Seminarium
Wydziału Fizyki i Informatyki Stosowanej AGH
16 kwietnia 2010

STATYSTYKA DANYCH SAMOSKORELOWANYCH

Andrzej Zięba

Plan:

1. Wstęp
- formalizm standardowy i jego ograniczenia
- matematyczny opis danych samoskorelowanych

2. Teoria analityczna, gdy funkcja autokorelacji jest znana *a priori*

3. Przypadek funkcji autokorelacji estymowanej z analizowanej próby losowej – teoria & MC

4. Konkluzje i podziękowania

ad 1a: najczęściej stosowane wzory statystyki matematycznej:

♣ wynik pomiaru
≡ średnia arytmetyczna $$x \equiv \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

♦ odchylenie st. pojedynczego pomiaru $$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

♥ niepewność pomiaru
≡ odchylenie st. średniej $$u(x) \equiv s(\bar{x}) = \frac{s}{\sqrt{n}}$$

♠ względny rozrzut estymatorów odchylenia st. $$\frac{s(s)}{s} = \frac{s(s(\bar{x}))}{s(\bar{x})} \cong \frac{1}{\sqrt{2(n-1)}}$$

Ww. estymatory są zgodne, nieobciążone i najefektywniejsze jeżeli obserwacje są

(i) równoważne,
(ii) wzajemnie nieskorelowane,
(iii) obarczone błędem przypadkowym o rozkładzie normalnym.

Co robić, jeżeli $\{x_i\}$ są skorelowane?

(Pozostałe dwa założenia obowiązują.)

1b. MATEMATYCZNY OPIS

DANYCH

SAMOSKORELOWANYCH

Ciąg n skorelowanych obserwacji $\{x_i\}$ można opisywać przy użyciu trzech formalizmów teorii prawdopodobieństwa

(i) – realizacja **wielowymiarowej zmiennej losowej** $\{X_1, X_2, \ldots, X_n\}$ o *równoważnych* składowych.

Parametry:
- wartość oczekiwana μ,
- odchylenie st. σ,
- macierz wsp. korelacji ρ_{ij}

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & \rho_n \end{bmatrix}$$

Struktura macierzy autokorelacji, wynikająca z założenia równoważności, umożliwia zastąpienie jej przez jednowymiarową dyskretną

funkcję autokorelacji: $\rho_0, \rho_1, \rho_2, \ldots, \rho_{n-1}$

(ii) *n*-elementowa próba ze

stacjonarnego szeregu czasowego

→ stacjonarność ≡ równoważność

→ modele szeregów czasowych, np.:

autoregresyjny AR(1) $x_n = ax_{n-1} + (1-a)u_n$

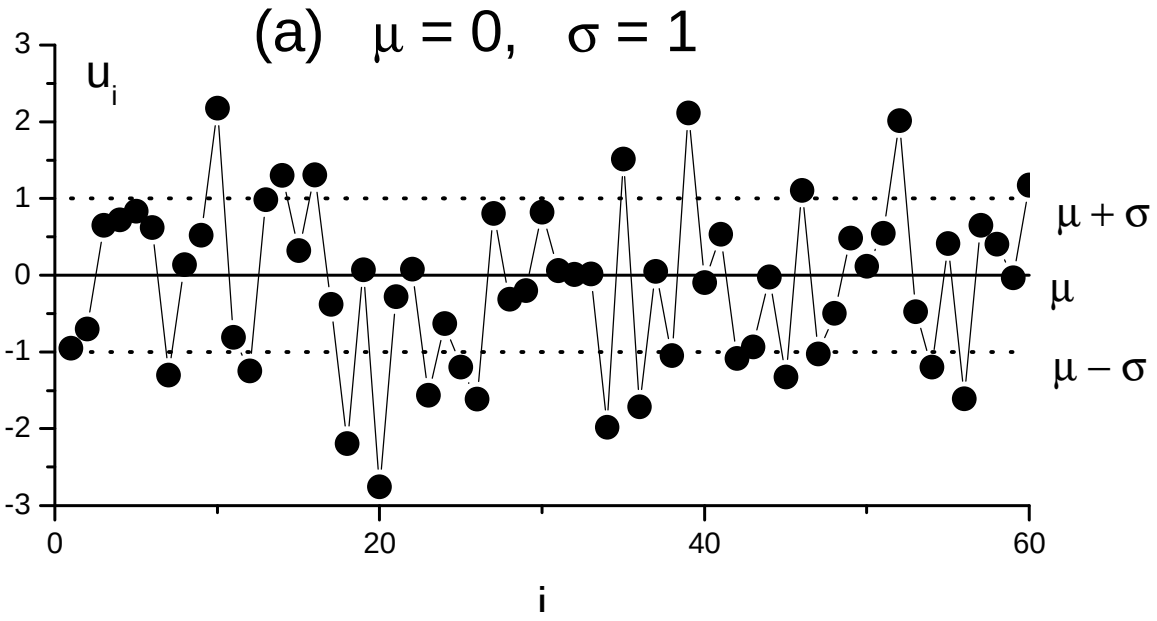
średnia ruchoma SMA $x_n = (u_n + u_{n-1} + \ldots + u_{n-m})/m$

