Metody repróbkowania dla niestacjonarnych modeli stochastycznych

mgr Rafał Synowiecki

Jubileusz 10-lecia Wydziału Matematyki Stosowanej AGH
10 października 2008

Moja grupa badawcza

dr hab. Jacek Leśkow,
WMS AGH i WSB-NLU

mgr Anna Dudek, WMS AGH
mgr Łukasz Lenart, WSB-NLU
mgr Rafał Synowiecki, WMS AGH

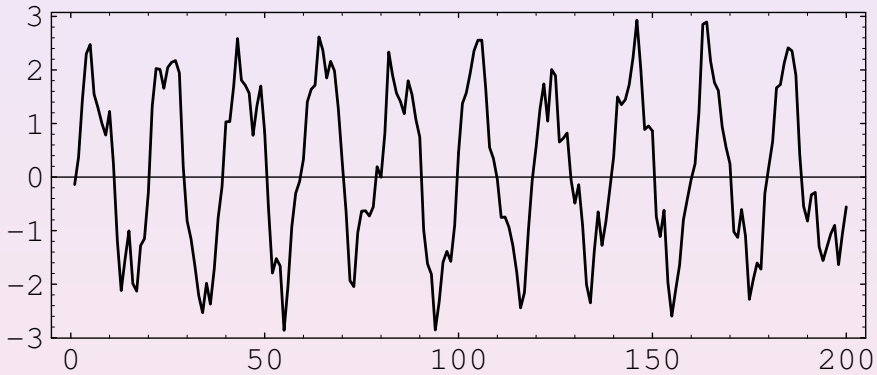
Plan prezentacji

- Niestacjonarne modele stochastyczne
  - modele OS i POS, dziedzina czasu
  - modele OS i POS, dziedzina częstotliwości
  - modele oparte na okresowych procesach liczących
- Rezultaty graniczne
- Dlaczego repróbkowanie?
- Wybrane rezultaty
- Bibliografia

Modelowanie stochastyczne

$\{X_t : t \in \mathbb{R}\}$ - proces stochastyczny (rodzina zmiennych losowych parametryzowana parametrem ciągłym)

$\{X_t : t \in \mathbb{R}\}$ - szereg czasowy (rodzina zmiennych losowych parametryzowana parametrem ciągłym)



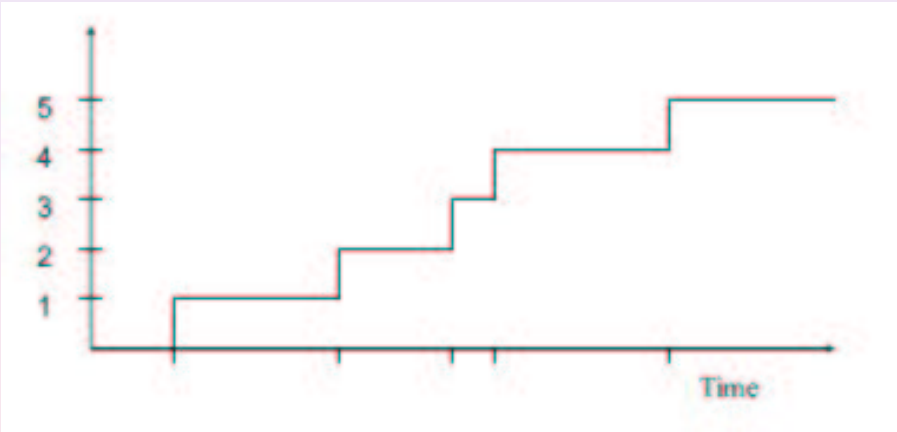
Przykłady:

- temperatura powietrza
- ceny spot energii elektrycznej
- sygnały telekomunikacyjne

Modelowanie stochastyczne

$\{N_t : t \geqslant 0\}$ - proces liczący to proces stochastyczny spełniający warunki:

- $N_t \in \mathbb{N} \cup \{0\}$
- $s < t \Rightarrow N_s \leqslant N_t$



Przykłady:

