Maiņas punkta noteikšana laikrindu analīzē

Agris Vaselāns

Latvijas Universitāte

2011. gada 10.novembrī.

Maiņas punkta analīze

Galvenie jautājumi

- Vai ststistiskajā modelī notikušas izmaiņas?
- Kad modelī notikušas izmaiņas?
- Vai maiņas punkts ir tikai viens?
- Cik maiņas punkti bijuši?
- utt.

Definīcija

Laika momentu, kad statistiskajā modelī notikušas izmaiņas sauc par maiņas punktu.

Spektrs (spektrālā blīvuma funkcija)

Teorēma

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

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

Periodogramma un DFT

Definīcija

Ja doti dati $x_1, x_2, \ldots, x_n$, tad par diskrēto Furjē transformāciju (DFT) sauc

$$d(\omega_j) = n^{-1/2} \sum_{t=1}^{n} x_t e^{-2\pi i \omega_j t} = a(j) + ib(j)$$

$$= n^{-1/2} \sum_{t=1}^{n} V(t)(\cos(2\pi\omega_j t) + i\sin(2\pi\omega_j t))),$$

kur $j = 0, 1, \ldots, n-1$ un frekvences $\omega_j = j/n$ tiek sauktas par Furjē jeb fundamentālajām frekvencēm, a(j) un b(j) sauc par Furjē koeficientiem.

Periodogramma un DFT

Definīcija

Ja doti dati $x_1, x_2, \ldots, x_n$, tad par periodogrammu sauc

$$I(\omega_j) = |d(\omega_j)|^2 = \frac{n}{2}(a(j)^2 + b(j)^2)$$

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

Īpašības

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

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

Tātad, ja stohastiskais process ir Gausa, tad a(j) un b(j) ir asimptotiski i.i.d. un sadalīti pēc $N(0, f(\omega_j)/2)$.

Kodolu gludinātā periodogramma

Tā kā periodogramma nav spektra būtisks novērtējums, tad nepieciešams definēt gludināto periodogrammu.

Definīcija

Par kodolu gludināto periodogrammu sauc

$$\hat{\mathrm{f}}(\omega) = \frac{\sum_{\mathrm{j}\in\mathcal{Z}} \mathrm{K}(\frac{\omega-\omega_{\mathrm{j}}}{\mathrm{h}})\mathrm{I}(\mathrm{j})}{\sum_{\mathrm{j}\in\mathcal{Z}} \mathrm{K}(\frac{\omega_{\mathrm{j}}}{\mathrm{h}})\mathrm{I}(\mathrm{j})},$$

kur $\mathrm{K}(\cdot)$ ir kodolsl, $\mathrm{I}(\mathrm{j})$ ir periodogramma frekvencē ω_{j}.

TFT (Time Frequency Toggle Bootstrap) Kirch. C., Politis D.N., 23.12.2010.

Procedūras apraksts

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

- Aprēķina Furjē koeficientus, izmantojot FFT (fast fourier transform)
- Apzīmē $a(T)^* = b(T)^* = 0$ (Ja T pāra, tad $a(T/2)^* = b(T/2)^* = 0$.
- Izmantojot kādu no butstrapa procedūrām, iegūst 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 $z_t^* = x_t^* - \mu_t^*$.

Procedūras apraksts

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

- Ģenerē standarta normālā sadalījuma 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, ka

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

Maiņas punkta analīzes piemērs

Hipotēžu pārbaude vidējai vērtībai (skalācijas parametram)

$$H_0 : X_i = a + e_i,\ i = 1, 2, \ldots, M$$

$$H_1 : \exists m \in \{1, 2, \ldots, M\},$$

ka

$$X_i = a + e_i,\ i = 1, 2, \ldots, m,$$

$$X_i = a + \delta + e_i,\ i = m + 1, 2, \ldots, M,$$

turklāt e_i ir iid.

Piezīme

Šādu m tad sauc par procesa maiņas punktu.

