmu likumu* (vai arī 6 *sigmu likumu*), proti, kā zināms no matemātiskās statistikas, ka datiem ar normālo sadalījumu ir spēkā šāda teorēma [13, 63.lpp]

Teorēma 19. *Ja* $X \sim N(\mu, \sigma^2)$, *tad* $Z = (X-\mu)/\sigma \sim N(0,1)$ *un, ja* $Z \sim N(0,1)$, *tad* $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$. *Citiem vārdiem, ja* X *ir gadījuma lielums ar normālo sadalījumu, tad tā standartizētā versija vienmēr ir gadījuma lielums ar standarta normālo sadalījumu. Turklāt*

$$X \sim N(\mu, \sigma^2),\ P(X < x) = \Phi\left(\frac{x-\mu}{\sigma}\right), \forall x,$$

kur

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

Līdz ar to, izmantojot centrālo robežteorēmu,

$$P(\mu - \sigma \leq x \leq \mu + \sigma) = \Phi(1) - \Phi(\text{-}1) \approx 0.6827;$$

$$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = \Phi(2) - \Phi(\text{-}2) \approx 0.9772 - (1 - 0.9772) \approx 0.9545;$$

$$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) = \Phi(3) - \Phi(\text{-}3) \approx 0.9973.$$

Pēdējo no vienādībām sauc par trīs sigmu likumu. Analoģiski var iegūt kvalitātes kontroles metodēs bieži izmantoto sešu sigmu likumu

$$P(\mu - 6\sigma \leq x \leq \mu + 6\sigma) = \Phi(6) - \Phi(\text{-}6) \approx 0.999999998027.$$

Šī vienādība norāda, ka ticamības intervāls neiekļaus vidēji tikai 1 no 506 797 345.897 gadījumiem, kas kvalitātes kontroles sistēmās ir pieņemts kā piemērots rezultāts.
Diemžēl šāds variants neder, ja apskatītais process nav baltā trokšņa process. Tādēļ tālāk aplūkosim maiņas punkta analīzes metodes stacionāriem procesiem, kuru autokovarāciju funkcija nav identiska nullei.

2.2. Piemēri CUMSUM grafiki

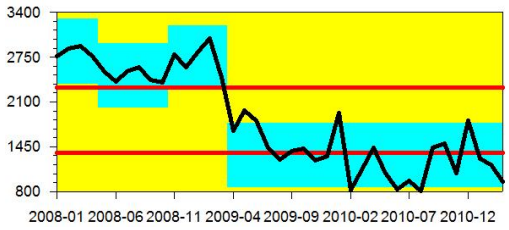
Pielietosim vienkāršo CUMSUM algoritmu laikrindai, kur atspoguļoti elektroenerģijas patētiņa rādījumi, izmantojot datorprogrammu paketi *Change-Point analyzer*.

1. tabula Datu piemērs.

Mēnesis	2008-01	2008-02	2008-03	2008-04	2008-05	2008-06	2008-07	2008-08	2008-09
kWh	2768	2872	2918	2761	2538	2394	2547	2603	2422
2008-10	2008-11	2008-12	2009-01	2009-02	2009-03	2009-04	2009-05	2009-06	2009-07
2384	2790	2610	2822	3020	2461	1679	1980	1832	1432
2009-08	2009-09	2009-10	2009-11	2009-12	2010-01	2010-02	2010-03	2010-04	2010-05
1268	1390	1423	1249	1310	1940	820	1110	1430	1070
2010-06	2010-07	2010-08	2010-09	2010-10	2010-11	2010-12	2011-01	2011-02	2011-03
830	960	810	1430	1500	1070	1830	1270	1190	940

Analizējot tabulā apkopotos datus ar datorprogrammu paketi, kas nodrošina kumulatīvo summu algoritma realizāciju, iegūsim šādus rezultātus, kas apkopoti ekrānizdrukā. (Tabulā paskaidrojošie teksti latviskoti.)

Kā redzams ekrānizdrukās, tad šī piemēra laikrindai ir 3 maiņas punkti, no kuriem visātrāk tika noteikts novērojums, kas datēts ar $2009 - 04$, bet piektajā zarošanās posmā noteikti pārējie divi. Laikrindas grafikā zaļās joslas norāda ticamības intervālu, kas veidots izmantojot 3 sigmu likumu. Tāpat var redzēt, ka katra josla norāda apgabalu vienas



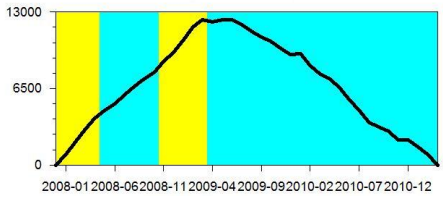
13. att. Novērojumi ar noteiktiem maiņas punktiem

Nozīmīgo maiņas punktu tabula

Nozīmības līmenis, lai iekļautu tabulā=90%; Ticamības intervāla nozīmības līmenis =95%;
Butstrapojumu skaits 10000 (ar atkārtojumiem); MSE novērtējums

Laika moments	Ticamības intervāls	Noz. līm.	No	Uz	Līmenis
2008-05	(2008-05, 2008-05)	96%	2829.8	2481.3	5
2008-11	(2008-09, 2009-03)	91%	2481.3	2740.6	3
2009-04	(2009-04, 2009-04)	100%	2740.6	1323.5	5

14. att. Nozīmīgo maiņas punktu ekrānizdruka (latviskota)



15. att. Kumulatīvo summu jeb CUMSUM grafiks

vidējās vērtības ietvaros. Maiņas punktu tabulā vēl var redzēt, kā mainījusies vidējā vērtība un cik nozīmīgas bijušas izmaiņas (Šeit gan jāteic, ka ekrānizdrukā nozīmības līmenis, kas izteikts procentos, ir noapaļots līdz veseliem procentpunktiem.). Vēl varam aplūkot *tipisku CUMSUM* grafiku, kurā ļoti labi redzams kā veidojas kumulatīvās summas, kā arī to, ka nosakot maiņas punktu pēc otras iespējamās metodes, tas arī sākotnēji būtu novērojums, kas datēts ar $2009 - 04$. Vēl datorprogrammu paketē *SPSS* varam pārliecināties, ka katrā intervālā, kurā aprēķināta vidējā vērtība, dati ir neatkarīgi. Šeit izmantosim jau iepriekš aprakstīto metodi ar Boksa-Ljunga-Pīrsa statistiku. Ilustrācijai apskatīsim tikai pirmo apgabalu, jo pārējos rezultāti ir analoģiski. Veicot pārbaudi, iegūstam jau sagaidāmo, proti, ka nevaram noraidīt hipotēzi, ka novērojumi ir neatkarīgi.

Autokorelācijas

Lags	Autoko-relācijas vērt.	Stand. kļ.[a]	Boksa - Ljunga statistika		
			Vērtība	Brīvības pak.	p-vērtība
1	,301	,234	1,655	1	,198
2	,028	,226	1,671	2	,434
3	,040	,217	1,706	3	,636
4	-,304	,208	3,846	4	,427
5	-,216	,198	5,039	5	,411
6	-,184	,188	6,000	6	,423
7	-,311	,177	9,075	7	,247
8	-,255	,166	11,451	8	,177
9	,017	,153	11,463	9	,245
10	,181	,140	13,127	10	,217
11	,173	,125	15,046	11	,180
12	,072	,108	15,485	12	,216
13	-,006	,089	15,490	13	,278

a. Nulles hipotēzē pieņemts, ka process ir baltais troksnis

16. att. Boksa-Ljunga-Pīrsa statistika.

Maiņas punkta analīze stacionāru stohastisku procesu gadījumā

Iepriekš tika aprakstīts modelis, kurā būtisks bija pieņēmums, ka novērojumi ir neatkarīgi savā starpā, bet lietojumos bieži vien tā var arī nebūt, tāpēc nepieciešams zināms *mehānisms* arī šādu problēmu risināšanai. Lai risinātu šo problemātiku, aplūkosim šādu modeli, kas piemērots, ja aplūkojam vidējās vērtības izmaiņas laikrindās (*angliski* AMOC - *at most one change* modelis). Šeit novērojumus apzīmēsim ar Y_i, $i = 1, \ldots, T$. Līdz ar to

$$Y(i) = \begin{cases} \mu_1 + V(i), & 1 \leq i \leq \tilde{k}, \\ \mu_2 + V(i), & \tilde{k} \leq i \leq T \end{cases}, \tag{2.4}$$

kur $V(\cdot)$ ir stacionārs process ar $EV(\cdot) = 0$, turklāt k, μ_1, μ_2 nav zināmi un problemātika izsakāma šādā hipotēžu pārbaudes veidā

$$H_0 : \quad \tilde{k} < T, \mu_1 \neq \mu_2 \quad H_1 : \quad \tilde{k} = T \ . \tag{2.5}$$

Tad vienkāršākā, arī saukta par CUMSUM, statistika ir [3, 22.-24. lpp.]

$$C_T = \max_{1 \leq k \leq T} \left| \frac{1}{\sqrt{T}} \sum_{j=1}^{k} (Y(j) - \overline{Y}_T) \right| . \tag{2.6}$$

Kritisko vērtību iegūst izmantojot butstrapa metodes frekvenču domēnā (spektrālo analīzi). Zināms, ka

$$\frac{C_T}{\tau} \to \sup_{0 \le t \le 1} |B(t)|,$$

kur $B(\cdot)$ ir Brauna tilta process un $\tau^2 = 2\pi f(0)$, kur $f(\cdot)$ ir spektrālā blīvuma funkcija procesam $\{V(\cdot)\}$. Nesen izstrādāta efektīva TFT butstrapa metode šai problemātikai [3], kuru apskatīsim nākošajā sadaļā. Pielikumā pievienots izveidotais programmas R kods šai problemātikai.