- liczba urodzeń w danym szpitalu
- samochody przejeżdżające przez skrzyżowanie

Podejście w dziedzinie czasu

Szereg czasowy $\{X_t : t \in \mathbb{Z}\}$ jest (prawie) okresowo skorelowany jeśli

$$\mu_X(t) = E(X_t)$$

i funkcja autokowariancji

$$B_X(t, \tau) = \mathrm{cov}(X_t, X_{t+\tau})$$

są (prawie) okresowe względem zmiennej t dla każdego $\tau \in \mathbb{Z}$. Wtedy

$$\mu_X(t) = \sum_{\gamma \in \Gamma} b(\nu) e^{i\gamma t},$$

$$B_X(t, \tau) = \sum_{\lambda \in \Lambda} a(\lambda, \tau) e^{i\lambda\tau}.$$

Podejście w dziedzinie czasu:

$$\hat{b}_n(\gamma) = \frac{1}{n-\tau} \sum_{t=1}^{n} X(t) e^{-i\gamma t},$$

$$\hat{a}_n(\lambda, \tau) = \frac{1}{n-\tau} \sum_{t=1}^{n-\tau} X(t+\tau) X(t) e^{-i\lambda t}.$$

Badamy wlasności nieznanych funkcji μ_X, B_X poprzez estymatory $\hat{b}_n(\gamma), \hat{a}_n(\lambda, \tau)$.

Podejście w dziedzinie częstotliwości

Harmonizowalny szereg czasowy $\{X(t) : t \in \mathbb{Z}\}$

$$X(t) = \int_0^{2\pi} e^{i\xi t} Z(d\xi).$$

Miara bispektralna jest zdefiniowana jako

$$R((a,b] \times (c,d]) = E[(Z(b) - Z(a))(Z(d) - Z(c))],$$

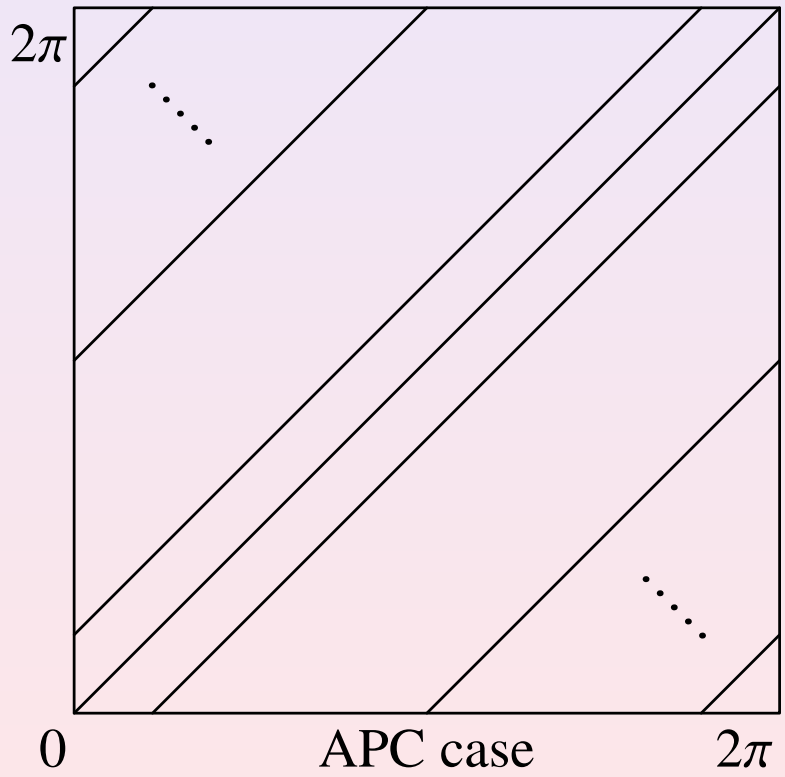
z nośnikiem

$$S = \bigcup_{\lambda \in \Lambda} \{(\xi_1, \xi_2) \in (0, 2\pi]^2 : \xi_2 = \xi_1 \pm \lambda\}.$$