Piemērs TFT butstrapam

Vidējās vērtības izmaiņas laikrindās (AMOC - at most one change modelis)

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

kur $V(\cdot)$ ir stacionārs process ar $EV(\cdot) = 0$, turklāt k, μ_1, μ_2 nav zināmi.

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

Tad vienkāršākā, arī saukta par CUMSUM, statistika ir

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

Vidējās vērtības izmaiņas laikrindās (AMOC - at most one change modelis)

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

$$\frac{\mathrm{C_T}}{\tau} \rightarrow^{\mathrm{L}} \sup_{0 \leq \mathrm{t} \leq 1} |\mathrm{B(t)}|,$$

kur $\mathrm{B}(\cdot)$ ir Brauna tilta process un $\tau^2 = \mathrm{f}(0)$, kur $\mathrm{f}(\cdot)$ ir spektrālā blīvuma funkcija procesam $\{\mathrm{V}(\cdot)\}$. Nesen izstrādāta efektīva TFT butstrapa metode šai problemātikai (2011, Kirch C., Politis D.N.).

Lietojot TFT-bootstrapu

Novērtējam $Z(t) = V(t) - \overline{V_T}$ un
$\hat{Z}(t) = Y(t) - \hat{\mu}_1 I_{\{t \le \hat{k}\}} - \hat{\mu}_2 I_{\{t > \hat{k}\}}$,
kur $\hat{k} = \text{argmax}\{|\sum_{j=1}^{k}(Y(j) - \overline{Y_T})| : 1 \le k \le T\}$,
$\hat{\mu}_1 = 1/\hat{k} \sum_{j=1}^{\hat{k}} Y(j)$, $\hat{\mu}_2 = 1/(T - \hat{k}) \sum_{j=\hat{k}+1}^{T} Y(j)$.

Vidējo vērtību problemātika (AMOC - at most one change model)

Tad butstrapa statistiku definējam

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

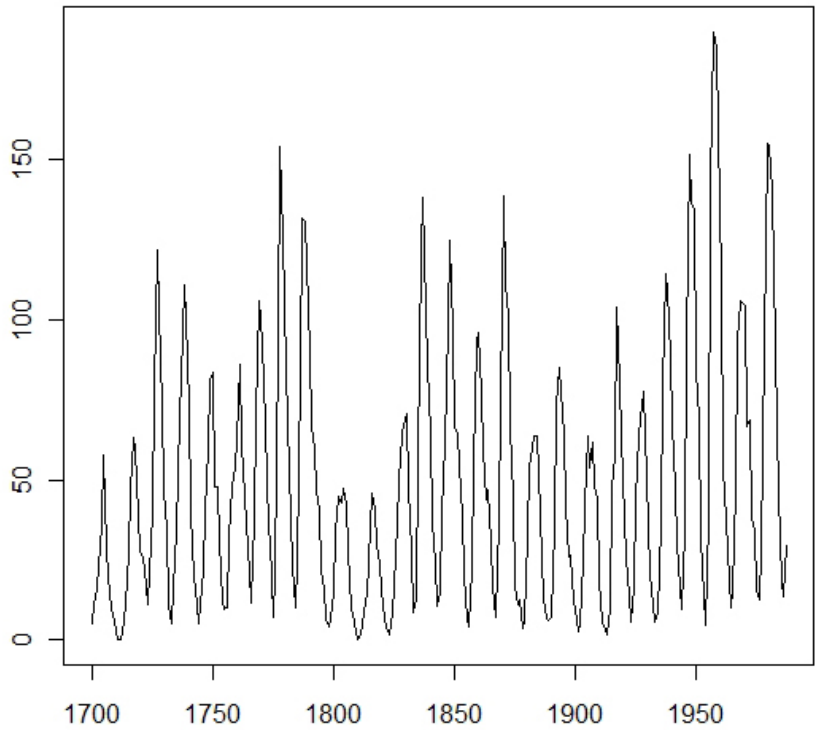
kur $\{Z^*(\cdot)\}$ir bootstrapotie g.l., kas iegūti ar boostrapa shēmas palīdzību.