2.3. TFT butstrapa metode

Šajā darba sadaļā apskatīsim maiņas punkta noteikšanas metodi stacionāriem procesiem, kas izmanto aplūkotos stacionāru procesu spektrālās analīzes rezultātus. Metodes pamatā ir izmantota procesa spektrālā reprezentācija, kurā tiek veikts butstraps, tad iegūtais rezultāts tiek reprezentēts atkal laika domēnā. Šāda procesa dēļ arī metode iemantojusi TFT (angliski Time-Frequency-Toggle bootstrap) butstrapa nosaukumu, jo notiek pārslēgšanās no laika domēna uz frekvenču domēnu, kurā tiek veikta datu pārkārtošana, izmantojot butstrapa metodi. Šāda veida pieeja, kad tiek butstrapoti tieši paši Furjē koeficienti tika ieviesta pavisam nesen [3]. Iepriekš tika piedāvāts tikai butstrapot uzreiz novērtēto periodogrammu. Šeit apskatīsim tikai nelielu šīs metodes izklāstu, sīkākai analīzei skatīt Kirhas un Politis publikāciju [3].

Sākumā nepieciešams atcerēties Periodogrammas definīciju (1.4) un Furjē koeficientu $a(j)$ un $b(j)$ vienādojumus (1.1), (1.2) un (1.3). Tagad uzsvērsim Furjē koeficientu īpašības.

Apgalvojums 20. *[3, 5.lpp] Furjē koeficienti (1.2) apmierina šādus nosacījumus:*

$$Ea(j) \to 0,\ Eb(j) \to 0,\ n \to \infty$$

$$Da(j) \to f(\omega_j)/2,\ Db(j) \to f(\omega_j)/2,\ n \to \infty,$$

kur $f(\omega)$ ir spektrālā blīvuma funkcija. Pie tam, ja stohastiskais process ir Gausa, tad $a(j)$ un $b(j)$ ir asimptotiski i.i.d. un sadalīti pēc $N(0, f(\omega_j)/2)$.

Procedūras apraksts

Apskata stacionāru procesu x_t, $t = 1, \ldots, T$.

- Aprēķina Furjē koeficientus, izmantojot FFT (ātro furjē transformāciju) (1.2)
- Apzīmē butstrapotos koeficientus $a(T)^* = b(T)^* = 0$ (Ja T pāra, tad $a(T/2)^* = b(T/2)^* = 0$.
- Izmantojot kādu no butstrapa procedūrām, iegūst pārējos jaunos Furjē koeficientus $a(j)^*$ un $b(j)^*$, $j = 1, \ldots, N$, $N = \lfloor (T-1)/2 \rfloor$.
- Iegūst pārējos koeficientus, no sakarībām

$$a(j)^* = a(T-j)^*,\ b(j)^* = b(T-j)^*.$$

- Izmanto inverso FFT, lai iegūtu centrēto laikrindu $z_t^* = x_t^* - \mu_t^*$.

Tad kā vienkāršāko izmatoto butstrapu var ņemt mežonīgo (angliski *wild*) butstrapu, kur tiek izmantotas procesa spektra komponentes.

Mežonīgais butstraps

Procedūras apraksts

Apskata stacionāru procesu x_t, $t = 1, \ldots, T$.

- Ģenerē standarta normālā sadalījuma $N(0,1)$ gadījuma lielumus $\{s_j : 1 \leq j \leq 2N\}$
- Iegūst gadījuma lielumu butstrapa izlasi s_j^*un jaunos koeficientus

$$a(j)^* = \sqrt{\widehat{f(\omega_j)/2}}s_j^*,\ b(j)^* = \sqrt{\widehat{f(\omega_j)/2}}s_{N+j}^*,$$

kur $\widehat{f(\cdot)}$ ir gludinātais spektrs (skat. (1.5)), ka

$$\sup_{\omega} |\widehat{f}(\omega) - f(\omega)| \to^p 0.$$

Vidējās vērtības izmaiņas laikrindās

Šeit apskatīsim AMOC jeb vienlaikus vienas izmaiņas modeli (angliski at most one change model). Tas nozīmē, ka pieņemam, ka pastāv tikai viens maiņas punkts, kuru vēlamies noskaidrot. Tas neizslēdz iespēju, ka tad, kad atrasts viens maiņas punkts, meklēt nākamo, jo atkal var aplūkot katru no sadalītā procesa (laikrindas) daļām atsevišķi un atkārtot analīzi šim daļām.

Tātad apskatam jau aplūkoto hipotēžu pārbaudi (2.4), (2.5) un (2.6). Tā kā nav zināms, vai spēkā nulles hipotēze vai alternatīva, mēs novērtējam $Z(t) = V(t) - \overline{V_T}$ by $\hat{Z}(t) = Y(t) - \hat{\mu}_1 I_{\{t \leq \hat{k}\}} - \hat{\mu}_2 I_{\{t > \hat{k}\}}$,
kur $\hat{k} = argmax\{|\sum_{j=1}^{k}(Y(j) - \overline{Y_T})| \ : \ 1 \leq k \leq T\}$, $\hat{\mu}_1 = 1/\hat{k}\sum_{j=1}^{\hat{k}} Y(j)$, $\hat{\mu}_2 = 1/(T-\hat{k})\sum_{j=\hat{k}+1}^{T} Y(j)$.

Tad TFT butstapotā statistika definējam

$$C_T = \left| \frac{1}{\sqrt{T}} \sum_{j=1}^{k} Z^*(j) \right|,$$

kur $\{Z^*(\cdot)\}$ is butstrapa virkne, kas iegūta pēc TFT-butstrapa shēmas. Tad atliek tikai iegūt statistikas butstrapoto sadalījumu un kritisko vērtību.

3. DIVU IZLAŠU LOKĀCIJAS PARAMETRU VIENĀDĪBAS TESTS AR EL, TĀ PIELIETOJUMS MAIŅAS PUNKTA ANALĪZĒ

3.1. Motivācija un īss vēsturisks ieskats

Šās nodaļas pamatideju var izteikt formulējot divu izlašu lokācijas parametru starpības hipotēžu pārbaudi. Pieņemsim, ka dotas divas izlases $X = X_1, X_2, \ldots, X_{n_1}$ un $Y = Y_1, Y_2, \ldots, Y_{n_2}$, tad varam formulēt divu izlašu lokācijas parametru vienādības hipotēžu pārbaudes nulles hipotēzi

$$H_0 : \ \mu_2 - \mu_1 = 0 \tag{3.1}$$

pret alternatīvu

$$H_1 : \ \mu_2 - \mu_1 \neq 0. \tag{3.2}$$

Apskatīsim, kā izmatojot empīriskās ticamības funkcijas metodi divu izlašu lokācijas parametru starpības pārabudei un atkarīgu datu bloku veidošanu, iespējams veikt maiņas punkta noteikšanu lokācijas parametram.

Nesen (2011.g.) interneta versijā pieejamajā, bet žurnālā apstiprinātajā publikācijā [14, 56.-57.lpp.] apskatīts empīriskās ticamības funkcijas pielietojums atkarīgu datu gadījumā, kurā apskatīta ticamības intervāla konstruēšana divdimensionālas stacionāras laikrindas $X_1 = (X_1^1, X_1^2), X_2 = (X_2^1, X_2^2), \ldots, X_n = (X_n^1, X_n^2)$ korelācijas koeficientam

$$\rho = cov(X^1, X^2)/\sqrt{D(X^1)D(X^2)}.$$

Šajā darbā arī veiktas atbilstošas simulācijas. Sīkākai informācijas iegūšanai skatīt publikāciju. Interesi saista šajā publikācijā minētais, ka kopš Ovens (Owen) iepazīstināja ar empīrisko ticamības funkciju un ticamības intervāla konstruēšanu vidējajai vērtībai, šis virziens ir strauji attīstījies vairākos virzienos: regresiju analīzē, aditīvajos riska modeļos, kopulu analīzē un divu izlašu lokācijas problemātikā. Turklāt, ja apskatāmie funkcionāļi ir nelineāri, empīriskās ticamības funkcijas metodes kļūst vēl komplicētākas, jo tad Vilksa (Wilks) teorēma vispārīgā gadījumā vairs nav spēkā, t.i., empīriskās ticamības funkciju attiecības testa asimptotiskais sadalījums vispārīgā gadījumā vairs nav χ_q^2- sadalījums ar atbilstošām q brīvības pakāpēm.

Savukārt motiváciju apskatīt problemātiku divu izlašu lokācijas parametru starpībai at-

karīgu datu gadījumā dod R.Frīda (Roland Fried) publikācija [15] par robustu lokācijas parametra noteikšanu autoregresīviem procesiem, kur šī problemātika cieši sasaistās ar maiņas punkta noteikšanu vāji atkarīgu (stacionāru) procesu gadījumā.

3.2. Problēmas nostādne

Kopš Y. Kitamuras publikācijas [16] 1997. gadā zināms kā, izmantojot datu pārkārtošanu blokos jeb datu blokošanu, iespējams pielietot dažādus empīriskās ticamības funkcijas metodi atkarīgu datu gadījumā. Sākumā definēsim vāji atkarīgus stohastiskus procesus vispārējā gadījumā, ko raksturo jauktie koeficienti.

Definīcija 11. [14, 57], [16, 2086] Process$\{X_i, 1 \leq i \leq n\}$ ir $\mathbb{R}^d$ stacionārs process ar stingro atkarības koeficientu $\alpha_x(k)$, ja

$$\alpha_x(k) = \sup_{A \in \mathcal{F}_\infty^0, B \in \mathcal{F}_k^{+\infty}} |P(AB) - P(A)P(B)| \to 0,\ k \to \infty,$$

kur $\mathcal{F}_a^b = \sigma(X_i, a \leq i \leq b)$ ir σ-algebra, kas balstīta uz X_i, $a \leq i \leq b$. Kā zināms, šādu nosacījumu apmierina arī ARMA procesi [14].