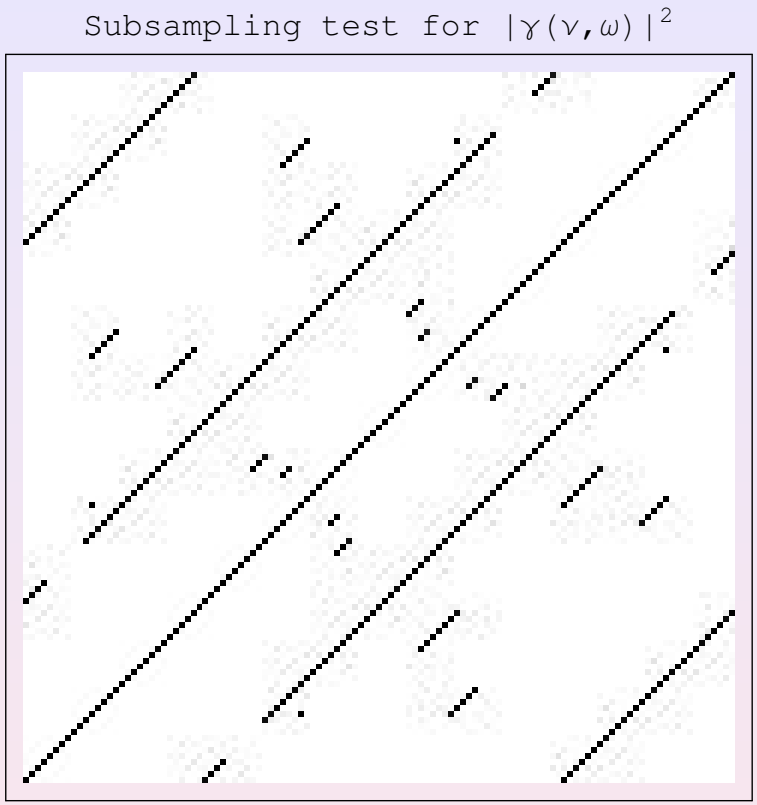
Estymator gęstości spektralnej

$$\hat{G}_n(\nu,\omega) = \frac{1}{2\pi n}\sum_{t=1}^{n}\sum_{s=1}^{n} K_n(s-t)X_tX_s e^{-i\nu t}e^{i\omega s}.$$

Linie nośnika



Przykład symulacyjny



$$X_t = (2 + \sin(2\pi t/4))\, Y_{t-1} + Y_t,$$

gdzie Y_t są niezależne z rozkładu $N(0, 1)$.

Niestacjonarny proces liczący

N - proces liczący na przedziale $[0, A]$. Definujemy funkcję intensywności λ

$$P\left(N(t+dt) - N(t) = 1 | \mathcal{F}_t\right) = \lambda(t)dt.$$

W modelu multiplikatywnej intensywności

$$\lambda(t) = \lambda_0(t)Y(t), \qquad t \in [0, A]$$

- λ_0 – nieujemna funkcja okresowa
- Y – nieujemmny prognozowalny proces stochastyczny

Przykłady:

- $Y(t)$ - liczba osób narażonych na ryzyko, $\lambda_0(t)$ - intensywność śmierci (analiza przeżycia)
- $Y(t)$ - liczba włączonych komputerów, $\lambda_0(t)$ - intensywność wysyłania pakietów

Estymator sitowy funkcji $\lambda_0(t)$

Histogramowy estymator największej wiarygodności funkcji okresowej $\lambda_0(t)$ ma postać:

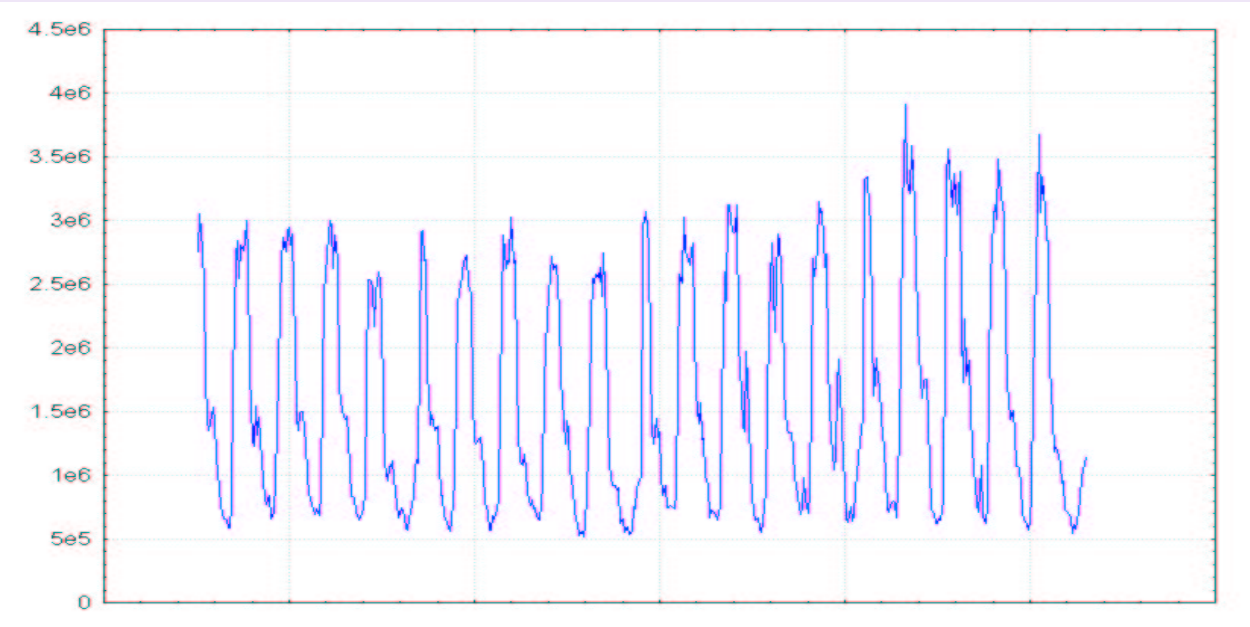
$$\widehat{\lambda}_n(s) = \frac{\sum_{k=1}^{n} X_k(B_n^s)}{\sum_{k=1}^{n} \int_{B_n^s} Y_k(u)du} \mathbf{1}_{D_n}(s), \quad s \in [0, T],$$

