LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

MAIŅAS PUNKTA NOTEIKŠANA MATEMĀTISKĀS STATISTIKAS PROBLĒMĀS

DIPLOMDARBS

Autors: **Vineta Vītola**

Stud. apl. vv08073

Darba vadītājs: doc. Dr. math. Jānis Valeinis

RĪGA 2013

ANOTĀCIJA

Laika momentu, kad modelis ir mainījies, sauc par maiņas punktu. Darba mērķis ir apskatīt maiņas punkta noteikšanu matemātiskās statistikas problēmās ar programmā *R* iebūvētajām paketēm. Tiek apskatītas paketes *changepoint* un *strucchange* maiņas punktu noteikšanai. Tiks apskatītas dažas galvenās metodes, kuras ir iebūvētas paketēs un tiks parādīta to pielietošana gan uz simulētiem datiem, gan arī uz reāliem datu piemēriem.

Atslēgvārdi: Maiņas punkts, CUSUM statistika, F statistika, PELT, Binārā segmentācija, Segmenta kaimiņu metode.

ABSTRACT

A time moment when the model has changed is called a change point. The goal of this work is to analyze change point detection in problems of the mathematical statistics using different packages in program *R*. More specifically we will analyze packages *changepoint* and *strucchange*. We will discuss some of the key methods which are implemented in packages and highligh their applications to simulated and real data examples.

Keywords: change point, CUSUM statistic, F statistic, PELT, binary segmentation, Segment Neighbourhoods

Saturs

APZĪMĒJUMI

OLS - mazākā kvadrātu metode

$\longrightarrow_d$ - konverģence pēc sadalījuma

$=_{asy}$ - asimptotiskā vienādība

CSS (angliski *Cumulative Sums of Squares*) - kvadrātu kumulatīvās summas

CUSUM - kumulatīvo summu metode

AMOC (angliski *At Most One Change*) - AMOC (visma viena izmaiņa) metode

PELT (angliski *Pruned Exact Linear Time*) - saīsinātais precīza lineārā laika

IEVADS

Maiņas punkta modeļi sākotnēji tika izstrādāti saistībā ar kvalitātes kontroli, kur izmaiņa no atbilstoša uz neatbilstošu tiek noteikta balstoties uz pieejamiem gadījuma novērojumiem. Page (1955)[1] izstrādāja kumulatīvo summu diagrammu (CUSUM), kas ir sekojošu datu analīzes tehnika. To parasti izmanto izmaiņu noteikšanas kontrolē. Page (1955) ar ''kvalitātes numuru'' θ apzīmēja parametru varbūtības sadalījumam, piemēram, vidējai vērtībai. Viņa izstrādāja CUSUM kā metodi, lai noteiktu izmaiņas parametram, un ierosināja kritēriju, lai izlemtu, kad veikt korektīvus pasākumus.

Kopš tā laika maiņas punkta problēma ir izveidojusies par fundamentālu problēmu statistiskās vadības teorijā, laikrindas stāvokļa novērtēšanā, regresijas modeļa izmaiņas pārbaudē un novērtēšanā, un pavisam nesen DNS secības saskaņošanā un salīdzināšanā mikrovektoru datu analīzē.

Statistiskiem modeļiem vispirms tiek pārbaudīts, vai tiem ir notikušas izmaiņas. Nulles hipotēzē pieņem, ka modelis ir tāds pats visā novērošanas laikā. Alternatīvā hipotēze pieņem, ka kādā laika momentā modelis mainās. Gadījumā, kad tiek noraidīta nulles hipotēze, tad var pārbaudīt kādu no jautājumiem:

- kurā vietā modelī notikušas izmaiņas;
- ir tikai viena izmaiņa, vai vairākas;
- kāds ir kopējais izmaiņu skaits, utt.

Laika momentu, kad modelis ir mainījies, sauc par maiņas punktu.

Daudzas metodoloģiskās pieejas tiek izmantotas maiņas punktu modeļu pārbaudē; kā piemēram, maksimālās ticamības novērtējums, Beijesa novērtējums, izotoniskā regresija un kvazi - ticamības funkcija ir starp metodēm, kuras tiek piemērotas, lai atrisinātu maiņas punktu problēmas.

Lielākajā daļā zinātniskos pētījumos pieņem, ka izmaiņas parametru modeļos rodas reti un ir pēkšņas. Eksistē daudzi testi izmaiņu klātbūtnes noteikšanai, sākot ar Chow (1960) darbu [2], kurā pieņēma, ka laika moments, kad notikušas izmaiņas, ir zināms. Citi testi, kas mazina šo pieņēmumu ir izstrādāti Brown, Durbin un Evans (1974)[3], Ploberger un Kramer (1992) [4] un citos pētījumos.

Maiņas punkta noteikšana vienmēr attīstās mijiedarbojoties ar dažādām jomām. Piemēram, klimata datu kopās maiņas punkts ir laiks, kurā klimats mainās dramatiski; fi-

nanšu ekonometrijā maiņas punkta analīze palīdz noteikt tirgus vai ekonomikas virzienu; inženierzinātnēs, ir svarīgi uzzināt nepārtrauktam ražošanas procesam, vai ir punkts, kurā kvalitāte pasliktinās. Jaunākie piemēri ir bioinformātiskos pielietojumos (Erdman and Emerson, 2008)[5], ļaunprogrammatūru noteikšana programmatūras ietvaros (Yan et al., 2008)[6], tīkla satiksmes analīze (Kwon et al., 2006) [7], klimataloģijā (Jaxk et al., 2007) [8] un okeanogrāfijā (Killick et al, 2010) [9].

Maiņas punkta noteikšana dispersijai ir svarīga finanšu datiem. Kā piemēram, svārstīgumam, kas ir statistikas termins, kas parāda sagaidāmās attiecīgā pamataktīva cenas svārstības. To izmanto finanšu jomā, lai novērtētu riska līmeni konkrētā finanšu instrumentā. Augsts svārstīgums nozīmē, ka vērtspapīra cena var ļoti mainīties īsā laika posmā abos virzienos. Noteikti indeksi kontrolē attiecīgos svārstīgumus, piemēram, FTSE 100 ir akciju indekss no 100 kompānijām ar vislielāko tirgus kapitalizāciju. Tas ir viens no visplašāk izmantotajiem akciju indeksiem un tiek uzskatīts par augstas konjunktūras mērinstrumentu. Augsta konjunktūra ceļ uzņēmuma un tā aktīvu vērtību tirgū. Darbā tiks aplūkota dispersija maiņas uz simulētiem datiem, kā arī uz iepriekš pieminētā FTSE 100 indeksa.

Par diplomdarba mērķi tiek izvirzīts:

- apskatīt paketes programmā R, kuras risina maiņas punktu problēmas,
- aprakstīt teoriju, uz kuru balstās iepriekš apskatītās paketes,
- pielietot tās dažādiem datu piemēriem.

Darbs sastāv no divām nodaļām. Pirmajā nodaļā aprakstīta maiņas punkta problemātika regresijai. Otrajā - laikrindām. Trešajā nodaļā ir aprakstīta vispārējā maiņas punkta teorija. Ceturtajā nodaļā tiek praktiski aplūkoti iepriekšējās nodaļās aprakstītās problēmas, izmantojot R programmā iebūvētās paketes *changepoint* un *strucchange*.

1. MAIŅAS PUNKTA NOTEIKŠANA REGRESIJĀ

Pamatproblēma maiņas punkta noteikšanā regresijā attiecas uz lēmumu, vai mainījušās attiecības starp mainīgajiem novērošanas laikā. Vienkāršākais gadījums, ko var aplūkot ir vienkāršā lineārā regresija, kur var aplūkot daudzas dažādas situācijas, piemēram, [10]:

- vai nu viens, vai abi parametri ir mainījušies;
- vai nu parametri pirms maiņas punkta ir zināmi, vai arī tie nav zināmi;
- vai nu regresijas funkcijai maiņas punktā tiek pieņemta nepārtrauktība, vai arī var būt pārtraukums, utt.

Testus maiņas punktu noteikšanai var iedalīt divās klasēs, kas ir atšķirīgi piemēroti konkrēta modeļa testēšanai. Pirmā klase ir testi no vispārināto svārstību testa rāmja, kas var atklāt dažādas strukturālas izmaiņas. Otra klase ir testi no F testa rāmja, kuri pieņem, ka ir viens pārtraukumpunkts pie alternatīvas.

Vispārējie svārstību testi piemēro parametrisku modeli datiem, izmantojot mazāko kvadrātu (OLS) metodi vai līdzvērtīgi maksimālo ticamības varbūtību (ML). Tās izmanto parasto tuvināšanu un iegūst procesu, kas atspoguļo svārstības rekursīvajiem vai OLS atlikumiem, vai rekursīvajiem vai kustīgajiem novērtējumiem un noraida, ja to svārstības ir neticami lielas.

Tiek aplūkots standarta lineārās regresijas modelis

$$y_i = x_i^\top \beta_i + u_i \quad (i = 1, ..., n), \tag{1.1}$$

kur laikā i, y_i ir atkarīgā mainīgā novērojums, x_i ir $k \times 1$ regresoru vektors parasti ar pirmo komponenti vienādu ar viens, un β_i ir $k \times 1$ regresijas koeficienti, kuri var mainīties laika gaitā.

Tiek pārbaudīta hipotēze, ka regresijas koeficienti paliek nemainīgi

$$H_0 : \beta_i = \beta_0 \quad (i = 1, ..., n) \tag{1.2}$$

pret alternatīvu, ka vismaz viens koeficients mainās laika gaitā. Ir pamatoti pieņemt, ka ir m pārtraukumpunkti, kur koeficienti pāriet no vienas stabilas regresijas attiecības uz

citu. Tātad ir $m+1$ segmenti, kuros regresijas koeficienti ir konstanti, un modeli 1.1 var uzrakstīt kā

$$y_i = x_i^\top \beta_j + u_i \quad (i = i_{j-1}+1, .., i_j, \quad j = 1, ..., m+1), \tag{1.3}$$

kur j ir segmenta indekss, $I_{m,n} = i_1, ..., i_m$ apzīmē pārtraukumpunktu kopas ($I_{m,n}$ sauc arī par m - sadalīšanu), un pēc vienošanās $i_0 = 0$ un $i_{m+1} = n$ [11].

Izmaiņu noteikšana

Praksē pārtraukumpunkti reti ir doti, lielākoties tie ir nezināmi un ir jānovērtē no datiem.

Pieņemsim, ka doti m segmenti $i_1, ..., i_m$, β_j mazāko kvadrātu novērtējumus viegli var iegūt. Atlikumu kvadrātu summas ir dotas ar

$$RSS(i_1, ..., i_m) = \sum_{j=1}^{m+1} rss(i_{j-1}+1, i_j), \tag{1.4}$$

kur $rss(i_{j-1}+1, i_j)$ ir parastā minimālā atlikumu kvadrātu summa segmentā j. Problēma strukturālo izmaiņu noteikšanā ir atrast pārtraukumpunktus $\hat{i}_1, ..., \hat{i}_m$, kuri minimizē mērķa funkciju

$$(\hat{i}_1, ..., \hat{i}_m) = argmin_{(i_1,...,i_m)} RSS(i_1, ..., i_m) \tag{1.5}$$

pa visiem segmentiem $(i_1, ..., i_m)$ ar $i_j - i_{j-1} \geq n_h \geq k$.

Globālo minimizētāju noteikšanai 1.5 ar plašo režģi būtu $O(n^m)$ kārta un skaitļošana būtu apgrūtinoša, ja $m > 2$ (un katram saprātīgam izlases lielumam n). Tāpēc daudzi hierarhiskie algoritmi tika piedāvāti, kas veic rekursīvo sadalīšanu vai apakšizlašu pievienošanu, bet tie ne vienmēr atradīs globālos minimizētājus. Tos var vieglāk atrast ar dinamiskās programmēšanas pieeju, kam ir $O(n^2)$ kārta jebkāda daudzuma pārtraukumpunktiem m. Pamatideja ir no Bellmana principa: optimālā segmentācija apmierina rekursiju:

$$RSS(I_{m,n}) = \min_{mn_k \leq i \leq n-n_k} [RSS(I_{m-1,i}) + rss(i+1, n)]. \tag{1.6}$$

Tāpēc pietiek zināt katram punktam i ''optimālo iepriekšējo partneri", ja i bija pēdējais pārtraukumpunkts m segmentā. To var iegūt no $rss(i,j)$ trijstūra matricas ar $j - i \geq n_h$ aprēķināšanu, kuru atvieglo rekursīvā attiecība $rss(i,j) = rss(i,j-1) + rss(i,j)^2$, kur $r(i,j)$ ir rekursīvie atlikumi laikā j, izlasi sākot no i (Brown 1975) [3]. Par dinamiskās programmēšanas algoritmu vairāk skatīties Bai un Perron (2003) darbā [12].

1.1. Vispārīgais svārstību tests

Svārstību testi ir balstīti uz novērtējumiem, vai atlikumiem. Ideja testam balstītam uz novērtējumu ir, ka, ja ir izmaiņas datos, regresijas koeficientu novērtējumam balstītam uz visiem datiem būtu ievērojami jāatšķiras no novērtējumiem no datu apakšizlases, kura nesatur izmaiņas. Bet šiem novērtējumiem būtu jābūt diez gan līdzīgiem, ja īstie koeficienti paliek konstanti laika gaitā. Tāpēc šajā gadījumā empīrisku procesu var aprēķināt ar apakšizlases novērtējumu atšķirībām no kopējā novērtējuma. Apakšizlasi izvēlas rekursīvi, t. i., sākot ar pirmajiem k novērojumiem un iekļauj soli pa solim nākošo novērojumu, vai arī ar logu ar konstantu platumu, kas kustās pāri visam izlases periodam. Izrietošajiem procesiem nevajadzētu svārstīties (novirzīties no nulles) pārāk daudz pie nulles hipotēzes un tā, kā asimptotiskie sadalījumi šiem procesiem ir labi zināmi, robežas var aprēķināt, tās šķērso tikai ar noteiktu kontrolētu varbūtību α. Bet, ja empīriskais process uzrāda lielas svārstības un šķērso robežu, ir pierādījumi, ka dati satur izmaiņas. Tādā gadījumā rekursīvajam novērtējuma procesam vajadzētu būt pīķim (maksimumam) ap maiņas punktu.