Sākumā aplūkosim jau zināmus rezultātus [17, 3.-5. lpp] par empīriskās ticamības funkcijas attiecību testu divu izlašu gadījumā neatkarīgiem datiem. Apskatīsim neatkarīgus un vienādi sadalītus datus no divām izlasēm $X_1, \ldots, X_{n_1}$ un $Y_1, \ldots, Y_{n_2}$ ar attiecīgi sadalījuma funkcijām F_1 un F_2. Apskatīsim Δ un parametru θ. Pieņemsim, ka visa nepieciešamā informācija par patiesajiem parametriem θ_0 un Δ_0 no nenovirzītiem vienādojumiem

$$E_{F_1} g_1(X, \theta_0) = 0, \tag{3.3}$$

$$E_{F_2} g_2(Y, \theta_1) = 0. \tag{3.4}$$

Ja $\Delta_0 = \theta_1 - \theta_0$, kur $\theta_0 = \int x dF_1(x)$ un $\theta_2 = \int x dF_2(x)$ tad izvēlamies

$$g_1(X, \theta_0, \Delta_0) = X - \theta_0,$$

$$g_2(Y, \theta_0, \Delta_0) = Y - \theta_0 - \Delta_0,$$

kas apmierina (3.3) un (3.4). Tālāk apskatīsim empīriskās ticamības funkcijas attiecības testa statistikas konstruēšanu divu izlašu gadījumā. [17, 4. -6. lpp] Iegūsim ticamības

intervālu parametram Δ. Definēsim empīrisko ticamības funkciju attiecību funkciju

$$R(\Delta,\theta)=\sup_{p,q}\prod_{i=1}^{n_1}(n_1p_i)\prod_{j=1}^{n_2}(n_2q_j),$$

kur $p=(p_1,p_2,\ldots,p_{n_1})$ un $q=(q_1,q_2,\ldots,q_{n_2})$ ir sadalījuma vektori, t.i. , nenegatīvi lielumi ar summu viens, turklāt $\sum_{i=1}^{n_1}p_ig_1(X_i,\theta_0)=0$ un $\sum_{j=1}^{n_2}q_jg_1(Y_j,\theta_1)=0$.
Viens vienīgs atrisinājums eksistē, ja 0 piederētu izliektajai čaulai, ko veido $g_1(X_i,\theta,\Delta)$ un $g_2(Y_j,\theta,\Delta)$. Maksimumu atrod, izmantojot Lagranža reizinātājus, kas dod

$$p_i=\frac{1}{n_1(1+\lambda_1(\theta))g_1(X_i,\theta,\Delta)},\ i=1,2,\ldots,n_1$$

$$q_j=\frac{1}{n_2(1+\lambda_2(\theta))g_2(Y_j,\theta,\Delta)},\ j=1,2,\ldots,n_2,$$

kur Lagraža reizinātāji nosakāmi no vienādojumiem

$$\sum_{i=1}^{n_1}\frac{g_1(X_i,\theta,\Delta,t)}{1+\lambda_1(\theta)g_1(X_i,\theta,\Delta)}=0 \tag{3.5}$$

$$\sum_{j=1}^{n_2}\frac{g_2(Y_j,\theta,\Delta,t)}{1+\lambda_2(\theta)g_2(Y_j,\theta,\Delta)}=0. \tag{3.6}$$

Visbeidzot mēs definējam empīrisko *log* - ticamības funkciju attiecību, kas pareizināta ar mīnuss divi

$$W(\Delta,\theta)=-2logR(\Delta,\theta)=2\sum_{i=1}^{n_1}(1+\lambda_1(\theta)g_1(X_i,\theta,\Delta))+2\sum_{j=1}^{n_2}(1+\lambda_2(\theta)g_2(Y_j,\theta,\Delta)).$$

Lai iegūtu novērtējumu $\hat{\theta}=\widehat{\theta(\Delta)}$ nepieciešamajam parametram θ, kas maksimizē $R(\Delta,\theta)$ fiksētam Δ, jāveido vienādojumu sistēma (3.5), (3.6) un (3.7).

$$\frac{\partial W(\Delta,\theta)}{\partial\theta}=\sum_{i=1}^{n_1}\frac{\lambda_1\alpha_1(X_i,\theta,\Delta)}{1+\lambda_1(\theta)g_1(X_i,\theta,\Delta)}+\sum_{j=1}^{n_2}\frac{\lambda_2\alpha_2(Y_j,\theta,\Delta)}{1+\lambda_2(\theta)g_2(Y_j,\theta,\Delta)}=0, \tag{3.7}$$

kur

$$\alpha_1(X_i,\theta,\Delta)=\frac{\partial g_1(X_i,\theta,\Delta)}{\partial\theta},\ \alpha_2(Y_j,\theta,\Delta)=\frac{\partial g_2(Y_j,\theta,\Delta)}{\partial\theta}.$$

Teorēma 21. *Izpildoties gluduma nosacījumiem funkcijām $g_1,g_2,\alpha_1,\alpha_2$, eksistē tāds $\hat{\theta}$,*

ka $\hat{\theta}$ *ir būtisks* θ_0 *novērtējums un* $R(\Delta, \theta)$ *sasniedz lokālo maksimumu pie* $\hat{\theta}$*, turklāt*

$$\sqrt{n_1}(\hat{\theta} - \theta_0) \to_d N\left(0, \frac{\beta_1\beta_2}{\beta_2\beta_{10}^2 + k\beta_1\beta_{20}^2}\right),$$

kur $k < \infty$ *ir pozitīva konstante, kurai izpildās* $n_2/n_1 \to k$*, kad* $n_1, n_2 \to \infty$*, turklāt*

$$-2\log R(\Delta_0, \hat{\theta}) \to_d \chi_1^2,\ n_1, n_2 \to \infty, \forall t \in T,$$

kur

$$\beta_1 = E_{F_1} g_1^2(X, \theta_0, \Delta_0),\ \beta_2 = E_{F_2} g_2^2(Y, \theta_0, \delta_0),$$

$$\beta_{10} = E_{F_1} \alpha_1(X, \theta_0, \Delta_0),\ \beta_{20} = E_{F_2} \alpha_2(Y, \theta_0, \delta_0).$$

Tad punktveida ticamības intervāls parametram Δ nosakāms no $\Delta : \{R(\Delta, \hat{\theta}) > c\}$ visiem Δ_0, kur c nosakāms no šīs teorēmas.

3.3. Datu bloku veidošana jeb blokošana

Tā kā praksē bieži vien nākas saskarties ar atkarīgām datu struktūrām, tad nosacījums, ka dati ir neatkarīgi, bieži vien neizpildās, līdz ar to iepriekš aplūkotie rezultāti vairs nav spēkā. Šādās situācijās nepieciešams izmantot nedaudz citu pieeju. Kitamura savā publikācijā [16] piedāvā izmantot datu bloku veidošanu speciālā veidā vienas izlases gadījumā, kur pēc tam varētu piemērot iepriekš aplūkoto teorētisko rezultātu. Savā publikācijā Kitamura izmanto datu bloku veidošanu vienas izlases gadījumā, kad izlases dati ir ar atkarības struktūru, bet šeit mēs apskatīsim bloku veidošanai, kas analoģiska Kitamuras [16] un [14] publikācijās. Apskatīsim funkcionāļus $G(x, \theta)$ tādus, ka $EG(x, \theta) = 0$. Ar M un L apzīmēsim tādus veselus skaitļus, kas atkarīgi no n, $M \to \infty$, $M = o(n^{1/2})$ un $L = O(M)$, kad $n \to \infty$. Izvēlamies $Q = [(n - M)/L] + 1$, kur ar $[\cdot]$ apzīmēta veselā daļa. Tātad tiek izveidoti Q bloki no M novērojumiem katrā blokā un L atstatumu starp bloku sākumpunktiem $B_i = (X_{(i-1)L+1}, \ldots, X_{(i-1)L+M})$, $i = 1, \ldots, Q$. Tālāk Kitamura piedāvā jau šos blokus izmantot kā novērojumus attiecīgi statistikā, proti, katram no blokiem tiek aprēķināts $T_i = \phi_M(B_i, \theta)$, kur B_i ir i-tais bloks un attēlojumu $\phi_M : \mathbb{R}^M \times \theta \to \mathbb{R}$ uzdod šādā formā

$$\phi_M(B_i, \theta) = \sum_{j=1}^{M} g(X_{j-1)L+n}, \theta)/M,\ i = 1, 2, \ldots, Q.$$

Tātad tiek ņemti M slīdošie vidējie ar dažādiem svariem.

Pielietosim šo bloku veidošanas paņēmienu katrai no izlasēm un tad aplūkosim iepriekš apskatīto teoriju katras izlases bloku funkciju rezultātiem, tātad divu izlašu gadījumā Visbeidzot varam veikt pārbaudi par divu izlašu lokācijas parametru starpību. Tādēļ veiksim šādu hipotēžu pārbaudi:

$$H_0 : \Delta_0 = 0,\ H_1 : \Delta_0 \neq 0$$

tikai šoreiz jāņem vērā, ka mēs apskatīsim nevis statistiku pašām izlasēm, bet gan to bloku veidotajām izlasēm T_i, jo nosacījums par izlasēm, t. i., to elementu neatkarību un vienādo sadalījumu pašām izlasēm nav spēkā, un līdz ar to asimptotiskais sadalījums vispārīgā gadījumā nebūtu hī-kvadrāta sadalījums ar atbilstošām brīvības pakāpēm.
Tā kā šajā gadījumā funkcionālim

$$g_1(T_1, \theta_0, \Delta_0, t) = X - \theta_0,$$

mums jānovērtē parametrs θ_0, kas nepieciešams attēlojumā ϕ_M, jo savādāk nevaram aprēķināt T_i, kuri definēti iepriekš (nevaram veikt aprēķinu blokos). Šoreiz θ_0 ir pirmās izlases vidējā vērtība. Šoreiz lietosim tā saukto ievietošanas jeb *plug-in* metodi un parametra novērtējumam izmantosim $\hat{\theta}_0 = 1/n_1 \sum_{i=1}^{n_1} X_i$ un līdz ar to arī

$$g_2(T_2, \theta_0, \Delta_0, t) = Y - \theta_0 - \Delta_0$$

izmantosim šo pašu novērtējumu. Definēsim statistiku

$$-2 \log R(\Delta_0, \hat{\theta}) \to_d \chi_1^2,\ Q_1, Q_2 \to \infty,$$

Tālāk rīkojamies tāpat, kā noteikts nodaļā 3.2.Tā kā mēs pārbaudām hipotēzi par divu izlašu vidējo vērtību vienādību, tad konstruējot ticamības intervālu parametram Δ_0, tam jāpārklāj 0. Līdz ar to varam veikt simulācijas, lai pārbaudītu pārklājuma precizitāti izmantojot dažādus bloku garumus M un dažādus atstatumus starp bloku sākumpunktiem L.