gdzie $s \in B_n^s$, B_n^s jest przedziałem długości T/b który zawiera s i

$$D_n = \{\sum_{k=1}^{n} \int_{B_n^s} Y_k(u)du > 0\}.$$

Przykład danych rzeczywistych

Liczba przechodzących pakietów podczas jednej godziny pomiędzy siecią Uniwersytetu Waikato i dostawcą Internetu.



Rezultaty graniczne

Asymptotyczna normalność, dziedzina czasu

Zachodzi zbieżność

$$\sqrt{n}\,(\hat{a}_n(\lambda,\tau) - a(\lambda,\tau)) \xrightarrow{d} \mathcal{N}_2(0,\Sigma),$$

gdzie

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix},$$

$$\sigma_{11} = \frac{1}{T}\sum_{s=1}^{T}\sum_{k=-\infty}^{\infty} B_{Z_\tau}(s,k)\cos(\lambda s)\cos(\lambda k),$$

$$\sigma_{22} = \frac{1}{T}\sum_{s=1}^{T}\sum_{k=-\infty}^{\infty} B_{Z_\tau}(s,k)\sin(\lambda s)\sin(\lambda k),$$

$$\sigma_{12} = \sigma_{21} = \frac{1}{T}\sum_{s=1}^{T}\sum_{k=-\infty}^{\infty} B_{Z_\tau}(s,k)\cos(\lambda s)\sin(\lambda k),$$

oraz $Z(t,\tau) = X(t)X(t+\tau) - B_X(t,\tau)$,
$B_{Z_\tau}(s,k) = Cov\,(Z(s,\tau), Z(s+k,\tau))$.

Asymptotyczna kowarinacja, szeregi POS

Lemat (Lenart, 2008)

Jeśli

(i) istnieje $\delta > 0$ taka, że $\sup_{t\in\mathbb{Z}} \|X_t\|_{6+3\delta} \leqslant \Delta < \infty$,

(ii) $\sum_{k=1}^{\infty} k^2 \alpha(k)^{\frac{\delta}{2+\delta}} \leqslant K < \infty$[a] ,

(iii) $K_n(s-t) = I\{|s-t| \leqslant w_n\}$ + dodatkowe warunki reg.

to mamy zbieżność

$$\lim_{n\to\infty} \frac{n}{w_n} cov\left(\hat{G}_n(\nu_1,\omega_1), \hat{G}_n(\nu_2,\omega_2)\right) = P(\nu_1,\nu_2)\overline{P(\omega_1,\omega_2)} + P(\nu_1, 2\pi-\omega_2)\overline{P(\nu_2, 2\pi-\omega_1)},$$

dla każdych $(\nu_1,\omega_2),(\nu_2,\omega_2) \in (0,2\pi]^2$.

---

[a] ciąg α-mieszania definiujemy jako $\alpha_X(\tau) = \sup_{t\in\mathbb{Z}} \sup_{\substack{A\in\mathcal{F}_X(-\infty,t)\\ B\in\mathcal{F}_X(t+\tau,\infty)}} |P(A\cap B) - P(A)P(B)|$, gdzie $\mathcal{F}(t_1,t_2)$ oznacza σ-ciało generowane przez obserwacje $\{X_t : t_1 \leqslant t \leqslant t_2\}$.