Līdzīgi svārstību procesus var aprēķināt pamatojoties uz kumulatīvajām vai kustīgajām summām divu veidu atlikumiem: parastajiem OLS atlikumiem vai rekursīvajiem atlikumiem, kas ir viena soļa uz priekšu prognozes kļūdas. Testu balstītu uz kumulatīvo summu rekursīvajiem atlikumiem (CUSUM tests) pirmo reizi ieviesa Brown et al. (1975)[3]. Ja ir tikai viens strukturālais pārtraukums koeficientos ceļš sāks atstāt nulles vidējo vērtību pie maiņas punkta, jo viena soļa uz priekšu prognozes kļūda būs liela [13].

CUSUM testi

Standarta CUSUM tests ir balstīts uz kopējo summu rekursīvajiem atlikumiem

$$\tilde{u}_t = \frac{y_t - x_t^\top \hat{\beta}^{(t-1)}}{\sqrt{1 + x_t^\top (X^{(t-1)\top} X^{(t-1)})^{-1} x_t}} \quad (t = k+1, .., n), \tag{1.7}$$

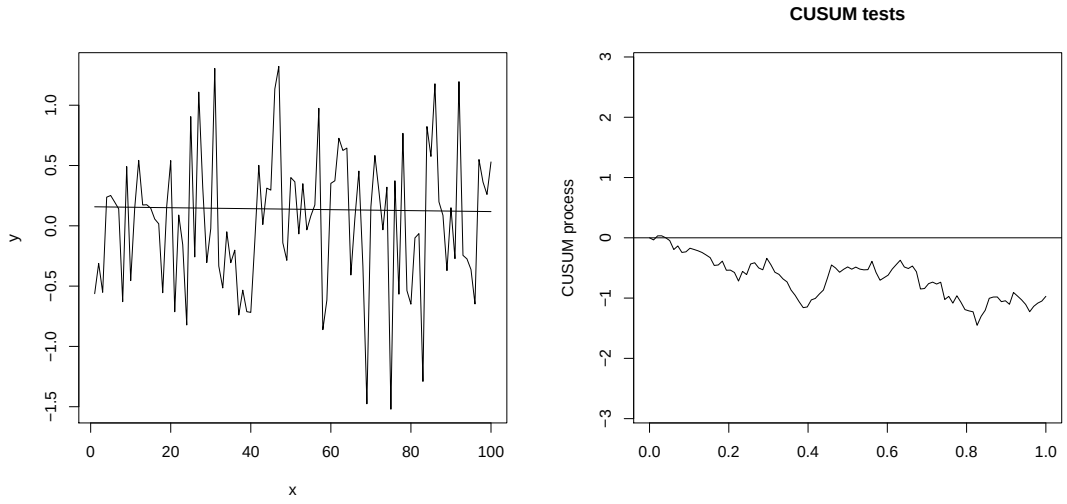
kuriem vidējā vērtība ir nulle un dispersija σ^2 pie H_0. $\hat{\beta}^{(t-1)}$ ir parastais mazāko kvadrātu novērtējums no regresijas koeficientiem balstītiem uz novērojumiem līdz $t-1$. Līdzīgi $X^{(t-1)}$ ir regresoru matrica visiem novērojumiem līdz $t-1$.

CUSUM ceļš ir definēts kā

$$W_n(t) = \frac{1}{\tilde{\sigma}\sqrt{n-k}} \sum_{i=k+1}^{\lfloor k+t(n-k) \rfloor} \tilde{u}_i \quad (0 \le t \le 1), \tag{1.8}$$

kur $\tilde{\sigma} = \sqrt{\frac{1}{n-k}\sum_{t=k+1}^{n}(\tilde{u}_t - \bar{\tilde{u}})^2}$. Mainīgā t nozīme mainās mazliet, tas ir standartizēts intervālā $[0, 1]$.

CUSUM statistika simulētiem datiem 1.(a). bez maiņas punkta attēlota 1.(b). grafikā.

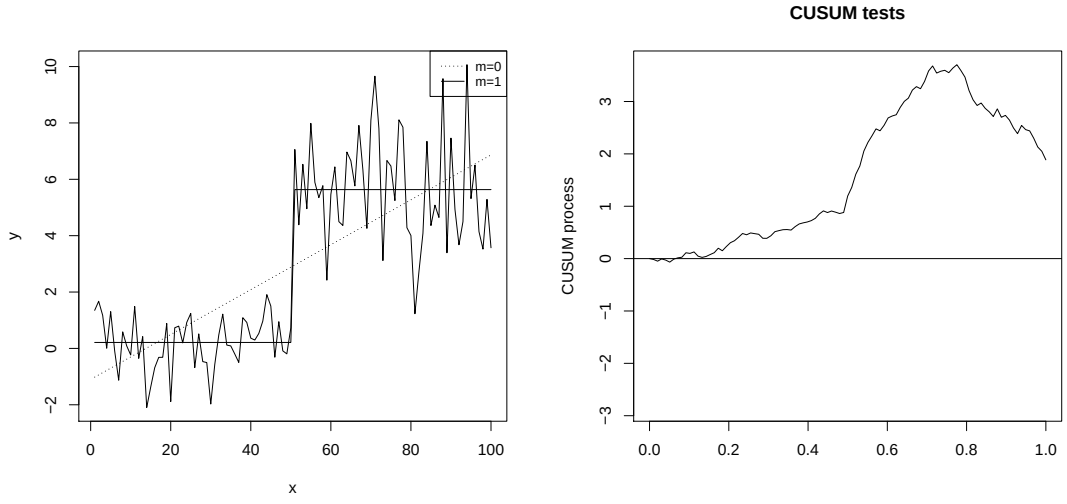


(a) Simulētie dati bez maiņas punkta ar regresijas līniju

(b) CUSUM

1. att. CUSUM statistikas uzvedība simulētiem datiem bez maiņas punktiem

CUSUM statistiku simulētiem datiem 2.(b). grafikā ar vienu maiņas punktu var apskatīt 2.(b). grafikā.



(a) Simulētie dati ar regresijas līnijām

(b) CUSUM

2. att. CUSUM statistikas uzvedība simulētiem datiem

Salīdzinot 1.(b) un 2.(b) grafikus, var redzēt, ka 2.(b) grafikā CUSUM statistika sāk atstāt nulles vidējo vērtību, kas norāda uz maiņas punktu.

Ja ir tikai viena izmaiņa fiksētā laikā $t_0 < 1$, tad rekursīvajiem atlikumiem vidējā vērtība būs nulle tikai līdz t_0 un atšķirīga pēc tam. Tātad CUSUM ceļš $W_n(t)$ sāks atstāt vidējo vērtību nulle pie t_0. H_0 tiek noraidīts tad, kad $W_n(t)$ šķērso vai nu $c(t)$, vai arī $-c(t)$ ar $c(t) = \lambda + 2\lambda t$, kurš ir līdzvērtīgs noraidīt nulles hipotēzi, kad testa statistika

$$S = \sup_{0 \le t \le 1} \left| \frac{W_n(t)}{1+2t} \right| \tag{1.9}$$

ir lielāka par λ, kas ir atkarīga no nozīmības līmeņa testam.

Kramer, Ploberger, Alt (1988) [14] parādīja, ka $n \to \infty$

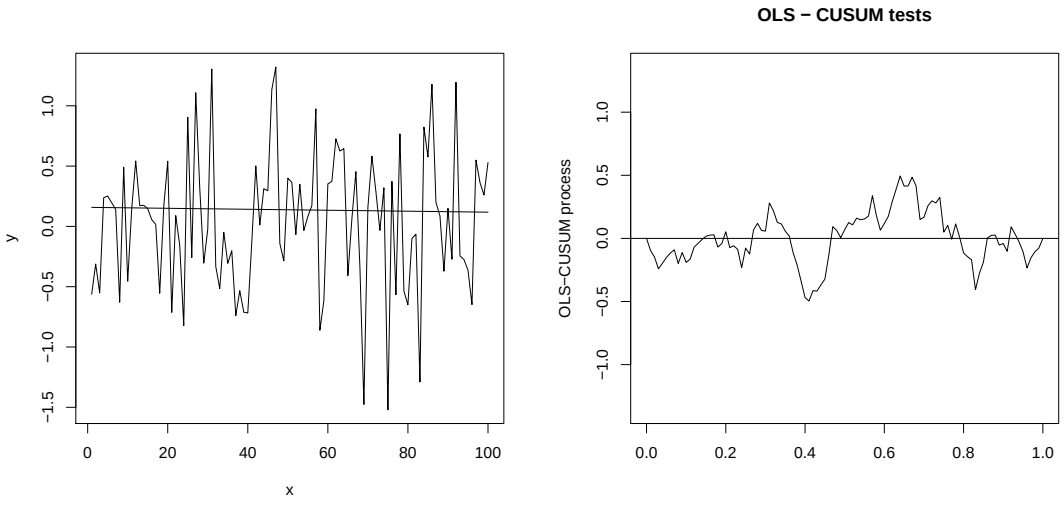
$$W_n(t) \longrightarrow^d B(t), \tag{1.10}$$

kur $\longrightarrow^d$ apzīmē konverģenci pēc sadalījuma un kur $B(t)$ ir Brauna kustība.

CUSUM tests balstīts uz OLS tiek definēts analoģiski, izmantojot OLS atlikumus $\hat{u}_i = y_i - x_i^\top \hat{\beta}$ rekursīvo atlikumu vietā. CUSUM balstīts uz OLS ir definēts ar t no $[0, 1]$ kā

$$W_n^0(t) = \frac{1}{\hat{\sigma}\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \hat{u}_i, \tag{1.11}$$

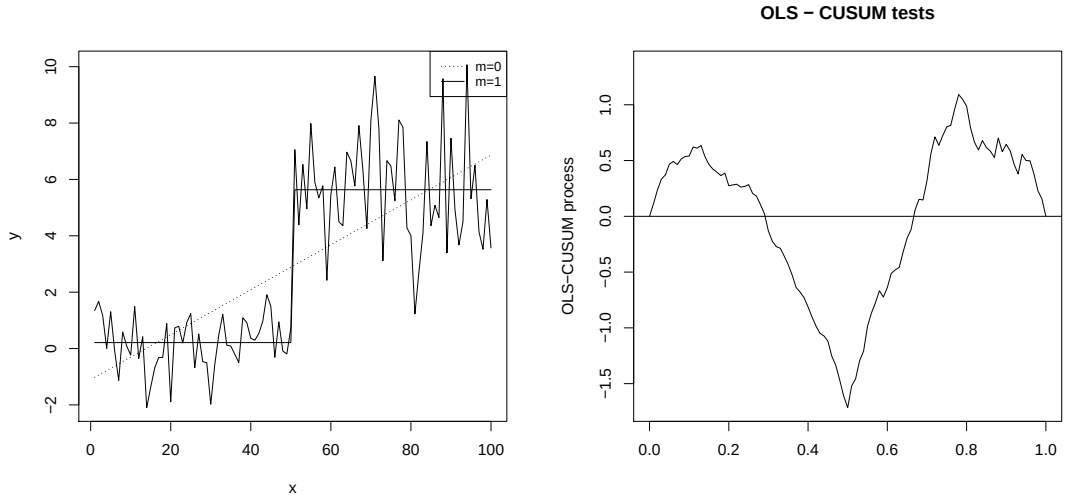
kur $\hat{\sigma} = \sqrt{\frac{1}{n-k} \sum_{i=1}^{n} \hat{u}_i^2}$. Ceļš ne tikai sāksies nullē, bet arī atgriezīsies, bet ja ir izmaiņas punktā t_0, tad tam jābūt virsotne tuvu pārtraukuma punktam t_0.



(a) Simulētie dati bez maiņas punkta ar regresijas līniju

(b) OLS - CUSUM

3. att. OLS - CUSUM statistikas uzvedība simulētiem datiem bez maiņas punktiem



(a) Simulētie dati ar regresijas līnijām (b) OLS - CUSUM

4. att. OLS - CUSUM statistikas uzvedība simulētiem datiem ar vienu maiņas punktu

Salīdzinot OLS - CUSUM grafikus simulētajiem datiem ar vienu maiņas punktu un bez maiņas punkta, var redzēt 4.(b). grafikā izteiktu pīķi uz leju, kas norāda uz maiņas punktu.

H_0 noraida, ja ceļš šķērso λ vai $-\lambda$, kas ir ekvivalenti noraidīt, kad testa statistika

$$S^0 = \sup_{0 \leq t \leq 1} |W_n^0(t)| \tag{1.12}$$

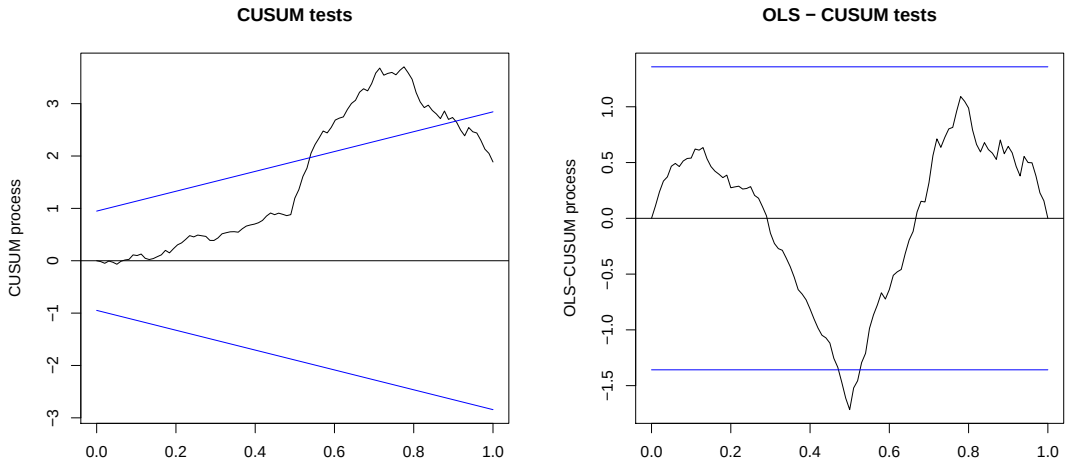
ir lielāka par λ, kuru nosaka testa nozīmības līmenis.

Plobergers un Kramers (Ploberger and Kramer) (1992) [4] parādīja, ka $n \longrightarrow \infty$

$$W_n^0 \longrightarrow^d B^0(t), \tag{1.13}$$

kur $B^0(t)$ ir Brauna tilts.