Divu izlašu t-tests

Viens no populārākajiem testiem divu izlašu vidējo vērtību starpībai ir divu izlašu t-tests [18, 400. - 403. lpp.]. Lai gan tas ir stingri ierobežots ar nosacījumu pa abu izlašu

normalitāti un izlašu datu neatkarību, tomēr samērā bieži tas tiek izmantots izlasēm, kas neapmierina šos nosacījumus. Tādēļ apskatīsim divu izlašu t-testu.

Teorēma 22. *Pieņemsim, ka* $X_1, X_2, \ldots, X_m \sim^{iid} N(\mu_1, \sigma^2)$ *un* $Y_1, Y_2, \ldots, Y_n \sim^{iid} N(\mu_2, \sigma^2)$ *, kur visi* $m+n$ *novērojumi ir neatkarīgi un vienādi sadalīti. Tad divu izlašu* t*-testa statistika ir*

$$T_{m,n} = \frac{\overline{X} - \overline{Y}}{s\sqrt{m^{-1} + n^{-1}}}\,,\; s^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}.$$

Turklāt pie $H_0 : \mu_1 = \mu_2$ *zināms, ka* $T_{m,n} \sim t_{m+n-2}$*, kur* t_{m+n-2} *ir studenta sadalījums ar* $m+n+2$ *, kā arī* $T_{m,n} \to N(0,1)$*, ja* $m, n \to \infty$*.*

Piezīme 23. *[18, 401.lpp] Vispārīgā gadījumā, kad* $X_1, \ldots, X_m \sim F$ *un iid. un* $Y_1, \ldots, Y_n \sim G$ *un iid., un* F, G *ir ar vienādu vidējo vērtību un dispersiju, tad pēc centrālās robežteorēmas un Slutska lemmas tāpat kā normālā sadalījuma gadījumā* $T_{m,n} \to^d N(0,1)$*, ja* $m, n \to \infty$*, turklāt ar tādu pašu konverģences ātrumu un jaudu.*

3.4. Metodes empīriskā analīze

Llai pārbaudītu iepriekšējā darba sadaļās apskatītās teorētisko metodi, kur izmaojam EL divu izlašu testam par lokācijas parametru vienādību, kā arī, lai gūtu labāku priekšstatu par to, kā izvēlēties piemērotu bloka loga platumu M un novērtētu tā ietekmi, veiksim metodes pārklājuma precizitātes pārbaudi.

3.5. pārklājuma precizitātes analīze

Sākumā apskatīsim izlases ar novērojumu skaitu $n_1 = n_2 = N = 1000$. Veicot 1000 simulācijas izlasēm, kuru ģenerējošais process ir autoregresīvais process $AR(1)$,

$$X_t = aX_{t-1} + \varepsilon_t$$

ar dažādām parametra a vērtībām $0.3; -0.3; 0.9; -0.9$. Aplūkojot tabulā apkopotos datus, var redzēt, ka pie izvēlētās precizitātes 0.95 (vai nozīmības līmeņa $\alpha = 0.05$), izvēlētā metode strādā samērā labi, vienīgi parametra vērtībai $a = 0.9$ rezultāti nav apmierinoši. Kā redzams tad arī bloku platuma izvēle M ir būtiska. Kā jau tas neparametriskā statistikā bieži sastopams, arī šeit būtiski ir izvēlēties piemērotu bloku platumu M, bet, diemžēl, šī problemātika nav triviāla. Literatūrā visbiežāk bloku platuma izvēle tiek atstāta lietotāja ziņā, jo šī metode ir jauna, un iesaka apskatīt līdzīgas problemātikas risināšanu bloku butstrapa gadījumā. Šajās simulācijās bloku platums ņemts kā ieteikts publikācijā [14], t.i., $M = N^{i/5}, i = 1, 2, 3$. Pirmajos trijos gadījumos aplūkoti nešķeļošie bloki, bet nākošajā tabulā apkopoti bloki, kas savstarpēji šķeļas. Kā redzams, tad šajā gadījumā viennozīmīgi labāku rezultātu dod nešķeļošo bloku gadījums. Sīkākai analīzei par bloku platumu un skaitu skatīt Y. Kitamuras publikāciju [19, 35.-44. lpp]

2. tabula: Nešķeļošo bloku EL novērtējuma ticamības intervālu pārklājuma precizitāte, $n_1 = n_2 = N = 1000$, simulāciju skaits $N_{sim} = 1000$

	$a = 0.3$	$a = -0.3$	$a = 0.9$	$a = -0.9$
L= 15 M= 15 Q= 66	0.937	0.958	0.814	0.990
L= 63 M= 63 Q= 15	0.935	0.946	0.912	0.948
L= 3 M= 3 Q= 333	0.930	0.963	0.537	0.997

3. tabula: Šķeļošu bloku EL novērtējuma ticamības intervālu pārklājuma precizitāte, $n_1 = n_2 = N = 1000$, simulāciju skaits $N_{sim} = 1000$

	$a = 0.3$	$a = -0.3$	$a = 0.9$	$a = -0.9$
L= 7 M= 15 Q= 142	0.799	0.819	0.623	0.903
L= 31 M= 63 Q= 31	0.807	0.822	0.778	0.806
L= 1 M= 3 Q= 998	0.693	0.814	0.334	0.994

Šādā pat veidā apskatīsim rezultātus, kas iegūti, ja izlašu apjomi ir par kārtu mazāki $n_1 = n_2 = N = 100$. Šajā gadījumā, tāpat kā iepriekš varam redzēt, ka tikai pie parametra $a = -0.9$ metode strādā samērā slikti. Pat par kārtu samazinot izlases apjomu, varam redzēt, ka rezultāti būtiski nepasliktinās, līdz ar to varam teikt, ka metode strādā samērā labi apskatītajā gadījumā.

4. tabula: Nešķeļošo bloku EL novērtējuma ticamības intervālu pārklājuma precizitāte, $n_1 = n_2 = N = 100$, simulāciju skaits $N_{sim} = 1000$

	$a = 0.3$	$a = -0.3$	$a = 0.9$	$a = -0.9$
L= 6 M= 6 Q= 16	0.925	0.941	0.626	0.977
L= 15 M= 15 Q= 6	0.861	0.888	0.731	0.966
L= 2 M= 2 Q= 50	0.898	0.973	0.465	0.993

Kā redzams, tad arī šeit rezultāti būtiski atkarīgi no bloka platuma M izvēles, kā arī no L. Ne mazāku interesi izraisa rezultāti, kurus varam iegūt, ja izmantojam divu izlašu t-testu. Šo simulāciju rezultāti apkopoti tabulā. Kā redzams, tad, lai arī neizpildās nosacījums par datu neatkarību, tomēr t-tests arvien strādā labi, ļoti tuvu ar apskatīto metodi.

5. tabula: Šķeļošu bloku EL novērtējuma ticamības intervālu pārklājuma precizitāte, $n_1 = n_2 = N = 100$, simulāciju skaits $N_{sim} = 1000$

	$a = 0.3$	$a = -0.3$	$a = 0.9$	$a = -0.9$
L= 6 M= 3 Q= 32	0.780	0.846	0.488	0.927
L= 15 M= 7 Q= 13	0.749	0.783	0.577	0.855
L= 2 M= 1 Q= 99	0.768	0.883	0.330	0.933

6. tabula: t-testa ticamības intervālu parametru starpībai pārklājuma precizitāte atbilstošiem $n_1 = n_2 = N$, simulāciju skaits $N_{sim} = 1000$

	$a = 0.3$	$a = -0.3$	$a = 0.9$	$a = -0.9$
$N = 1000$	0.862	0.986	0.333	0.989
$N = 100$	0.846	0.985	0.354	0.997

Kā varam secināt no izdarītajām simulācijām, tad ir divi būtiski aspekti, ko jāņem vērā. Pirmkārt rezultātus samērā būtiski ietekmē bloka platuma M izvēle kā arī tas, vai izvēlas šķeļošus vai nešķeļošus blokus. Otrkārt var pamanīt, ka apskatītās metodes efektivitāte nesamazinās tik būtiski, ja apskata arī ievērojami mazākas izlases. Treškārt, ja salīdzina ar rezultātiem, ko iegūst ar divu izlašu t-testa palīdzību, varam secināt, ka šie rezultāti būtiski neatšķiras, vienīgi var redzēt, ka šajā gadījumā daudz sliktāk abas metodes strādā pie koeficienta 0.9.

3.6. Maiņas punkta noteikšana ar divu izlašu testu un logiem

Kad aplūkota metode, kur, izmantojot EL veikta hipotēžu pārbaude par lokācijas parametru vienādību, šo ideju tālāk varam attīstīt un pielietot kāda stacionāra procesa laikrindas x_T maiņas punkta atrašanai. Proti, izvēlēsimies logu

$$W_h(x_T) = (X_{T+1}, X_{T+2}, \ldots, X_T + h)$$

un tā garumu $h = 1, 2, \ldots$, tad, izmantojot divus šādus logus $W_h(x_T)$ un $W_h(x_{T+h})$, kas ir sekojoši un nepārklājas, veiksim hipotēžu pārbaudi par šo logu veidoto laikrindu (apakšlaikrindu) vidējo vērtību, izmantojot iepriekš aprakstīto metodiku gadījumā, kad logi nepārklājas. Tad, konstruējot p-vērtību grafiku, var veikt grafisku maiņas punkta noteikšanas pārbaudi, t.i., intervālā, kur šis grafiks pieņem nulles vērtību meklējams laikrindas patiesais maiņas punkts, precīzāk maiņas punkts ir tieši šā intervāla viduspunkts. Tādējādi, variējot ar loga garumu, iespējams noteikt laikrindas maiņas punktus un to nozīmīgumu.

Apskatīsim nelielu datu piemēru, kur apskatīts $AR(1)$ process, kas konstruēts tā, lai lokācijas parametra maiņas punkts būtu tieši pie $k = 500$.