Asymptotyczna normalność, szeregi POS

Twierdzenie (Lenart, 2008)

Jeśli

(i) istnieje $\delta > 0$ taka, że $\sup_{t\in\mathbb{Z}} \|X_t\|_{6+3\delta} \leqslant \Delta < \infty$,

(ii) $w_n = O(n^\kappa)$ dla pewnego $\kappa \in (0, \delta/(4+4\delta))$,

(iii) $\sum_{h=1}^{\infty} h^{2r}\alpha(h)^{\frac{\delta}{2(r+1)+\delta}} < \infty$, gdzie r jest naturalne i $r > \max\left\{1+\delta, \frac{1-\kappa}{4\kappa}, \frac{\kappa(1+\delta)}{\delta-2\kappa(1+\delta)}\right\}$,

to

$$\sqrt{\frac{n}{w_n}}\left(\hat{G}_n(\nu,\omega) - P(\nu,\omega)\right) \longrightarrow N(0, \Sigma(\nu,\omega)),$$

gdzie macierz $\Sigma(\nu,\omega)$ może być otrzymana z poprzedniego lematu.

Asymptotyczna normalność, przypadek procesów liczących

Twierdzenie (Dudek, 2008)

Jeśli

(i) proces $\{Y_s\}$ jest okresowo skorelowany z okresem T i funkcja $E(Y_s)$ jest odcięta od 0,

(ii) $\|Y_t\|_3 \leqslant \Delta < \infty$,

(iii) proces $\{Y_s\}$ jest α-mieszający i $\alpha(k) = o(k^{-3})$,

(iv) każdy kolejny okres T jest dzielony na b części, gdzie $b = O(\sqrt{n})$

(v) funkcja λ_0 jest okresowa z okresem T i EY_s spełnia warunki Lipschitza na $[0, T]$,

to

$$\sqrt{\frac{n}{b}}\left(\widehat{\lambda}_n(s) - \lambda_0(s)\right) \Rightarrow N\left(0, \frac{\lambda_0(s)}{E(Y_s)}\right).$$

Dlaczego repróbkowanie?

* szeregi czasowe POS: skomplikowana macierz kowariancji w rozkładzie asymptotycznym estymatorów

* okresowe procesy liczące: powolna zbieżność, potrzeba konstrukcji jednoczesnych obszarów ufności

Subsampling dla współczynników Fouriera funkcji autokowariancji

Zgodność zachodzi dla estymatora $\hat{\theta}_n = |\hat{a}_n(\lambda, \tau)|$. Oznaczmy

$$J_n(x, P) = Prob_P(\sqrt{n}(|\hat{a}_n(\lambda, \tau)| - |a(\lambda, \tau)|) \leqslant x).$$

Z CTG dla $\hat{a}_n(\lambda, \tau)$ i metody delta mamy

$$J_n(P) \xrightarrow{d} J(P).$$

Definiujemy analogicznie subsamplingowy rozkład jako

$$L_{n,b}(P) = \frac{1}{n-b+1} \sum_{t=1}^{n-b+1} \mathbf{1}\{\sqrt{b}(|\hat{a}_{n,b,t}(\lambda, \tau)| - |\hat{a}_n(\lambda, \tau)|) \leqslant x\}$$

Twierdzenie (Lenart, Leśkow, Synowiecki, 2008)

Niech $\{X(t) : t \in \mathbb{Z}\}$ będzie prawie okresowo skorelowanym szeregiem czasowym. Jeśli

(i) $b \to \infty$ ale $b/n \to 0$,

(ii) $\sup_t E|X(t)|^{4+4\delta} < \infty$,

(iii) $\sum_{k=0}^{\infty} (k+1)^2 \alpha(k)^{\frac{\delta}{4+\delta}} < \infty$,

(iv) funkcja $V(t, \tau_1, \tau_2, \tau_3) = E(X(t)X(t+\tau_1)X(t+\tau_2)X(t+\tau_3))$ jest prawie okresowa,

to subsampling jest zgodny, czyli

$$\sup_x |J_n(x, P) - L_{n,b}(x)| \xrightarrow{P} 0.$$

Zastosowanie prcedury subsamplingu dla okresowo skorelowanych szeregów czasowych

Problem testowania:

$$H_0 : B(\cdot, \tau) \text{ jest okresowa z okresem } T_0,$$
$$H_1 : B(\cdot, \tau) \text{ jest okresowa z okresem } T_1.$$