Standarta robežas lielākajai daļai ierobežojošu procesu ir $d(t) = 1$, bet tikai Brauna kustībai tās ir $d(t) = 1 + 2t$, jo procesam ir pieaugoša dispersija [15] . CUSUM testa robežas var apskatīt 5.(a). grafikā kopā ar CUSUM statistiku, bet OLS CUSUM testa robežas ir 5.(b). grafikā kopā ar atbilstošo testa statistiku.



(a) CUSUM statistikas ar robežām (b) OLS CUSUM statistikas ar robežām

5. att. Testa statistikas ar robežām simulētiem datiem

1.2. F tests

Diezgan atšķirīga pieeja, lai izpētītu, vai nulles hipotēze (nav strukturālu izmaiņu) ir spēkā, ir izmantot F testa statistiku. Būtiska atšķirība ir tāda, ka alternatīva ir norādīta: tā kā vispārējais svārstību tests ir piemērots dažādiem strukturālu izmaiņu modeļiem, F tests ir paredzēts, lai pārbaudītu vienu izmaiņu pie alternatīvas. Tādējādi alternatīvu var formulēt, pamatojoties uz modeli (1.1),

$$\beta_i = \begin{cases} \beta_A & (1 \leq i \leq i_0) \\ \beta_B & (i_0 \leq i \leq n) \end{cases}, \tag{1.14}$$

kur i_0 ir kāds maiņas punkts intervālā $(k, n-k)$. Chow (1960) bija pirmais, kurš ieteica testu strukturālām izmaiņām gadījumā, ja (potenciālais) maiņas punkts ir zināms. Nulles hipotēzi noraida, ja

$$F_i = \frac{\hat{u}^\top \hat{u} - \hat{e}^\top \hat{e}}{\hat{e}^\top \hat{e}/(n-2k)} \tag{1.15}$$

ir pārāk liels, kur $\hat{e} = (\hat{u}_A, \hat{u}_B)^\top$ ir pilnā modeļa atlikumi, kuriem koeficienti apakšizlasē ir novērtēti atsevišķi, un $\hat{u}$ ir ierobežotā modeļa atlikumi, kur parametri ir piemēroti visiem novērojumiem uzreiz. Testa statistikai F_i ir asimptotisks χ^2 sadalījums ar k brīvības pakāpēm un (pie normalitātes pieņēmuma) F_i/k ir precīzs F sadalījums ar k un $n-2k$ brīvības pakāpēm. Lielākai šķērslis *Chow* testam ir tāds, ka maiņas punktam jābūt jau zināmam iepriekš, bet ir arī testi balstīti uz F statistiku (Chow statistika), kas neprasa specifikāciju konkrēta maiņu punktā.

Ir vairāk nekā viena iespēja, lai paplašinātu F statistiku. Ideja *Chow* testa palielināšanai ir aprēķināt F statistikas visiem potenciālajiem maiņas punktiem vai visiem potenciālajiem maiņas punktiem intervālā $[i_-, i^-]$ un noraidīt, ja kaut viena statistika kļūst pārāk liela. i_- un i^- attiecīgi ir apakšintervāla ar izvēlētu garumu kreisā un labā robeža [16].

Andrews (1993) [17] un Andrews and Ploberger (1994)[18] attiecīgi ieteica trīs dažādas testa statistikas un pārbaudīja to asimptotisko sadalījumu:

$$\sup F = \sup_{i_- \leq i \leq i^-} F_i, \tag{1.16}$$

$$aveF = \frac{1}{i^- - i_- + 1} \sum_{i=i_-}^{i^-} F_i, \tag{1.17}$$

$$\exp F = \log \left(\frac{1}{i^- - i_- + 1} \sum_{i=i_-}^{i^-} \exp(0.5 * F_i) \right). \tag{1.18}$$

2. MAIŅAS PUNKTA NOTEIKŠANA LAIKRINDĀM

Maiņas punkta analīzē daudzi ir ieinteresēti atklāt izmaiņas laikrindās kā, piemēram, vidējās vērtības izmaiņa AMOC modelī

$$Y(i) = \begin{cases} \mu_1 + V(i), & 1 \leq i \leq \tilde{k}, \\ \mu_2 + V(i), & \tilde{k} < i \leq T, \end{cases} \tag{2.1}$$

kur $V(.)$ ir stacionārs process ar $EV(0) = 0$, μ_1, μ_2 un $\tilde{k}$ ir nezināmi. Jautājums ir, vai vidējās vērtība ir izmainījusies kādā nezināmā laikā $\tilde{k}$. Problemātika izsakāma

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

Tad attiecīgā CUSUM statistika [19]

$$C_T = \max_{1 \leq k \leq T} \left| \frac{1}{\sqrt{T}} \sum_{j=1}^{k} (Y(j) - \bar{Y}_T) \right|. \tag{2.2}$$

Tālāk tiks apskatīta vienkāršākā situācija, kura rodas, ja tiek pieņemts, ka dažas raksturīgās pazīmes (piemēram, ražošanas procesa) svārstās ap noteiktu konstanti a_0, kas ir dota. Pieņemsim, ka procesa sākums tiek kontrolēts. Tomēr var gadīties, ka sakarā ar ražošanas iekārtas kļūdu, novērotā pazīme pēkšņi sāk svārstīties ap citu konstanti $a_1 \neq a_0$.

Vienkāršākajā gadījumā ar zināmu a_0 vērtību un dispersiju σ^2, var standartizēt novērojumus, lai iegūtu standartizētus mainīgos Y_i, $i = 1, ..., n$, kuriem sākumā vidējā vērtība ir 0 un dispersija 1, pārbaudīt sekojošu nulles hipotēzi H_0 pret alternatīvu H_1, t. i.,

$$H_0 : Y_i = e_i, \qquad i = 1, ..., n,$$

$$H_1 : \exists m \in \{0, ..., n-1\} \quad \text{tāds, ka}$$

$$Y_i = e_i \qquad i = 1, ..., m,$$

$$Y_i = a + e_i, \qquad i = m+1, ..., n,$$

kur $a \neq 0$ un $\{e_i\}$ ir neatkarīgi un vienādi sadalīti (turpmāk, iid) gadījuma mainīgie (kļūdas). Lielumu m sauc ar maiņas punktu.

Metode testa statistikas iegūšanai

Lēmumi par nulles hipotēzes noraidīšanu, ir balstīti uz testa statistiku. Tiek apskatīta *maksimālās ticamības metode*, kuru var izmantot, lai iegūtu testa statistiku. Vienkāršībai pieņemsim, ka $\{e_i\}$ ir neatkarīgi un sadalīti atbilstoši standarta normālajam sadalījumam $N(0,1)$ ar blīvuma funkciju $\phi(x)$.

Vislielākās ticamības attiecību metode. Sakumā tiek pieņemts, ka maiņas punkts m ir zināms un $m = k$. Ja $a \neq 0$, tad logaritmiskā ticamības attiecība H_0 pārbaudei pret H_1 ir

$$\Lambda_k = \sup_a \log \frac{\prod_{i=1}^{k} \phi(Y_i) \prod_{i=k+1}^{n} \phi(Y_i - a)}{\prod_{i=1}^{n} \phi(Y_i)}$$

$$= \sup_a \{-\frac{1}{2} \sum_{i=k+1}^{n} (Y_i - a)^2 + \frac{1}{2} \sum_{i=k+1}^{n} Y_i^2\} = \frac{1}{2(n-k)} \left(\sum_{i=k+1}^{n} Y_i \right)^2 .$$

Nulles hipotēzi noraida, ja $\Lambda_k > C_\alpha$, ko var ekvivalenti izteikt kā

$$\left| \frac{1}{\sqrt{n-k}} \sum_{i=k+1}^{n} Y_i \right| > \sqrt{2C_\alpha},$$

kur C_α ir konstante, kura izvēlēta, lai tā atbilstu fiksētam nozīmības līmenim α.

Vienkāršojot ievietojam

$$\bar{Y}_k = \frac{1}{k} \sum_{i=1}^{k} Y_i \quad \text{un} \quad \bar{Y}_k^0 = \frac{1}{n-k} \sum_{i=k+1}^{n} Y_i.$$

Jāievēro, ka $\bar{Y}_k^0$ ir mazākā kvadrāta novērtējums nezināmai konstantei a un $\sqrt{n-k}\bar{Y}_k^0$ ir standarta normālais sadalījums $N(0,1)$. Vienas puses alternatīvai ar $a > 0$, iegūst testa statistiku

$$\frac{1}{\sqrt{n-k}} \sum_{i=k+1}^{n} Y_i,$$

bet divpusējai alternatīvai ar $a \neq 0$, izmanto absolūto vērtību

$$\left| \frac{1}{\sqrt{n-k}} \sum_{i=k+1}^{n} Y_i \right|.$$

Kad maiņas punkts m nav zināms (tātad abi a un m ir nezināmi), jāņem suprēms logaritmiskajai ticamības atiecībai, t. i.,

$$\max_{0 \leq k \leq n-1} \sup_a log \frac{\prod_{i=1}^{k} \phi(Y_i) \prod_{i=k+1}^{n} \phi(Y_i - a)}{\prod_{i=1}^{n} \phi(Y_i)} = \max_{0 \leq k \leq n-1} \frac{1}{2(n-k)} \left(\sum_{i=k+1}^{n} Y_i \right)^2 ,$$

un testa statistika, ko izmanto nezināma maiņas punkta m gadījumā, ir formā

$$\max_{0\le k\le n-1}\{\frac{1}{\sqrt{n-k}}\sum_{i=k+1}^{n}Y_i\}$$

un

$$\max_{0\le k\le n-1}\{|\frac{1}{\sqrt{n-k}}\sum_{i=k+1}^{n}Y_i|\}.$$

Ja apskata divpusēja alternatīvu ar $a\neq 0$, tad nulles hipotēzi H_0 noraida, ja pienācīgi izvēlētai konstantei $C_{1\alpha}$

$$\max_{0\le k\le n-1}\{|\frac{1}{\sqrt{n-k}}\sum_{i=k+1}^{n}Y_i|\}>C_{1\alpha},$$

kur $C_{1\alpha}$ ir konstante, izvēlēta atbilstoši fiksētam nozīmības līmenim α. Tāds nosacījums ir saprātīgs, jo tas noraida H_0, ja kaut vienam $k, 0\le k\le n-1$,

$$\{\frac{1}{\sqrt{n-k}}\sum_{i=k+1}^{n}Y_i\}>C_{1\alpha}.$$

Līdzīgi vienpusējai alternatīvai ar $a>0$, nulles hipotēzi noraida, ja

$$\max_{0\le k\le n-1}\{\frac{1}{\sqrt{n-k}}\sum_{i=k+1}^{n}Y_i\}>C_{2\alpha},$$

kur $C_{2\alpha}$ ir atkal pienācīgi izvēlēta konstante.

Kritiskās vērtības. Lai pieņemtu lēmumu par nulles hipotēzes noraidīšanu, ir jāzina testa statistikām kritiskās vērtības. Tas nozīmē, ka jāzina to sadalījums pie H_0.

Par piemēru apskatīsim testa statistiku

$$\max_{0\le k\le n-1}\{\left|\frac{1}{\sqrt{n-k}}\sum_{i=k+1}^{n}Y_i\right|\}.$$

Pieņemot, ka $\{Y_i\}$ ir *iid* ar standarta normālo sadalījumu $N(0,1)$, tad statistikai

$$\frac{1}{\sqrt{n-k}}\sum_{i=k+1}^{n}Y_i,\qquad k=0,...,n-1, \tag{2.3}$$

ir $N(0,1)$ sadalījums. Ja m ir nezināms un vienāds ar k, varētu noraidīt H_0 pie nozīmības līmeņa α, ja

$$\left|\frac{1}{\sqrt{n-k}}\sum_{i=k+1}^{n}Y_i\right|>u_{1-\alpha/2},$$

kur $u_{1-\alpha/2}$ ir $100(1-\alpha/2)\%$ $N(0,1)$ kvantile [10].

3. VISPĀRĒJĀ MAIŅAS PUNKTU TEORIJA

Pieņemsim, ka ir sakārtota datu kopa $y_{1:n} = (y_1, ..., y_n)$. Saka, ka ir notikusi izmaiņa kopā, ja eksistē tāds $\tau \in \{1, ..., n-1\}$, ka statistiskās īpašības $\{y_1, ..., y_\tau\}$ un $\{y_{\tau+1}, ..., y_n\}$ ir atšķirīgas kaut kādā veidā. Paplašinot ideju no viena maiņas punkta līdz vairākiem, būs skaitā m maiņas punktu kopā ar to pozīciju $\tau_{1:m} = (\tau_1, ..., \tau_m)$. Katra maiņas punkta pozīcija ir skaitlis starp 1 un $n-1$ iekļaujoši. Definē $\tau_0 = 0$ un $\tau_{m+1} = n$, un pieņem, ka maiņas punkti ir sakārtoti tā, ka $\tau_i < \tau_j$, tad un tikai tad, ja $i < j$. Tātad m maiņas punkti sadalīs datus $m+1$ segmentā, i-tais segments satur $y_{(\tau_{i-1}+1):\tau_i}$. Katrs segments tiks apkopots ar parametru kopu. Parametri atbilstoši i-tajam segmentam būs $\{\theta_i, \phi_i\}$, kur θ_i ir (iespējams nulle) traucējošo parametru kopa un ϕ_i ir parametru kopa, kuri var saturēt izmaiņas.

Vienas izmaiņas noteikšanu var balstīt uz hipotēžu testu. Nulles hipotēze H_0, ka nav izmaiņu ($m = 0$) un alternatīva hipotēze H_1, ka ir viena ($m = 1$).

Izmantojot pieeju balstītu uz vislielāko ticamības funkciju, maiņas punkta noteikšanai pirmais ierosināja Hinkley (1970)[20], kurš ieguva asimptotisko sadalījumu vislielākās ticamības attiecības testa statistikai vidējās vērtības izmaiņai normāli sadalītiem novērojumiem. Gupta un Tang (1987) [21] paplašināja uz vislielākās ticamība balstīto pieeju uz izmaiņām dispersijai ar normāli sadalītiem novērojumiem.