(iii) $\{x_i\}$ wynikiem próbkowania (w równych przedziałach czasu Δt) ciągłego

stacjonarnego procesu stochastycznego $x(t)$

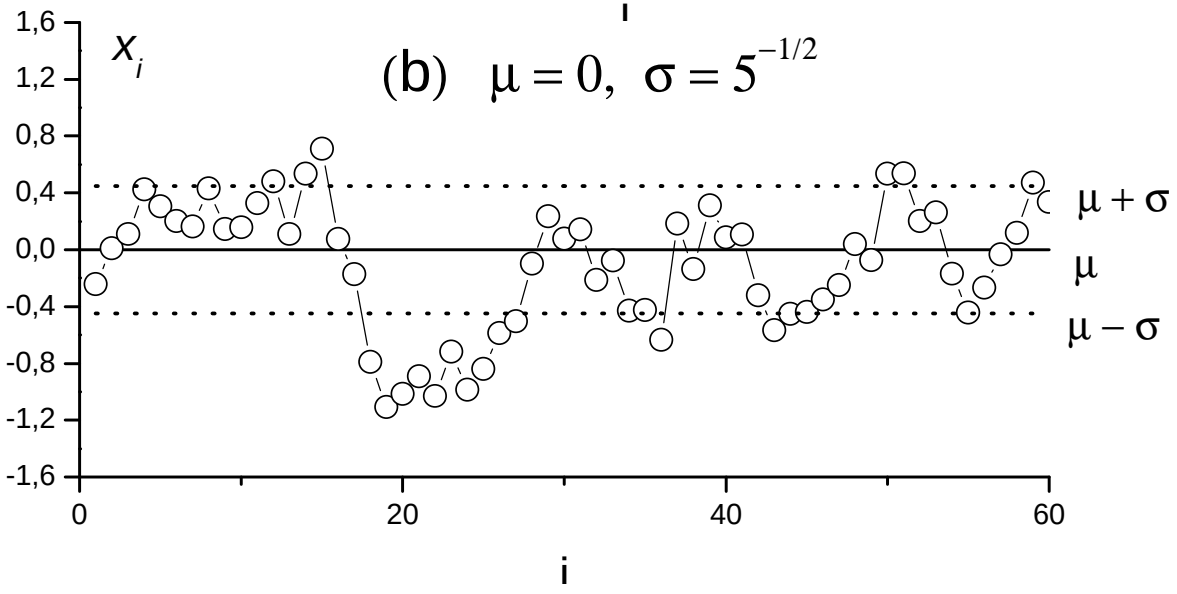
Przykład: prosta średnia ruchoma (SMA)

u_i - liczby nieskorelowane



x_i - liczby skorelowane

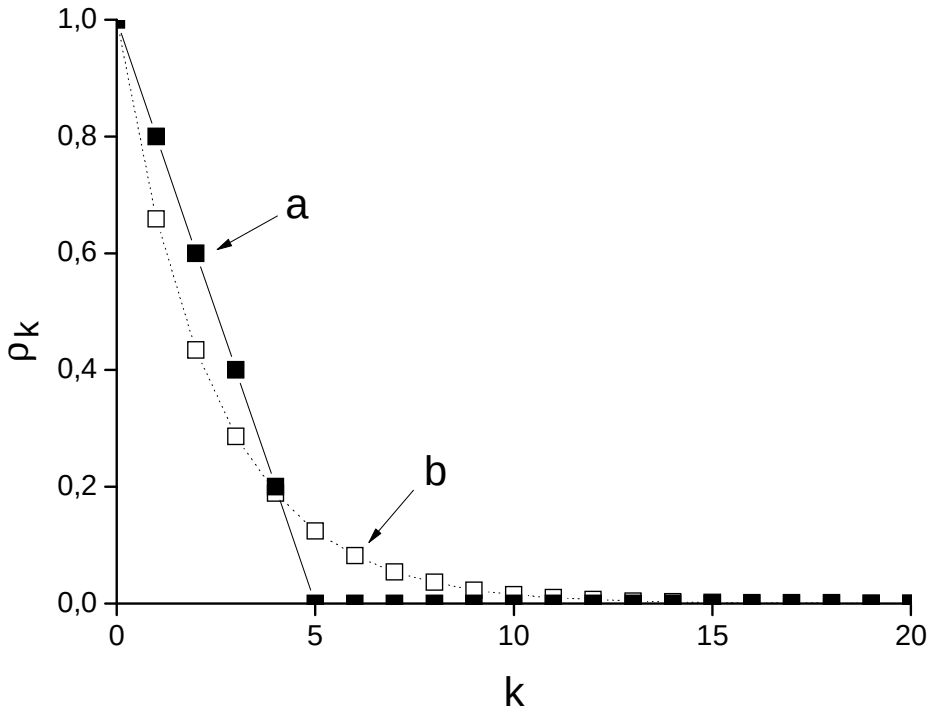
$$x_i = \frac{u_{i-2} + u_{i-1} + u_i + u_{i+1} + u_{i+2}}{5}$$



Funkcje autokorelacji dla SMA, $m = 5$:

$\rho_0 = 1, \rho_1 = 0{,}8 \quad \rho_2 = 0{,}6 \quad \rho_3 = 0{,}4 \quad \rho_4 = 0{,}2 \quad$ pozostałe $= 0$

AR(1): $\rho_k = a^{-k}$



2. FORMALIZM DLA PRZYPADKU, GDY FUNKCJA AUTOKORELACJI JEST ZNANA a priori

Średnia arytmetyczna

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

pozostaje najlepszym estymatorem wartości oczekiwanej bo:

- jako zmienna losowa jest liniową kombinacją zmiennych X_i
  (ze współczynnikami $c_i = 1/n$)

- twierdzenie o wartości oczekiwanej kombinacji liniowej zmiennych losowych jest takie samo
  dla zmiennych nieskorelowanych i skorelowanych

Ad ♥

ZWIĄZEK MIĘDZY WARIANCJĄ I WARIANCJĄ ŚREDNIEJ.

EFEKTYWNA LICZBA OBSERWACJI

Intuicyjnie: na przykładzie danych skorelowanych wygenerowanych jako średnia ruchoma z *m* elementów łatwo pojąć, że wzór

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

nie może obowiązywać dla danych skorelowanych.
Gdyby był słuszny, obliczenie średniej ruchomej z *m* elementów, a następnie średniej z tejże, dostarczało by „cudownego triku" na zmniejszenie niepewności pomiaru do wartości circa

$(m\ n)^{1/2}$ razy mniejszej!