Statystyka testowa (Lenart, Leśkow, Synowiecki, 2008):

$$U_n(\tau) = \sqrt{n} \left(\sum_{\lambda \in \Lambda_{T_1} \setminus \Lambda_{T_0}} |\hat{a}_n(\lambda, \tau)| \right).$$

Przy hipotezie H_0:

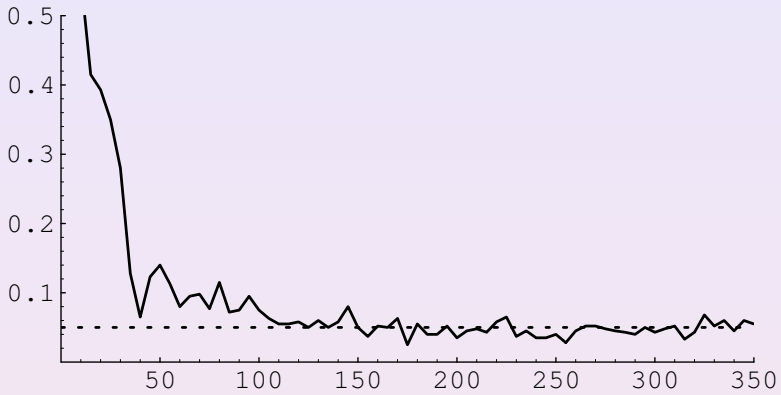
$$U_n(\tau) \xrightarrow{d} J.$$

Przy hipotezie H_1:

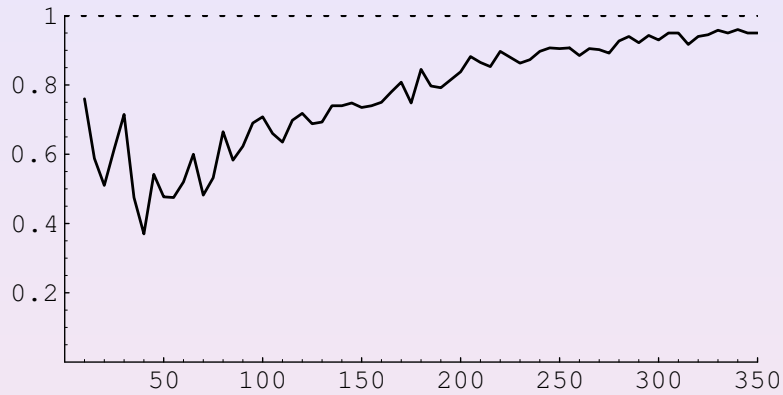
$$U_n(\tau) \longrightarrow \infty.$$

Duże wartości statystyki $U_n(\tau)$ świadczą na korzyść hipotezy H_1. Obszar odrzucenia jest postaci $[c_{1-\alpha}, \infty)$. Aby znaleźć kwantyl $c_{1-\alpha}$ można stosować subsampling.

Przykład symulacyjny



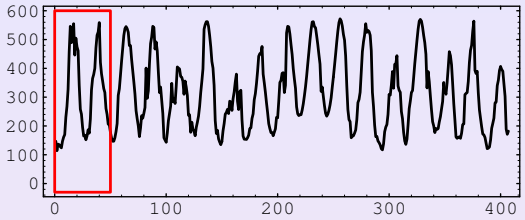
(a) Prawdopodobieństwo odrzucenia H_0 pod warunkiem, że hipoteza H_0 jest prawdziwa.



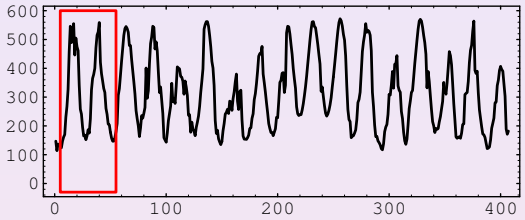
(b) Prawdopodobieństwo odrzucenia H_0 pod warunkiem, że hipoteza H_1 jest prawdziwa.

Rysunek: Aproksymacje Monte Carlo błedów testu subsamplingowego.