Vislielākās ticamības attiecību metode prasa rēķināt maksimālo logaritmisko ticamības vērtību gan pie nulles, gan pie alternatīvas hipotēzes. Nulles hipotēzei maksimālās logaritmiskās ticamības vērtība ir $\log p(y_{1:n}|\hat{\theta})$, kur $p(.)$ ir varbūtības blīvuma funkcija un $\hat{\theta}$ ir maksimālās ticamības parametru novērtējums. Pie alternatīvas apskata modeli ar maiņas punktu pie τ_1 ar $\tau_1 \in \{1, 2, ..., n-1\}$. Tad maksimālā logaritmiskā ticamība dotam τ_1 ir

$$ML(\tau_1) = \log p(y_{1:\tau_1}|\hat{\theta}_1) + \log p(y_{(\tau_1+1):n}|\hat{\theta}_2).$$

Maksimālās logaritmiskās ticamības vērtība pie alternatīvas ir $\max_{\tau_1} ML(\tau_1)$. Testa statistika ir

$$\lambda = 2\left[\max_{\tau_1} ML(\tau_1) - \log p(y_{1:n}|\hat{\theta})\right]. \tag{3.1}$$

Tests ietver iespēju izvēlēties slieksni c tādu, ka tiek noraidīta nulles hipotēze, ja $\lambda > c$. Ja tiek noraidīta nulles hipotēze, attiecīgi noteikts maiņas punkts, tad novērtējam tā pozīciju $\hat{\tau}_1$ kā, vērtību τ_1, kas maksimizē $ML(\tau_1)$.

Ir skaidrs, ka ticamības testa statistiku var paplašināt uz vairākām izmaiņām vienkārši summējot ticamības katram m segmentam. Problēma rodas identificējot $ML(\tau_{1:m})$ maksimumu visām iespējamām $\tau_{1:m}$ kombinācijām. Tālāk tiek aprakstītas meklēšanas metodes, kas tiek galā ar to.

Lai identificētu vairākus maiņas punktus gadījumos, kad maiņas punktu skaits palielinās, savācot vairāk datus, piemēram, ģenētikā, analizējot lielākus genomu reģionus, vai finansēs, aplūkojot laikrindas ilgākā laika periodā, viena no pieejām ir minimizēt

$$\sum_{i=1}^{m+1}[C(y_{(\tau_{i-1}+1):\tau_i})] + \beta f(m), \tag{3.2}$$

kur C ir izmaksu funkcija segmentam un $\beta f(m)$ ir sods, kas nodrošina pret pārmērīgu piemērošanu. Parasti par izmaksu funkciju maiņas punktu literatūrā izmanto divreiz negatīvo logaritmisko ticamības funkciju, skatīt Chen un Gupta (2000) [22], tiek izmantotas arī citas izmaksu funkcijas, kā piemēram, kumulatīvās summas, skatīt Inclan un Tiao (1994) [23]).

Visbiežāk praksē izvēlētais sods ir lineārs maiņas punktam, t. i., $\beta f(m) = \beta m$. Piemēri tādiem sodiem ir AIC ($\beta = 2p$) un BIC ($\beta = p \log n$), kur p ir papildus ieviesto parametru skaits, pievienojot maiņas punktu. Eksistē trīs algoritmi, kas minimizē 3.2:

- binārā segmentācija,
- kaimiņu segmentu,
- PELT (angliski *Pruned Exact Linear Time*).

Binārā segmentācija

Binārās segmentēšanas algoritms ir viens no izplatītākajiem meklēšanas algoritmiem maiņas punkta literatūrā. Būtībā algoritms paplašina jebkuru viena maiņas punkta metodi uz vairākiem maiņas punktiem, iteratīvi atkārtojot dažādām secības apakškopām. Sākotnēji tiek piemērota viena maiņas punkta metode visiem datiem, t. i., tiek pārbaudīts

$$C(y_{1:\tau}) + C(y_{(\tau+1):n}) + \beta < C(y_{1:n}). \tag{3.3}$$

Ja 3.3 ir nepatiess, tad maiņas punkts ir atrasts un algoritms apstājas. Pretējā gadījumā sadala datus divos segmentos, pirms un pēc optimālā maiņas punkta τ_a, un piemēro

noteikšanas metodi abiem segmentiem

$$C(y_{1:\tau}) + C(y_{(\tau+1):\tau_a}) + \beta < C(y_{1:\tau_a}) \tag{3.4}$$

un

$$C(y_{\tau_{a+1}:\tau}) + C(y_{(\tau+1):n}) + \beta < C(y_{\tau_{a+1}:n}). \tag{3.5}$$

Ja viens no diviem, vai abi ir patiesi tad abi segmenti tiek sadalīti tālāk segmentos pie optimālā atrastā maiņas punkta, un tiek piemērota metode katram segmentam. Procedūra tiek atkārtota, kamēr maiņas punkti netiek vairāk konstatēti nevienā segmentā.

Procedūra ir aptuvena 3.2 minimizācija ar $f(m) = m$. Binārās segmentācijas algoritma priekšrocība ir tā tiek uzskatīta par ātri skaitļojamu $O(n \log n)$. Datu kopai ar acīmredzamiem maiņas punktiem, aproksimācija būs niecīga, bet kad maiņas punktus kļūst grūtāk noteikt (vai nu mazas izmaiņas vai izmaiņas, kuras ir tuvumā), tad aproksimācija var izraisīt maiņas punktu pazušanu vai neprecīzu noteikšanu.

Segmenta kaimiņu

Auger and Lawrence (1989) aplūkoja alternatīvu maiņas punkta noteikšanas algoritmu, sauktu par Segmenta kaimiņu algoritmu. Pamatprincips pieejai ir meklēt visu segmentācijas telpu efektīvā veidā. Lietotājs definē augšējo robežu segmentācijas telpas izmēram, kas ir maksimālais segmentu skaits Q. Lai sāktu algoritmu aprēķina izmaksu funkciju visiem iespējamiem segmentiem. Tad visi iespējamie segmenti tiek aplūkoti, sākot ar vienu maiņas punktu un beidzot ar $Q-1$ maiņas punktiem. Meklēšana tiek veikta, izmantojot dinamisko algoritmu tā, ka segmentāciju atrod ar vienu maiņas punktu, informē meklēt divus maiņas punktus un tā tālāk.

Segmenta kaimiņu algoritma priekšrocības ietver iespēju iekļaut patvaļīgu soda funkciju $\beta f(m)$ un faktu, ka tas ir precīzs, jo visas iespējas tiek aplūkotas. Sekas tādai izsmeļošai meklēšanai ir tādas, ka algoritmam skaitļošanas izmaksas ir $O(Qn^2)$, kas ir lēnas salīdzinot ar Bināro segmentāciju. Parasti datu kopai palielinot izmēru, maksimālais maiņas punktu skaits Q arī palielinās un algoritms uzvedās vairāk kā $O(n^3)$.

PELT

Saīsinātā precīza lineārā laika metode Killick et al. (2011) darbā [24], turpmāk tekstā - PELT, dod precīzus rezultātus ātrāk nekā Segmenta kaimiņa un Binārās segmentācijas algoritms. PELT algoritms aplūko datus secīgi un katra solī optimālā segmentācija

līdz tam solim ir ierakstīta. Algoritms sākas, aplūkojot optimālo segmentāciju pirmajiem diviem datu punktiem $y_{1:2}$, vai nu ir maiņas punkts pie 1 un y_1, y_2 ir ar dažādiem sadalījumiem, vai arī nav maiņas punkta un tie ir ar vienādu sadalījumu. Nākošajā iterācijā PELT aplūko optimālo segmentāciju $y_{1:3}$, tagad ir 4 iespējas

- maiņas punkts pie 1,
- maiņas punkts pie 2,
- maiņas punkti pie 1 un 2,
- nav maiņas punktu.

Datiem $y_{1:2}$ ir jau nolemts, vai ir maiņas punkts pie 1. Tādējādi iespējas ir jāsamazina līdz

- pēdējais maiņas punkts ir pie 1,
- pēdējais maiņās punkts ir pie 2,
- nav maiņas punktu.

Iespēju skaits katrā solī ir lineārs, jo vienmēr tiek aplūkots pēdējais maiņas punkts esot katrā iepriekšējā vērtībā. Neatkarīgi no tā, cik daudz maiņas punkti ir bijuši. Lai samazinātu skaitļošanas pūles, PELT algoritms vēl katrā solī nolem, vai pēdējais punkts nevar būt ar nākotnes maiņas punktu. Tas var samazināt iespēju skaitu. Iterācijas algoritms turpinās, kamēr tiek sasniegts n, kur tiek aprēķināta galīgo maiņas punktu kopa.

4. MAIŅAS PUNKTA NOTEIKŠANA PRAKSTISKI

4.1. Izmantojot programmas *R* paketi *changepoint*

changepoint pakete mēģina nodrošināt gan ar vispāratzītām, gan jaunām metodēm lietotājam draudzīgu paketi. Tā ir izstrādāta, lai lietotājam dotu pieeju daudziem paņēmieniem maiņas punkta analīzei no dažām pamatfunkcijām. Pakete ietver meklēšanas algoritmu izvēli vairāku maiņas punktu identificēšanai. Tie ietver bināro segmentēšanu (Scott and Knott, 1974), kaimiņu segmentu (Auger and Lawrence, 1989), un nesen ierosināto aptuveni lineāra laika, Pruned Exact Linear Time (PELT), meklēšanas algoritmu (Killick et al., 2011). *changepoint* pakete satur trīs primārās izsaukšanas funkcijas izmaiņu noteikšanai vidējai vērtībai, dispersijai un gan vidējai vērtībai, gan dispersijai. Katrai no tām ir līdzīga argumentu struktūra un katrai izvadē ir *cpt* klases objekts. Tāda pieeja ir iepriekš nodomāta, lai radītu vieglu pārzināšanu un lietošanas ērtumu.

changepoint pakete ievieš jaunu objektu klasi *cpt*, lai uzglabātu maiņas punktu analīzes objektus. Katra no galvenajām funkcijām izvada *cpt S4* objektus. Klase ir izveidota tā, lai *cpt* objekti saturētu galvenās īpašības, kas nepieciešamas maiņas punktu analīzei.

Ir trīs galvenās metodes, kas saistītas ar *cpt* klasi un, kuras dažiem lietotājiem var būt noderīgas. *sumary* un *print* metodes parāda standarta informāciju par *cpt* klases objektu. *summary* funkcija izvada īsu konspektu par maiņas analīzes rezultātiem, tostarp maiņas punktu skaitu un, ja tas ir mazs, tad arī maiņas punktu atrašanās vietu. Turpretī *print* funkcija izdrukā informāciju, kas attiecas uz *S4* klasi ieskaitot **slot** nosaukumus un to, kad *S4* objekti tika izveidoti.

Veicot maiņas punktu analīzi, bieži vien ir noderīgi attēlot maiņas punktus uz sākotnējiem datiem, lai noteiktu, vai novērtētie maiņas punkti ir pamatoti. Lai to izdarītu, ir izveidota *plot* metode *cpt* klasei. Metode pielāgojas pieņemtajam maiņas punkta veidam, nodrošinot atšķirīgu izvadi atkarīgu no izmaiņas veida. Piemēram, izmaiņa dispersijai ir apzīmēta ar vertikālu līniju pie maiņas punkta atrašanās vietas, bet vidējās vērtības izmaiņu parāda ar horizontālu līniju attēlojot vidējo vērtību dažādos segmentos. [25]

Izmaiņas vidējā vērtībā

Agrākos darbos par maiņas punktu problēmām koncentrējās uz vidējās vērtības izmaiņas noteikšanu un ietver Page (1954) [26] un Hinkley (1970) [20] darbus, kuri attiecīgi izveidoja varbūtību attiecību un kumulatīvo summu (CUSUM) testa statistikas. *changepoint* paketes ietvaros visas vidējās vērtības izmaiņu metodes ir pieejamas izmantojot *cpt.mean* funkciju. Funkcijai struktūra ir sekojoša:

```
cpt.mean(data,penalty="SIC",value=0,method="AMOC",Q=5,dist="Normal")
```

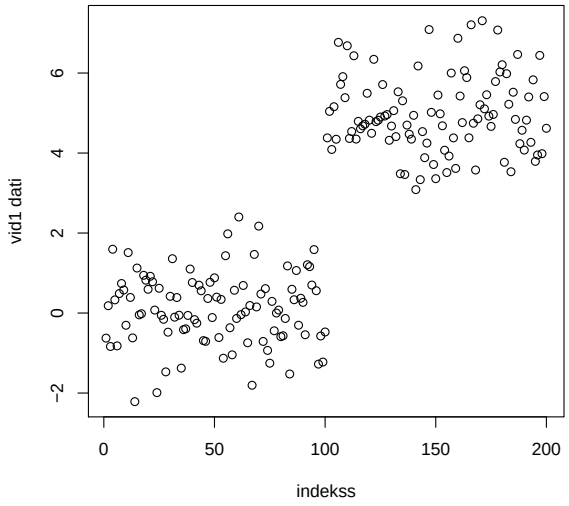
Funkcijas argumenti:

- `data` - vektors, kas satur datus, kuriem jāmeklē izmaiņas vidējai vērtībai, ja vairākām datu kopām nepieciešama analīze, tad ta var būt matrica, kurai katra rinda tiek uzskatīta par atsevišķu datu kopu.
- `penalty` - sodu ‘‘`None`’’, ‘‘`SIC`’’, ‘‘`BIC`’’, ‘‘`AIC`’’, ‘‘`Hannan-Quinn`’’, ‘‘`Asymptotic`’’ and ‘‘`Manual`’’ izvēle. Ja ‘‘`Manual`’’ ir izvēlēts, tad `value` satur manuālo sodu. Ja ir izvēlēts ‘‘`Asymptotic`’’, tad `value` satur teorētisko `I` tipa kļūdu.
- `value` - teorētiskā `I` tipa kļūda, piemēram, 0.05, kad izmanto ‘‘`Asymptotic`’’ sodu.
- `method` viena vai vairāku maiņas punktu metode. ‘‘`AMOC`’’ (vismaz viena izmaiņa), ‘‘`PELT`’’, ‘‘`SegNeigh`’’ vai ‘‘`BinSeg`’’. Noklusējumā ir ‘‘`AMOC`’’.
- `Q` - maksimālais maiņas punktu skaits, ko meklēt, izmantojot ‘‘`BinSeg`’’ metodi. Maksimālais segmentu skaits (maiņas punktu skaits +1) ko meklēt, izmantojot ‘‘`SegNeigh`’’ metodi. Tas nav nepieciešams ‘‘`PELT`’’ metodei, jo tas automātiski izvēlas segmentu skaitu.
- `dist` - pieņemtais datu sadalījums. Izvēle starp ‘‘`Normal`’’ vai ‘‘`CUSUM`’’.