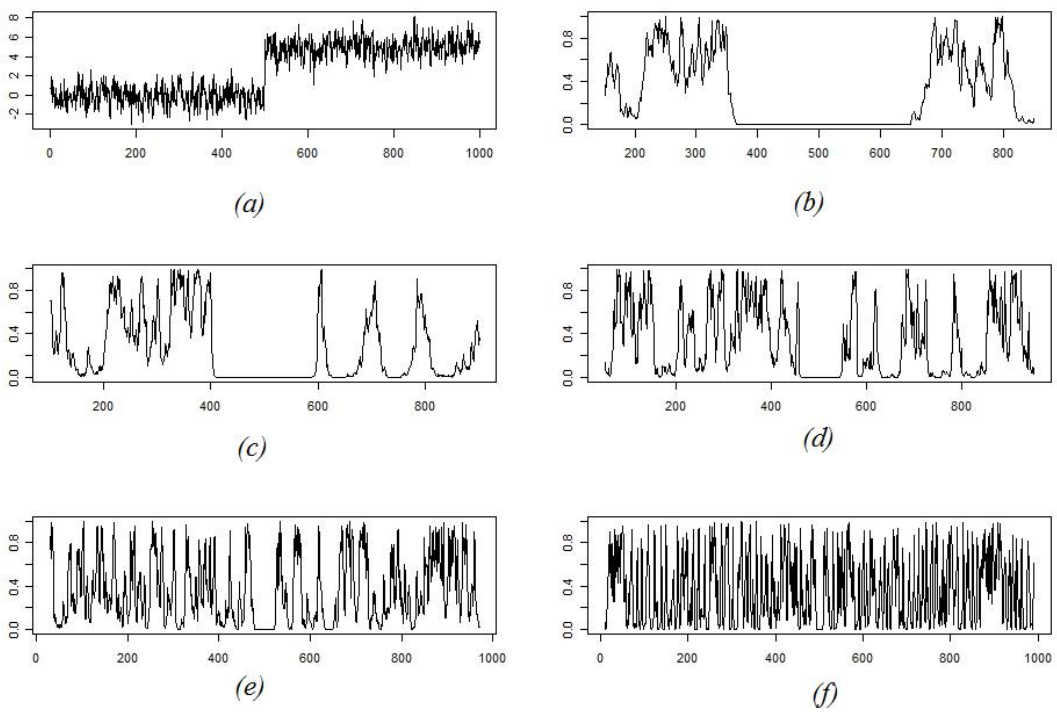
$$X_i = \begin{cases} \mu_1 + V(i), & 1 \leq i \leq 500, \\ \mu_2 + V(i), & 500 < i \leq 1000 \end{cases},$$

kur $V(\cdot)$ ir $AR(1)$ process, turklāt ņemsim $\mu_1 = 0, \mu_2 = 5$ un tātad $k = 500$.

Izmantojot šo pieeju, mēģināsim novērtēt $\tilde{k}$. Veicot piemēra konstruēšanu paketē R, iegūsim p-vērtību grafikus, no kuriem varam secināt, ka arī ar metodi iegūtais maiņas punkt ir tuvas patiesajam $k = 500$. Precīzāku novērtējumu $\tilde{k}$, iegūsim no aprēķinātajām p-vērtībām, proti ņemot laikrindas laika momentus, kuri ir šī intervāla galapunkti, tad šī intervāl viduspunkts ir tieši meklētais maiņas punkts. Kā redzams attēlā, tad izmantojot lielāku loga platumu, var noteikt, kur meklējams maiņas punkts un tad samazinot loga platumu, to var noteikt precīzāk.

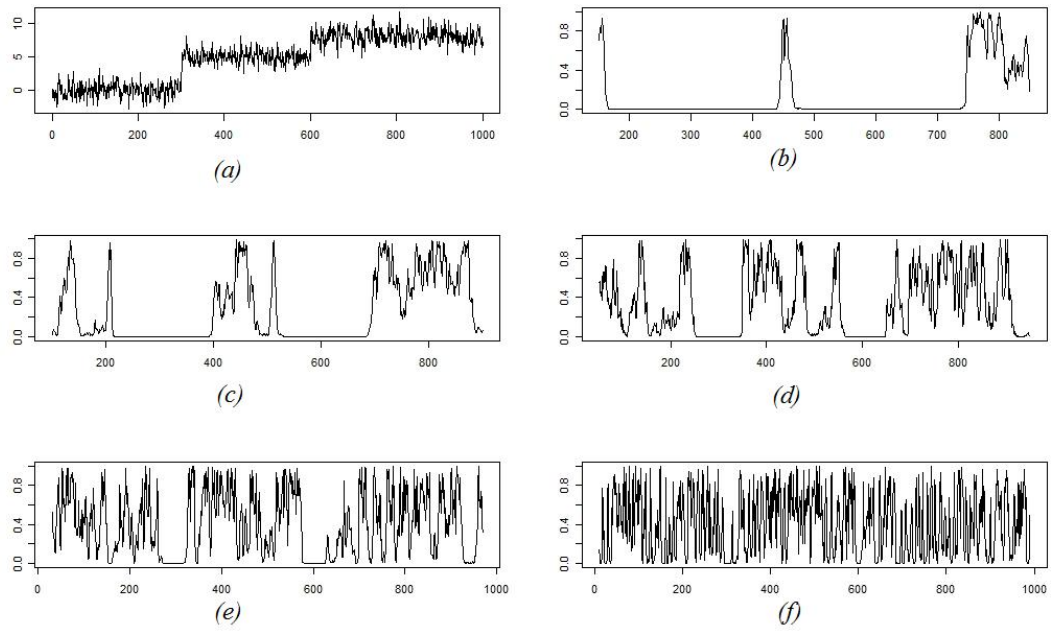
Interesanti aplūkot gadījumu, kad laikrindai ir divi maiņas punkti k_1 un k_2. Tad aplūkosim šādu piemēru.

$$X_i = \begin{cases} \mu_1 + V(i), & 1 \leq i \leq 300, \\ \mu_2 + V(i), & 300 < i \leq 600, \\ \mu_3 + V(i), & 600 < i \leq 1000 \end{cases},$$



17. att.: Laikrindas attēlojums (a) un p-vērtību grafiki dažādiem loga garumiem $(b) - (f)$ $(h = 150; 100; 50; 30; 10)$.

kur $V(\cdot)$ ir $AR(1)$ process, turklāt ņemsim $\mu_1 = 0, \mu_2 = 5, \mu_3 = 8$ un tātad $k_1 = 300$ un $k_2 = 600$. Tātad viena no šīs metodes priekšrocībām ir tā, ka vienlaicīgi, izvēloties atbilstošu loga garumu, var noteikt vairākus maiņas punktus.



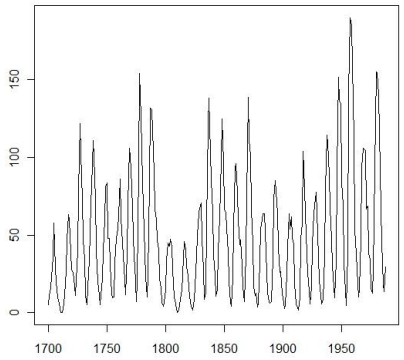
18. att.: Laikrindas ar 2 maiņas punktiem attēlojums (a) un p-vērtību grafiki dažādiem loga garumiem $(b) - (f)$ $(h = 150; 100; 50; 30; 10)$.

4. METOŽU PIELIETOJUMS

4.1. Reālu datu piemēri.

Šajā darba daļā aplūkosim reālu laikrindu datus. Šeit aplūkosim tieši laikrindu datus, jo tie ir visbiežāk sastopamie praktiskos lietojumos un, ja laikrinda atspoguļo kāda fizikāla procesa vai ekonomikas norises procesu, iegūtos rezultātus viegli interpretēt, tādejādi saprotot, cik atbilstošs ir iegūtais rezultāts. Apskatīsim jau pieminēto *sunspot* laikrindu, kā arī *respiration* datus, kur apkopoti dati par smadzeņu darbības aktivitāti atkarībā no cilvēka nomoda stāvokļa, tas ir, vai cilvēks ir aizmidzis vai nomodā. Šo datu analīze palīdzēs saprast, vai ar šim metodēm iegūtie rezultāti ir līdzīgi un to, cik labi strādā šīs metodes aplūkotajiem gadījumiem.

Sākumā aplūkosim rezultātu, kādu var iegūt ar TFT butstrapa palīdzību, ja aplūkojam datus par saules aktivitāti. Kā jau iepriekš minēts, *sunspot* datu kopa satur informāciju par saules aktivitātes datiem, kas bieži tiek izmantoti laikrindu analīzē (datus var iegūt, piemēram, `http://www.napscience.com/astro/sunspots/sunspotdata.htm`, ņemot datus par 1700.-1987. gadu).

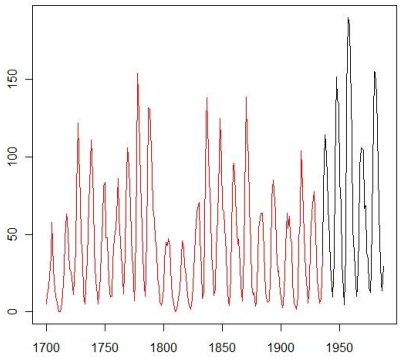


19. att.: Laikrindas ar 2 maiņnas punktiem attēlojums un p-vērtību grafiki dažādiem loga garumiem ($h = 150; 100; 50; 30; 10$).

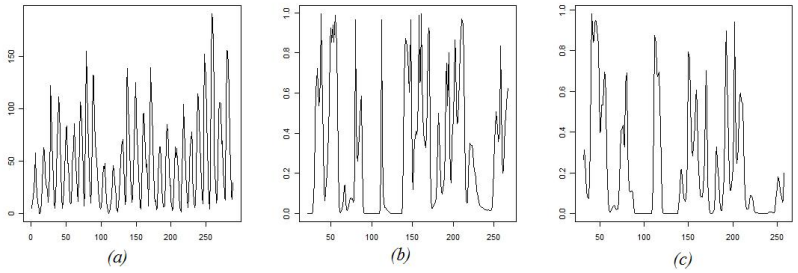
Ja izmantojam TFT butstrapa metodi, tad iegūsim, ja laikrindai ir maiņas punkts ir laikrindas 236 punkts jeb 1935. gads.

Tagad apskatīsim, kādu rezultātu iegūsim ar jaunaplūkoto metodi. Kā redzams , tad ar šo metodi atradām ne tikai jau iepriekšējo maiņas punktu, bet arī vēl divus citus maiņas punktus (1800.gads un 1840.gads), kas aplūkojot laikrindu arī šķiet gana iespējami.