TFT-bootstrapa pielietojums

Maiņas punkts saules aktivitātes datos(AMOC - at most one change model)

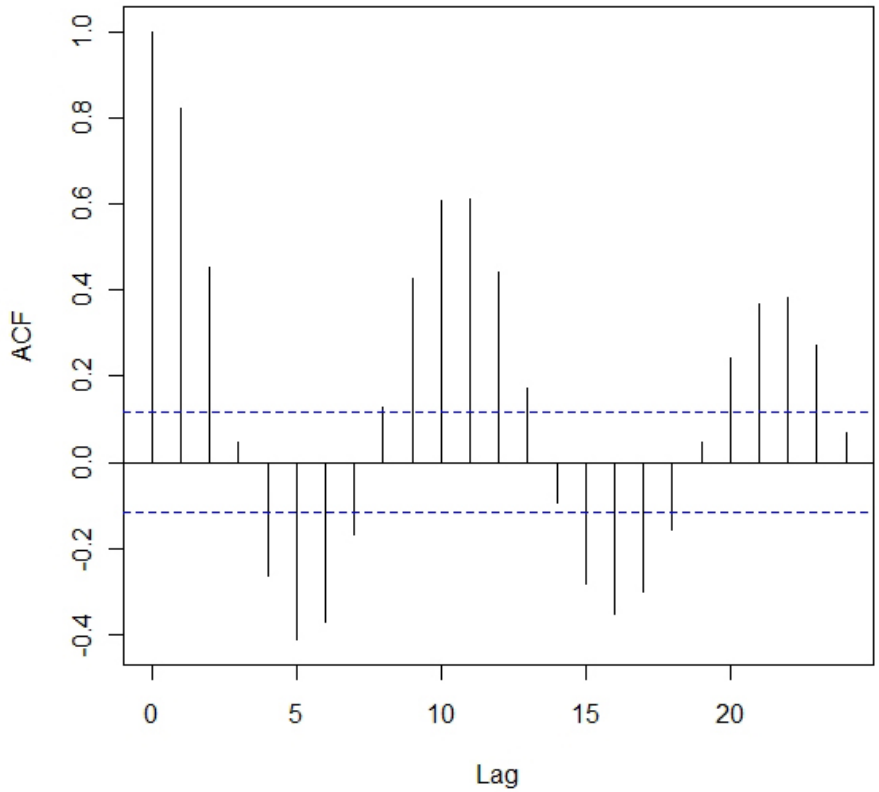
Saules plankumu (aktivitātes) dati (laikaposmā no 1700. līdz 1987.g.). Pielietojot TFT redzams, ka pastāv maiņas punkts 1935. gadā.