Vairākas standarta soda funkcijas, izmantotas maiņas punkta analīzē, ir iekļautas *cpt.mean* funkcijā. Tās ir: SIC (Švarca informācijas kritērijs), BIC (Beijesa informācijas kritērijs), AIC (Akaike informācijas kritērijs) un Hanna-Kvinna. Lietotājs var arī manuāli ievadīt soda lielumu ar skaitlisku vērtību vai formulu. Meklēšanas iespējas sastāv no precīzām metodēm: PELT ($O(n)$), kaimiņu segmenti ($O(Qn^2)$) un aptuvenām metodēm: Binārā segmentācija ($O(nlogn)$).

Piemērs ar vienu maiņas punktu vidējai vērtībai

Tika simulēta datu kopa *vid1*, kas ir secīga ar garumu 200 ar maiņas punktu pie 100, sākotnējā vidējā vērtība ir 0, pēc izmaiņas tā ir 5.



6. att. Datu kopas *vid1* grafiks

Pirmais jautājums maiņas punkta analīzē ir, vai datiem pastāv izmaiņa vidējai vērtībai. Vizuāli no datiem 6. attēlā tiek sagaidīta viena izmaiņa. Sākumā tiek pieņemts, ka datiem ir normālais sadalījums.

```
> library(changepoint)
> vid1.amoc=cpt.mean(vid1)
> plot(vid1.amoc, ylab="vid1 dati", xlab="indekss")
> cpts(vid1.amoc)
cpt
100 200
```

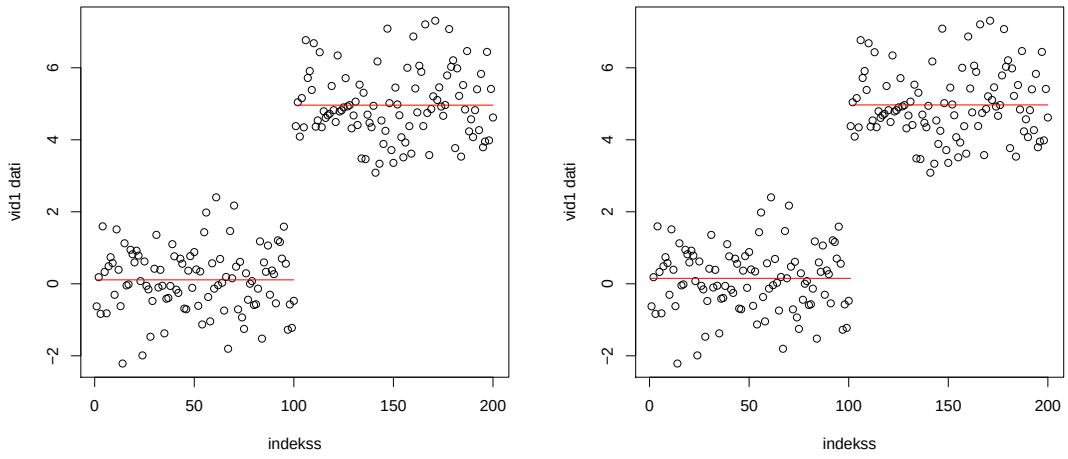
vid1.amoc objekts satur maiņas punktus, kas tiek konstatēti, izmantojot noklusējuma vērtības, kas ir SIC sods, normalitātes pieņēmumu un AMOC metodi. *cpts* funkcija izvada identificētos maiņas punktus. Atgrieztā vērtība parāda, ka maiņas punkts ir pie 100 un datu kopas garums ir 200. 7.(a) attēlā redzams oriģinālās datu kopas grafiks kopā ar novērtētajām vidējām vērtībām, iegūts izsaucot iepriekšminēto funkciju *plot*. Kad maiņas punkti ir noteikti, tad tiek noteiktas arī vidējās vērtības katrā segmentā. Novērtējumi tiek iegūti, izmantojot *param.est* funkciju.

```
> param.est(vid1.amoc)
$mean
[1] 0.1088874 4.9621919
```

Sākumā tika pieņemts, ka datiem ir normālais sadalījums. Ja tāds pieņēmums tiek uzskatīts par neiespējamu, tā vietā var veikt neparametrisku testu maiņas punkta noteikšanai, izmantojot CUSUM metodi.

```
>vid1.cusum=cpt.mean(vid1,dist='CUSUM',penalty='Manual',value=1)
> plot(vid1.cusum, ylab="vid1 dati", xlab="indekss")
> cpts(vid1.cusum)
cpt
101 200
> param.est(vid1.cusum)
$mean
[1] 0.151172 4.968076
```

Testa statistika iegūta ar CUSUM testa statistiku ir parasti daudz mazāka nekā ar varbūtību attiecības testa statistiku. Tādējādi iepriekšējai analīzei tiek izmantots manuālais sods, jo SIC sods ir pārāk liels un tradicionāli pieņemtās īpašības vairāk nepastāv. 7.(b) attēls parāda CUSUM rezultātus datu kopai *vid1*. Salīdzinot abas pieejas, var redzēt, ka normalitātes pieņēmumam un CUSUM metodei rezultāti ir ļoti līdzīgi; maiņas punkts pie 100 vai attiecīgi pie 101.



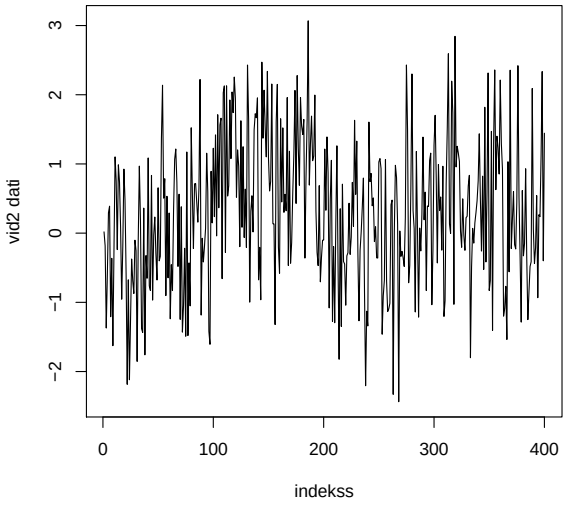
(a) vid1 pie normalitātes pieņēmuma (b) vid1 ar CUSUM

7. att. Datu kopas vid1 grafiki

Piemērs ar vairākiem maiņas punktiem vidējai vērtībai

Ja datu kopai ir vairākas izmaiņas un tiek izvēlēts analizēt to ar AMOC (vismaz viena izmaiņa) pieeju, tad maiņas punktu atgriezīs to, kuram ir lielākā starpība starp pirms maiņas un pēc maiņas testa statistiku.

Tiek simulēta datu kopa *vid2* ar garumu 400 ar vairākiem maiņas punktiem pie 100, 200 un 300. Secīgi ir četri segmenti un vidējā vērtība katram ir 0, 1, 0, 0.2.



8. att. Datu kopas *vid2* grafiks

changepoint pakete satur 3 dažādas maiņas punkta meklēšanas metodes: PELT, kaimiņu segmentu un bināro segmentu.

No datu *vid2* grafika tiek minēts, ka varētu būt vairāki maiņas punkti. Tātad jebkuru no trim metodēm varētu izmantot. Tiks izmantota PELT metode, jo tai ir precīzs algoritms un aprēķina maiņas punktu lineārā laikā. Sākumā tiek pieņemts, ka datiem ir normālais sadalījums.

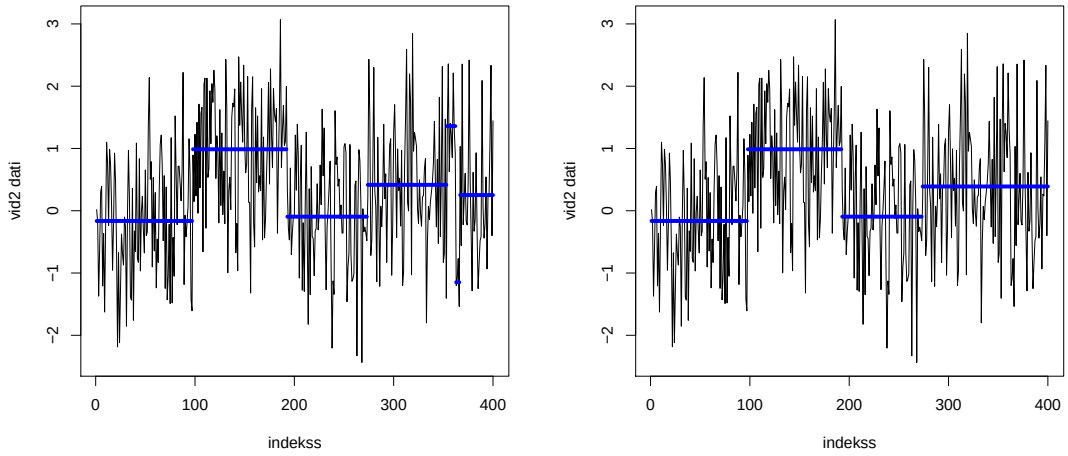
```
> vid2.pelt=cpt.mean(vid2,method='PELT')
>plot(vid2.pelt,type='l',cpt.col='blue',ylab="vid2 dati", xlab="indekss",
 cpt.width=4)
> cpts(vid2.pelt)
[1]  97 192 273 353 362 366 400
```

Tādā gadījumā, ja tiek izmantots noklusējuma SIC sods, *cpts* funkcija atgriež 6 maiņas punktus. Pēc konstrukcijas zināms, ka datiem ir jābūt trīs maiņas punktiem. Var ticēt, ka

ir 6 maiņas punkti, vai uzskatīt, ka metode ir pārāk jūtīga un palielināt sodu. Atbilstoša soda izvēle atkarīga no daudziem faktoriem, ieskaitot izmaiņu skaita un segmentu garuma, no kuriem abi nav zināmi pirms analīzes. Praksē tas bieži vien tiek novērtēts no datu un maiņas punktu grafika, lai redzētu vai tie šķiet saprātīgi.

9.(a) attēla ir paradīti `vid2.pelt` maiņas punkti. Tuvu datu kopas beigām ir divi maiņas punkti, kuriem ir ļoti mazi segmenti. Maz ticams, ka tie ir īsti maiņas punkti pamatdatiem. Lai novērstu šķietami mazos maiņas punktus, var palielināt sodu uz $1.5 * log(n)$ nevis $log(n)$ (SIC). To var izdarīt nomainot soda tipu uz ‘‘`Manual"` un `value` argumentu norādīt $1.5 * log(n)$. 9.(b) attēls parāda rezultātu, kas šķiet daudz ticamāks.

```
ual=cpt.mean(mean2,method='PELT',penalty='Manual',value='1.5*log(n)')
>plot(vid2.manual,type='l',cpt.col='blue',xlab='indekss', ylab="vid2 dati",
 cpt.width=4)
> cpts(vid2.manual)
[1]  97 192 273 400
```



(a) vid2 ar noklusējuma sodu (b) vid2 ar manuālu sodu

9. att. Datu kopas vid1 grafiki ar maiņas punktiem

Ja normalitātes pieņēmums ir nepamatots, tad var izmantot CUSUM metodi, kam nav sadalījuma pieņēmuma.

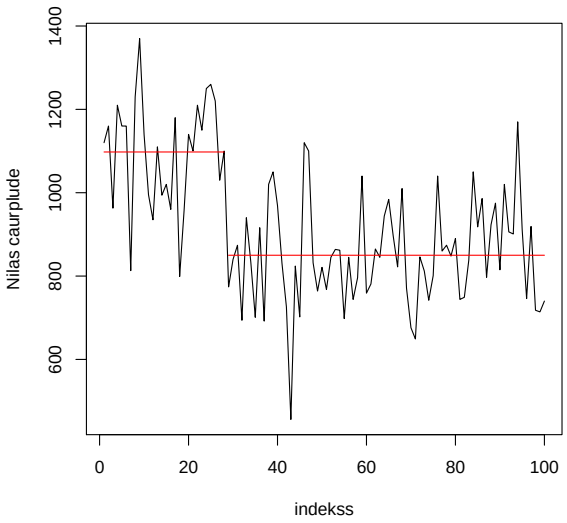
```
>vid2.cusum=cpt.mean(vid2,method='BinSeg',dist='CUSUM',penalty='Manual',value=0.11)
>plot(vid2.cusum,type='l',cpt.col='blue',xlab='indekss', ylab="vid2 dati",cpt.width=4)
> cpts(vid2.cusum)
[1]  97 192 273 400
```

Iegūtie aprēķini ir pamatoti. Īpaši tāpēc, ka rezultāti ir tādi paši, kā *cpt.manual* analīzē.

Reāls datu piemērs

Cobb (1978) [27] pētīja upes Nīlas pie Asuānas ikgadējo plūsmas tilpumu no 1871. līdz 1970. gadam. Meteoroloģiskā interese mērījumos ir pierādījums par iespējamu strauju izmaiņu nokrišņu režīmā tuvu deviņpadsmitajam gadsimtam. Dati ir pieejami *R* datu kopu pakete, rakstot *Nile*.

```
> nile=cpt.mean(Nile[1:100])
> plot(nile,type='l', ylab="Nīlas caurplūde", xlab="indekss")
> cpts(nile)
cpt
 28 100
> param.est(nile)
$mean
[1] 1097.7500  849.9722
```



10. att. Upes Nīlas caurplūdes grafiks

10. attēls norāda, viena maiņas punkta pieņēmums ir pamatots, un analīze identificē 28. (t. i., 1898. gads) novērojumu kā novērtēto maiņas punktu. Tas atspoguļo arī Cobb (1978) rezultātus. Turklāt Cobb (1978) pieņēma pirms un pēc izmaiņās vidējas vērtības

1100 un 850, kas ir ļoti līdzīgi `Nile` novērtējumam un parāda, ka normalitātes pieņēmums ir pamatots.