Otra datu kopa, ko aplūkosim atspoguļo smadzeņu aktivitātes mērījumus dažādos cilvēka fizioloģiskajos stāvokļos (nomoda, miega, dziļa miega utml.). Šie dati analizēti

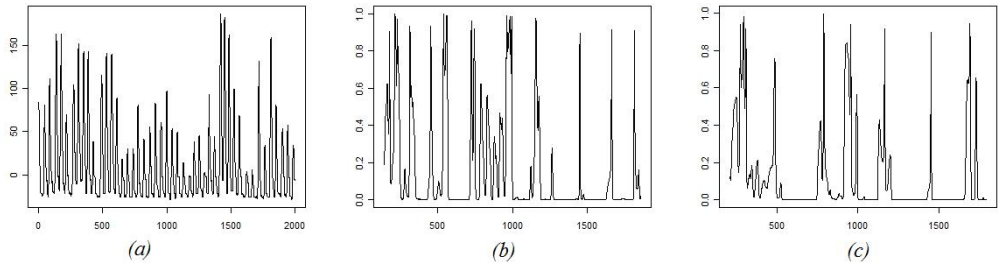


20. att. Saules aktivitātes datu analīze ar TFT butstrapu.



21. att.: Saules aktivitātes datu laikrindas attēlojums (a) un p-vērtību grafiki dažādiem loga garumiem $(b), (c)$ $(h = 20; 30)$.

arī publikācijā [1], kur izmantota vēl cita laikrindu maiņas punkta noteikšanas metode, kuru šajā darbā neaplūkosim. Datos akopotās svārstības raksturo aktivitāti, proti, kad cilvēks ir nomodā svārstības ir biežķas un ar lielāku amplitūdu, turklāt citu lokācijas parametru, ko tad arī mēģināsim pārbaudīt. Šīs datu masīvs satur nedaudz vairāk nekā 18 tūkstošus ierakst, bet mēs analizēsim tikai daļu no tās (8000. līdz 9500.ierakstu). Kā redzams, tad vienlaicīgi var labi novērtēt maiņas punktus ar jauno metodi, kamēr ar TFT butstrapu, varētu iegūt tikai vienu vienīgu, ar kura palīdzību var sadalīt laikrindu atkal divās daļās un atkārtot analīzi katai no daļām. Salīdzinot rezultātus ar tiem, kas pieejami [1], var secināt, ka abas metodes šajā gadījumā strādā vienlīdz labi. Tagad



22. att.: NPRS datu laikrindas attēlojums (a) un p-vērtību grafiki dažādiem loga garumiem $(b), (c)$ $(h = 20; 30)$.

arī praktiskos piemēros redzējām, kā darbojas apskatītās metodes. Vienīgais būtiskākais trūkums jaunaplūkotajai metodei ir tas, ka tā nespēj identificēt maiņas punktus, kas ir laikrindas sākumposmā un beigu posmā, ja šie intervāli ir mazāki par loga garumu, līdz ar to loga garumu nevajadzētu ņemt pārāk lielu, savukārt pārāk maza loga izvēle noved pie tā, ka metode zaudē savu efektivitāti, jo to vairāk iespaido gadījuma kļūdas. Tātad jārod kompromiss par loga platuma izvēli. Šobrīd šī darba ietvaros loga platuma izvēle tiek atstāta pētnieka ziņā.

5. NOBEIGUMS

Diplomdarbā apskatījām galvenos spektrālās analīzes jēdzienus un uzmanību vērsām uz periodogrammas gludināšanu, kas arvien ir aktuāls temats daudzās publikācijās. Aplūkojām gan neparametriskas, gan parametriskas gludināšanas metodes. Kad aplūkoti spektrālās analīzes pamatjautājumi un problēmu risinājumi attiecībā uz periodogrammas gludināšanu, šos rezultātus var izmantot maiņas punkta analīzes metodēs, kam tika veltīta šī darba galvenā sadaļa.

Diplomdarbā tika aplūkota maiņas punkta analīzes vispārējā forma un maiņas punkta analīzes iespējamie virzieni. Šajā darbā vislielāko vērību pievēsām tieši maiņas punkta analīzes metodēm, kas apskata maiņas punkta noteikšanu lokācijas parametram. Sākotnēji tika apskatīts visvienkāršākais kumulatīvo summu algoritms, bet pēc tam plašāk analizētas tika maiņas punkta analīzes metodes, kas piemērotas laikrindu datiem, tātad atkarīgām datu struktūrām. Tika apskatīta metode, kas izmanto TFT butstrapu, kā arī ieviesta un realizēta metode, kas izmanto divu izlašu testa par lokācijas parametru vienādību atkarīgiem datiem modifikāciju. Šis divu izlašu tests tika aplūkots jau zinātniski - pētnieciskās prakses laikā, savukārt tagad tika izstrādāts lietojums maiņas punkta noteikšanai. Ar šo metodi tika parādīts, ka vienlaicīgi iespējams noteikt vairākus maiņas punktus, kas ir būtiska priekšrocība attiecībā pret testiem, kur katrā etapā tiek pieņemts, ka eksistē tikai viens vai nav neviens maiņas punkts.Tā kā metožu apskatam tika izveidots arī atbilstošs paketes R kods, kas pievienots darba pielikumā, tad bija iespējams veikt arī praktisku metožu pielietojumu.

Darba praktiskajā daļā aplūkotā datu analīze ilustrē priekšrocības, ko sniedz trešā aplūkotā metode. Šeit redzams, ka metode labi darbojas tieši laikrindām, kas satur daudzus maiņas punktus kā datos par smadzeņu aktivitāti. Savukārt, pārliecinājāmies, ka, ja izteikts viens maiņas punkts, tad metodes strādā līdzīgi.

Kopsavilkumā jāteic, ka aplūkotās metodes un it īpaši jaunaplūkoto metodi iespējams vēl analizēt plašāk, tādējādi paplašinot lietojumu un teorijas apskata virzienu, jo maiņas punkta analīzes iespējas un tās lietojumi arvien vēl strauji attīstās. Tāpat iespējams vēl aplūkot līdzīgu metodi kādai citai problemātikai, ne tikai lokācijas parametru maiņas punkta noteikšanai. Nobeigumā gribu teikt arī paldies prakses vadītājam par nepieciešamajām konsultācijām atvēlēto laiku un savstarpējo sadarbību.

Izmantotā literatūra un avoti

[1] Kawahara Y. and Sugiyama M. Change-point detection in time-series data by direct density-ratio estimation. In *SDM'09*, pages 389–400, 2009.

[2] Stoffer D.S Shumway R.H. *Time Series Analysis and Its Applications with R Examples 2nd edition.* Springer Science+Business Media, LLC, New York, 2006.

[3] Politis N.D. Kirch C. Tft-boostrap: Resampling time series in the frequency domain to obtain replicates in the time domain.

[4] Metcalfe A.V. Cowpertwait P.S.P. *Introductory Time Series with R.* Springer Science+Business Media, LLC, New York, 2009.

[5] Hamilton J.D. *Time series analysis.* Priceton University Press, New Jersy, 1994.

[6] Chan K.S. Cryer J.D. *Time Series Analysis with Applications in R 2nd edition.* Springer Science+Business Media, LLC, New York, 2008.

[7] Kreuztzberger E. Fan J. Automatic local smoothing for spectral density estimation. *Scandinavian Journal of Statistics*, 25:359–369, June 1998.

[8] Sperlich S. Werwatz A. Hardle W., Muller M. *Nonparametric and semiparametric models.* Springer-Verlag Berlin Heidelberg, Heidelberg, 2004.

[9] Loader C. *Local Regression and Likelihood.* Springer-Verlag New York Inc., New York, 1999.

[10] Wasserman L. *All of Nonparametric Statistics.* Springer Science + Business Media, LLC, New York, 2006.

[11] Jaruskova D. Antoch J., Huskova M. *Change point detection.* Lecture Notes of the 5th IASC Summer School, 2000.

[12] Taylor W.A. *Change-Point Analysis: A Powerful New Tool For Detecting Changes.* Taylor enterprises, Libertyville, Illinois, 2000.

[13] DasGupta A. *Probability for Statistics and Machine Learning.* Springer Science + Business Media, LLC, New York, 2011.

[14] Qi Y. Zhang R., Peng L. Jackknife - blockwise empirical likelihood methods under dependence. *Journal of Multivariate Analysis*, 104:56–72, 2012.

[15] Fried R. Robust shift detection in autoregressive models. *Proceedings of 8th international conference computer data analysis and modeling*, pages 60–67, 2007.

[16] Kitamura Y. Empirical likelihood methods with weakly dependent processes. *The Annals of Statistics*, 25:2084–2102, 1997.

[17] Cers E. Valeinis J. Extending the two-sample empirical likelihood. *preprint*, 2011.

[18] DasGupta A. *Asymptotic Theory of Statistics and Probability.* Springer Science + Business Media, LLC, New York, USA, 2008.

[19] Kitamura Y. Empirical likelihood methods in econometrics theory and practice. *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, 3, June 2006.

A PIELIKUMS.