Sunspot Data un TFT



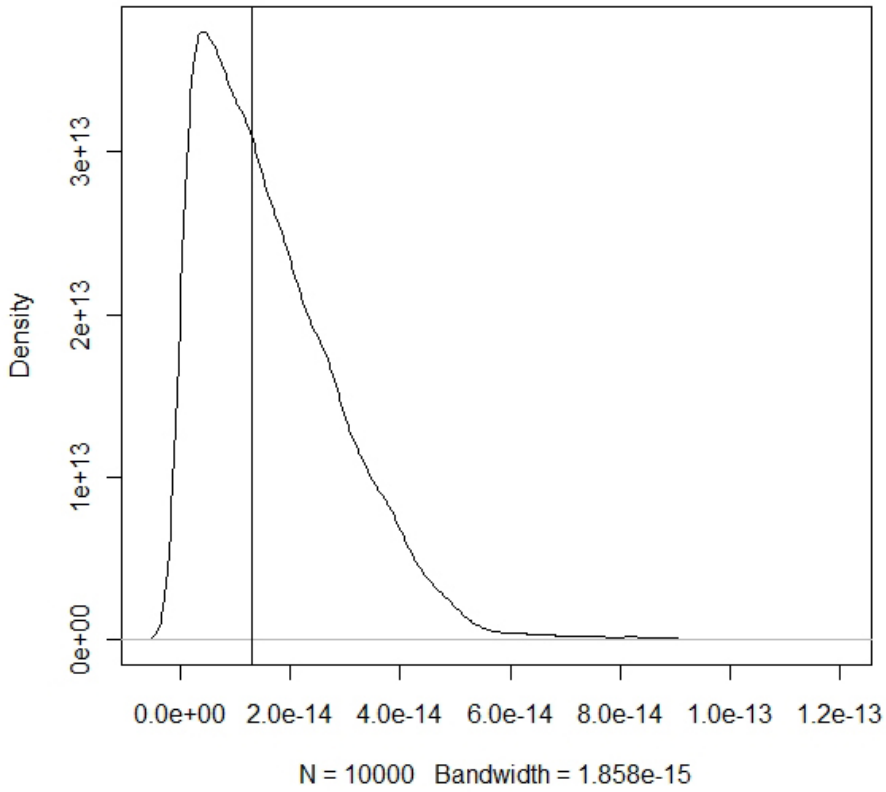
att.: Sunspot data

Sunspot Data un TFT



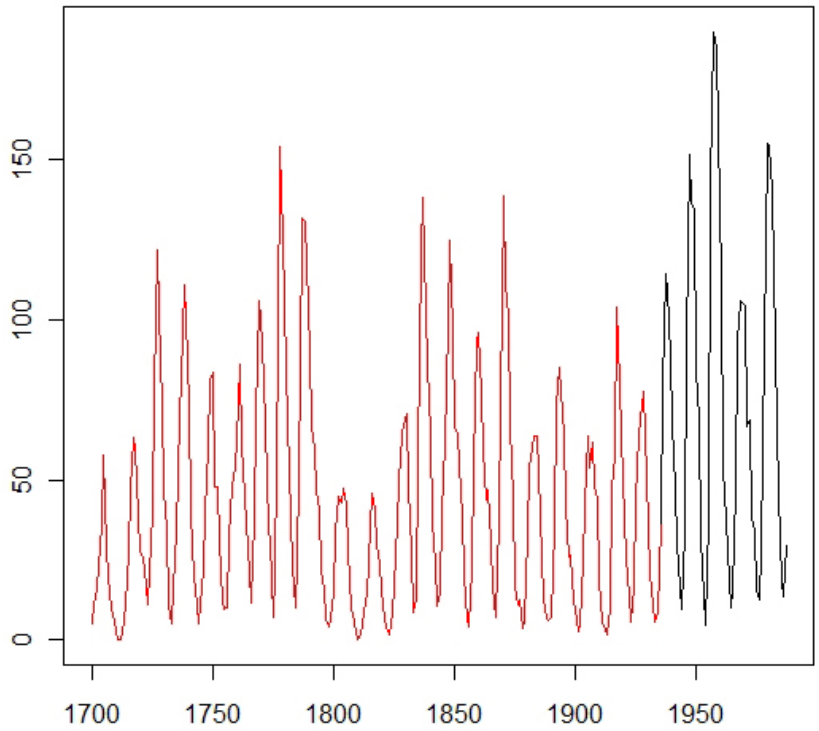
att.: Autocorrelations

Sunspot Data un TFT



att.: Test statistic

Sunspot Data un TFT



att.: Seperated data series.

Citas iespējas

Maiņas punkta analīze normāla sadalījuma gadījumā

Iegūtas kritiskās vērtības un sarežģīti asimptotiskie sadalījumi. Beijesa metode.

Maiņas punkta analīze regresijas modeļos

Izmaiņas regresijas koeficientos

Maiņas punkta analīze vispārīgā gadījumā ar korelāciju atlikumos

Izmaiņas vidējā vērtībā un/vai skalācijas parametros.