Teoria prawdopodobieństwa: wzór na wariancję sumy – oraz

kombinacji liniowej – jest dla zmiennych skorelowanych inny.

Wzór prawidłowy:

$$\sigma^2(\bar{x}) = \frac{\sigma^2}{n^2}\left[n + 2\sum_{k=1}^{n-k}(n-k)\rho_k \right]$$

Box G. E. P., Jenkins G. M, Reinsel G. C.: *Time Series Analysis: Forecasting and Control*, 3rd Ed. Prentice Hall 1994

Wyprowadzenie: średnia to liniowa kombinacja zm. losowych: $\bar{x} = \sum_{i=1}^{k} c_k X_k, \; c_k = \frac{1}{n}$

jej wariancja:

$$\sigma^2(\bar{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_j c_j \sigma_i \sigma_j \rho_{ij} =$$

$$= \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\rho_{i,j} =$$

$$\rho_{ij} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}$$

$= (\sigma^2/n^2) \times$ suma elementów macierzy kowariancji

Wzór ten

$$\sigma_{\bar{x}}^2 = \left[n + \sum_{k=1}^{n-k} (n-k)\rho_k \right] \frac{\sigma^2}{n^2}$$

zapisać można wprowadzając pojedynczy parametr

efektywna liczba obserwacji n_{eff}.

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n_{eff}}} \qquad n_{eff}' = \frac{n}{1 + 2\sum_{k=1}^{n-k} (1 - k/n)\,\rho_k}$$

Własności n_{eff}

- liczba rzeczywista z przedziału $[1, \infty)$

- próbkowanie procesu stochastycznego: $n \to \infty$
  (w ustalonym przedziale, $\Delta t \to 0$) $n_{eff} \to const$

Przykład:

model SMA(5), n = 60 $\to$ n_{eff} = 12,33

Wyniki poszukiwań literaturowych, 2009:

- reduced number of coordinates (Bartels 1935)
- effective number of independent observations (Bayley & Hammersley 1946)
- equivalent number of independent data (Bagrov 1969)
- equivalent number of uncorrelated samples (Lubman 1969)
- effective independent sample size (Leith 1973)
- effective sample size (Taubenheim 1974)
- equivalent number of independent observations (Priestley 1981)
- equivalent independent process effective number (Zen 1998)
- effective number of uncorrelated observations (Dorozhovets & Warsza 2007)

Ad ♦ i ♥

NIEOBCIĄŻONE ESTYMATORY WARIANCJI

Wyprowadzenie dla
danych nieskorelowanych:

(i) definiujemy estymator wariancji: $$s_b^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

(ii) obliczamy wart. oczekiwaną
(7 linijek algebry): $$\mathrm{E}\left(s_b^2\right) = \sigma^2 - \sigma^2(\bar{x})$$

(iii) dla danych nieskorelowanych $$\sigma^2(x) = \sigma^2/n$$

(iv) obciążenie estymatora s_b^2 $$\rightarrow \quad \mathrm{E}(s_b^2) = (1-1/n)\,\sigma^2$$

(iv) czynnik korekcyjny $$(1-1/n)^{-1} = n/(n-1)$$

(v) estymator s_b^2 mnożymy przez
czynnik korekcyjny uzyskując
nieobciążony estymator wariancji: $$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Wyprowadzenie dla
danych skorelowanych:

(i) definiujemy estymator wariancji: $$s_b^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

(ii) obliczamy wart. oczekiwaną
(7 linijek algebry): $$\mathrm{E}\left(s_b^2\right) = \sigma^2 - \sigma^2(\bar{x})$$

(iii) dla danych skorelowanych $$\sigma^2(x) = \sigma^2/n_{eff}$$

(iv) obciążenie estymatora s_b^2 $$\rightarrow \quad \mathrm{E}(s_b^2) = (1 - 1/n_{eff})\,\sigma^2$$

(iv) czynnik korekcyjny $$(1 - 1/n_{eff})^{-1} = n_{eff}/(n_{eff} - 1)$$

(v) estymator s_b^2 mnożymy przez
czynnik korekcyjny uzyskując
nieobciążony estymator wariancji: $$s_a^2 = \frac{n_{eff}}{n_{eff} - 1} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

Wzór ten:

$$s_a^2 = \frac{n_{eff}\,(n-1)}{n(n_{eff}\,-1)}\; s^2$$

podany był bez wyprowadzenia w pracy:

Bayley and Hammersley, The “Effective” Number of Independent Observations in an Autocorrelated Time-Series. *J. Roy. Stat. Soc. Suppl.* **8**, 184-197 (1946)

zapomniany (?) przez 60 lat.

It follows that

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \qquad (8)$$

$$m_2' = \frac{1}{n}\sum_{j=1}^{n} \{x_j - E(x)\}^2 \qquad (9)$$

$$s^2 = \left\{\frac{n_b^*(n-1)}{n(n_b^*-1)}\right\}\left\{\frac{1}{n-1}\sum_{j=1}^{n}(x_j-\bar{x})^2\right\} \qquad (10)$$