A1. EL pārklājuma precizitātes programpaketes R kods

```
#Programma #
#Programma #
################################################################
######### SākuMĀ JAPALAIŽ EL_noC3.R ########################
################################################################




#2 izlašu tests atkarīgiem datiem - izmantojot data blocking

N<-1000 #datu apjoms
M<-trunc(N^(2/5)) #bloka garums -  window with no publikācijas
L<-M #bloku skaits, ja non-overlaping
Q<-trunc((N-M)/L)+1  #- bloku skaits

N.sim<-10000  #sim. skaits
delta0<-0 #īstā vidējo vērtību starpība
el<-0
set.seed(1)

X.data<-arima.sim(model=list(ar=0.3),n=N) #sākotnējie dati prog. pārbaudei
Y.data<-arima.sim(model=list(ar=0.3),n=N) #sākotnējie dati prog. pārbaudei
mu1=mean(X.data)
mu2=mean(Y.data)

blocking<-function(X.data, Y.data)
{
        X.block<-c()
        mu1=mean(X.data)
        for(i in 1:Q) X.block[i]<- mean(X.data[((i-1)*L+1):((i-1)*L+M)]-mu1)
        #transformētie X bloku datiti
        Y.block<-c()
        mu2=mean(Y.data)
        for(i in 1:Q) Y.block[i]<-mean(Y.data[((i-1)*L+1):((i-1)*L+M)]-mu1-delta0)
         #transformētie X bloku datiti
        (EL.means(X.block, Y.block)$conf.int[1]-delta0)*(EL.means(X.block, Y.block)$conf.int[2]-delta0)
         #Tail bound inequality
        #patiesās vidējās vērtības ir nulle un starpība delta līdz ar to arī ir nulle
 }



rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=0.3),n=N), arima.sim(model=list(ar=0.3),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte

EL.means(X.data,Y.data)
```

```
##################################################
##### bllocking - coverage accuracy ##############
##################################################

#simulāciju sākšana I
N.sim<-1000
N<-1000 #datu apjoms
M<-trunc(N^(2/5)) #bloka garums -  window with no publikācijas
L<-M #bloku sākumpunktu starpība
Q<-trunc((N-M)/L)+1  #- bloku skaits
N;M;L;Q;
theta<-0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<-0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte


#simulāciju sākšana II
N.sim<-1000
N<-1000 #datu apjoms
M<-trunc(N^(3/5)) #bloka garums -  window with no publikācijas
L<-M #bloku sākumpunktu starpība
Q<-trunc((N-M)/L)+1  #- bloku skaits
N;M;L;Q;
theta<-0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<-0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte


#simulāciju sākšana III
N.sim<-1000
N<-1000 #datu apjoms
```

```
M<-trunc(N^(1/5)) #bloka garums -  window with no publikācijas
L<-M #bloku sākumpunktu starpība
Q<-trunc((N-M)/L)+1  #- bloku skaits
N;M;L;Q;
theta<-0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<-0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte



#simulāciju sākšana IV
N.sim<-1000
N<-1000 #datu apjoms
M<-trunc(N^(2/5)) #bloka garums -  window with no publikācijas
L<-trunc(M/2) #bloku sākumpunktu starpība
Q<-trunc((N-M)/L)+1  #- bloku skaits
N;M;L;Q;
theta<-0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<-0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte


#simulāciju sākšana V
N.sim<-1000
N<-1000 #datu apjoms
M<-trunc(N^(3/5)) #bloka garums -  window with no publikācijas
L<-trunc(M/2) #bloku sākumpunktu starpība
Q<-trunc((N-M)/L)+1  #- bloku skaits
N;M;L;Q;
theta<-0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.3
```

```
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<-0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte


#simulāciju sākšana VI
N.sim<-1000
N<-1000 #datu apjoms
M<-trunc(N^(1/5)) #bloka garums -  window with no publikācijas
L<-trunc(M/2) #bloku sākumpunktu starpība
Q<-trunc((N-M)/L)+1  #- bloku skaits
N;M;L;Q;
theta<-0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<-0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte


######################################################################
### samazinam datu apjomu izlasēm N=100 ##############################################
######################################################################
N<-100

#simulāciju sākšana I
N.sim<-1000
M<-trunc(N^(2/5)) #bloka garums -  window with no publikācijas
L<-M #bloku sākumpunktu starpība
Q<-trunc((N-M)/L)+1  #- bloku skaits
N;M;L;Q;
theta<-0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<-0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
```

```
theta<- -0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte


#simulāciju sākšana II
N.sim<-1000
M<-trunc(N^(3/5)) #bloka garums -  window with no publikācijas
L<-M #bloku sākumpunktu starpība
Q<-trunc((N-M)/L)+1  #- bloku skaits
N;M;L;Q;
theta<-0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<-0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte


#simulāciju sākšana III
N.sim<-1000
M<-trunc(N^(1/5)) #bloka garums -  window with no publikācijas
L<-M #bloku sākumpunktu starpība
Q<-trunc((N-M)/L)+1  #- bloku skaits
N;M;L;Q;
theta<-0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<-0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte



#simulāciju sākšana IV
N.sim<-1000
M<-trunc(N^(2/5)) #bloka garums -  window with no publikācijas
L<-trunc(M/2) #bloku sākumpunktu starpība
Q<-trunc((N-M)/L)+1  #- bloku skaits
```

```
N;M;L;Q;
theta<-0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<-0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte


#simulāciju sākšana V
N.sim<-1000
M<-trunc(N^(3/5)) #bloka garums -  window with no publikācijas
L<-trunc(M/2) #bloku sākumpunktu starpība
Q<-trunc((N-M)/L)+1  #- bloku skaits
N;M;L;Q;
theta<-0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<-0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte


#simulāciju sākšana VI
N.sim<-1000
M<-trunc(N^(1/5)) #bloka garums -  window with no publikācijas
L<-trunc(M/2) #bloku sākumpunktu starpība
Q<-trunc((N-M)/L)+1  #- bloku skaits
N;M;L;Q;
theta<-0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.3
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<-0.9
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.9
```

```
rez<-replicate(N.sim,blocking(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte



##############t-tests 2 izlasēm  ############################
N.sim<-1000
N<-1000
Ttest<-function(X.data, Y.data)
{
       (t.test(X.data, Y.data)$conf.int[1])*(t.test(X.data, Y.data)$conf.int[2])
         #Tail bound inequality
        #patiesās vidējās vērtības ir nulle un starpība delta līdz ar to arī ir nulle
 }
theta<- 0.3
rez<-replicate(N.sim,Ttest(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.3
rez<-replicate(N.sim,Ttest(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- 0.9
rez<-replicate(N.sim,Ttest(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.9
rez<-replicate(N.sim,Ttest(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte

####n=100
N<-100
theta<- 0.3
rez<-replicate(N.sim,Ttest(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.3
rez<-replicate(N.sim,Ttest(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- 0.9
rez<-replicate(N.sim,Ttest(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
theta<- -0.9
rez<-replicate(N.sim,Ttest(arima.sim(model=list(ar=theta),n=N), arima.sim(model=list(ar=theta),n=N) ))
 length(rez[rez<0])/N.sim # pārklājuma precizitāte
```

A2. TFT metodes programpaketes R kods

```
###change point TFT
n<-1000
#
dati<-read.table(file="sunspot.txt",header=F)
n<-length(dati$V2)
simulacija<-function(){
theta<-0.5
N<-1000
xx<-arima.sim(n = N, list(ar = c(theta), ma=0 ))
mu1<-3
mu2<-15
spec.pgram(xx,log="no")
h.nulles<-c(xx[1:500]+mu1,xx[501:1000]+mu2)

h.altern<-xx+mu1

Y<-h.nulles
#Y<-h.altern
#plot(1:1000,Y)
###saakuma stat.
Y_k_vid<-c()
S<-0
for (i in 1:n){
S<-S+Y[i]
Y_k_vid[i]<-S-i*mean(Y)
}



C.T.vect<-((sqrt(n))^(-1))*(abs(Y_k_vid))
C.T<-max(C.T.vect)

##ar z-tiem

k.arg.max<-which(abs(Y_k_vid) ==max( abs(Y_k_vid) ))
k<-k.arg.max
Z<-c()
Z<-c(Y[1:k]-mean(Y[1:k]),Y[(k+1):n]-mean(Y[(k+1):n]))
Z.vid<-0
Z.vid<-abs((sqrt(n))^(-1)*(sum(Z[1:k])))



C.T.vect<-Z.vid
C.T<-max(C.T.vect)
C.T
return(list(stat=C.T))
}
####simulacijas
simulacijas.stat<-c()
```

```
simulacijas.stat<-replicate(n,simulacija()$stat)

plot(density(simulacijas.stat))
hist(simulacijas.stat, prob=T)
lines(density(simulacijas.stat),col="green")

###laikrindas grafiks
plot(dati$V1,dati$V2,type="lty")
acf(dati$V2)
###iepr.rezult.
#alter<-simulacijas.stat
lines(density(alter),col="red")
##########################################################################
###JAASAAK bootstrapot!!!

#dati
#theta<-0.5
N<-length(dati$V2)
# xx<-arima.sim(n = N, list(ar = c(theta), ma=0 ))
# mu1<-3
# mu2<-15
# spec.pgram(xx,log="no")
# h.nulles<-c(xx[1:500]+mu1,xx[501:1000]+mu2)

# h.altern<-xx+mu1

Y<-dati$V2
#Y<-h.altern

Y_k_vid<-c()
S<-0
for (i in 1:n){
S<-S+Y[i]
Y_k_vid[i]<-S-i*mean(Y)
}
k.arg.max<-which(abs(Y_k_vid)==max( abs(Y_k_vid) ))
k<-k.arg.max
Z<-c()
Z<-c(Y[1:k]-mean(Y[1:k]),Y[(k+1):n]-mean(Y[(k+1):n]))
Z.vid<-0
Z.vid<-abs((sqrt(n))^(-1)*(sum(Z[1:k])))
C.T.vect<-Z.vid
C.T<-max(C.T.vect)
C.T




xx<-Y
spektrs<-spec.pgram(Y,taper=0,kernel("daniell", c(30,30)),
```

```
 fast=FALSE, detrend=FALSE, log="no",
main=expression("Raw perodogramm, N=100"))


###y<-fft(x[,i])/(sqrt(N))
##bootstr.paraugs
#n<-500
#izl<-rnorm(n,5,1)
#B<-10000
#mean.boot<-replicate(B,mean(sample(izl,replace=TRUE)))
#Z<- sqrt(n)*(mean.boot-mean(izl))/sqrt(var(izl))
#hist(Z,prob=TRUE, main="3",ylim=c(0,0.5))
##xx<-seq(-5,5,len=1000)
#points(xx,dnorm(xx,0,1),type="l")

###turpinajums
gludinasana<-function(x,ht,freq,spec)
{
sum(dnorm((x-freq)/ht)*spec)/sum(dnorm(freq/ht))

}

frkv<-spec.pgram(Y,taper=0,fast=FALSE, detrend=FALSE, log="no",
main=expression("Raw perodogramm, N=100"))$freq
spektr<-spec.pgram(Y,taper=0,fast=FALSE, detrend=FALSE, log="no",
main=expression("Raw perodogramm, N=100"))$spec
#izmeginasana
gludinasana(0.4,0.01,frkv,spektr)
#vektorizesana
Gludinasana<-Vectorize(gludinasana)

x.j<-Re(fft(xx)/sqrt(n))
y.j<-Im(fft(xx)/sqrt(n))
##apakseja dalja
NN<-floor((n-1)/2)

x.j.NN<-x.j[1:NN]
y.j.NN<-y.j[1:NN]

lambda.j<-(1:(NN+1))/n
s.j<-c()
s.j[(N/2)]<-0
s.j[N]<-0
glud<-c()
glud[(N/2)]<-0
for (i in 1:NN){

s.j[i]<-x.j.NN[i]/sqrt(2*gludinasana(lambda.j[i],0.01,frkv,spektr))
s.j[i+NN+1]<-y.j.NN[i]/sqrt(2*gludinasana(lambda.j[i],0.01,frkv,spektr))
glud[i]<-gludinasana(lambda.j[i],0.01,frkv,spektr)
```

```
}
plot(lambda.j,glud)
lines(frkv,spektr)

##Normēšana
ss.j<-c()
ss.j<-(s.j-(1/2*NN)*sum(s.j) ) / (sqrt( (1/2*NN)*sum((s.j-(1/2*NN)*sum(s.j))^2)  ))
NNN<-length(ss.j)
#izlase<-ss.j
izlase<-rnorm(N)
B<-1000
reverss.furje<-function(dati1,dati2)
{
x.star<-sqrt(glud/2)*sample(dati1,replace=TRUE)
y.star<-sqrt(glud/2)*sample(dati2,replace=TRUE)
return(list(m=x.star,k=y.star))
}
atlikusie.koef<-function(x.star,y.star,k)
{
x.zvaigznite.pilns<-c(x.star,x.star[((ceiling((n-1)/2))-1):1])
y.zvaigznite.pilns<-c(y.star,-y.star[((ceiling((n-1)/2))-1):1])
skaitlis.pilns<-x.zvaigznite.pilns+y.zvaigznite.pilns*1i
skaitlis.pilns[(N/2)]<-0
skaitlis.pilns[N]<-0
Z.laikrinda<-fft(skaitlis.pilns, inverse = TRUE)/sqrt(n)
Y<-abs(Z.laikrinda)
Y_k_vid<-c()
S<-0
for (i in 1:n){
S<-S+Y[i]
Y_k_vid[i]<-S-i*mean(Y)
}
k.arg.max<-which(abs(Y_k_vid)==max( abs(Y_k_vid) ))
k<-k.arg.max
Z<-c()
Z<-c(Y[1:k]-mean(Y[1:k]),Y[(k+1):n]-mean(Y[(k+1):n]))
Z.vid<-0
Z.vid<-abs((sqrt(n))^(-1)*(sum(Z[1:k])))


C.T.vect<-Z.vid
C.T<-max(C.T.vect)
C.T
return(list(stat.vertiba=C.T))
}

###
#plot((1:1000),Z,type="lty",main="dati")
#plot((1:1000),h.altern,type="lty")
#str<-Z
#TFT<-str
```

```
#acf(Z)
###
mean(Z+mean(h.altern))
mean(h.altern)


atlikusie.koef(reverss.furje(izlase[1:(NN+1)],izlase[(NN+2):NNN])$m,
reverss.furje(izlase[1:(NN+1)],izlase[(NN+2):NNN])$k,k)$stat.vertiba
gg<-c()
hh<-c()
#ss.boot<-replicate(B,reverss.furje(izlase[1:NN+1],izlase[(NN+2):NNN]))

##gala stat. sadalijums
B<-10000
ss.boot<-replicate(B,atlikusie.koef(reverss.furje(izlase[1:(NN+1)],izlase[(NN+2):NNN])$m,
reverss.furje(izlase[1:(NN+1)],izlase[(NN+2):NNN])$k,k)$stat.vertiba)

plot(density(ss.boot))

###
saglab<-ss.boot
lines(density(saglab),col="red")
abline(v=C.T)
####
x.star<-reverss.furje(izlase[1:(NN+1)],izlase[(NN+2):NNN])$m
y.star<-reverss.furje(izlase[1:(NN+1)],izlase[(NN+2):NNN])$k
###
plot(dati$V1,dati$V2)
lines(dati$V1[1:236],dati$V2[1:236],type="lty",col="red")

par(mfrow=c(1,2))
```