Literatūra

- Antoch J. et al. Change point detection
- Kirch C., Politis D.N. TFT-bootstrap: Resampling time series in the frequency domain to obtain replicates in the time domain
- Kirch C. Resampling in the frequency domain of time series to determine critical values for change point tests.
- Taylor W.A. Change point analysis: a powerful new tool for detecting changes

Cita pieeja

Motivācija

1. Eksistē samērā daudz risinājumu divu izlašu lokācijas parametra starpības pārbaudei neatkarīgu datu gadījumā (Divu izlašu t-tests, Mana-Whitneya tests u.c.)
2. Jauni pētījumi un risinājumi neparametriskajā statistikā atkarīgu datu struktūru gadījumā (Zhang, 2011.)
3. Maiņas punkta analīze atkarīgiem datiem (R.Fried, 2011)
4. Maiņas punkta analīzes un empīriskās ticamības funkcijas (EL) metožu salīdzinājums divu izlašu lokācijas parametru starpības pārbaudei.

EL divu izlašu gadījumā

Nosacījumi

- Apskatīsim $X_1, X_2, \ldots, X_{n1} \sim^{iid} F_1(x, \theta_0, t)$ un $Y_1, Y_2, \ldots, Y_{n2} \sim^{iid} F_2(y, \theta_1, t)$
- Apskatīsim $t \to \Delta(t)$, $t \in T \subset \mathbb{R}$, $\theta(t)$ un, fiksējo t, $\theta(t) := \theta$, $\Delta(t) := \Delta$
- Visu nepieciešamo informāciju par parametriem iegūsim no nenovirzītiem funkcionāļiem:

$$E_{F_1} g_1(X, \theta_0, t) = 0, \quad E_{F_2} g_2(Y, \theta_1, t) = 0.$$

Apgalvojums

Ja $\Delta_0 = \theta_1 - \theta_0$, tad iegūstam tieši problemātiku vidējo vērtību starpībai. Apzīmējot $\theta_0 = \int x dF_1(x)$ un $\Delta_0 = \int y dF_2(y) - \int x dF_1(x)$ iegūsim funkcionāļus

$$g_1(X, \theta_0, \Delta_0, t) = X - \theta_0, \quad g_2(Y, \theta_0, \Delta_0, t) = Y - \theta_0 - \Delta_0.$$

EL divu izlašu gadījumā - TESTA STATISTIKA

Definīcija

Definēsim empīrisko ticamības funkciju attiecību funkciju

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

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 $\sum_{i=1}^{n_1} p_i g_1(X_i, \theta_0, t) = 0$ un $\sum_{j=1}^{n_2} q_j g_2(Y_j, \theta_1, t) = 0$.
Viens vienīgs atrisinājums eksistē, lai 0 piederētu izliektajai čaulai, ko veido $g_1(X_i, \theta, \Delta, t)$ un $g_2(Y_j, \theta, \Delta, t)$. Maksimumu atrod, izmantojot Lagranža reizinātājus

EL divu izlašu gadījumā - TESTA STATISTIKA

Definīcija

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

$$\mathrm{W}(\Delta,\theta) = -2\log \mathrm{R}(\Delta,\theta) =$$

$$= 2\sum_{\mathrm{i}=1}^{\mathrm{n}_1}(1+\lambda_1(\theta)\mathrm{g}_1(\mathrm{X}_\mathrm{i},\theta,\Delta,\mathrm{t})) + 2\sum_{\mathrm{j}=1}^{\mathrm{n}_2}(1+\lambda_2(\theta)\mathrm{g}_2(\mathrm{Y}_\mathrm{j},\theta,\Delta,\mathrm{t})).$$

Lai iegūtu parametru novērtējumu , maksimizē $\mathrm{R}(\Delta,\theta)$

Apgalvojums

Tad punktveida ticamības intervāls parametram Δ katram fiksētam $\mathrm{t} \in \mathrm{T}$ nosakāms no $\Delta : \{\mathrm{R}(\Delta,\hat{\theta}) > \mathrm{c}\}$ visiem Δ_0, kur c nosakāms no teorēmas.