2. If τ is the time interval between successive observations, so that $\sigma^2\rho(j\tau) = E(x_a x_{a+j})$ then

$$\text{var}\,(\bar{x}) = \frac{\sigma^2}{n} + \frac{2\sigma^2}{n^2}\sum_{j=1}^{n-1}(n-j)\rho(j\tau) \qquad (11)$$

$$\text{var}\,(m_2') = \frac{2\sigma^4}{n} + \frac{4\sigma^4}{n^2}\sum_{j=1}^{n-1}(n-j)\rho^2(j\tau) \qquad (12)$$

$$\text{var}\,(s^2) = \frac{2\sigma^4\{n^2(n-1) - 4n\Sigma_1 + 2n\Sigma_3 - 8\Sigma_4 - 4n\Sigma_6 - 8n\Sigma_7\}}{n^2(n-1)^2 - 4n(n-1)\Sigma_1 + 2\Sigma_2 - 2\Sigma_3 + 8\Sigma_4 + 8\Sigma_5} \qquad (13)$$

Ponadto:

- wzory na obciążenie estymatorów macierzy kowariancji podał T.W. Anderson: *The Statistical Analysis of Time Series*. Wiley, New York, 1971, ale nie wykorzystał do wyprowadzenia nieobciążonego estymatora wariancji,

- Şen (1998): wyprowadzenie dla modelu ARMA(1)

Wariancja średniej

(AZ, PPM 2008):

$$s_a^2(\bar{x}) = \frac{s_a^2}{n_{eff}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n(n_{eff} - 1)}$$

Konwencja GUM: niepewność pomiaru typu A jest pierwiastkiem kwadratowym z nieobciążonego estymatora wariancji

zatem niepewność pomiaru
dla *n* obserwacji skorelowanych:

$$u(x) \equiv s(\bar{x}) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n\,(n_{eff} - 1)}}$$

Ad ♥

DOKŁADNOŚĆ OCENY NIEPEWNOŚCI

Próba los. nieskorelowana: $u(s)/s = (2\nu)^{-1/2}$ gdzie $\nu = n - 1$

Próba samoskorelowana: $u(s_a)/s_a = (2\nu_{eff})^{-1/2}$
gdzie

ν_{eff} - **efektywna liczba stopni swobody**

(różna od $\boldsymbol{n_{eff} - 1}$, własność: $\mathbf{0} < \nu_{eff} < \boldsymbol{n}$)

Bayley & Hammersley 1946: wzór dokładny:

$$\nu_{eff} = \frac{n^3(n-1) - 4n^2\Sigma_1 + 2\Sigma_2 + 8\Sigma_5 - 4n\Sigma_6 - 8n\Sigma_7}{n^2(n-1) - 4n\Sigma_1 + 2\Sigma_3 - 8\Sigma_4 - 4n\Sigma_6 - 8n\Sigma_7} - 1$$

wzory dla określonych modeli: Priestley 1981, Taubenheim 1974, Fortus 1999)

Wzór przybliżony,
poprawny efekt uproszczenia
wzoru dokładnego (AZ 2009)

$$\nu_{eff} \cong \frac{n}{1 + 2\sum_{k=1}^{n-1} \rho_k^2} - 1$$

Publikacja:

A. Zięba

EFFECTIVE NUMBER OF OBSERVATIONS AND UNBIASED ESTIMATORS OF VARIANCE FOR AUTOCORRELATED DATA – AN OVERVIEW

***Metrology and Measurement Systems*,**
Vol. 17, nr 1 (2010), str. 3-16

obejmuje podane do tego miejsca wyniki,

Formalizm można stosować bezpośrednio do analizy danych, jeżeli znamy funkcję autokorelacji (z metod typu B)

Jest punktem wyjścia do obliczeń wykorzystujących tylko zbiór obserwacji $\{x_i\}$ tj. pełnej metody typu A przedstawionych w dalszej części referatu.

3. ESTYMOWANIE n_{eff}
I
ESTYMATORÓW WARIANCJI
WYŁACZNIE
Z PRÓBY LOSOWEJ

Zasada tworzenia estymatora n_{eff}:

parametr: estymator:

$$n_{eff} = \frac{n^2}{n + 2\sum_{k=1}^{n-k}(n-k)\,\rho_k} \quad \Rightarrow \quad \hat{n}_{eff} = \frac{n^2}{n + 2\sum_{k=1}^{n_c}(n-k)\,r_k}$$