A3. EL maiņas punkta programpaketes R kods

```
#############################################################
#########sunspot +EL_noC3.R ##########################
#############################################################
dati<-read.table(file="sunspot.txt",header=F)
NN<-length(dati$V2); NN; #sunspot datu garums
delta0<-0 #īstā vidējo vērtību starpība
el<-0
sun.dati<-dati$V2

#dati
par(mfrow = c(1,3))
plot(1:NN,sun.dati, type='lty')



#2 izlašu tests atkarīgiem datiem - izmantojot data blocking
for (k in 2:3){
N<-k*10; N; #datu apjoms logos
#N<-30 20 10
M<-trunc(N^(3/5));M; #bloka garums -  window with no publikācijas
L<-M; L; #bloku skaits, ja non-overlaping
Q<-trunc((N-M)/L)+1; Q; #- bloku skaits

#mu1=mean(X.data)
#mu2=mean(Y.data)


blocking<-function(X.data, Y.data)
{
        X.block<-c()
        mu1=mean(X.data)
        for(i in 1:Q) X.block[i]<- mean(X.data[((i-1)*L+1):((i-1)*L+M)]-mu1)
        #transformētie X bloku datiti
        Y.block<-c()
        mu2=mean(Y.data)
        for(i in 1:Q) Y.block[i]<-mean(Y.data[((i-1)*L+1):((i-1)*L+M)]-mu1-delta0)
         #transformētie X bloku datiti
        #(EL.means(X.block, Y.block)$conf.int[1]-delta0)*(EL.means(X.block, Y.block)$conf.int[2]-delta0)
         #Tail bound inequality
        #patiesās vidējās vērtības ir nulle un starpība delta līdz ar to arī ir nulle
  #p-value grafikam
  EL.means(X.data,Y.data)$p.value
 }
pp.values<-c()
for (i in 1:(NN-2*N)){
F1.block<-sun.dati[i:(i+N-1)];
F2.block<-sun.dati[(i+N):(i+2*N-1)];
#mean(F1.block);mean(F2.block);
```

```
pp.values[i]<-blocking(F1.block,F2.block);
}

plot((N+1):(length(pp.values)+N),pp.values, type='lty')
pp=round(pp.values,2);pp


}


#length(p.values)

#####################################################
## Piemeeriem
#########################################
#################################################################

NN<-1000; NN; #ārējo datu garums
#N.sim<-1000  #sim. skaits
delta0<-0 #īstā vidējo vērtību starpība
el<-0
#set.seed(6)

#dati
X.data<-arima.sim(model=list(ar=0.3),n=NN/2) #sākotnējie dati prog. pārbaudei
Y.data<-5+arima.sim(model=list(ar=0.3),n=NN/2) #sākotnējie dati prog. pārbaudei
XY.data<-c(X.data, Y.data)
par(mfrow = c(3,2))
plot(1:NN,XY.data, type='lty')



#2 izlašu tests atkarīgiem datiem - izmantojot data blocking

N<-10; N; #datu apjoms logos
#N<-150 100 50 30 10
M<-trunc(N^(3/5));M; #bloka garums -  window with no publikācijas
L<-M; L; #bloku skaits, ja non-overlaping
Q<-trunc((N-M)/L)+1; Q; #- bloku skaits

#mu1=mean(X.data)
#mu2=mean(Y.data)


blocking<-function(X.data, Y.data)
{
        X.block<-c()
        mu1=mean(X.data)
        for(i in 1:Q) X.block[i]<- mean(X.data[((i-1)*L+1):((i-1)*L+M)]-mu1)
        #transformētie X bloku datiti
        Y.block<-c()
```

```
        mu2=mean(Y.data)
        for(i in 1:Q) Y.block[i]<-mean(Y.data[((i-1)*L+1):((i-1)*L+M)]-mu1-delta0)
         #transformētie X bloku datiti
        #(EL.means(X.block, Y.block)$conf.int[1]-delta0)*(EL.means(X.block, Y.block)$conf.int[2]-delta0)
         #Tail bound inequality
        #patiesās vidējās vērtības ir nulle un starpība delta līdz ar to arī ir nulle
  #p-value grafikam
  EL.means(X.data,Y.data)$p.value
 }
p.values<-c()

# parbaudei:   xx<-blocking(X.data,Y.data)

#WW<-(trunc(NN/N));WW  # logu skaits
#for (i in 1:(WW-1)){
#F1.block<-XY.data[(1+(i-1)*N):(N+(i-1)*N)];
#F2.block<-XY.data[(1+(i)*N):(N+(i)*N)];
#mean(F1.block);mean(F2.block);
#p.values[i]<-blocking(F1.block,F2.block);
#}

#plot(1:(WW-1),p.values, type='lty')
#pp=round(p.values,2);pp

#acf(XY.data[1:500])

WW<-(trunc(NN/N));WW  # pilnu logu skaits, kas ietilpst vienā (ārējais cikls)
for (i in 1:(WW-1)){
for (j in 1:(N-1)){
F1.block<-XY.data[(j+(i-1)*N):(N+(j-1)+(i-1)*N)];
F2.block<-XY.data[(j+(i)*N):(N+(j-1)+(i)*N)];
mean(F1.block);mean(F2.block);
p.values[(i-1)*(N-1)+j]<-blocking(F1.block,F2.block);
}
}

plot(1:length(p.values),p.values, type='lty')
pp=round(p.values,2);pp
```

Diplomdarbs “Maiņas punkta noteikšana laikrindu analīzē” 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: Agris Vaselāns

_____ _____
(paraksts) (datums)

Rekomendēju darbu aizstāvēšanai.
Vadītājs: doc. Dr. math. Jānis Valeinis

_____ _____
(paraksts) (datums)

Recenzente: doc. Dr. math. Nadežda Siņenko

Darbs iesniegts Matemātikas nodaļā _____
(datums)

_____
(darbu pieņēma)

Darbs aizstāvēts valsts pārbaudījuma komisijas sēdē

_____ prot. Nr. _____, vērtējums_____
(datums)

Komisijas sekretāre: doc. Dr. math. Ingrīda Uļjane _____
(paraksts)