Izmaiņas dispersijā

changepoint paketes ietvaros visas metodes dispersijas izmaiņu noteikšanai ir pieejamas, izmantojot funkciju `cpt.var`. Funkcijas struktūra ir sekojoša:

```
cpt.var(data,penalty,value,know.mean=FALSE,mu=-1000,method,Q,dist="Normal")
```

Argumenti `data, penalty, value, method, Q` ir tādi paši kā `cpt.mean` funkcijai. Trīs atlikušie argumenti tiek interpretēti tā:

- `know.mean` - loģiskais arguments, kas ir nepieciešams tikai pie `dist=‘‘Normal"`. Ja TRUE, tad vidējā vērtība ir pieņemta kā zināma un `mu` tiek pieņemta, kā tās vērtība. Ja FALSE un `mu=NA` (noklusējuma vērtība), tad vidējo vērtību novērtē ar maksimālās ticamības metodi. Ja FALSE un `mu` vērtība ir piegādāta, tad `mu` nenovērtē, bet tā tiek izskaitļota kā novērtēto parametru lēmumiem.

- `mu` - vajadzīgs tikai `dist=‘‘Normal"`. Datu patiesās vidējās vērtības skaitliskā vērtība (ja zināma). Vai nu viena vērtība, vai vektors garumā `nrow(data)`. Ja dati ir matrica un `mu` ir viena vērtība, tad tā pati tiek izmantota katrā rindā.

- `dist` - Pieņemtais datu sadalījums. Izvēle starp “Normal‘’ un “CSS”. Testa statistikas sadalījuma variantiem ietver Chen un Gupta (2000) [22] “Normal” variantam un Chen un Gupta (1997) [28] “CSS” variantam.

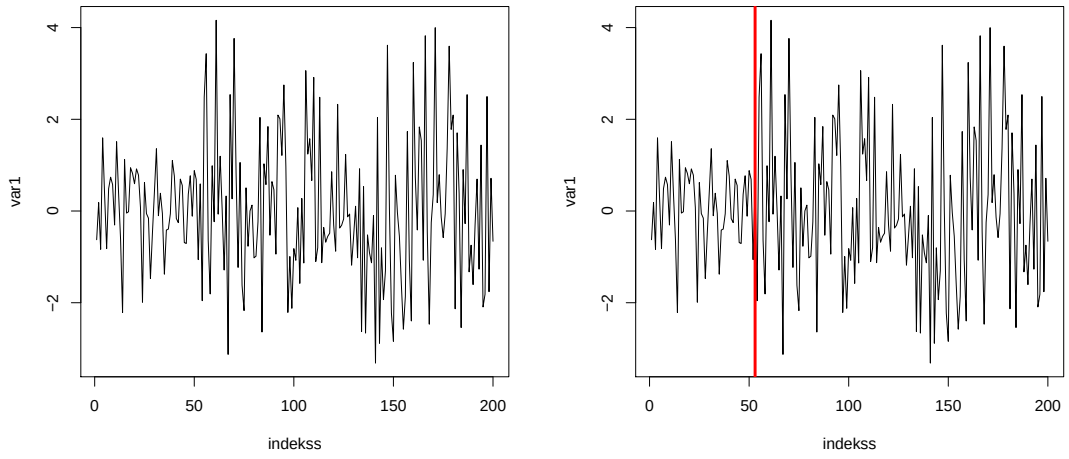
Viena izmaiņa dispersijā

Tika simulēti dati *var1* ar garumu 200 ar vienu izmaiņu dispersijai pie 50. Pirms maiņas dispersija ir 1 un pēc maiņas - 3.

No 11.(a). grafika ir redzams, ka ir notikušas izmaiņas. Sākumā tiek veikta maiņas punkta analīze pie normalitātes pieņēmuma.

```
> var1.amoc=cpt.var(var1)
> plot(var1.amoc,type='l',cpt.width=3,xlab="indekss", ylab="vid2")
> cpts(var1.amoc)
cpt
```

```
 53 200
> param.est(var1.amoc)
$variance
[1] 0.6885335 2.8147428
```



(a) var1 dati (b) var1 ar manuālu sodu

11. att. Datu kopas var1 grafiki ar maiņas punktiem

Izvadē redzams, ka izmaiņas konstatētas 53. novērojumā. Turklāt pirms un pēc izmaiņas novērtējumi attiecīgi ir 0.6885335 un 2.8147428. Grafiks 11.(b) parāda, ka tas šķiet saprātīgs secinājums.

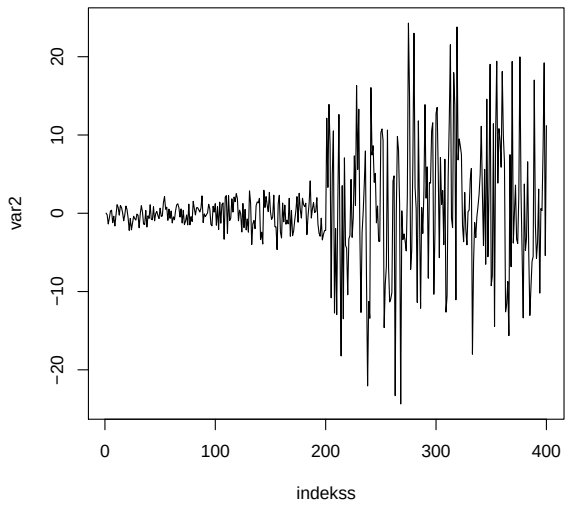
Ja šķiet, ka normalitātes pieņēmums nav atbilstošs, tad var izmantot CSS metodi.

```
> var1.css=cpt.var(var1,dist='CSS')
> cpts(var1.css)
cpt
 53 200
> param.est(var1.css)
$variance
[1] 0.6885335 2.8147428
```

Dotajam piemēram, pieņemot normalitāti un bez sadalījuma pieņēmuma rezultātu dod to pašu.

Vairākas izmaiņas dispersijā

Tiek simulēta datu kopa garumā 400 ar izmaiņām dispersija pie 100, 200 un 300. Dispersija katram segmentam ir 1, 4, 100 un 27.



12. att. var2 dati

12. grafikā redzams, ka maiņas punktu pie 200 ir vieglāk noteikt nekā pie 100 un 300.

```
> var2.pelt=cpt.var(var2,method='PELT')
> plot(var2.pelt,type='l',cpt.width=3,cpt.col='blue',xlab="indekss", ylab="var2")
> cpts(var2.pelt)
[1] 102 200 323 328 330 394 396 400
```

Attiecīgajā piemērā vairāki ļoti īsi segmenti ir identificēti, tas ir rādītājs, ka soda vērtība varētu būt par mazu. Tātad ir tā ir jāpalielina no $logn$ uz $2logn$.

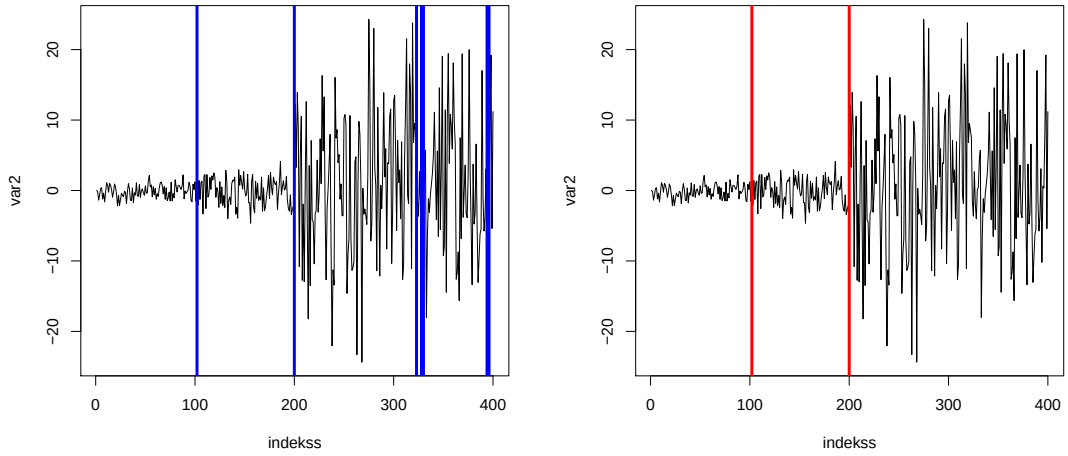
```
> var2.manual=cpt.var(var2,method='PELT',penalty='Manual',value='2*log(n)')
> plot(var2.manual,type='l',cpt.width=3,xlab="indekss", ylab="var2")
> cpts(var2.manual)
[1] 102 200 400
> param.est(var2.manual)
$variance
[1]  0.8968322  3.8077424 91.3573302
```

Izmaiņu pie 300 ir grūti noteikt pat ar aci. Tas nav pārsteidzoši, jo tas ir pirmais maiņas punkts, kas pazūd, palielinot sodu.

Ja normalitātes pieņēmums nav piemērots, tad var izmantot CUSUM testa statistiku kā alternatīvu.

```
> var2.css=cpt.var(var2,method='BinSeg',penalty='Manual',
value='log(2*log(n))',dist='CSS')
> cpts(var2.css)
[1] 107 200 400
> param.est(var2.css)
$variance
[1]  0.9602248  3.8911524 91.3573302
```

Maiņas punkti metodei bez sadalījuma pieņēmuma ir līdzīgi, kad pieņemta normalitāte, bet nav identiski.



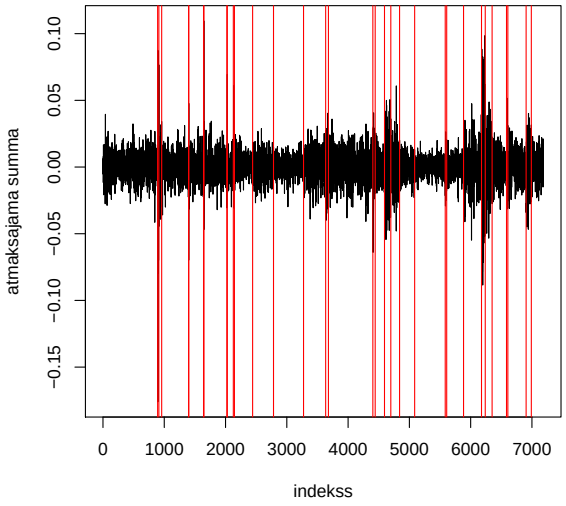
(a) var2 ar noklusējuma sodu (b) var2 ar manuālu sodu

13. att. Datu kopas var2 grafiki ar dažādiem sodiem

Reāls datu piemērs ar vairākām izmaiņām dispersijā

Tiek apskatīta *FTSE100* datu kopa. Tā sastāv no datumiem un katras dienas atgrieztām summām no UK FTSE 100 indeksiem, kas ir definētas kā $c_{t+1}/c_t - 1$, kur c_t noslēguma cena t dienā laika periodā no 1984. gada 2. aprīļa līdz 2012. gada 13. septembrim. Lielas atšķirības atgrieztajām summām rāda nenoteiktību tirgos. Tāpat pēkšņas izmaiņas ir tipiski saistītas ar ārējiem notikumiem tirgos, kas ietekmē uzticību tirgiem, piemēram, uzņēmuma sabrukums. Dati ir pieejami *changepoint* paketes ietvaros.

```
> ftse.pelt=cpt.var(ftse100[,2],method='PELT')
> cpts(ftse.pelt)
 [1]   87  223  257  410  676  697  844  847  892  912  958 1398 1400 1641 1648
[16] 1797 1863 1867 2021 2034 2095 2134 2145 2300 2437 2674 2848 3017 3020 3237
[31] 3264 3340 3497 3501 3634 3679 3685 3743 3979 4086 4148 4280 4325 4404 4416
[46] 4452 4594 4697 4785 4789 4840 5147 5434 5456 5585 5609 5652 5655 5787 5800
[61] 5888 5907 6013 6017 6080 6084 6169 6238 6338 6508 6512 6585 6607 6674 6905
[76] 6951 7034 7051 7172 7175 7187
```



14. att. FTSE 100 dati kopā ar atrastajiem maiņas punktiem

Atkal ir vairāki ļoti īsi segmenti, tādēļ piemērotāks būtu lielāks sods.

```
> ftse.manual=cpt.var(ftse100[,2],method='PELT',penalty='Manual',value='2*log(n)')
> plot(ftse.manual,type='l',ylab="atmaksājamā summa", xlab="indekss")
> ftse100[cpts(ftse.manual),1]
 [1] "1987-10-13" "1987-11-10" "1988-01-18" "1989-10-12" "1989-10-16"
 [6] "1990-09-28" "1990-10-09" "1992-03-30" "1992-04-09" "1992-09-01"
[11] "1992-09-25" "1993-11-26" "1995-04-04" "1997-03-12" "1998-08-17"
[16] "1998-10-20" "2001-09-05" "2001-10-30" "2002-06-11" "2002-11-04"
[21] "2003-06-02" "2004-05-20" "2006-05-11" "2006-06-15" "2007-07-17"
[26] "2008-09-12" "2008-12-08" "2009-05-21" "2010-04-26" "2010-05-27"
[31] "2011-08-02" "2011-11-30" "2012-09-13"

> param.est(ftse.manual)
$variance
 [1] 8.991143e-05 3.430965e-03 2.903557e-04 8.224182e-05 6.903331e-03
 [6] 9.475767e-05 2.724758e-03 8.008394e-05 9.378226e-04 8.122508e-05
[11] 6.429225e-04 5.090071e-05 9.888649e-05 4.088542e-05 1.149468e-04
[16] 4.552452e-04 1.310309e-04 5.398544e-04 7.602984e-05 6.047071e-04
[21] 2.538047e-04 5.660594e-05 3.283803e-05 2.662113e-04 4.832210e-05
[26] 2.150463e-04 1.627570e-03 3.422710e-04 1.039255e-04 4.395527e-04
[31] 9.116181e-05 3.740327e-04 9.070072e-05
```

Īsi segmenti garumā 2 joprojām pastāv, taču tas atspoguļo neziņu tieši pirms 1992. gada vispārējām vēlēšanām. Tika arī atklātas paaugstinātas svārstības ap Melno pirmdienu (1987. gada 19. oktobrī), kad akciju tirgū bija ļoti liela samazināšanās vienā dienā un Kabulas krišana 2001. gada 12. novembrī. Maiņas punkts 2008. gada 11. septembrī sakrīt ar Lehman Brother bankrotu un var redzēt svārstību samazināšanos, kad ASV koalīcijas valdība atvēra parlamentu 2010. gada 26. maijā.