EL divu izlašu gadījumā - TESTA STATISTIKA

Teorēma

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

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

kur $k < \infty$ ir pozitīva konstante, kura apmierina $n_2/n_1 \to k$, kad $n_1, n_2 \to \infty$, turklāt

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

EL divu izlašu gadījumā - Bloku veidošana atkarīgiem datiem

Bloku izveide

- M un L veseli skaitļi, 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
  $B_i = (X_{(i-1)L+1}, \ldots, X_{(i-1)L+M})$, $i = 1, \ldots, Q$.
- tad 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 attēlojams $\phi_M : \mathbb{R}^M \times \theta \to \mathbb{R}$ uzdots šādā formā

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

Vispārīgā problēmas nostādne

Hipotēžu pārbaude divu izlašu lokācija parametru starpībai

$$\mathrm{H}_0 : \Delta_0 = 0,\ \mathrm{H}_1 : \Delta_0 \neq 0$$

un statistiku

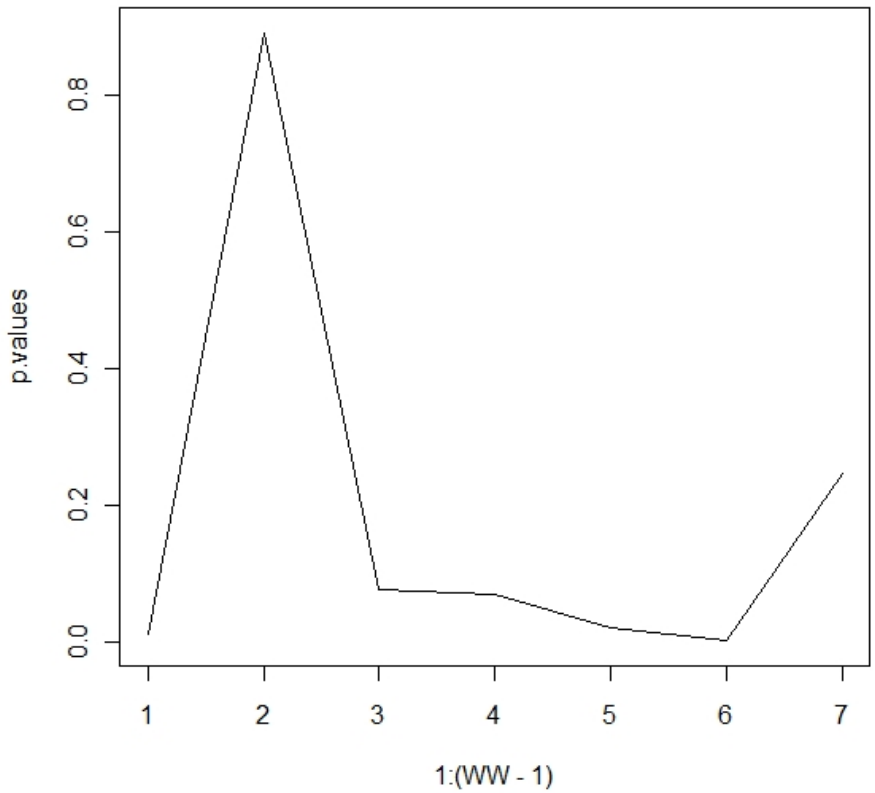
$$-2 \log \mathrm{R}(\Delta_0, \hat{\theta}) \to_{\mathrm{d}} \chi^2_1,\ \mathrm{Q}_1, \mathrm{Q}_2 \to \infty, \forall \mathrm{t} \in \mathrm{T},$$

jāņem vērā, ka mēs apskatīsim statistiku bloku izlasēm.

Ideja

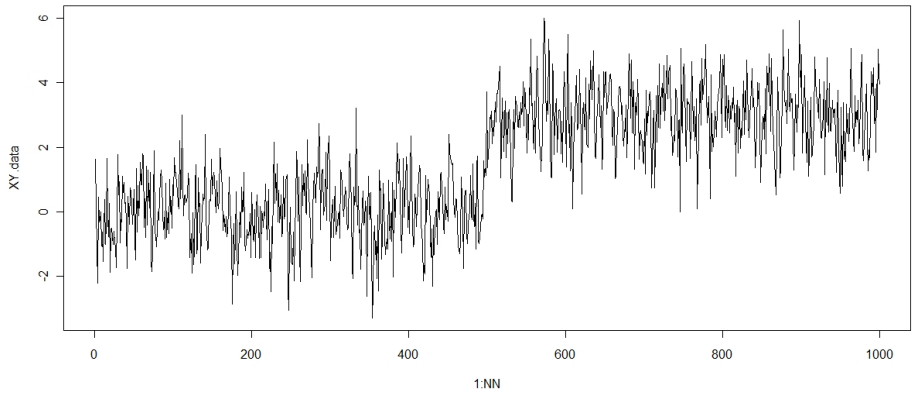
Divu izlašu vietā aplūkosim laikrindu un divus logus. Slīdošos logus uzskatīsim par izlasēm un apskatīsim p-vērtību grafikus. Punktā, kur grafiks sasniedz pietiekami mazas vērtības, laikrindā var būt maiņas punkts.