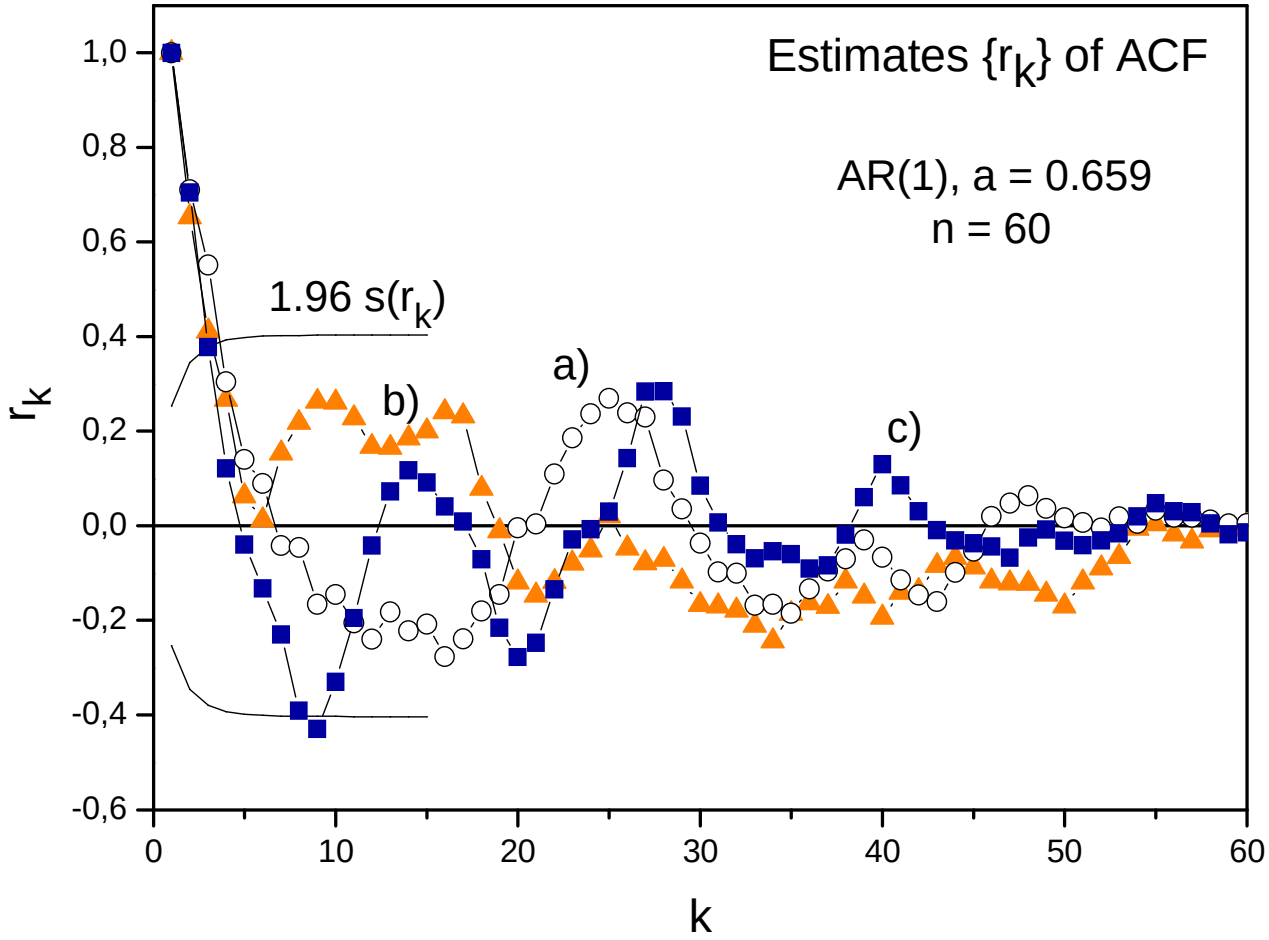
(a) ograniczenie sumowania do wskaźnika $n_c \le n-1$

(b) zastąpienie funkcji $\{\rho_k\}$ przez jej estymator, tu: $\{r_k\}$

obliczony z tejże próby losowej $\{x_i\}$

Standardowy estymator funkcji autokorelacji

$$r_k = \frac{\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$



Metoda ostatniej statystycznie niezerowej r_k (LSN)

N. F. Zhang, „Calculation of the uncertainty of the mean of autocorrelated measurements”, *Metrologia* (2006) S276-S281

parametr:

$$n_{eff} = \frac{n^2}{n + 2\sum_{k=1}^{n-k}(n-k)\,\rho_k}$$

⇨

estymator:

$$\hat{n}_{eff} = \frac{n^2}{n + 2\sum_{k=1}^{n_c}(n-k)\,r_k}$$

Sumowanie ograniczone na wartości odstępu n_c odpowiadającemu ostatniej niezerowej wartości r_k
dla prawdopodobieństwa objęcia 95%"

$$n_c = \max\left\{k \,\middle|\, |r_k| > 1.96\, s(r_k)\right\}$$

gdzie

$$s(r_k) = \sqrt{\frac{1 + 2\sum_{j=1}^{k-1} r_k^2}{n}}$$

Metoda pierszego przejścia przez zero (FTZ):

nieformalnie: Dorozhovets i Warsza, 2007

sformułowanie metody& badanie własności metodą MC
Zięba & Ramza, PPM'09:

graniczny odstęp n_c wyznacza pierwsze przejście funkcji autokorelacji przez zero, formalnie

$$n_c = \min\left\{k \,\middle|\, (r_k > 0 \ \wedge \ r_{k+1} < 0)\right\}$$

Konieczne założenie: wszystkie ρ_k nieujemne

Własności: - $0 < n_{eff} < n$ dla każdych danych
- mniejsze obciążenie i rozrzut

Badanie estymatorów n_{eff} metodą Monte Carlo, n = 60:

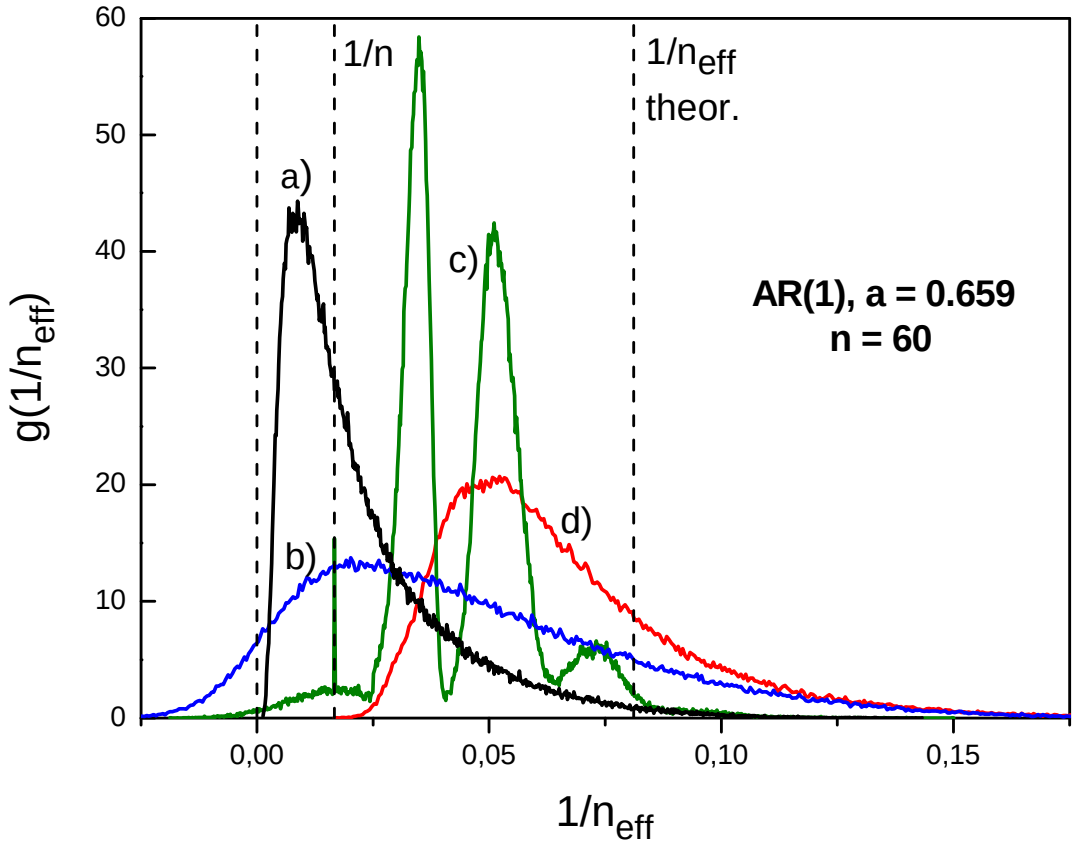
Sumowanie:

a) wszystkich wyrazów: $n_c = n - 1$

b) obcięte na: $n_c = n/4$

c) do ostatniego r_k statystycznie niezerowego (LSN)

d) do pierszego przejścia $\{r_k\}$ przez zero (FTZ)



Inne opcje dla estymatora ACF:

- estymator standardowy

$$r_k = \frac{c_k}{c_0} = \frac{\frac{1}{n}\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

2 przyczyny obciążenia:

- nierówna liczebność sum
- użycie średniej zamiast wartości oczekiwanej

- „z gwiazdką": z usuniętym obciążeniem od nierównej liczebności sum

$$r_k^* = \frac{c_k^*}{c_0^*} = \frac{\frac{1}{n-k}\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Opinia z podręcznika:

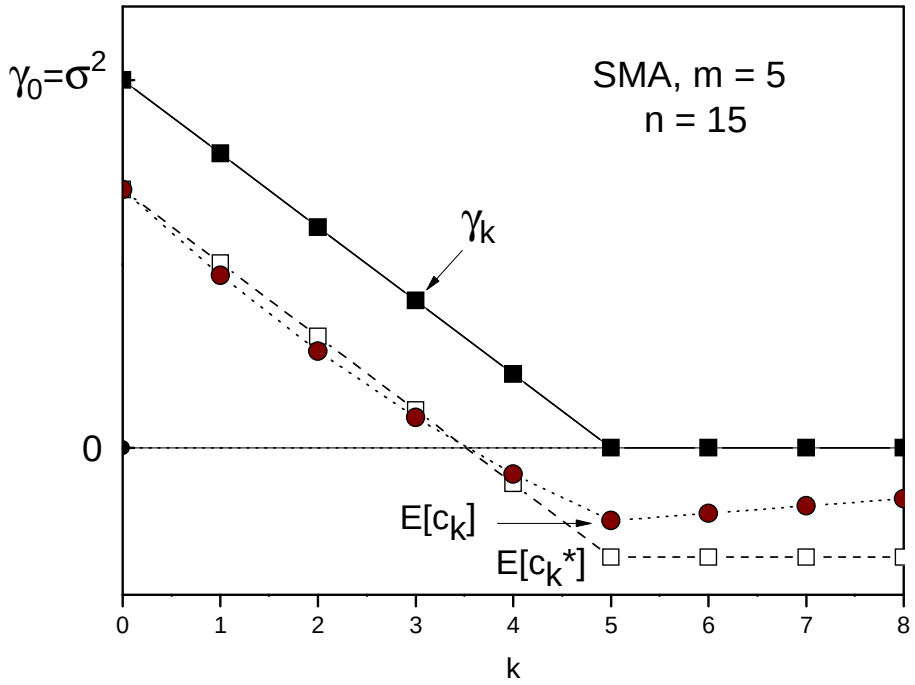
Priestley,
Spectral analysis and time series,
Elsevier 1981

A plot of $\hat{R}(r)$ (or $\hat{R}^*(r)$) against r, $(r = 0, \pm 1, \ldots \pm (N-1))$ is usually called the "*sample autocovariance function*". To distinguish between $\hat{R}^*(r)$ and $\hat{R}(r)$ we will stretch the usual terminology and refer to the former as the "*unbiased estimate*", and to the latter as the "*biased estimate*".

Choice of estimates

It would be fair to say that nowadays most time series analysts prefer to use the biased estimate $\hat{R}(r)$ rather than the unbiased estimate $\hat{R}^*(r)$ and this is reflected in the fact that the majority of computer time series packages also use the biased estimate. To statisticians this may seem surprising since, in general, there is a natural tendency to use unbiased rather than biased estimates, particularly when, as in this case, an unbiased estimate can be constructed so easily. To understand the reason for this apparent departure

W istocie:



T.W. Anderson: *The Statistical Analysis of Time Series*. Wiley, 1971.