Idea MBB

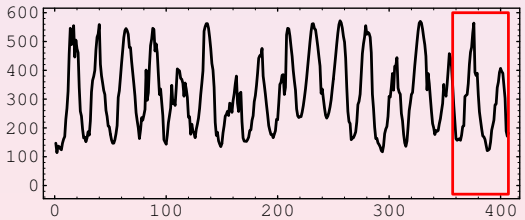


$$B_{1,b} = (X_1, \ldots, X_b)$$



$$B_{2,b} = (X_2, \ldots, X_{b+1})$$

$$B_{i,b} = (X_i, \ldots, X_{i+b-1})$$



$$B_{n-b+1,b} = (X_{n-b+1}, \ldots, X_n)$$

Idea MBB

$B^*_{1,b}, B^*_{2,b}, \ldots, B^*_{k,b}$ - i.i.d. z $U(\{B_{1,b}, \ldots, B_{n-b+1,b}\})$

Łączymy wylosowane bloki $\mapsto (X^*_1, \ldots, X^*_n)$.

$\hat{\theta}_n = T(F_n)$ - estymator, który jest funkcjonałem rozkładu empirycznego próbki $(X_1, \ldots, X_n)$.

$\hat{\theta}^*_n = T(F^*_n)$ - wersja MBB estymatora, F^*_n rozkład empiryczny próbki MBB $(X^*_1, \ldots, X^*_n)$.

Rozkład $\hat{\theta}^*_n$ przybliżamy za pomocą Monte Carlo.

Zgodność procedury MBB dla szeregów POS

Twierdzenie (Synowiecki, 2007)

Niech $\{X_t : t \in \mathbb{Z}\}$ będzie POS i α-mieszający, $(X_1^*, \ldots, X_n^*)$ oznacza próbkę MBB, niech $b \to \infty$ ale $b/n \to 0$. Załóżmy, że

(i) $\Lambda = \{\lambda : [0, 2\pi) : M_t(EX_t e^{-i\lambda t}) \neq 0\}$ jest skończony,

(ii) autokowariancja jest jednostajnie sumowalna,

(iii) $\sup_{s=1,\ldots,n-b+1} E\left(\frac{1}{\sqrt{b}} \sum_{t=s}^{s+b-1}(X_t - EX_t)\right)^4 < K$

(iv) zachodzi CTG, tzn. $\sqrt{n}\left(\overline{X}_n - M_t(EX_t)\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$

Wtedy procedura MBB jest zgodna, czyli

$$Var^*(\sqrt{n}\,\overline{X}_n^*) \xrightarrow{P} \sigma^2$$

oraz

$$\sup_{x \in \mathbb{R}} \left| P\left(\sqrt{n}\left(\overline{X}_n - \mu\right) \leqslant x\right) - P^*\left(\sqrt{n}\left(\overline{X}_n^* - E^*\overline{X}_n^*\right) \leqslant x\right)\right| \xrightarrow{p} 0.$$

Zgodność subsamplingu dla szeregów POS - koherencja spektralna

Twierdzenie (Lenart, 2008)

Przy warunkach regularności subsamplingowe przedziały ufności dla koherencji są zgodne, czyli

$$P\left(\sqrt{n/w_n}\,(|\hat{\gamma}_n(\nu,\omega)| - |\gamma(\nu,\omega)|) \leqslant c^{\gamma}_{n,b}(1-\alpha)\right) \longrightarrow 1-\alpha,$$

gdzie $b = b(n) \to \infty$, oraz $b/n \to 0$,

$$c^{\gamma}_{n,b}(1-\alpha) = \inf\{x : \; L^{\gamma}_{n,b}(x) \geqslant 1-\alpha\}.$$

$$L^{\gamma}_{n,b}(x) = \frac{1}{n-b+1} \sum_{t=1}^{n-b+1} \mathbf{1}\{\sqrt{b/w_b}(|\hat{\gamma}_{n,b,t}(\nu,\omega)| - |\hat{\gamma}_n(\nu,\omega)|) \leqslant x\}.$$

Zgodność bootstrapu dla procesów liczących

Twierdzenie (Dudek, 2008)