Sunspot Data un EL



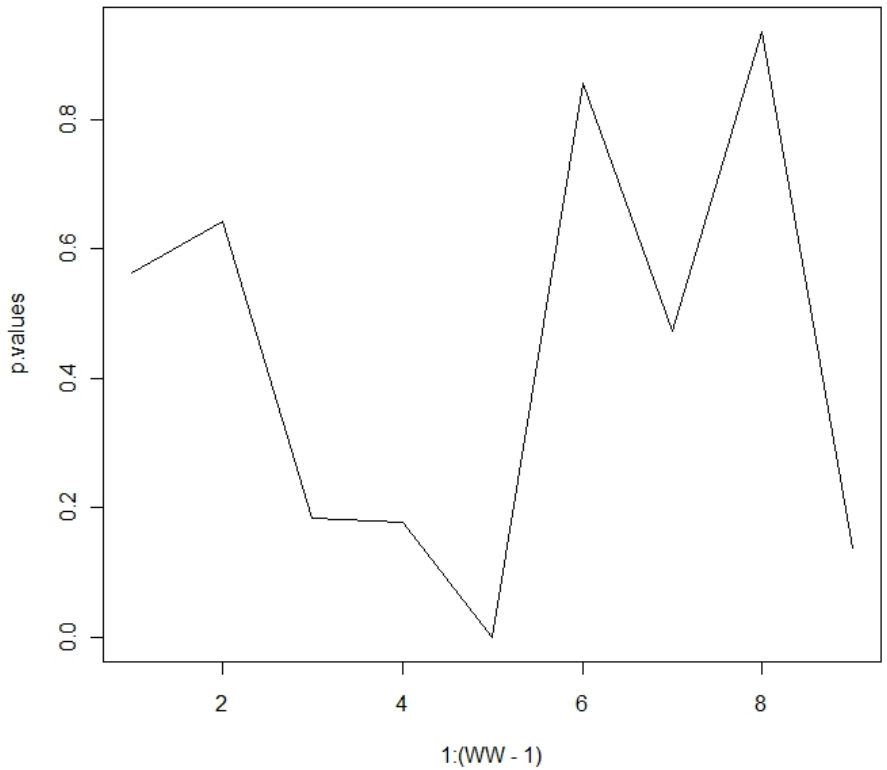
att.: Test statistic

Simulācijas AR1 ($\theta = 0.3$)