Straightforward, but laborious, calculation yields

$$(51)\quad \mathscr{E}C_0^* = \sigma(0) - \frac{1}{T}\left\{\sigma(0) + 2\sum_{r=1}^{T-1}\left(1 - \frac{r}{T}\right)\sigma(r)\right\},$$

$$(52)\quad \mathscr{E}C_h^* = \sigma(h) - \frac{1}{T}\left\{\sigma(0) + 2\sum_{r=1}^{h}\left[1 - \frac{rh}{T(T-h)}\right]\sigma(r) + 2\sum_{r=h+1}^{T-h-1}\left[1 - \frac{rh}{T(T-h)} - \frac{r-h}{T-h}\right]\sigma(r) + 2\sum_{r=T-h}^{T-1}\frac{(T-r)h}{T(T-h)}\sigma(r)\right\},$$

$$1 \le h < T - h - 1,$$

po uporządkowaniu:

$$E\left(c_k^*\right) = \gamma_k - \frac{\sigma^2}{n_{eff}} + O(1/n^2)$$

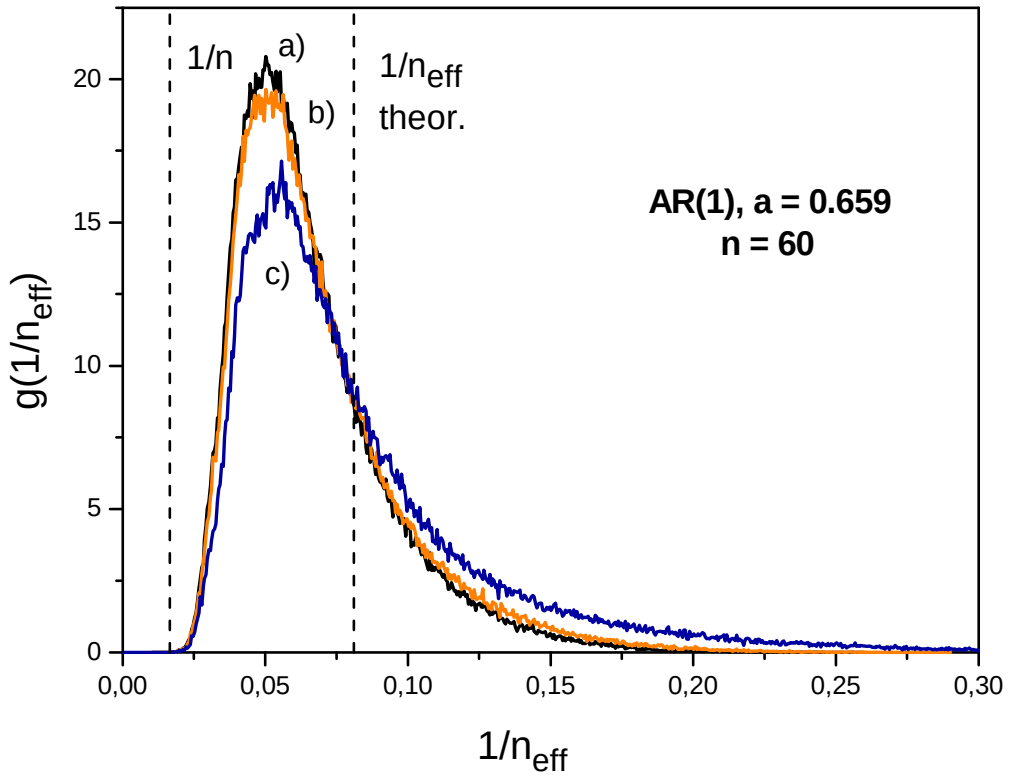
Obciążenie polega na przesunięciu funkcji $\{c_k^*\}$
o stały składnik

Obliczenie obciążenia umożliwia jego kompensację.

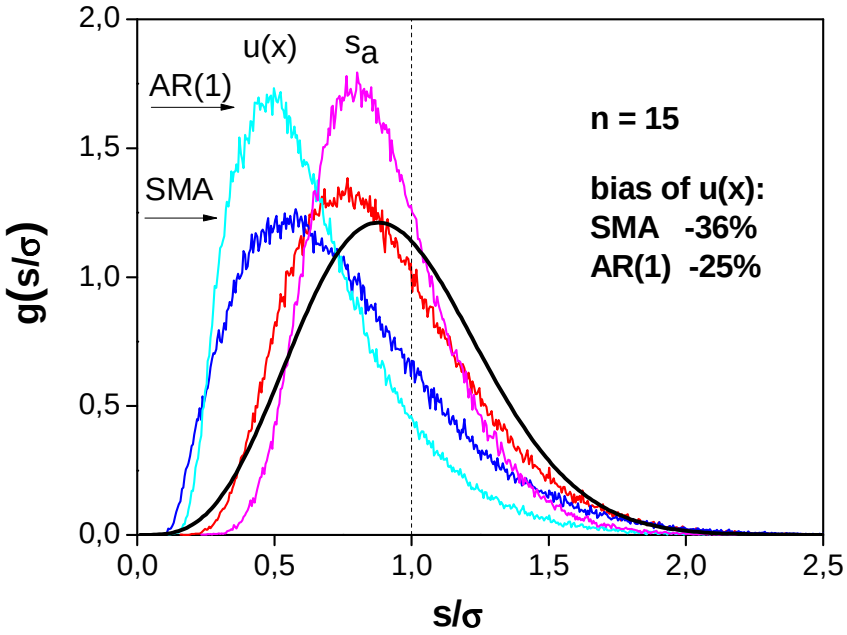
wyniki:

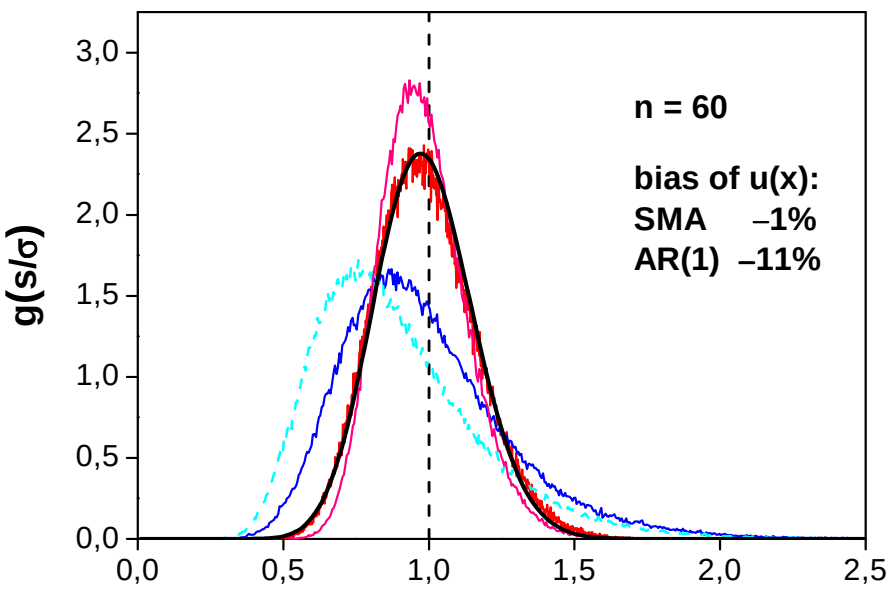
$$n_{eff}^{+} = \frac{n - 2n_c - 1 + n_c(n_c+1)/n}{1 + 2\sum_{k=1}^{n_c} r_k} + 1$$

$$r_k^{+} = \begin{cases} (1 - 1/n_{eff}^{+}) r_k^{*} + 1/n_{eff}^{+} & k \le n_c \\ 0 & k > n_c \end{cases}$$



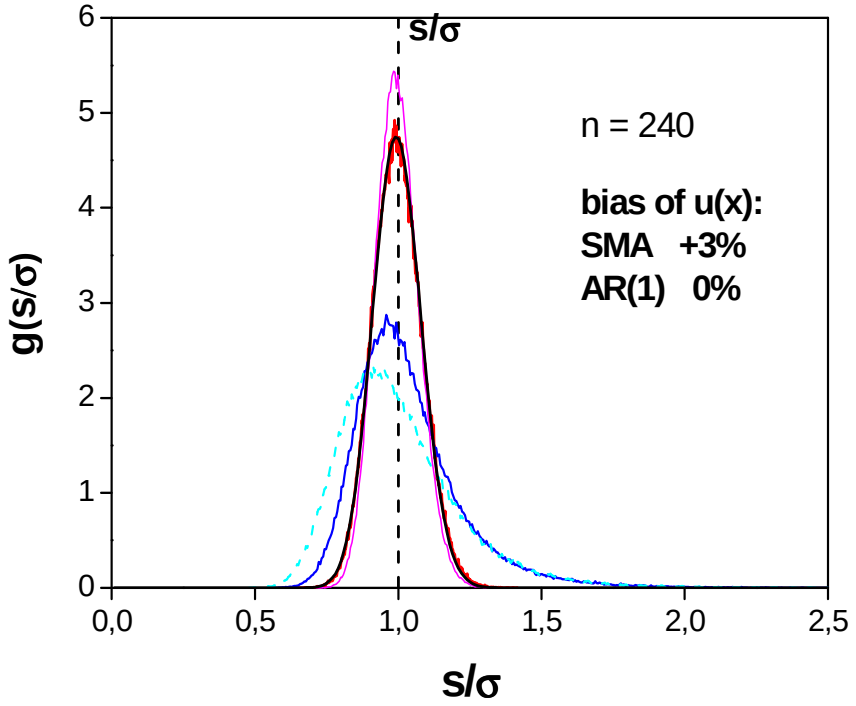
ESTYMATORY ODCHYLENIA STANDARDOWEGO BADANE METODĄ MC





Odchylenie standardowe:
pojedynczego pomiaru s_a
i średniej czyli $u(x)$

(ν_{eff} obliczane metodą FTZ)



Wnioski z symulacji MC:

- potwierdzenie przydatność formalizmu dla osób nieufnych w stosunku do nowych wzorów

- algorytm pracuje również dla zupełnie małej liczebności próby wbrew „wyssanej z palca" regule (ang. rule of thumb) że minimalna wartość n = 50 (podręcznik Box et al.)

- rozrzuty estymatorów odchylenia standardowego

w przybliżeniu równe $\frac{1}{\sqrt{2\nu_{eff}}}$

- rozrzut dla odchylenia standardowego średniej większy, ale nie bardziej niż dwa razy

- obciążenie estymatorów pomijalne w porównaniu z rozrzutem

4. PODSUMOWANIE

Konkluzja całościowa

Opracowano algorytmy
będące ścisłym odpowiednikiem wzorów dla
nieskorelowanej zmiennej gaussowskiej

umożliwiają obliczenie niepewności typu *A*
dla danych samoskorelowanych

Uporządkowanie tematu & nowe wyniki
w tym dziale statystyki matematycznej

Zagadnienia otwarte:

- przydatność formalizmu do różnych danych rzeczywistych

- obliczanie niepewności rozszerzonej

- obserwacje nierównoważne

- niegaussowskie funkcje rozkładu

- procesy stochastyczne nie posiadające ustalonej wariancji

- dopasowanie prostej i innych funkcji do danych skorelowanych

- itd., itp.

Podziękowania:

Zygmunt Warsza – za inspirację i info

Polskim metrologom poznanym na konferencjach,
Podstawowe Problemy Metrologii 2004-2009
i Sympozjach Niepewności Pomiaru 2008 i 2010

Anonimowi recenzenci *Metrologii* – za uporczywą krytykę,
która umożliwiła lepsze zrozumienie tematu przez
autora i zapobiegła przedwczesnym publikacjom

Piotr Ramza – obliczenia MC & współpraca
2008 - 2010

Dziękuję za uwagę