Izmaiņas dispersijā un vidējā vērtībā

changepoint paketē ietilpst trīs sadalījuma izvēles: eksopencialā, gammas un normālā. Katra sadalījuma izvēle ir pieejama `cpt.meanvar` funkcijas ietvaros. Pamata izsaukšanas veids:

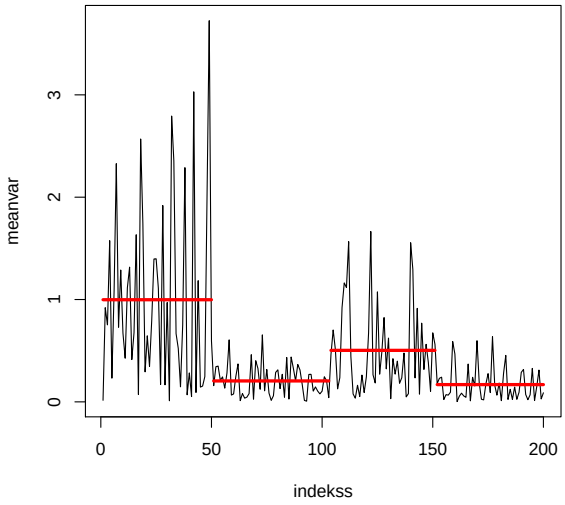
```
cpt.meanvar(data,penalty,value,method,Q,dist="Normal",shape=0)
```

`data, penalty, value, method, Q` ir tādi paši kā aprakstīts *cpt.mean* funkcijai. Atlikušie argumenti ir:

- `dist` - Pieņemtais sadalījums datiem. Izvēle starp ``Normal", ``Gamma", ``Exponential".
- `shape` - pieņemtās zināmās formas vērtība, nepieciešams tad, kad `dist=``Gamma"`.

Piemērs 1. Tiks apskatītas vairākas izmaiņas eksponenciāli sadalītiem datiem. Tiek simulēti eksponenciāli sadalīti dati garumā 200, kur raksturlielumi katrā segmentā ir 1, 5, 2 un 7. Tad tiek analizēti dati, izmantojot PELT metodi vairākām izmaiņām.

```
> meanvar.pelt=cpt.meanvar(meanvar1,dist='Exponential',method='PELT')
> plot(meanvar.pelt,type='l',cpt.width=3, xlab="indekss", ylab="meanvar")
> cpts(meanvar.pelt)
[1]  50 103 151 200
> param.est(meanvar.pelt)
$rate
[1] 1.002004 4.890951 1.987128 5.930341
```



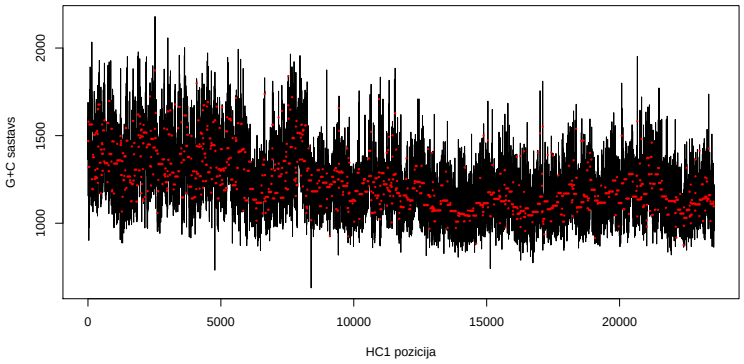
15. att. meanvar dati ar novērtētajām vidējām vērtībām

Tika atrastas 3 izmaiņas, tās redzamas 15. grafikā, tās šķiet saprātīgas.

Reāls datu piemērs

Par pamatdatiem tiek ņemts C+G īpatsvars cilvēka 1. hromosomā. Tiek modelēts katrs segments kā normālais sadalījumus ar savu vidējo vērtību un dispersiju. Tiek analizēts C+G īpatsvars 3kb logos un katram logam atgriezts G un C apjoms atbilstošajā logā. Sākotnējiem datiem bija dažas trūkstošās vērtības. Toties *changepoint* paketē pieejamie dati ir garākais datu posms bez trūkstošajām vērtībām, proti no 10Mb līdz 33Mb. Tiek analizēti dati ar PELT metodi ar manuāli ievadītu sodu.

```
>hc1.pelt=cpt.meanvar(HC1,method='PELT',penalty='Manual',value=14)
> plot(hc1.pelt,type='l',ylab='G+C sastāvs',xlab='HC1 pozīcija',cpt.width=3)
> ncpts(hc1.pelt)
[1] 805
```



16. att. C+G sastāvs cilvēka hromosomā 1 no 10Mb līdz 33Mb

Datu kopai ar garumu 23553 ir ļoti daudz maiņas punktu (805), tādēļ tie netiks parādīti, bet tos var iegūt kā iepriekš. 16. grafikā ir attēloti sākotnējie dati ar vidējām vērtībām.

4.2. Izmantojot programmas *R* paketi *strucchange*

Empīrisko svārstību procesu no vispārināta svārstību testa rāmja var iegūt ar funkciju

```
efp(formula, data, type, ...)
```

kur `formula` definē regresijas modeli, kuru jāpārbauda, piemēram, `y~x`. Arguments `data` ir datu rāmis, kas varētu saturēt mainīgos y un x un arguments `type` norāda svārstību procesa tipu, kas būtu jāpiemēro, piemēram, `''OLS-CUSUM''`, lai piemērotu uz OLS balstītu CUSUM procesu. Empīrisko svārstību procesa objektu, kas iegūts ar `epf` var attēlot grafiski kopā ar tā (asimptotiskajām) robežām, izmantojot funkciju `plot` un atbilstošo nozīmības testu var veikt ar funkciju `sctest`. Līdzīgi F statistiku var aprēķināt ar

```
Fstats(formula, data, cov.type, from = 0.15, ...)
```

kur `from` precizē atdalīšanas parametru h. Arguments `cov.type` atļauj rēķināt F statistiku, balstītu uz heteroskedastiski robustu kovariācijas matricas novērtējumu (noklusējumā ir sfēriskas kļūdas). Iegūto objektu var attēlot grafiski kopā ar tā (asimptotiskajām) robežām, un var veikt formālas nozīmes testus *supF, aveF, expF*.

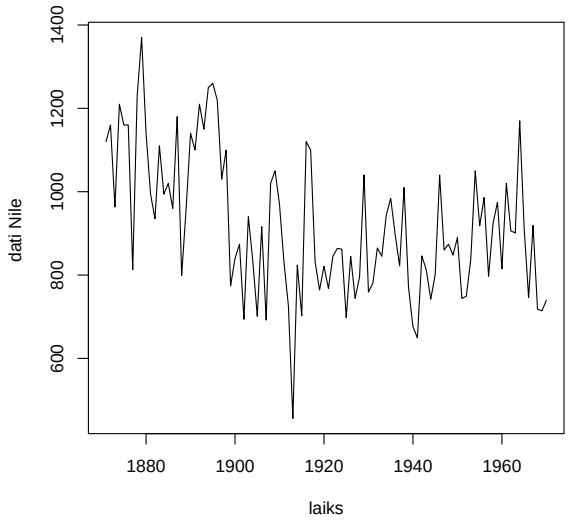
Ja ir pierādījumi par izmaiņām regresijas attiecībā, to skaitu var noteikt ar funkciju

```
breakpoints(formula, data, breaks, h = 0.15, ...)
```

kas izmanto dinamisko programmēšanu. Kopumā tā aprēķina trijstūra $rss(i, j)$ matricu. Parametrs `breaks` ir maiņas punktu m skaits, noklusējumā ir lielākais skaitlis, kuru pieļauj parametrs h.

Piemēri

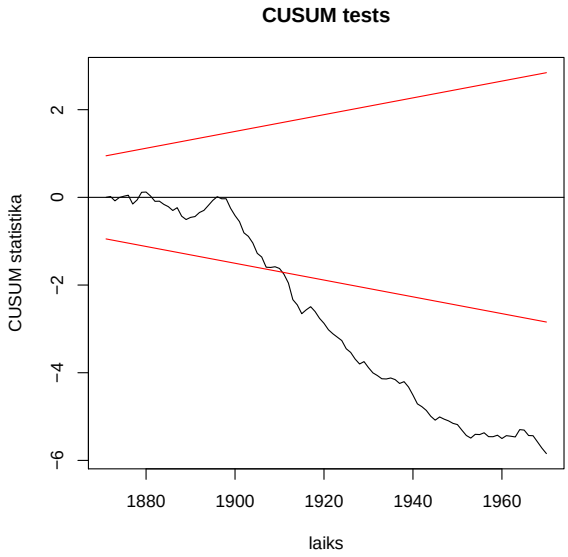
Viens no vienkāršākajiem piemēriem ir laikrinda, kuras vidējā vērtība mainās vienā pārtraukumpunktā. Tāda laikrinda ir, piemēram, upes Nīlas gada plūsmas pie Asuānas no 1871. gada līdz 1970. gadam. Tā mēra ikgadējo caurplūdi pie Asuānas $10^8 m^3$ un ir attēlota 17.. attēlā. Grafiks parāda, ka gada plūsma svārstās ap konstantu vidējo vērtību katrā segmentā - pirms 1989 un pēc, bet, ka tai ir viens pārtraukums, kurā vidējā gada plūsma samazinās sakarā ar Asuānas aizsprosta atvēršanu. Tiks pārbaudīts, vai vidējā



17. att. Gada plūsmas Nīlas upei pirms un pēc pirmā Asuānas aizsprosta atvēršanas

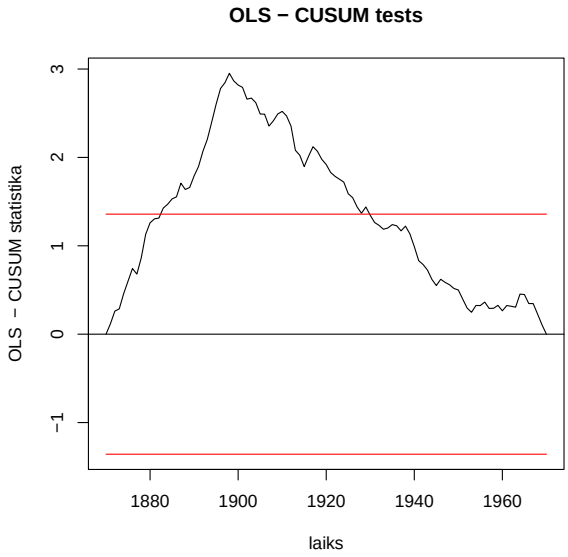
gada plūsma mainās laika gaitā, t.i., tiks pielāgota konstante datiem. Programmā R tas ir uzrakstāms kā $Nile \sim 1$, kur $Nile$ ir laikrindas objekts, kas satur datus. Lai to izdarītu, tiek izmantota R programmā pakete *strucchange*.

18. attēlā attēlots CUSUM process $Nile$ datiem ar tā robežām pie 5% nozīmības līmeņa. Pēc attēla redzams, ka CUSUM process atstāj nulles vērtību aptuveni pie 1900.



18. att. CUSUM process Nile datiem

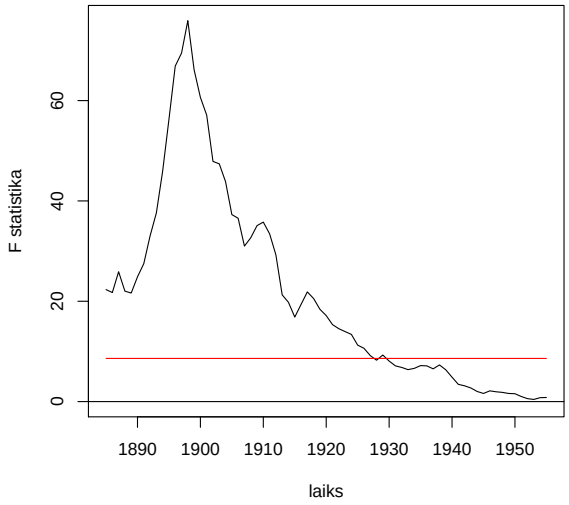
19.. attēlā attēlots OLS CUSUM process *Nile* datiem ar tā robežām pie 5% nozīmības līmeņa. Kā redzams attēlā, procesam ir pīķis (Maksimums) aptuveni pie 1900, kas



19. att. OLS CUSUM process Nile datiem

pārsniedz robežas, un tātad norāda strukturālu maiņu tajā punktā. Acīmredzams iemesls ir Asuānas aizsprosts, kas tika uzcelts 1898. gadā.

To pašu rezultātu var iegūt ar testu, izmantojot F statistiku 1.15, tas attēlots 20. Secīgi no F statistikas var iegūt optimālo maiņas punktu 2 segmentu gadījumā, jo tas ir



20. att. F statistika Nile datiem

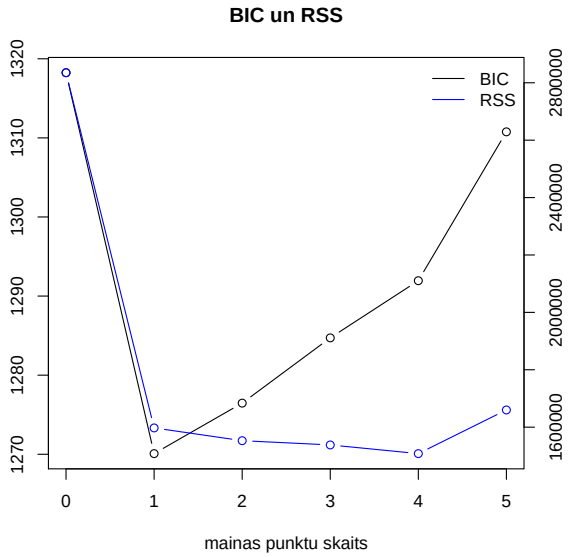
ekvivalenti maksimizēt F statistiku vai minimizēt atlikumu kvadrātu summas. Maiņas punkta novērtējumu viegli var iegūt ar `breakpoints(F_Nile)`. Kaut 2 segmentu modelis šķiet diez gan intuitīvs šiem datiem, arī tiek salīdzināts ar modeļiem ar papildus maiņas punktiem. Sekojošā komanda aprēķina patvaļīgu m segmentu modeli balstītu uz $rss(i, j)$ trijstūra matricas (noklusējmā $h = 0.15$.)

```
> bp.nile<-breakpoints(Nile~1)
```

`summary` iegūtajam objektam ziņo par maiņas punktiem $m + 1$ segmenta modelim ar $m = 0, ..., 5$ (maksimāli iespējamais ar $h = 0.15$), kā arī saistītajiem RSS un BIC. Tāds informācijas kritērijs tiek bieži izmantots modeļa izvēlei, kas šajā gadījumā nozīmē maiņas punktu m izvēli. Bai un Perron 2003 apgalvo, ka AIC parasti pārvērtē maiņas punktu skaitu, bet BIC ir piemērota izvēles procedūra daudzās situācijās. *Nile* datiem 21. grafikā parāda, ka BIC izvēlas modeli ar $m = 1$ maiņas punktu, kas apstiprina iepriekšējo testu rezultātus. Modelim maiņas punkts ir 28. novērojums vai attiecīgi 1989. gads, to var iegūt ar

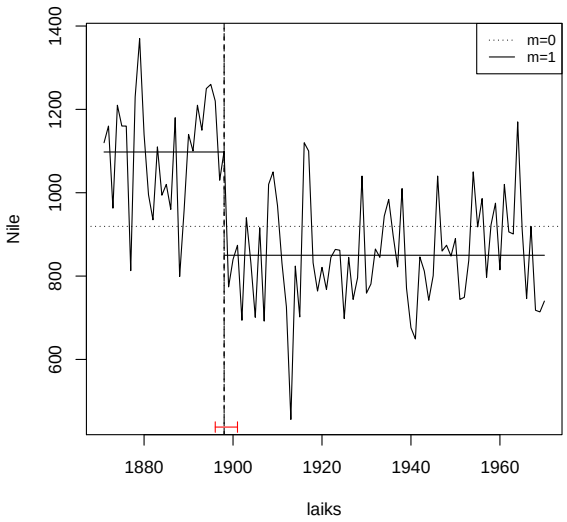
```
> bp1 <- breakpoints(bp.nile, breaks = 1)
```

Lai apkopotu rezultātus, tiek piemēroti divi lineārie modeļi datiem. Pirmais modelis ir pie nulles hipotēzes bez maiņas punktiem un otrs ir novērtētais 2 segmentu modelis. Rezultāts tiek attēlots grafiski, skatīt 22. Tajā redzamie piemērotie modeļi ar $m = 0, 1$ kopā



21. att. BIC un RSS modelim ar m maiņas punktiem

ar vertikāli raustītu līniju novērtētajam maiņas punktam un tā 90% ticamības intervālu apakšā.