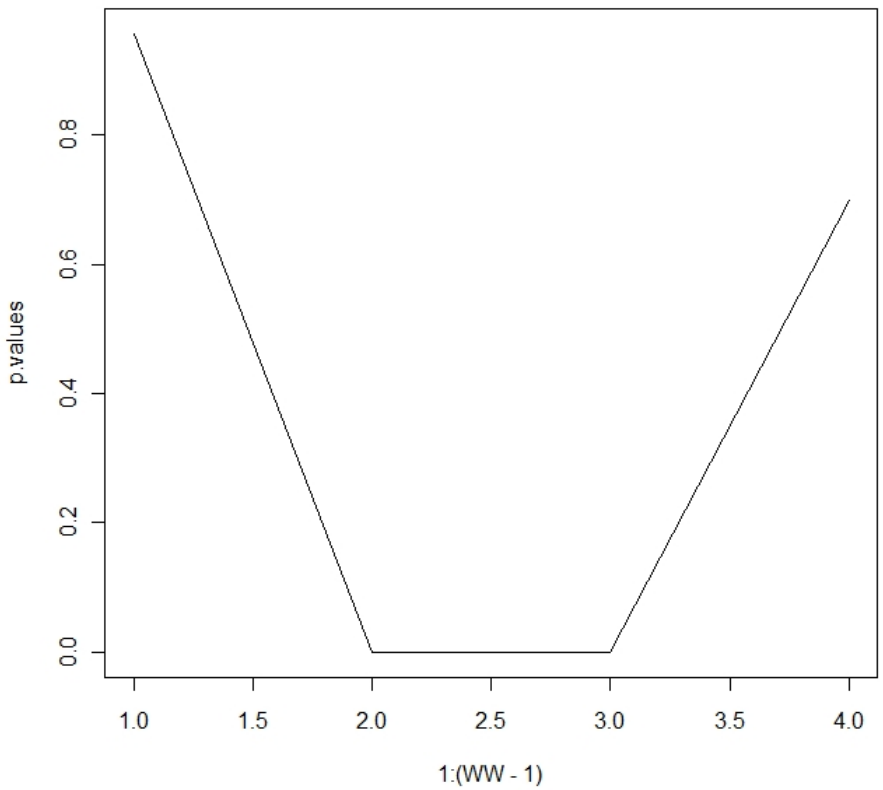
att.: Laikrinda ar shiftu

Simulācijas AR1 ($\theta = 0.3$)



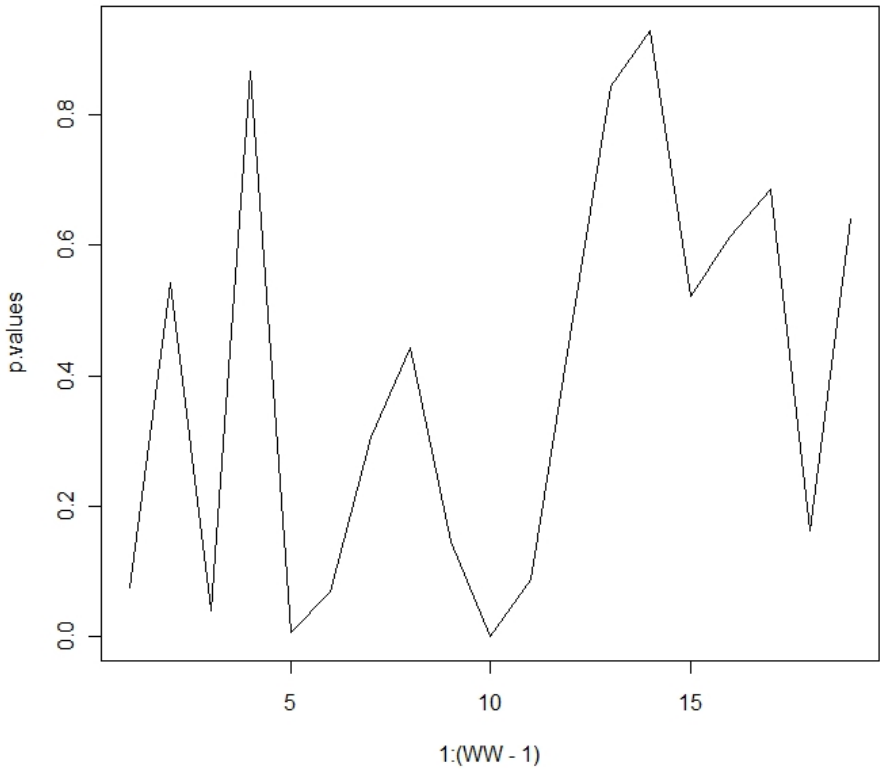
att.: Samērā laba izvēle

Simulācijas AR1 ($\theta = 0.3$)



att.: Pārāk liela loga platuma izvēle

Simulācijas AR1 ($\theta = 0.3$)



att.: Pārāk maza loga platuma izvēle

Paldies par uzmanību!

Jautājumi?