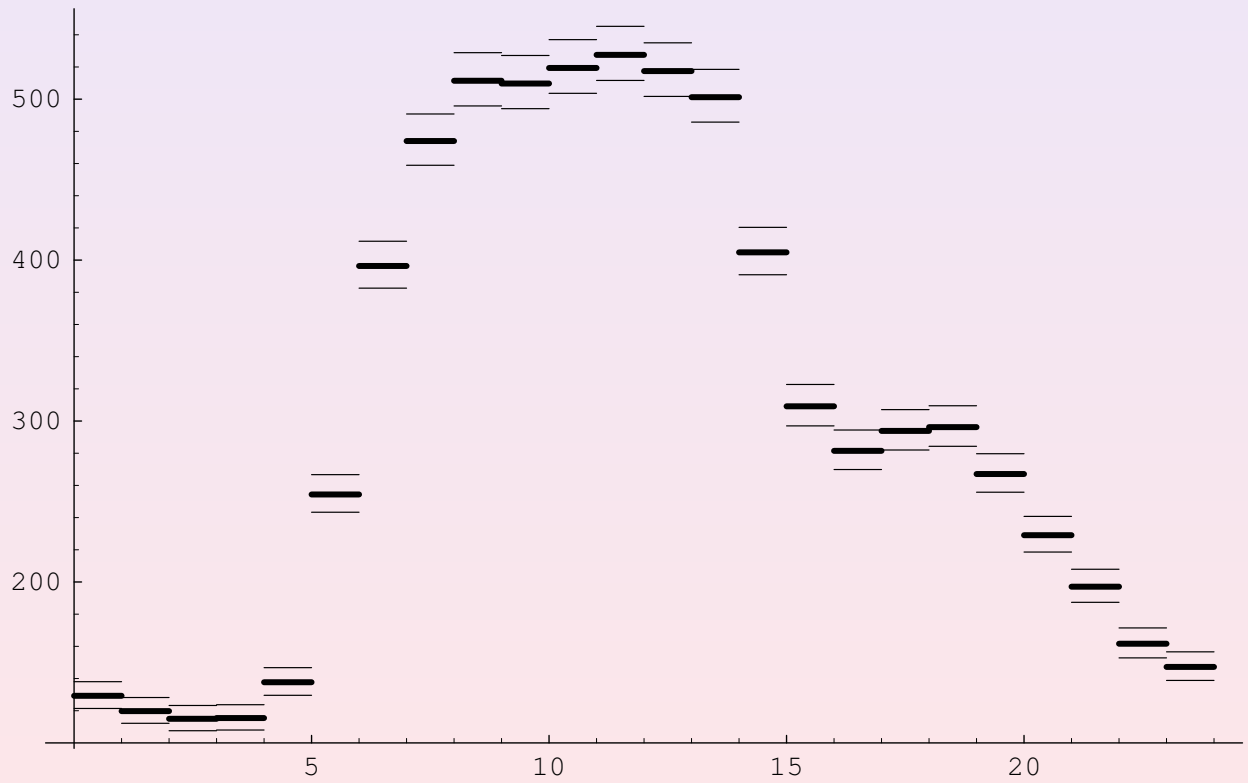
$$\sup_{u\in R}\left|P^*\left(\sqrt{\frac{n}{b}}(\widehat{\lambda}_n^*(s)-\widehat{\lambda}_n(s))\leqslant u\right)\right.$$
$$\left.-P\left(\sqrt{\frac{n}{b}}(\widehat{\lambda}_n(s)-\lambda_0(s))\leqslant u\right)\right| = o_P(1),$$

gdzie

$$\widehat{\lambda}_n^*(s)=\frac{\sum_{k=1}^n X_k^*(B_n^s)}{\sum_{k=1}^n \int_{B_n^s} Y_k(u)du}1_{D_n}(s).$$

Przykład danych rzeczywistych - proces liczący

Estymator intensywności ilości pakietów, które są wysyłane przez jeden komputer wraz 90% regionem ufności:



Bibliografia


Dudek A., Goćwin M., Leśkow J., (2008)
Simultaneous confidence bands for the integrated hazard function
Computational Statistics


Lenart Ł., Leśkow J., Synowiecki R. (2008)
Subsampling in testing autocovariance for periodically correlated time series
Journal of Time Series Analysis


Leśkow J., Synowiecki R. (2008)
On bootstrapping periodic random arrays with increasing period
submitted

References


Lenart Ł., (2008)
Asymptotic properties of periodogram for almost periodically correlated time series
Probability and Mathematical Statistics


Synowiecki R., (2007)
Some results on the subsampling for φ-mixing periodically strictly stationary time series
Probability and Mathematical Statistics


Synowiecki R., (2007)
Consistency and application of MBB for nonstationary time series with periodic and almost periodic structure
Bernoulli

Dziękuję za uwagę!