22. att. Piemērotie modeļi *Nile* datiem

SECINĀJUMI

Darbā tika apskatīta maiņas punkta noteikšana gan laikrindu, gan regresijas problēmām. Tika paskatītas divas *R* programmā iebūvētas paketes *changepoint* un *strucchange*. Pirmā pakete vairāk domāta laikrindām, bet otrā regresijai. Laikrindām tiek aplūkoti gadījumi, ka izmaiņa ir tikai vidējai vērtībai, un dispersijai. Regresijai tika aplūkoti testi gan no vispārīgo svārstību testa rāmja, t. i., CUSUM un OLS CUSUM testi, gan no F testa rāmja, balstīts uz F statistiku. Metodes tika attēlotas grafiski, gan uz simulētiem datiem, gan uz reāliem datu piemēriem.

Izmantotā literatūra un avoti

[1] E. S. Page. A test for a change in aparameter occuring at an unknown point. *Biometrika*, 42, 1955.

[2] G. C. Chow. Test of equality between sets of coeficients in two linear regressions. *Econometrica*, 28, 1960.

[3] R. L. Brown and J. M. Evans. Techniques for testing the constancy og regression relationships over time. *Journal of the Royal Statistical Society B*, 37, 1975.

[4] W. Ploberger and W. Kramer. The cusum test with ols residuals. *Econometrica*, 60, 1992.

[5] C. Erdman and J. W. Emerson. A fast bayesian change point analysis for the segmentation of microarray data. *Bioinformatics*, 24(19), 2008.

[6] G. Yan, Z. Xiao, and S. Eidenbenz. Catching instant messaging worms with change-point detection techniques.in proceedings of the usenix workshop on large-scale exploits and emergent threats. 2008.

[7] D. W. Kwon, M. Vannucci, A. L. N. Reddy, and S. Kim. Wavelet methods for detection of anomalies and their application too network traffic analysis. *Quality and Reliability Engineering International*, 22, 2006.

[8] R. Jaxk, J. Chen, R. Lund X. L. Wang, and L. QiQi. a review and comparsion of chanepoint detection techniques for climate data. *Journal of Applied Meterology and Climatology*, 6, 2007.

[9] R. Killick, I. A. Jonathan, P. Jonathan, and K. Ewans. Detection of changes in the characteristics of oceanographic time-series using chanhe point analysis. *Ocean Engineering*, 37(13), 2010.

[10] Jaromir Antoch, Marie Huskova, and Daniela Jaruskova. Change point detection. *Lecture Notes of the 5th IASC Summer School*, 2000.

[11] Achim Zeileis, Christian Kleiber, Walter Krmer, and Kurt Hornik. Testing and dating of structural changes in practice. *Computational Statistics Data Analysis*, 44(1–2):109–123, 2003.

[12] Bai J and P. Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18, 2003.

[13] Achim Zeileis. strucchange: Testing for structural change in linear regression relationships. *R News*, (1/3), 2001.

[14] W. Kramer, W. Ploberger, and R. Alt. Testing for structural change in dynamic models. *Econometrica*, 56, 1988.

[15] Achim Zeileis. p values and alternative boundris for cusum tests. 2000.

[16] Achim Zeileis. *Testing for Structural Change. Theory, Implementation and Applications.* dissertation, Dortmumdes Universitāte, 2003.

[17] D. W. K.Andrews. Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61, 1993.

[18] D. W. K. Andrews and W. Ploberger. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, 62, 1994.

[19] Claudia Kirch and Dimitris N. Politis. Tft-bootstrap:resampling time series in the frequency domain to obtain replicates in the time domain. *The Annals of Statistics*, 39, 2011.

[20] D. V. Hinkley. Inference about the change-point in a sequence of random variables. *Biometrika*, 57, 1970.

[21] A. K. Gupta and J. Tang. On testing homogeneity of variances for gaussian models. *Journal of Statistical Computation and Simulation*, 27, 1987.

[22] *Parametric statistical change point analysis.* Birkhauser, 2000.

[23] C. Inclan and G. C. Tiao. Use of cumulative sums of squares for retrospective detection of changes of variance. *JASA*, 89(427), 1994.

[24] R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107, 2012.

[25] Rebecca Killick and Idris A. Eckley. changepoint: An r package for changepoint analysis. 2011.

[26] E. S. Page. Continuous inspection schemes. *Biometrika*, 41, 1954.

[27] G. W. Cobb. The problem of the nile: Conditional solution to a changepoint problem. *Biometrika*, 65, 1978.

[28] G. W. Cobb. Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association*, 92, 1997.

PIELIKUMS

R programmas kods, izmantojot *changpoint* paketi

```
# dipersijas izmaina viena punkta
# tiek simuleti dati
set.seed(1)
var1=c(rnorm(50,0,1),rnorm(150,0,sqrt(3)))
plot(var1,type='l', xlab="indekss")

library(changepoint)
# pie normalitates pienemuma tiek noteikts mainas punkts
var1.amoc=cpt.var(var1)
plot(var1.amoc,type='l',cpt.width=3,xlab="indekss", ylab="var1")
cpts(var1.amoc)

#dispersijas pie normalitates pienemuma
param.est(var1.amoc)

# ar CSS metode tiek noteikts mainas punkts
var1.css=cpt.var(var1,dist='CSS')
cpts(var1.css)

#ar CSS metodi dispersijas
param.est(var1.css)

#dispersijas maina vairakos punktos
#datu simulacija
set.seed(10)
var2=c(rnorm(100,0,1),rnorm(100,0,2),rnorm(100,0,10),rnorm(100,0,9))
plot(var2,type='l',xlab="indekss", ylab="var2")

#ar noklusejuma sodu
var2.pelt=cpt.var(var2,method='PELT')
plot(var2.pelt,type='l',cpt.width=3,cpt.col='blue',xlab="indekss", ylab="var2")
cpts(var2.pelt)

#ar manuali noraditu sodu
var2.manual=cpt.var(var2,method='PELT',penalty='Manual',value='2*log(n)')
plot(var2.manual,type='l',cpt.width=3,xlab="indekss", ylab="var2")
cpts(var2.manual)

#dispersija pie manuali noradita soda
param.est(var2.manual)

#ar CSS pie manuali noradita soda
var2.css=cpt.var(var2,method='BinSeg',penalty='Manual',value='log(2*log(n))',dist='CSS')
cpts(var2.css)

#dispersijas ar CSS metodi pie manuali noradita soda
param.est(var2.css)
```

```
# reals datu piemers
#tiek ieguti dati un mekleti mainas punkti ar PELT metodi
dati<-data(ftse100)
ftse.pelt=cpt.var(ftse100[,2],method='PELT')
cpts(ftse.pelt)

#tiek mekleti mainas punkti ar manuali ievaditu sodu
ftse.manual=cpt.var(ftse100[,2],method='PELT',penalty='Manual',value='2*log(n)')
plot(ftse.manual,type='l',ylab="atmaksājamā summa", xlab="indekss")
ftse100[cpts(ftse.manual),1]

#tiek aprekinatas dispersija
param.est(ftse.manual)

# dipersijas izmaina viena punkta
# tiek simuleti dati
set.seed(1)
var1=c(rnorm(50,0,1),rnorm(150,0,sqrt(3)))
plot(var1,type='l', xlab="indekss")

library(changepoint)
# pie normalitates pienemuma tiek noteikts mainas punkts
var1.amoc=cpt.var(var1)
plot(var1.amoc,type='l',cpt.width=3,xlab="indekss", ylab="var1")
cpts(var1.amoc)

#dispersijas pie normalitates pienemuma
param.est(var1.amoc)

# ar CSS metode tiek noteikts mainas punkts
var1.css=cpt.var(var1,dist='CSS')
cpts(var1.css)

#ar CSS metodi dispersijas
param.est(var1.css)

#dispersijas maina vairakos punktos
#datu simulacija
set.seed(10)
var2=c(rnorm(100,0,1),rnorm(100,0,2),rnorm(100,0,10),rnorm(100,0,9))
plot(var2,type='l',xlab="indekss", ylab="var2")

#ar noklusejuma sodu
var2.pelt=cpt.var(var2,method='PELT')
plot(var2.pelt,type='l',cpt.width=3,cpt.col='blue',xlab="indekss", ylab="var2")
cpts(var2.pelt)

#ar manuali noraditu sodu
var2.manual=cpt.var(var2,method='PELT',penalty='Manual',value='2*log(n)')
plot(var2.manual,type='l',cpt.width=3,xlab="indekss", ylab="var2")
cpts(var2.manual)
```

```
#dispersija pie manuali noradita soda
param.est(var2.manual)

#ar CSS pie manuali noradita soda
var2.css=cpt.var(var2,method='BinSeg',penalty='Manual',value='log(2*log(n))',dist='CSS')
cpts(var2.css)

#dispersijas ar CSS metodi pie manuali noradita soda
param.est(var2.css)

# reals datu piemers
#tiek ieguti dati un mekleti mainas punkti ar PELT metodi
dati<-data(ftse100)
ftse.pelt=cpt.var(ftse100[,2],method='PELT')
cpts(ftse.pelt)

#tiek mekleti mainas punkti ar manuali ievaditu sodu
ftse.manual=cpt.var(ftse100[,2],method='PELT',penalty='Manual',value='2*log(n)')
plot(ftse.manual,type='l',ylab="atmaksājamā summa", xlab="indekss")
ftse100[cpts(ftse.manual),1]

#tiek aprekinatas dispersija
param.est(ftse.manual)

# Izmainas dispersijai un videjai vertibai
# tiek simuleti dati
set.seed(10)
meanvar=c(rexp(50,rate=1),rexp(50,rate=5),rexp(50,rate=2),rexp(50,rate=7))

#tiek izmantota PELT metode
meanvar.pelt=cpt.meanvar(meanvar1,dist='Exponential',method='PELT')
plot(meanvar.pelt,type='l',cpt.width=3, xlab="indekss", ylab="meanvar")
cpts(meanvar.pelt)
# atrod videjas vertibas katra segmenta
param.est(meanvar.pelt)

# reals datu piemers
data(HC1)
hc1.pelt=cpt.meanvar(HC1,method='PELT',penalty='Manual',value=14)
plot(hc1.pelt,type='l',ylab='G+C sastāvs',xlab='HC1 pozīcija',cpt.width=3)
ncpts(hc1.pelt)
```

R programmas kods, izmantojot *strucchange* paketi

```
setwd('D:/Vineta/5. kurss/Diplomdarbs')

require(strucchange)
#dati Nile
#OLS CUSUM tests
OLS_CUSUM<-efp(Nile~1, type="OLS-CUSUM")
plot(OLS_CUSUM)
```

```
plot(OLS_CUSUM, alt.boundary=TRUE)

#CUSUM tests
CUSUM<-efp(Nile~1, type="Rec-CUSUM")
plot(CUSUM)
# F tests

F_Nile<-Fstats(Nile~1)
plot(F_Nile)

# pātraukumpunkti
breakpoints(F_Nile)
bp.nile<-breakpoints(Nile~1)
summary(bp.nile)
plot(bp.nile, xlab="maiņas punktu skaits", main="BIC un RSS")
bp1<-breakpoints(bp.nile,breaks=1)

# grafiks piemerotajiem modeliem Nile datiem
fm0.nile<-lm(Nile~1)
coef(fm0.nile)
nile.fac<-breakfactor(bp1)
fm1.nile<-lm(Nile~nile.fac - 1)
g<-coef(fm1.nile)
plot(Nile~1,xlab="laiks")
abline(v=1898,lty=5)
abline(fm0.nile,lty=3)
segments(1871,g[1],1898,g[1])
segments(1898,g[2],1970,g[2])
lines(confint(bp.nile, breaks = 1, level = 0.9))
legend("topright", c("m=0","m=1"), cex=0.8, lty=c(3,1))

sctest(Nile~1)
```

Diplomdarbs “Maiņas punkta noteikšana matemātiskās statistikas problēmās” 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: Vineta Vītola

_____ _____

(paraksts) (datums)

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

_____ _____

(paraksts) (datums)

Recenzents:

_____ _____

(paraksts) (datums)

Darbs iesniegts Matemātikas nodaļā _____

(datums)

_____

(darbu pieņēma)

Darbs aizstāvēts valsts pārbaudījuma komisijas sēdē

_____ prot. Nr. _____, vērtējums_____

(datums)

Komisijas sekretārs/-e: _____ _____

(Vārds, Uzvārds) (paraksts)